The Divergent Lexicon: Lexical Overlap Decreases With Age in a Large Corpus of Conversational Speech

Stephan C. Meylan (smeylan@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Susanne Gahl (gahl@berkeley.edu)
Department of Linguistics, University of California, Berkeley, CA 94720 USA

Abstract

Changes in language processing and production accompanying aging have most commonly been interpreted as evidence for age-related cognitive decline. A recent proposal (Ramscar et al., 2014) challenges that interpretation, asserting instead that such changes emerge as a consequence of—and in order to support—processes of lifelong learning like continued vocabulary growth. Under this account, the mechanisms of language processing and production do not deteriorate with age, but rather the computational complexity of the underlying information processing task increases as more data is observed over the lifespan. The current study examines whether spoken language displays properties consistent with the notion of lifelong learning by examining the relationship between age, within-speaker lexical diversity, and between-speaker lexical overlap in a conversational speech corpus, Switchboard I. We find older speakers exhibit more diverse lexicons, and that they share fewer words with interlocutors than younger speakers.

Keywords: lexicon; aging; corpus linguistics; speech

Introduction

Age heralds, at least on first glance, a decline in speakers' language abilities: older adults are slower to recognize words (Spieler & Balota, 2000), experience more tip-of-the-tongue states in which they cannot produce the correct word (Brown & Nix, 1996), have a higher rate of disfluencies in conversational speech than their younger counterparts (Horton et al., 2010), and perceive a decline in their own language abilities (Ryan et al., 1992). Many of these declines implicate changes in processing and production related specifically to words, necessitating a theory of how lexical processing and production may change as speakers age.

Established theories that account for slowing in lexical processing, as described in Thornton & Light (2006), include the *inhibition deficit hypothesis*, that older adults are less able to focus on relevant information (Hasher et al., 1997), and the *transmission density hypothesis*, that weakened connections in memory result in slower retrieval (Burke et al., 1991). However, an alternative overarching hypothesis is that observed decreases in performance in lexical processing and production are a natural consequence of the increasing difficulty of the information processing problem of recognizing and retrieving an ever-greater number of entities (spoken or written words, or recognizing objects themselves) as an individual ages (Ramscar et al., 2014). Decreases in observed performance may be mediated by an increased number of competitors or an increase in the size of the search space; under this view, the decrease in observed performance is not interpreted as an age-related pathology, but rather a trade-off that allows older adults to effectively deal with a computational challenge of increasing complexity. Consistent with a meta-analysis suggesting older adult have larger vocabularies than younger ones (Verhaeghen, 2003), Ramscar et al. demonstrate that tests of vocabulary in older adults may underrepresent the size of speakers' lexicons because they reach ceiling levels of performance at relatively early ages. Such tasks, they argue, have concealed lifelong increases in vocabulary size; consequently, while researchers have emphasized the slowing of language processing and production, they have overlooked the all-important caveat that older adults demonstrate a mastery over a considerably larger amount of linguistic information.

The reading simulations in Ramscar et al. (2014) yield two predictions regarding conversational speech. First, if older adults have larger productive vocabularies as a consequence of lifelong vocabulary growth, we should expect their speech to be more lexically diverse than that of younger speakers. Second, as speakers' vocabularies grow upon exposure to highly diverse inputs, fewer vocabulary items should be shared shared across speakers. While both predictions seem intuitive, a small speech sample containing an extremely limited subset of a speaker's vocabulary may not reveal an appreciable distinction in word choice. Likewise it remains an open question as to whether age-related divergence in lexicons, if empirically observable at all, manifests in brief samples of conversational speech. Nonetheless, precisely these brief conversational interactions make up a substantial part of language use. The properties of these speech samples may diverge significantly from the texts used in the reading simulations of Ramscar et al.; the sampling process in the model may also differ in important ways from real speakers. The current paper thus explores how within-speaker lexical diversity (how many different words are used by a speaker) and between-speaker lexical overlap (how many words are used by both speakers) varies in conversational speech as a function of age when gender, level of education, dialect, and topic of conversation are controlled.

