Supplemental Material S1. Models used in the main analyses and the Markov chain Monte Carlo (MCMC) sampling procedure.

Productivity (Modeling Count Data)

The productivity model described in the Methods section is specified as follows.

$$\begin{gathered}\left(U_{ij}-1\right) \sim Poisson\left(RUD_{ij}-1\right) \\ \log\left(RUD_{ij}-1\right)=\beta_{0j}^{(U)}+\left(\beta_{age}^{(U)} \times age_i\right)+\left(\beta_{sex}^{(U)} \times sex_i\right) \\ +\left(\beta_{ses}^{(U)} \times ses_i\right)+\left(\beta_{tbi}^{(U)} \times tbi_i\right)+\alpha_i^{(U)} \\ \alpha_i^{(U)} \sim Normal\left(0, \tau^{(U)}\right),\end{gathered} \tag{1}$$

where $\alpha_i^{(U)}$ represents the random intercept for subject i, and $\tau^{(U)}$ measures the subject-to-subject variability (standard deviation) in productivity that is not accounted for by the covariates. While we could choose another distribution for the random intercepts, a normal distribution is consistent with reasonable data assumptions. This is an example of Poisson regression with mixed effects, which is a type of GLMM that is appropriate for many situations in which the response is a count.

The interpretation of the β coefficients is similar to linear regression, but on the log scale. When we invert the log scale by exponentiating, the parameters introduced in Equation **Error! Reference source not found.**) become highly interpretable as the multiplicative effect of the explanatory variables. Because the covariates in our model are centered, the intercept terms $\beta_{0j}^{(U)}$ are also interpretable: $\exp\left(\beta_{0j}^{(U)}\right)=e^{\beta_{0j}^{(U)}}$ is the overall average number of utterances (beyond the first utterance) spoken by a person summarizing lecture type j. Instead of interpreting these coefficients directly, we transform them in order to report additive effects, with which our audience is perhaps more familiar. For completeness, estimates of the raw parameters for all four models are reported in Table 5. For more details on interpreting the coefficients on the multiplicative scale, see Agresti and Kateri (2011).

Whether interpreted on the multiplicative or additive scale, estimating the coefficients allows us to discover patterns in productivity relating to each of the demographic covariates, and detect the effect of lecture type, thereby answering our questions of interest. Indeed, the Poisson regression model is useful for a variety of situations in which counts are the response of interest.

Syntactic Complexity (Analyzing Ratios of Counts)

To deal with the multiple utterances per discourse sample, we make a simplifying assumption that for each individual (i.e., given each individual's speech characteristics) the number of clauses/words in a given utterance is independent of the number of clauses/words in the other utterances of the same discourse sample, which allows us to conclude that the sum of the utterance-specific supplemental clauses follows a Poisson distribution with mean equal to the sum $\sum_1^{U_{ij}}\left(RS_{ij}-1\right)=U_{ij}\left(RS_{ij}-1\right)$. The complete model is as follows.

$$\begin{gathered}\left(C_{ij}-U_{ij}\right) \sim Poisson\left(U_{ij}\left(RS_{ij}-1\right)\right) \\ \log\left(RS_{ij}-1\right)=\beta_{0j}^{(C)}+\left(\beta_{age}^{(C)} \times age_i\right)+\left(\beta_{sex}^{(C)} \times sex_i\right)+\left(\beta_{ses}^{(C)} \times ses_i\right) \\ +\left(\beta_{tbi}^{(C)} \times tbi_i\right)+\alpha_i^{(C)} \\ \alpha_i^{(C)} \sim N\left(0, \tau^{(C)}\right),\end{gathered} \tag{2}$$

where $\tau^{(C)}$ measures subject-to-subject variability (SD) in syntactic complexity (as measured by clauses per utterance) that is not accounted for with the covariates in our model. The model for RWU is similar and is specified below. The only difference is the inclusion of the addition term ϵ_{ij}, which accounts for overdispersion (i.e., unexpectedly large variation for Poisson-distributed variables) not detected in any of the other models.

