Language experience and prediction across age groups: evidence from diachronic fine-tuning of language models

Alton Chao (achao28@ucmerced.edu)

Ellis Cain (ecain@ucmerced.edu)

Rachel Ryskin (rryskin@ucmerced.edu)
Department of Cognitive & Information Sciences
University of California, Merced

Abstract

Humans predict upcoming language input from context, which depends on prior language experience. This suggests that older adults' predictions may differ from those of young adults, due to longer language exposure. Here we use sentence completion data from two age cohorts (YA = 18-35 y.o.; OA = 50-80 y.o.) and language models fine-tuned to particular decades of a diachronic corpus of American English to examine the relationship between changes in language statistics and differences in linguistic prediction across different age groups. We observed greater consistency in contextual probabilities within age groups compared to across age groups, indicating that YA and OA make subtly different predictions given identical context. Next-word prediction performance for the fine-tuned models decreased as the temporal distance between the fine-tuning and testing decade increased, indicating that language usage statistics changed over the span of a few decades. Further, GPT-2 surprisal values are more predictive of YA than OA contextual probabilities, suggesting that the language statistics, as captured by a model trained largely on internet text, aligns more with YA's internal model than OA's. However, both age groups' data are better fit by models fine-tuned on more recent corpus decades.

Keywords: language production; prediction; aging; language change; language models

Introduction

Prediction plays a pivotal role in language comprehension both in humans and in neural network language models (Dell & Chang, 2014; Elman, 1990; Federmeier, 2007; Kuperberg & Jaeger, 2016; Levy, 2008; Ryskin & Nieuwland, 2023). For example, the N400 event-related potential (ERP) component — a negative-going deflection of the ERP that peaks around 400ms after the onset of a meaningful stimulus in centro-posterior channels — is more negative in response to a word that is less predictable from the preceding sentence context compared to one that is more predictable (e.g., Kutas & Federmeier, 2000). As with a language model which generates predictions based on the statistics of the language encoded in its training data, humans base their predictions both on the local context (e.g., preceding words in the sentence) and their knowledge of the language (and world) accumulated over their lifespan. A straightforward implication of this is that, *if the language experience of two individuals encodes different statistics, those individuals will make different predictions, given the same context*. On an alternative view, where prediction is not a ubiquitous computation reflecting a person's cumulative language experience, two individuals may differ in their predictions due to other idiosyncrasies but these need not be directly tied to the statistics of their language experience.

Previous studies have shown that comprehenders adapt their expectations to the statistics of their local environment (e.g., Fine, Jaeger, Farmer, & Qian, 2013; Ryskin, Qi, Duff, & Brown-Schmidt, 2017), but these studies, where the statistics of the language are manipulated on the timescale of a few minutes, cannot speak to the role of differing statistics on the timescale of the lifespan (see Ryskin & Fang, 2021). Other studies have shown adaptation on a longer timescale (e.g., Rodd et al., 2016; Verhagen, Mos, Backus, & Schilperoord, 2018) but they involve specific domain-expertise (e.g., being a rower) changing expectations regarding a circumscribed set of words and phrases rather than the statistics of the language as a whole.

In the current work, we use the lens of aging to test the relationship between lifelong language experience and prediction. Older adults certainly have more language experience than younger adults but, crucially, the statistics that they extract from this experience may differ in multiple ways. The statistics of the language may themselves be changing over the course of an individual's lifetime (e.g., Michel et al., 2011). Language usage statistics are constantly changing over time, as the complex interactions between learning and usage in a population lead to slight shifts that accumulate over generations (Beckner et al., 2009; Chater & Christiansen, 2010; Davies, 2012; Kirby, 2017; Michel et al., 2011).For example, some terms that were popular during early language acquisition for an older adult might have fallen out of 'vogue' by the time the younger adults are acquiring language. If older adults are uniformly aggregating over all of their language experience, they would arrive at distinct distributions than younger adults who have experienced only the most recent statistics.