On a most basic level, lexical diversity can be thought of as the number of distinct word types present in a speech sample. However, any index of diversity must be robust to sample size confounds because a larger speech sample (containing more tokens, or discrete, realized instances of types) is likely to contain more types than a smaller one. Measures of lexical diversity seeking to overcome this problem have been developed for a variety of research areas, including first language acquisition (Durán et al., 2004), speech pathology (Watkins et al., 1995), and language teaching (Malvern & Richards, 2002). As for measures capturing the effects of aging on lexical diversity in neurotypical adults, Horton et al. (2010) found that the Uber index, a measure of lexical diversity generally robust to sample size variation, increased with speaker age in the Switchboard corpus. As a supporting analysis for their examination of age-related changes in conversational speaking rate, their published analysis did not control for contributions of other demographic characteristics to lexical diversity.

A theoretically related, but potentially independent, measurement of the lexical properties of a conversational speech sample is the degree to which speakers use the same lexical types as one another in a conversation. While seemingly intuitive that the number of shared types would decrease as within-speaker lexical diversity increases (Table 1, case 1 → 2), it is not a forgone conclusion. Speakers could, alternatively, draw from similar sets of additional types as their lexical diversity increases (Table 1, case 1 → 3). The current work investigates how lexical overlap changes as a function of lexical diversity. In addition to the problem of sample size already encountered in assessing lexical diversity, measuring the similarity of an individual subject's word choice to that of an interlocutor depends crucially on properties of both the speaker and the interlocutor's speech. For this reason, properties of conversational dyads, such as the ages or levels of education of *both* speakers, are investigated as predictors of the proportion of shared lexical types in conversations.

The principle objectives of the current paper are thus threefold: first, to replicate the finding of an age-related increase in lexical diversity found by Horton et al. (2010) after better controlling for other demographic factors that might influence lexical diversity; second to examine how the pattern of lexical diversity relates to the proportion shared lexical types between speakers; third, to assess whether predictions derived from reading simulations in Ramscar et al. (2014) are supported by the observed properties of conversational speech.

Data

The Switchboard I Corpus (Godfrey et al., 1992) contains the transcribed contents of 2,866 telephone conversations between 543 speakers, aged 17 to 68, collected in the late 1980's. Participants were randomly assigned conversational partners on the basis of shared interest in any of 70 speech topics. Conversations averaged 6 minutes, though many continued for longer periods. Participants were free to leave the assigned topic. Many speakers participated in several conversations with various interlocutors in the corpus. Conversations were transcribed into a standardized format by court stenographers.

Methods

The publicly-available aligned Switchboard I corpus was downloaded from `http://www.isip.piconepress.com/projects/switchboard/`. All word-level annotations were extracted to a single table and associated with corresponding speaker-level and conversation-level metadata. Ages were calculated from birth years and the reference year of collection, 1988. Approximately one million tokens containing bracketed markup (including 917,000 [silence] tokens) were excluded from further analyses. All token strings were converted to lowercase.

All function words, including determiners, quantifiers, pronouns, conjunctions, interjections, and auxiliary verbs, as well as all contractions, salutations, and discourse particles (affirmatives, negatives, and non-lexical particles like "um-hum") were excluded using a wordlist. This procedure yielded 1.17 million tokens across 4,862 speaker-conversation pairs, the distribution of which is shown in Figure 1.

The Uber index of lexical diversity (Dugast, 1980; see also Jarvis, 2002) was calculated for each speaker in each conversation:

$$U(Tokens, Types) = \frac{log(Tokens)^2}{log(Tokens) - log(Types)} \qquad (1)$$

Tweedie & Baayen (1998) warn of residual sample size effects in the Uber index, thus MTLD (McCarthy & Jarvis, 2010) and Yule's I (the inverse of Yule's K, Yule 1944) were also calculated. Given the relatively equal sample sizes from speakers in Switchboard, and a high correlation between these metrics—Pearson's $r = .81$ and $r = .79$ with the Uber Index, respectively—the current work follows Horton et al. (2010) in reporting the Uber index.