$$\begin{gathered}\left(W_{ij} - U_{ij}\right) \sim Poisson\left(U_{ij}\left(RWU_{ij} - 1\right)\right) \\ \log\left(RWU_{ij} - 1\right) = \beta_{0j}^{(W)} + \left(\beta_{age}^{(W)} \times age_i\right) + \left(\beta_{sex}^{(W)} \times sex_i\right) \\ + \left(\beta_{ses}^{(W)} \times ses_i\right) + \left(\beta_{tbi}^{(W)} \times tbi_i\right) + \alpha_i^{(W)} + \epsilon_{ij}^{(W)} \\ \alpha_i^{(W)} \sim N\left(0, \tau^{(W)}\right);\ \epsilon_{ij}^{(W)} \sim N\left(0, \sigma^{(W)}\right),\end{gathered} \tag{3}$$

where $\sigma^{(W)}$ measures observation-level variability (SD) in the total word count. Interpretation of the coefficients in both of these models is analogous to those for our productivity model.

Lexical Diversity (Analyzing Proportions)

The model for PDW is

$$\begin{gathered}D_{ij} \sim Binomial(W_{ij}, PDW_{ij}) \\ \text{logit}\left(PDW_{ij}\right) = \beta_{0j}^{(D)} + \left(\beta_{age}^{(D)} \times age_i\right) + \left(\beta_{sex}^{(D)} \times sex_i\right) + \left(\beta_{ses}^{(D)} \times ses_i\right) + \\ \left(\beta_{tbi}^{(D)} \times tbi_i\right) + \left(\beta_W^{(D)} \times \log\left(\frac{W_{ij}}{100}\right)\right) + \alpha_i^{(D)} \\ \alpha_i^{(D)} \sim N(0, \tau^{(D)}),\end{gathered} \tag{4}$$

where $\tau^{(D)}$ measures subject-to-subject variability (again measured in SD) not accounted for by the demographic, lecture type, and discourse length variables. Note that, as in the Poisson regression examples, the random effects $\alpha_i^{(D)}$ are account for the fact that we are observing repeated measurements on each subject.

The model looks a lot like Equations **Error! Reference source not found.**) and **Error! Reference source not found.**), and the binomial regression coefficients are interpreted in an analogous manner, with a couple of distinctions. Exponentiated coefficients have a similar multiplicative interpretation, but the multiplicative effect is on the ratio of the probability of a distinct to a repeated word, instead of on the counts directly. Again, we opt for an additive interpretation here, but further details on the multiplicative interpretation can be found here (Agresti & Kateri, 2011).

The term $\beta_W^{(D)}$ is included to account for the effect of the total word count W_{ij} on the proportion of distinct words $PDW_{ij}^{(D)}$. To maintain interpretability of the intercept terms β_{0j}, we are careful to center the covariate $\log(W_{ij})$, similar to how we centered the demographic covariates. To accomplish this we use $\log\left(\frac{W_{ij}}{100}\right)$, so that the baseline is a discourse containing 100 total words – a relatively typical word count in our data set. Thus, $\exp(\beta_{0j}^{(D)})$ represents the overall mean ratio of distinct to repeated words produced by persons summarizing discourse j, given that there are 100 words in the discourse sample. We expect $\beta_W^{(D)}$ to be negative, since our data suggests that the proportion of distinct words $PDW_{ij}^{(D)}$ generally decreases with larger values of W_{ij}.

Estimating the regression coefficients in Equation (4) allows us to discover patterns relating to each of the demographic covariates, and to detect the effect of lecture type, thereby answering our questions of interest relating to lexical diversity. In general, binomial regression is useful for a variety of situations in which proportions are of key interest.

We evaluate the assumptions of our model via posterior predictive checks. In doing so, we find that observed traditional data summaries such as group-specific mean U, SI, MLU, and TTR in our dataset fall well within their prediction intervals, and these intervals are largely centered on the observed data values, with the exception of some of the very small counts, which are difficult to estimate in an unbiased way due to a lower boundary effect.