Previous work using diachronic word embeddings trained on decade-subsets of the Google books corpus (Hamilton, Leskovec, & Jurafsky, 2018) indicates that, at the level of the language as a whole, word meanings (defined in terms of co-occurrences with other words) shift even over the span of a few decades (Cain & Ryskin, 2023). Similarly, Qiu and Xu (2022) compared a pre-trained BERT model to HistBERT, which was additionally trained on a diachronic genre-balanced corpus (COHA) and demonstrated that the additional training was able to improve the representations for

words that had undergone meaning changes. It is heavily dependent on the quality of the diachronic corpus used for the additional training (i.e., whether the corpus has enough relevant instances of the usage patterns from that time period), and Qiu and Xu (2022) point out that excessive training on previous usage patterns could hamper model performance when generalizing to novel instances.

Moreover, the sources of language experience may differ between younger and older adults. For example, older adults may read more books from 'classic' authors (those who have been widely read for several decades) than younger adults, who are more likely to read works from recent authors (Grolig, Tiffin-Richards, & Schroeder, 2020). Some studies have additionally shown that older adults are more likely to engage in reading informational texts, such as newspapers, periodicals, or newspapers (Rentfrow, Goldberg, & Zilca, 2011; Smith, 1993, 1996). As indicated by a few recent studies, young adults might be getting more language input from internet text as they shift away from traditional print books, though there are recent 'revivals' of book-reading among the youth (Chaudhry & Low, 2009; Loh & Sun, 2019). Different genres are also known to differ in their language statistics (Biber & Finegan, 1989).

In sum, the statistics of the language plausibly differ systematically between younger and older adults (due to language change or differences in genre distribution in the input) and this may be reflected in systematically different predictions given the same context between the two age groups. Indeed, some studies suggest that behavioral and neural patterns reflecting prediction from context differ between younger and older adults (e.g., Federmeier, Kutas, & Schul, 2010; Payne & Silcox, 2019). Whether these differences are primarily the result of systematic differences in language experience as opposed to other factors (e.g., cognitive decline) is an open question.

Present research

In the present study, we use sentence completion ("cloze") task data from two age cohorts, younger adults and older adults, and compare the distributions of their next-word completions (i.e., contextual probabilities). Younger and older adults appear to generate subtly different contextual probability distributions over the same set of context. We then fine-tune a GPT-2 model on the decade-level subsets of the Corpus of Historical American English (COHA, Davies, 2012) and find that next-word prediction performance decreases as a function of increasing temporal distance between the decade of fine-tuning and the decade of testing.

Finally, we use surprisal values (the negative log probability of a word, given its preceding context) from GPT-2 as well as the fine-tuned models to predict the contextual probabilities computed for each age group separately. Contextual probabilities from younger adults are better predicted by surprisal values from all models relative to contextual probabilities from older adults. Contextual probabilities from both age groups are better predicted by models trained on more recent corpus data. Whether changing language statistics directly explain differences in prediction across age groups remains equivocal.

Methods

Participants

The experimental data set consists of 336 self-reported native English speakers, including 166 young (age range: 18–35 years, $M = 24.7$ y.o., $SD = 2.17$) and 170 older adults (age range: 50–80 years, $M = 64.1$ y.o., $SD = 5.34$). Each subject participated in a sentence completion task through Amazon's Mechanical Turk. The data only include responses from participants who passed a bot check (identifying images containing a plant) and provided correct completions on 6 catch trials (e.g., The opposite of big is ...) included at the beginning and end of the experiment.

Sentence completion data

The cloze task consisted of 300 sentence preambles (e.g., 'A dog has a good sense of ...') from Wlotko and Federmeier (2012). The preambles had previously been normed to include a variety of levels of constraint. For high constraint preambles (e.g., 'Nora couldn't take the message because she didn't have a pencil or a piece of ... '), most participants respond with the same word (low entropy). For low constraint preambles (e.g., 'One thing the old man lacked was ...'), the entropy of the next-word completion distribution is higher. Each participant provided a single-word completion for 150 sentence preambles, for a total of approximately 50,200 completions.

Corpus data

A subset of the cleaned version of COHA corpus (Davies, 2012), the Cleaned Corpus of Historical American English (CCOHA)[1] (Alatrash, Schlechtweg, Kuhn, & Schulte im Walde, 2020) was used for fine-tuning GPT-models (Radford et al., 2019). Specifically, all categories (fiction, non-fiction, magazines & news) from the 20th and early 21st century (1900–2010) were included. Texts were grouped by decade (i.e. 1900s decade corresponds to all texts from 1900 to 1910). Each decade subset was shuffled and then randomly split for 6-fold cross validation. Each training dataset consisted of approximately 2.5 billion tokens, while the validation sets contained 500 million tokens.