Following the calculation of lexical diversity, a subsampling procedure was used to ensure matched sample sizes in

	Speaker 1	Speaker 2	Lex. Div 1	Lex. Div 2	Jaccard Index
Case 1	AA BBB CC D E	A BBB CC FF G	.56 (5/9)	.56 (5/9)	.43 (3/7)
Case 2	AA BB C D E H I	A BB C FF G J K	.78(7/9)	.78 (7/9)	.27 (3/11)
Case 3	AA BB C D E H J	A BB C FF E H K	.78 (7/9)	.78 (7/9)	.55 (5/9)

Table 1: This toy example demonstrates how an increase in lexical diversity (Uber Index) may co-occur with either a decrease in lexical overlap between speakers (Case 1→ Case 2) or an increase in lexical overlap (Case 1 → Case 3). Individual letters represent tokens; types are grouped by color. The Jaccard index, a measure of similarity between speakers, is calculated as the number of types in the intersection between speakers divided by the number of types in the union.

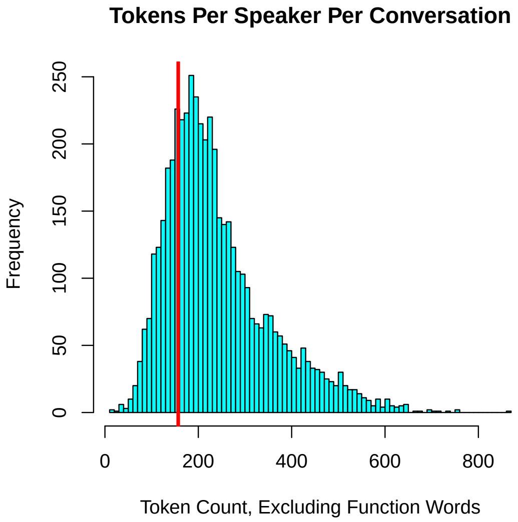

Figure 1: Distribution over the number of tokens per speaker per conversation in Switchboard I . Speakers to the left of the red line are excluded from the current analysis because of the subsampling procedure.

the comparison of speakers' type inventories. To permit comparison *across* conversations as well as within, all conversations above a fixed token count were repeatedly subsampled to a fixed token count before calculating overlap metrics, similar to the methodology described in Pine et al. (2013). A sample of 157 tokens was chosen in that it maximizes the total number of tokens analyzed (acceptable conversations × sample size). This yielded 1385 conversations out of the 2431 that survived function word exclusion (Figure 1). The distribution over ages of participants in the filtered set of conversations is shown in Figure 2. Subsampling to a fixed token count reflects a trade-off between completeness in sampling individual conversations and completeness in sampling the entire set of conversations: while a higher token threshold allows sampling more tokens from some of the individual conversations, fewer total conversations would have the requisite number of tokens for both speakers for inclusion in the analysis.

The Jaccard index, a set theoretic overlap measurement from early work in mathematical ecology (Jaccard, 1912), provides a conversation-level metric of the proportion of types appearing in the lexical inventories of *both* speakers out of the total number that appeared in *either*. If A and B are sets of lexical types, the Jaccard index is calculated as the cardinality of the set intersection divided by the cardinality of the set union:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

The Jaccard index is thus a symmetric measurement of overlap ranging from 0 (no shared lexical types) to 1 (only shared lexical types) between speakers: Speaker A cannot be closer to Speaker B than vice versa.