Prior Distributions

We would have preferred to elicit informative priors from subject matter experts in order to maximize our power to learn interesting parameter values. However, because this is a re-analysis, and our subject matter expert (JL) had already seen and analyzed the data before the present analysis, we resorted to diffuse minimally informative priors to minimize the influence of previous results on our current analyses. Our chosen priors are listed in Table 5, along with posterior mean and standard deviation. Posterior 95% credible intervals are shown in Table 6. The prior for the coefficient parameters are centered at zero, and all prior distributions are chosen to cover a substantially broader range than what our subject matter expert saw to be reasonable, in order to be on the conservative side. As a sensitivity analysis, we also tried using analogous priors whose 95% credible intervals had half the width of the more conservative ones, but the results were not at all different, indicating that our chosen priors were sufficiently diffuse.

Having little prior knowledge about the standard deviation of the random effects $\tau_U, \tau_C, \tau_W, \sigma_W,$ and τ_D for this application, we set each to follow a relatively diffuse half-Cauchy distribution with location 0 and scale $a = 2.5$. As a sensitivity analysis, we also tried using $a = 1$ and $a=10$, but the results were practically unchanged, so we kept the moderately diffuse value of $a = 2.5$.

Fitting the Model

To estimate posterior distributions of the parameters, we utilize a Markov chain Monte Carlo (MCMC) algorithm implemented via STAN software (Stan Development Team, 2018b) accessed via the RStudio interface (R Studio Team, 2020; Stan Development Team, 2018a) to obtain 30,000 random samples from the posterior distribution of the parameters, after removing 5,000 from each of three chains as burn-in. As per the default setting in STAN, initial values of the MCMC algorithm were selected uniformly at random in the interval (-2, 2) for parameters with an unrestricted domain; and for non-negative parameters, a uniform random number was chosen in the interval (-2, 2) and then exponentiated to obtain a random starting value.

To evaluate convergence, we examined trace plots and calculated the effective sample size and the Gelman Rubin potential scale reduction statistic for each posterior distribution in our model (Gelman & Rubin, 1992). The minimum effective sample size was 23,754 (less than our 30,000 total posterior samples), and the maximum potential scale reduction statistic was 1.015 (close to 1.0). All three metrics indicated that all parameters converged to the posterior distribution. We then used these samples to visualize and summarize the distribution of the parameters using R statistical software (R Core Team, 2019). In the main body of the paper we report all results as additive effects, as opposed to the multiplicative effects that are more natural in Poisson and logistic regression. We do this to ease the transition from the traditional approaches typically used in LSA to more advanced model-based approaches by keeping the interpretation on the same scale as the response variables of interest (as in ANOVA or SLR). In

Table 5 we report the prior distributions and posterior mean/standard deviation of each raw parameter from equations (1)-(4), for completeness. Because we utilize Bayesian methods, we are able to seamlessly transform between additive and multiplicative effects. We now briefly illustrate two examples of how this transformation is done:

Suppose that we have obtained an MCMC draw from the joint distribution of $\left(\beta_{CC}^{(U)}, \beta_{tbi}^{(U)}\right)$ from the RUD model given in Equation (1). The multiplicative effect on RUD for individuals of the highest SES (e.g. $SES_{ij} = 2.5$), compared to individuals of the lowest SES (e.g. $SES_{ij} = -2.5$), giving a discourse on *any* of the lectures is simply $\exp(5\beta_{ses}^{(U)})$, *where 5* is the difference between the SES score of the two groups. In other words, we would expect individuals of the high-SES group to exhibit an average RUD that is $\exp(5\beta_{ses}^{(U)})$ times that of the low-SES group, no matter the nature of the lecture prompting the discourse. The corresponding additive effect, specifically for a discourse given on the CC lecture, is $\left[\exp\left(\beta_{CC}^{(U)}\right) \times \left(\exp\left(2.5\beta_{ses}^{(U)}\right) - \exp\left(-2.5\beta_{ses}^{(U)}\right)\right)\right]$; and the overall additive effect is found by computing the arithmetic average of these effects across all three lecture types. The additive effect is obviously a bit more complicated to compute but is more easily interpretable for those who are more familiar with traditional approaches, as it directly describes changes on the original outcome variable scale.