Language model fine-tuning

6-fold cross validation was used to fine-tune the language models, following similar studies (Xu & Kemp, 2015; Hamilton, Leskovec, & Jurafsky, 2016). Therefore, each decade of COHA was aggregated, randomized, then split into six folds. One fold was left out for validation, while the other folds were used to fine-tune the model. 20 percent of each fold was reserved for testing.

[1] We will refer to the corpus as "COHA."

The default GPT-2 model and tokenizer are loaded and fine-tuned on the data using HuggingFace's Transformers library (Wolf et al., 2020). Each training input was padded, with a maximum length of 512 tokens.

Each of the 66 resulting (6 folds per decade, 11 decades from 1900 through 2010) fine-tuned models and tokenizers were evaluated by computing perplexity (a measure of how well a language model predicts the next word) of the testing data for all decades. As a baseline, the same process was performed with the default GPT-2.

Results

Contextual probabilities across age cohorts

To examine completion preferences across age cohorts, we calculated contextual probabilities, which represent the proportion of people who choose a certain word given a sentence stem, separately for each age cohort (as seen in Fig. 1). These probabilities are calculated by aggregating the data by age group and sentence stem and dividing the number of occurrences of a certain word by the total number of responses per sentence stem. Overall, the contextual probabilities were highly correlated between younger (YA) and older adults (OA) ($r = 0.934$, $p < 0.0001$). Responses from OA exhibited slightly lower average entropy ($M = 2.26$, $SD = 1.48$) compared to those from YA ($M = 2.53$, $SD = 1.43$, $t = -2.335$, $p < 0.020$).

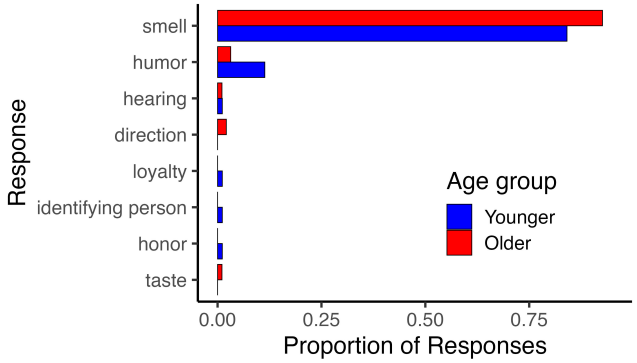

Figure 1: Example response distribution for sentence stem "A dog has a good sense of ..."

Next, we tested whether variation between YA and OA completions is greater than the variation that one might see within an age cohort (i.e., between any two subsets of YA or any two subsets of OA). The data from each age cohort was randomly divided in half, then aggregated by unique responses. Then, within and across each cohort (YA-YA, OA-OA, or YA-OA), the two halves were merged by sentence stem and response. The correlation was then calculated to determine how strongly correlated the contextual probabilities are across these independent halves of the data. For the calculations across young and older adults, the halves were randomly selected.

The contextual probabilities were robustly correlated across independent halves within an age group (Figure 2), with a Pearson's r of 0.966 and 0.968, respectively. The correlation across independent halves between age groups (samples of 50% of the young versus 50% of the older group), r decreased to 0.939.

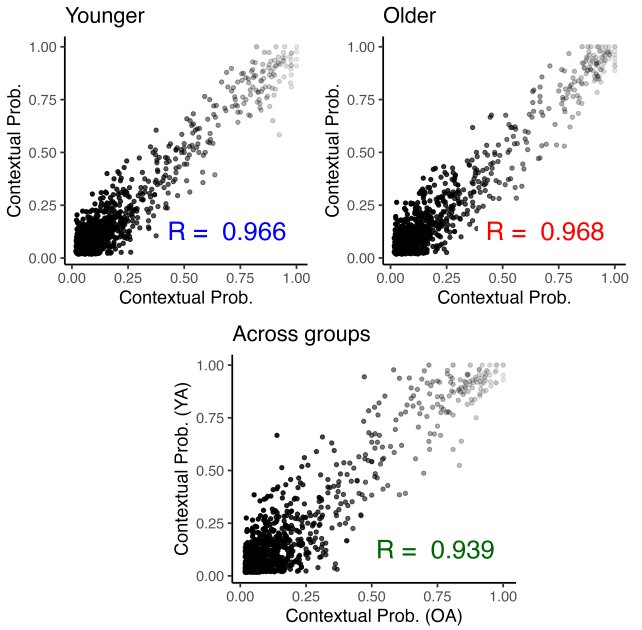

Figure 2: Correlations of contextual probabilities across two independent subsamples within and across age groups.