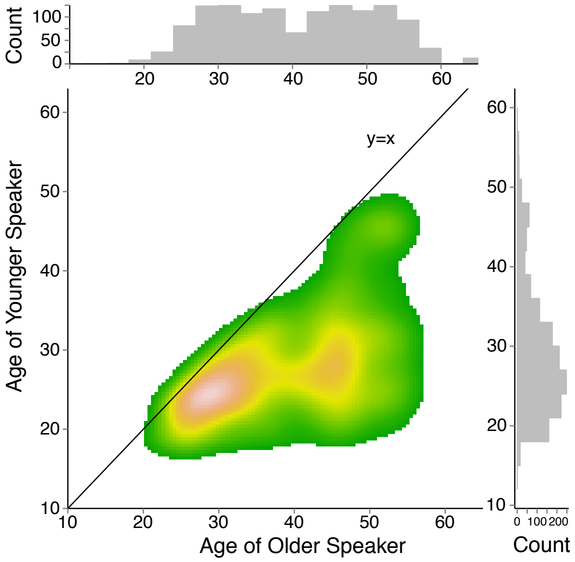

Figure 2: The smoothed joint distribution over ages of younger and older participants in conversations in Switchboard I meeting the sample size requirements for the calculation of lexical overlap. While most conversations occur between a younger speaker between 20 and 30 and an older speaker between 25 and 50, the corpus has appreciable coverage across speaker ages. Measurements are collapsed across the line of symmetry because a conversation between a 32 and a 48 year old speaker is identical to a conversation between a 48 and a 32 year old speaker.

For each conversation, we ran 100 independent simulations consisting of drawing 157 tokens with replacement from each speaker and calculating the Jaccard index. The reported Jaccard index was calculated by taking the mean value from these 100 runs.

The above data transformations and measurement procedures produce a dataset with measures of lexical diversity, the number of shared types in a size-controlled sample, and demographic properties for each speaker in each conversation. Given the dependence of observable behaviors in Switchboard on inherently dyadic communicative processes, we take special care to include demographic properties of the interlocutor along with those of speakers as predictors in our mixed effects models. By explicitly treating interlocutor properties as predictors, we can account for contributions of interlocutors to observed speaker behaviors.

Results

Lexical Diversity

Calculation of the Uber index for all speakers provides for a qualitative replication of the Horton, Shriberg, and Spieler (2010) finding of higher within-subject lexical diversity in the speech of older adults than that of younger adults (Figure 3, left). The proportion of lexical types used by a speaker that

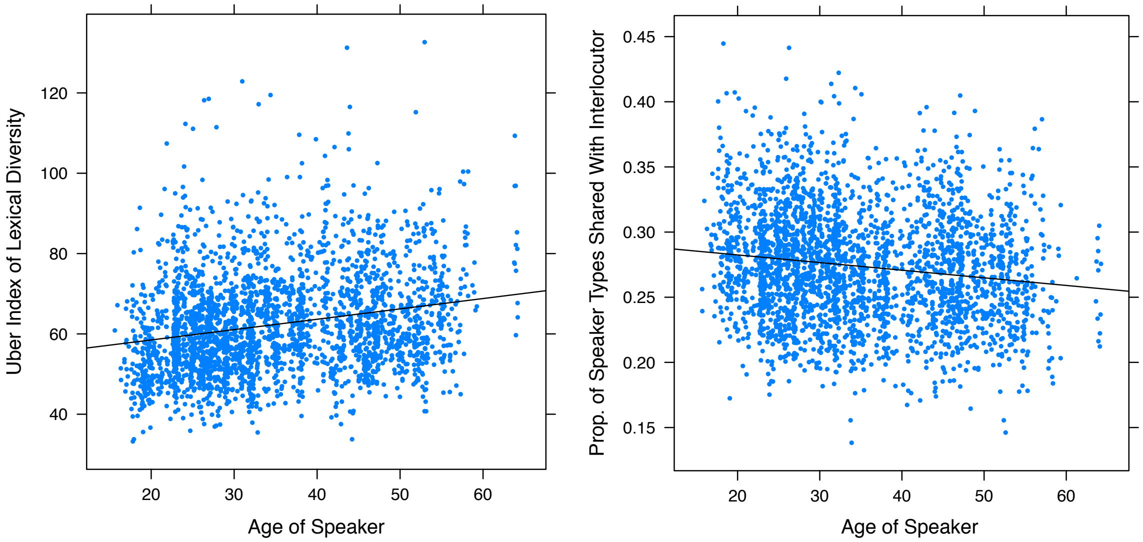

Figure 3: Left: the Uber index of lexical diversity as a function of speaker age. This measure of lexical diversity reveals that older adults use a more varied vocabulary than younger adults. Right: Fewer words used by older speakers are used by their interlocutors. The black line corresponds to the ordinary least squares regression using the single variable depicted. Age values are jittered in the range (+.5,-.5) to minimize overplotting.

are also used by his or her interlocutor, a possible but non-obligate correlate of the increase in lexical diversity (see Introduction), decreases with age (Figure 3, right).