As another example, suppose we have a draw from the distribution of $\left(\beta_{CC}^{(D)}, \beta_{tbi}^{(D)}\right)$ from the PDW model given in Equation (4). The multiplicative effect of TBI on the odds of a distinct word is simply $\exp(\beta_{tbi}^{(D)})$, while the additive effect on PDW for discourses given on the CC lecture is $\left[\text{logit}^{-1}\left(\beta_{CC}^{(D)} + \frac{\beta_{tbi}^{(D)}}{2}\right) - \text{logit}^{-1}\left(\beta_{CC}^{(D)} - \frac{\beta_{tbi}^{(D)}}{2}\right)\right]$. Again, the overall additive effect is the average of the effects across the CC, CE, and N lectures.

Reference

Gelman, A., & Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511. https://doi.org/10.1214/ss/1177011136

Tables

Table 5. Prior distributions and posterior mean/standard deviation for all parameters of interest. $\beta \sim N(\mu, \sigma)$ indicates that coefficient β has a normal prior distribution with mean μ and standard deviation σ, while $\tau \sim HC(0, \theta)$ indicates that τ has a half-Cauchy prior distribution with location 0 and scale θ. For each parameter, the notation "$\rightarrow (m, s)$" indicates that the posterior mean is m and the posterior standard deviation s.

Language construct	Prior Distribution → Posterior (Mean, Standard Deviation)
Productivity: RUD	
	$\beta_{tbi}^{(U)} \sim N(0, 0.45) \rightarrow (-0.520, 0.196)$
	$\beta_{sex}^{(U)} \sim N(0, 0.45) \rightarrow (0.182, 0.115)$
	$\beta_{ses}^{(U)} \sim N(0, 0.1) \rightarrow (0.056, 0.052)$
	$\beta_{age}^{(U)} \sim N(0, 0.1) \rightarrow (0.051, 0.033)$
	$\tau^{(U)} \sim HC(0, 2.5) \rightarrow (0.388, 0.055)$
	$\beta_{0j}^{(U)} \sim N(1.5, 1.2) \rightarrow CC: (2.060, 0.107), CE: (1.786, 0.109), N: (1.881, 0.108)$
Syntactic Complexity: RS	
	$\beta_{tbi}^{(C)} \sim N(0, 1.5) \rightarrow (-0.081, 0.212)$
	$\beta_{sex}^{(C)} \sim N(0, 1.5) \rightarrow (-0.179, 0.113)$
	$\beta_{ses}^{(C)} \sim N(0, 0.3) \rightarrow (0.166, 0.061)$
	$\beta_{age}^{(C)} \sim N(0, 0.3) \rightarrow (0.017, 0.032)$
	$\tau^{(C)} \sim HC(0, 2.5) \rightarrow (0.313, 0.053)$
	$\beta_{0j}^{(C)} \sim N(1.5, 1.2) \rightarrow CC: (-0.907, 0.119), CE: (-0.751, 0.121), N: (-0.479, 0.113)$
Syntactic Complexity: RWU	
	$\beta_{tbi}^{(W)} \sim N(0, 0.65) \rightarrow (-0.152, 0.108)$
	$\beta_{sex}^{(W)} \sim N(0, 0.725) \rightarrow (-0.079, 0.061)$
	$\beta_{ses}^{(W)} \sim N(0, 0.135) \rightarrow (0.078, 0.031)$
	$\beta_{age}^{(W)} \sim N(0, 0.135) \rightarrow (0.015, 0.018)$
	$\tau^{(W)} \sim HC(0, 2.5) \rightarrow (0.179, 0.030)$
	$\sigma^{(W)} \sim HC(0, 2.5) \rightarrow (0.201, 0.018)$
	$\beta_{0j}^{(W)} \sim N(1.5, 1.2) \rightarrow CC: (2.298, 0.059), CE: (2.475, 0.060), N: (2.352, 0.060)$
Lexical Diversity: PDW	
	$\beta_{tbi}^{(D)} \sim N(0, 2) \rightarrow (-0.180, 0.074)$
	$\beta_{sex}^{(D)} \sim N(0, 2.5) \rightarrow (0.126, 0.035)$
	$\beta_{ses}^{(D)} \sim N(0, 0.4) \rightarrow (0.056, 0.020)$
	$\beta_{age}^{(D)} \sim N(0, 0.5) \rightarrow (0.006, 0.010)$
	$\tau^{(D)} \sim HC(0, 2.5) \rightarrow (0.060, 0.027)$
	$\beta_{W}^{(D)} \sim N(-2, 1.3) \rightarrow (-0.689, 0.042)$
	$\beta_{0j}^{(D)} \sim N(1.5, 1.2) \rightarrow CC: (0.388, 0.042), CE: (0.289, 0.042), N: (0.318, 0.041)$