Further, we used a sampling approach to examine the probability that the "best" (highest probability) completion was the same both within and across age groups. The data from each age group were randomly divided into independent halves and contextual probabilities were calculated (by the same method as above) in each half. The proportion of stems with the same best response was calculated when comparing the two YA halves, the two OA halves, and the YA halves with the OA halves (randomly selecting one half from each age group). This procedure was repeated 1000 times in order to obtain distributions for the proportions of same best completions (Figure 3). Older adults' best completions matched an average of 86.05% of the time, compared to 81.36% for young adults, consistent with the observation of lower entropy in OA responses. The probability that the best completion was the same across two groups was lowest when comparing across age groups, 78.25%.

We additionally measure the idiosyncrasy of the responses in each age group by dividing the number of shared responses by the total number of unique responses for each sentence stem. This was calculated separately for YA and OA. The total number of unique responses per sentence stem was virtually the same across age groups, suggesting some consistency in the variety of responses. Thus, when comparing the two groups, a smaller ratio would indicate more idiosyncrasy

in responses (i.e., less shared responses for a given sentence stem). Consistent with the entropy measure, older adults demonstrated less idiosyncrasy in their responses. OA responses produced a higher average ratio ($M = 9.56$, $SD = 9.73$) compared to the ratio for the YA responses ($M = 7.24$, $SD = 6.95$, $t = 3.366$, $p < 0.0008$).

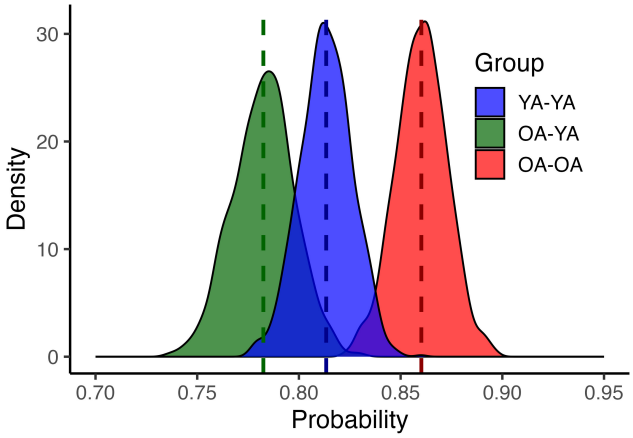

Figure 3: Distributions of the highest probability response matching across 1000 independent subsamples. Color indicates the comparison group, either within an age group (i.e., YA-YA) or across groups (OA-YA).

Fine-tuned language model performance across decades

Figure 4 summarizes the performance of each language model on left-out COHA documents from each decade. Each fine-tuned model achieved the lowest average perplexity against the decade it was trained on, indicating successful fine-tuning. On average, perplexity was 19.96 lower when the fine-tuning decade matched, compared to the model's performance on the other decades of corpus. As the temporal distance increased between the test and train decades, the difference in perplexity gradually increased. In general, these models were more attuned to the corpus text compared to the default GPT-2, which achieved significantly higher perplexity scores across decades, with the lowest perplexity being for the most recent test decade.