We constructed a mixed effects linear regression model in which a speaker's lexical diversity, as measured by centered Uber index, was predicted from demographic properties of both the speaker and the interlocutor (age, gender, level of education, and dialect) and the interactions between these properties (Speaker Age × Interlocutor Age, Speaker Dialect × Speaker Dialect, etc.) as fixed effects. Conversational topic was treated as a random intercept. This initial model was pruned by comparing the Bayesian Information Criterion (BIC) of the full model against versions of the model with each predictor omitted in turn. Through this procedure, all interaction terms as well as Interlocutor Dialect and Interlocutor Level of Education were removed; the resulting model is displayed in Table 2. Speaker age has a small positive β coefficient, but is highly reliable. Men tend to exhibit more diverse vocabularies than women in these brief conversations. Lexical diversity increases with speaker level of education, and diversity varies as a function of speaker dialect, though both predictors have high standard error. Interestingly, lexical diversity exhibited by a speaker is dependent on some demographic properties of their interlocutor, possibly because speakers increase or decrease their lexical diversity to match interlocutors in a form of accommodation. Alternatively, interlocutors may directly influence the properties of discourse; for example, an interlocutor may govern the rate at which the conversational dyad moves into new material.

Proportion of Shared Types

While age is not a strong predictor of shared lexical types in the absence of additional controls (Figure 4, left), conversations between speakers with high lexical diversity result in a smaller proportion of shared lexical types (Figure 4, right). This decrease in overlap with an increase in lexical diversity suggests speakers do not draw from the same set of types when they exhibit more diverse vocabularies within conversations. Similarly-aged speaker are no more likely to share types (a similarity benefit would manifest as higher values on the diagonal of Figure 4, left).

A linear mixed effects regression model predicting the proportion of shared lexical items (Jaccard's index) per conversation was constructed with topic as a random intercept and cumulative age of the dyad, the difference in age of the speakers, male-male, female-female, or mixed speakers, same vs. different dialect speakers, cumulative education, and difference in education, as fixed effects. Conversations with one or more speakers with Unknown education levels (*n*=40) were excluded from the analysis, while the remaining levels were treated as a scalar in the range 0-3. Stepwise model pruning on the basis of BIC supported the exclusion of difference terms and dialectal properties of speakers from the final model (Table 3). Older dyads exhibited marginally higher proportions of shared types than younger ones. Male and female dyads exhibit fewer shared types than female-female dyads; male-male dyads exhibit even lower type overlap. Higher levels of education were predictive of a lower Jaccard index.

To further leverage information from the bootstrap, we also constructed a set of 1,000 linear mixed effects models, each predicting one set of sampled proportion of shared types. The

		Coef β	SE(β)	Approx. *df*	t	$Pr(>\|t\|)$
	Intercept	−15.32	3.15	2765.75	−4.86	<**.0001**
Speaker	Age	0.27	0.02	2741.54	12.79	<**.0001**
	Gender: Male	3.47	0.47	2764.69	7.32	<**.0001**
	Education: Less than College	0.04	2.97	2743.82	0.01	>.9
	Education: College	3.51	2.84	2748.28	1.24	>0.2
	Education: Some College	4.53	2.85	2749.22	1.59	>0.1
	Speaker Education: Unknown	7.88	3.38	2746.50	2.33	<**.05**
	Dialect: New England	−2.60	1.33	2717.47	−1.96	<**.05**
	Dialect: North Midland	−2.96	1.04	2737.57	−2.83	<**.01**
	Dialect: Northern	−1.00	1.10	2732.89	−0.90	>0.4
	Dialect: NYC	−3.13	1.23	2733.32	−2.55	<**.05**
	Dialect: South Midland	−0.59	0.96	2736.46	−0.61	>0.5
	Dialect: Southern	−2.60	1.12	2732.45	−2.32	<**.05**
	Dialect: Western	−2.70	1.07	2750.17	−2.51	<**.05**
Interlocutor	Gender: Male	1.70	0.46	2764.48	3.73	<**.0001**
	Age	0.05	0.02	2745.18	2.34	<**.05**