Table 6. Posterior credible intervals for all parameters of interest.

Language construct	Posterior 95% Credible Interval	
Productivity: RUD		
$\beta_{0CC}^{(U)}: (1.85, 2.27)$	$\beta_{tbi}^{(U)}: (-0.91, -0.14)$	$\beta_{ses}^{(U)}: (-0.05, 0.16)$
$\beta_{0CE}^{(U)}: (1.57, 2.00)$	$\beta_{sex}^{(U)}: (-0.04, 0.41)$	$\beta_{age}^{(U)}: (-0.01, 0.11)$
$\beta_{0N}^{(U)}: (1.67, 2.09)$	$\tau^{(U)}: (0.29, 0.51)$	
Syntactic Complexity: RS		
$\beta_{0CC}^{(C)}: (-1.14, -0.68)$	$\beta_{tbi}^{(C)}: (-0.50, 0.33)$	$\beta_{ses}^{(C)}: (0.05, 0.29)$
$\beta_{0CE}^{(C)}: (-0.99, -0.52)$	$\beta_{sex}^{(C)}: (-0.40, 0.04)$	$\beta_{age}^{(C)}: (-0.05, 0.08)$
$\beta_{0N}^{(C)}: (-0.71, -0.26)$	$\tau^{(C)}: (0.22, 0.43)$	
Syntactic Complexity: RWU		
$\beta_{0CC}^{(W)}: (2.18, 2.41)$	$\beta_{tbi}^{(W)}: (-0.37, 0.06)$	$\beta_{ses}^{(W)}: (0.02, 0.14)$
$\beta_{0CE}^{(W)}: (2.35, 2.59)$	$\beta_{sex}^{(W)}: (-0.20, 0.04)$	$\beta_{age}^{(W)}: (-0.02, 0.05)$
$\beta_{0N}^{(W)}: (2.23, 2.47)$	$\tau^{(W)}: (0.12, 0.24)$	$\sigma^{(W)}: (0.17, 0.24)$
Lexical Diversity: PDW		
$\beta_{0CC}^{(D)}: (0.31, 0.47)$	$\beta_{tbi}^{(D)}: (-0.32, -0.03)$	$\beta_{ses}^{(D)}: (0.02, 0.09)$
$\beta_{0CE}^{(D)}: (0.21, 0.37)$	$\beta_{sex}^{(D)}: (0.06, 0.20)$	$\beta_{age}^{(D)}: (-0.01, 0.03)$
$\beta_{0N}^{(D)}: (0.24, 0.40)$	$\tau^{(D)}: (0.01, 0.11)$	$\beta_{W}^{(D)}: (-0.77, -0.61)$