Language model predictions and contextual probabilities across age cohorts

Using these various fine-tuned models and the base GPT-2 model, we calculated surprisal scores for each sentence stem and the corresponding best completion. In these analyses we focus on the best completions because the probability estimates for some of the most infrequent responses are likely to be highly noisy (e.g., some responses may be produced by one person only). Because the sentence preambles varied in constraint, there is still substantial variability in the contextual probabilities of these best completions. Six models represent each decade due to the 6-fold cross-validation; therefore,

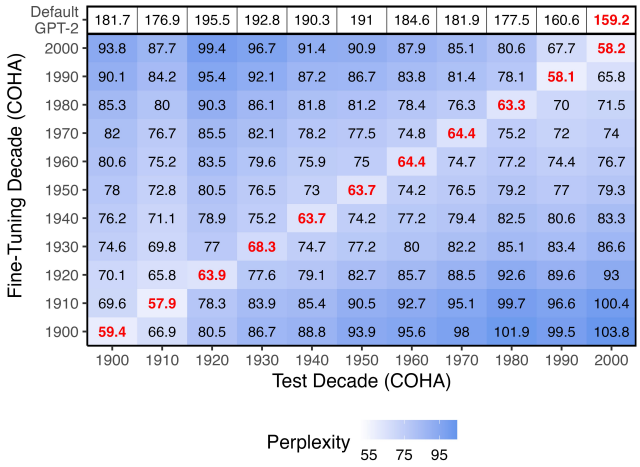

Figure 4: Average perplexity across all models, against all decades of corpus. Red text represents the decade with the lowest perplexity per model. The top row shows the default pre-trained GPT-2 model performance.

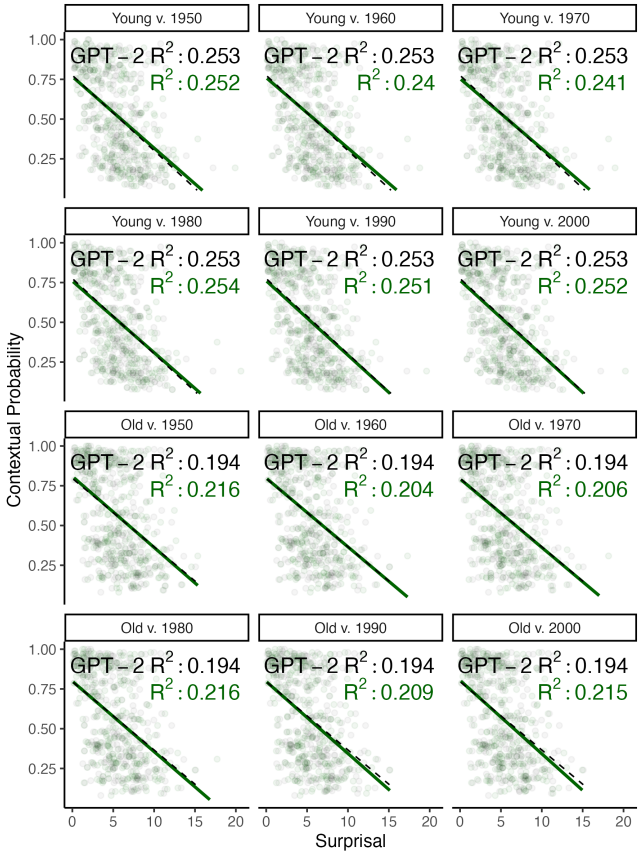

Figure 5: Contextual probability of highest probability responses plotted against surprisal. Black points represent scores calculated by the default GPT-2, and the green points represent those calculated by the fine-tuned models. The text labels represents the R^2 from a linear model predicting contextual cloze probability from model surprisal.

we averaged the surprisal of a given word in context across all same-decade models for analysis.

We compare these results with contextual probabilities calculated from the responses of the participants. Figure 5 summarizes the relationships between base GPT-2 and decade-level surprisal and the contextual probabilities from each age group.

Base GPT-2 surprisal values better predict YA contextual probabilities relative to OA contextual probabilities (R^2_{YA} = 0.253; R^2_{OA} = 0.194). Further, across all the fine-tuned models, surprisal values explain more variance in YA responses than OA responses. As the model's fine-tuning decade approaches the present (i.e., the model more closely reflects modern day language usage), participants' responses correlate more strongly with the surprisal scores (Figure 6) for both YA and OA. Additionally, for YA, the fine-tuned models from the most recent decades perform similarly to the default GPT-2. On the other hand for OA, the fine-tuned models from the most recent decades outperform the base GPT-2 model.