Table 2: Fixed effects from a linear mixed effects regression model for speaker lexical diversity in which topic was treated as a random effect. Degrees of freedom are calculated according to Satterthwaite's approximation (Satterthwaite, 1946).

full model was fit for each sample. 99% CIs from this analysis are presented in Table 3. Results of this analysis were consistent with the above mixed effects model.

Limitations And Future Work

The lexical inventory of a speaker in a brief telephone conversation is necessarily modulated by the particular communicative needs of that conversation. While the current work suggests that properties of speaker identity have a detectable effect on lexical diversity *despite* the brevity of conversations and discourse-specific effects, corpora with longer interactions between speakers and more conversations per speaker could allow for better decoupling of speaker-specific effects from discourse-specific effects.

The current work excluded function words from the analysis because of extremely high levels of overlap between subjects. "A," "I," and, "the," for example, were used by virtually every speaker in the sample. However, the blanket exclusion of function word, including relatively low frequency function words like "although" and "moreover," removes a potentially interesting source of lexical variability between speakers and age groups. Given previous work on gendered differences in the use of function words (Newman et al., 2008), we might expect even greater gender-based effects in lexical overlap than those observed here.

Another shortcoming of the current method is that it treats each conversation as a single temporal point, and neglects variability within the timecourse of the conversation. No strong conclusions may be drawn regarding processes of lexical accommodation, wherein speakers display increasing or decreasing levels of similarity in lexical choice over the course of the conversation. Comparison of size-matched temporal subsets would further reduce the number of analyzable tokens; as such new metrics for calculating between speaker lexical overlap may be required to elucidate within-conversation dynamics.

Conclusion

The current work leaves us with a consistent picture of lexical diversity and overlap in conversational speech. A speaker's lexical diversity is conditioned on the properties of his or her interlocutor, but age and higher levels of education predict increased lexical diversity for individual speakers. Within-speaker type inventories that are more diverse result in fewer

	Coef β	SE(β)	Approx. *df*	t	$Pr(>\|t\|)$	Bootstrapped 99% CI
Intercept	0.1654	0.00172	131.4648	96.37	<**.0001**	0.1616 – 0.1831
Cumulative Age	−0.0002	0.00004	1310.0294	−5.8097	<**.0001**	−0.0003 – −0.0002
Female - Male Dyad	−0.0087	0.00166	1321.4170	−5.2408	<**.0001**	−0.0118 – −0.0042
Male - Male Dyad	−0.0120	0.00197	1331.5834	−6.0959	<**.0001**	−0.0178 – −0.0089
Cumulative Education	−0.0026	0.00084	1306.4791	−3.1574	<**.01**	−0.0038 – −0.0002

Table 3: Fixed effects from a linear mixed effects regression model for lexical overlap in conversations in which topic was treated as a random effect. Degrees of freedom are calculated according to Satterthwaite's approximation (Satterthwaite, 1946).

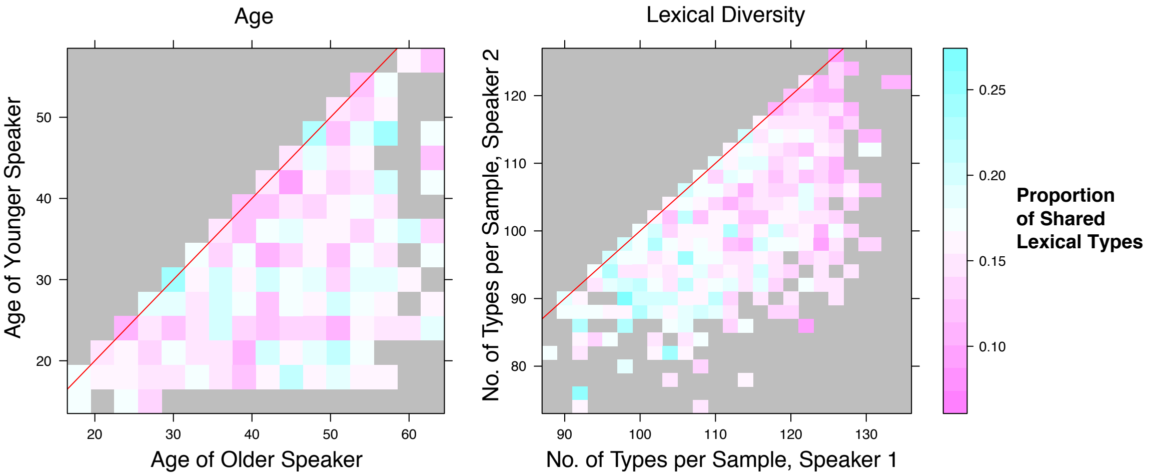

Figure 4: Left: proportion of lexical types shared between speakers in a conversation as a function of the average number of types in samples of a fixed token size. Conversations between speakers with more diverse (within-speaker) lexical inventories exhibit lower proportion of shared lexical types within a conversation. Right: the same metric of shared lexical types in a conversation as a function of age of speakers. Each pixel represents a mean of observed values in that domain.

shared lexical items in conversations. Consistent with predictions derived from Ramscar et al. (2014) regarding lifelong learning, older speakers use more word types than younger speakers, and their particular selection of words is more likely to diverge with other older speakers. That such patterns are identifiable even in brief samples of conversational speech suggests that lifelong changes in language production may be implicated even in short episodes of everyday language use.

References

Brown, A. S., & Nix, L. A. (1996). Age-related changes in the tip-of-the-tongue experience. *Am J Psychol*, *109*, 79–91.

Burke, D., MacKay, D., Worthley, J., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, *30*, 542-579.

Dugast, D. (1980). *La statistique lexicale*. Slatkine.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*, 220-242.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92* (Vol. 1, p. 517-520).

Hasher, L., Quig, M., & May, C. (1997). Inhibitory control over no-longer-relevant information: Adult age differences. *Memory and Cognition*, *25*, 286-295.

Horton, W. S., Spieler, D. H., & Shriberg, E. (2010). A corpus analysis of patterns of age-related change in conversational speech. *Psychol Aging*, *25*, 708–713.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*, 37-50.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*, 57-84.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*, 85-104. Retrieved from `http://ltj.sagepub.com/content/19/1/85.abstract`

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-d, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381-392.

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

Pine, J., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's Law and the case of the determiner. *Cognition*, *127*, 345-360.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Harald, B. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*, 5–42.

Ryan, E., Kwong See, S., Meneer, W., & Trovato, D. (1992). Age-based perceptions of language performance among young and older adults. *Communication Research*, *19*, 423-443.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

Spieler, D., & Balota, D. (2000). Factors influencing word naming in younger and older adults. *Psychol Aging*, *15*, 225–231.

Thornton, R., & Light, L. (2006). Language comprehension and production in normal aging. In J. Birren & K. Schaie (Eds.), *Handbook of the Psychology of Aging* (6th ed., p. 261-287). Elsevier.

Tweedie, F., & Baayen, R. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, *32*, 323-352.

Verhaeghen, P. (2003). Aging and vocabulary scores: a meta-analysis. *Psychol Aging*, *18*, 332–339.

Watkins, R., Kelly, D., Harbers, H., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Langauge, and Hearing Research*, *38*, 1349-1355.

Yule, G. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.