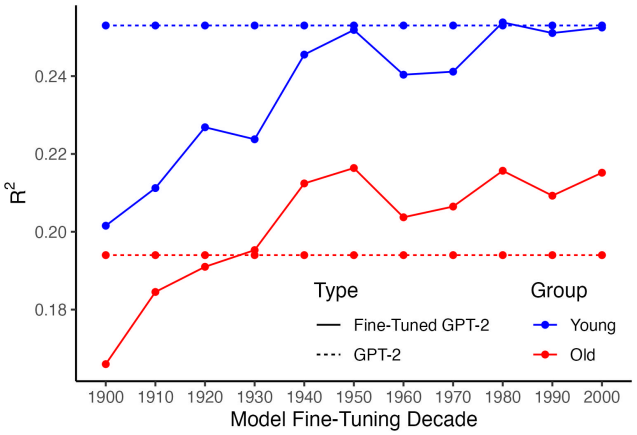

Figure 6: Temporal trends in the coefficient of determination between surprisal and contextual probability, plotted against decade of model fine-tuning. Each point represents the R^2 between the surprisal from a given model and the contextual probability from a given age group. Fine-tuned models increasingly perform similarly to the default GPT-2 when evaluated on young adults but outperform the default GPT-2 when evaluated on older adults.

In addition, we compare the distributions of the responses, using the contextual probability distributions in human and model evaluations. To achieve this, we normalize the model-calculated word-probability scores for a valid probability distribution. The Jensen-Shannon Divergence (JSD) was then calculated between the human and model distributions for each unique sentence stem, as shown in Figure 7. JSD is a measure (ranging from 0 to 1) of how similar two probability distributions are. JSDs closer to zero indicate higher similarity between two distributions.

The JSD between human and model response distributions gradually decreases from approximately 0.30 to 0.25 as the decade of fine-tuning gets more recent. The JSD for the base GPT-2 model was the lowest. The JSD distributions for YA and OA remained similar across models, with no statistically significant differences by age group.

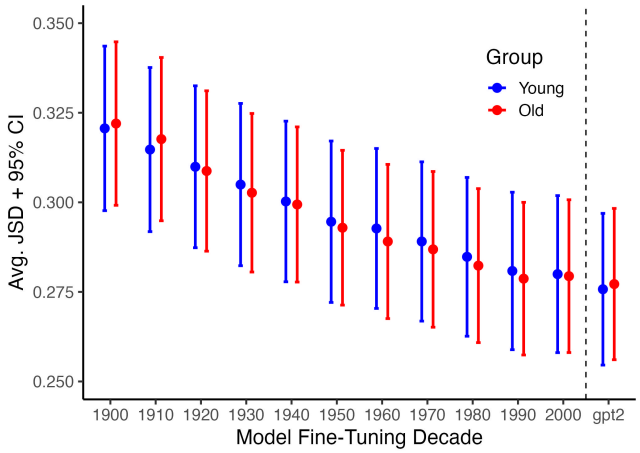

Figure 7: Alignment between human and model responses, as measured by Jensen-Shannon divergence. Color indicates age group. The default GPT-2 model alignment is shown on the far right.

Discussion

In a sentence completion ("cloze") study, we found broad agreement in next-word predictions given a sentence context between OA and YA as well as systematic differences. Distributions of contextual probabilities were more consistent within an age group than across age groups. YA tended to be more variable in their responses while OA were more consistent (within-group).

To investigate the extent of language change and how it affects next-word prediction, GPT-2 models were fine-tuned on texts from previous decades (1900-2000). As expected with successful fine-tuning, model performance was best when the decade of fine-tuning matched the decade of testing. Further, model performance gradually decreased as temporal distance between the decades of fine-tuning and testing increased, consistent with prior findings of meaningful changes in the statistics of the language over the span of decades (i.e., within the lifespan of an individual, Cain & Ryskin, 2023).

To test whether changes in language statistics over historical time explain differences in predictions between OA and YA, we used surprisal values from fine-tuned models and the base GPT-2 model to predict the contextual probabilities from each age group. YA contextual probabilities were better predicted by GPT-2 surprisal values than OA contextual probabilities, suggesting that the statistics of the language, as captured by a model trained largely on present-day internet text, are more in line with YA's internal model of the language than OA's. Both YA and OA data are better fit by models fine-tuned on more recent corpus decades, consistent with the

idea that, as the statistics of the input data become closer to the statistics experienced by the participants, the fit to the data improves. When predicting YA data, the best fine-tuned models approach GPT-2 performance, suggesting that fine-tuning on any data from earlier decades leads to the model's statistics being at best equally aligned, but mostly less aligned, with YA's internal model of the statistics of the language.

In contrast, models fine-tuned on more recent decades outperform the base GPT-2 model for predicting OA responses, indicating that fine-tuning on data from some earlier decades (1930s-1990s) leads to the model's statistics being better aligned with OA's internal model of the statistics of the language. However, the pattern of model fits across decades is strikingly similar for predicting both YA and OA data. It is not the case that decades that were experienced by OA but not YA (e.g., 1960s) are differentially predictive of OA but not YA data, suggesting that the performance boost in predicting OA data from fine-tuning on COHA may not be tied to OA specifically experiencing those time periods. It may be that limitations of the behavioral data or the models obscured such experience-driven patterns. The behavioral data were collected on a relatively small number of sentences (300). A larger dataset, with more variability in the contextual probabilities, may provide a better testbed. Similarly, the age groups are fairly close (minimum 15-year gap), and it may be that the experiences of participants near the age group cutoffs (e.g., 35 and 50) are similar enough to make differences too subtle to detect. The fine-tuning, though successful, may not sufficiently change the statistics of the input to override the disproportionate influence that comes from the GPT-2 training data. Additionally, the language found in COHA may differ from the kinds of sentences being used in the stimuli, leading to less accurate surprisal estimates. Further, surprisal values from fine-tuned language models reflect only a coarse approximation — based on text alone — of the language experienced by individuals. Differences between the language experience of younger and older adults may manifest primarily in oral communication or other forms of linguistic input not captured by CCOHA.

Alternatively, if the results are taken at face value, we can speculate that older adults may receive more language input from sources that are better represented in COHA, because of genre-balancing, than in the GPT-2 training data which consists of millions of web pages. In other words, differences between the predictions made by OA and YA may be better explained by differences in the genres of language input received by the two age groups rather than by different timespans of language input between the two age groups. We leave it to future work to test this speculation.

While some researchers have pointed out issues with using corpora to estimate 'everyday' language input, such as the inclusion of scientific texts biasing the lexicon (Pechenick, Danforth, & Dodds, 2015), it still serves as a practical way to estimate language exposure for the larger population, due to the relationship between print exposure and language input over the lifespan (Landauer, Kireyev, & Panaccione, 2011; Grolig et al., 2020; Mol & Bus, 2011; Payne, Gao, Noh, Anderson, & Stine-Morrow, 2012). The genre-balancing of COHA also likely minimizes the potential biases that might be introduced by over-representing a specific genre or type of text. Future studies could extend this analysis to other corpora to check whether the trends hold.

Finally, these data do not rule out the possibility that differences in prediction between OA and YA are not explained by any differences in language experience, but are caused by some other factor (e.g., cognitive decline). Since our sample was also cross-sectional, our analyses are not able to directly address this aspect of aging. One potential way to address this would be to utilize 'lifespan' corpora that have collect writings from across various authors' lifespan, such that the lifespan trends in usage patterns can be measured for the individual authors (i.e., Petré et al., 2019).

On a methodological note, this finding suggests that caution is warranted when researchers use large language models to estimate surprisal values (or other measures) for the purposes of relating them to human data. These models do not capture the predictions of all populations equally well, likely due to mismatches between the aggregate statistics present in the training data and some unique properties of the language experience of different populations. As a result, if estimates from these models are being used to draw conclusions about cognitive differences between populations (e.g., older adults don't predict as well as younger adults), they may be confounded with differences in accuracy of those estimates across populations.

Acknowledgments

This fine-tuning and evaluation of these models was conducted using the Pinnacles cluster (NSF MRI, #2019144) at University of California, Merced.

References

Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020, May). CCOHA: Clean corpus of historical American English. In N. Calzolari et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 6958–6966). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.859/

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., … Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*, 1–26. doi: 10.1111/j.1467-9922.2009.00533.x

Biber, D., & Finegan, E. (1989). Drift and the Evolution of English Style: A History of Three Genres. *Language*, *65*(3), 487–517. doi: 10.2307/415220

Cain, E., & Ryskin, R. (2023). Diachronic Language Change and Its Influence on Lexico-semantic Representations Across the Lifespan. In *Proceedings of the Cognitive Science Society.* Sydney, AUS.

Chater, N., & Christiansen, M. H. (2010). Language Acquisition Meets Language Evolution. *Cognitive Science*, *34*(7), 1131–1157. doi: 10.1111/j.1551-6709.2009.01049.x

Chaudhry, A. S., & Low, G. (2009). Reading preferences among different generations: A study of attitudes and choices in singapore. *Singapore Journal of Library & Information Management*, *38*(1), 27–48.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, *7*(2), 121–157.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394–20120394. doi: 10.1098/rstb.2012.0394

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*(2), 179–211. doi: $10.1207/s15516709cog1402_1$

Federmeier, K. D. (2007, July). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. doi: 10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161. doi: 10.1016/j.bandl.2010.07.006

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013, October). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS ONE*, *8*(10), e77661. doi: 10.1371/journal.pone.0077661

Grolig, L., Tiffin-Richards, S. P., & Schroeder, S. (2020). Print exposure across the reading life span. *Reading and Writing*, *33*(6), 1423–1441. doi: 10.1007/s11145-019-10014-3

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, August). Diachronic word embeddings reveal statistical laws of semantic change. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P16-1141/ doi: 10.18653/v1/P16-1141

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change* (No. arXiv:1605.09096). arXiv.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, *24*(1), 118–137. doi: 10.3758/s13423-016-1166-7

Kuperberg, G. R., & Jaeger, T. F. (2016, January). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. doi: 10.1080/23273798.2015.1102299

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. doi: 10.1016/S1364-6613(00)01560-6

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, *15*(1), 92–108. doi: 10.1080/10888438.2011.536130

Levy, R. (2008, March). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Loh, C. E., & Sun, B. (2019). "i'd still prefer to read the hard copy": Adolescents' print and digital reading habits. *Journal of Adolescent & Adult Literacy*, *62*(6), 663–672.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, ... Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. doi: 10.1126/science.1199644

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296. doi: 10.1037/a0021890

Payne, B. R., Gao, X., Noh, S. R., Anderson, C. J., & Stine-Morrow, E. A. L. (2012). The effects of print exposure on sentence processing and memory in older adults: Evidence for efficiency and reserve. *Aging, Neuropsychology, and Cognition*, *19*(1-2), 122–149. doi: 10.1080/13825585.2011.628376

Payne, B. R., & Silcox, J. W. (2019). Aging, context processing, and comprehension. In *Psychology of Learning and Motivation* (Vol. 71, pp. 215–264). Elsevier. doi: 10.1016/bs.plm.2019.07.001

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLOS ONE*, *10*(10), e0137041. doi: 10.1371/journal.pone.0137041

Petré, P., Anthonissen, L., Budts, S., Manjavacas, E., Silva, E.-L., Standing, W., & Strik, O. A. (2019). Early modern multiloquent authors (emma): Designing a large-scale corpus of individuals' languages. *ICAME journal: computers in English linguistics/International Computer Archive of Modern English; International Computer Archive of Modern and Medieval English.-[Bergen, Norway], 1987, currens*, *43*(1), 83–122.

Qiu, W., & Xu, Y. (2022). HistBERT: A Pre-trained Language Model for Diachronic Lexical Semantic Analysis. doi: 10.13140/RG.2.2.14905.44649

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. , 24.

Rentfrow, P. J., Goldberg, L. R., & Zilca, R. (2011). Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of personality*, *79*(2), 223–258.

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchin-

son, C., & Adler, A. (2016, April). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37. doi: 10.1016/j.jml.2015.10.006

Ryskin, R., & Fang, X. (2021, October). The many timescales of context in language processing. In *Psychology of Learning and Motivation.* Academic Press. doi: 10.1016/bs.plm.2021.08.001

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052. doi: 10.1016/j.tics.2023.08.003

Ryskin, R., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2017, May). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 781–794. doi: 10.1037/xlm0000341

Smith, M. C. (1993). The reading abilities and practices of older adults. *Educational Gerontology: An International Quarterly*, *19*(5), 417–432.

Smith, M. C. (1996). Differences in adults' reading practices and literacy proficiencies. *Reading Research Quarterly*, *31*(2), 196–219.

Verhagen, V., Mos, M., Backus, A., & Schilperoord, J. (2018, June). Predictive language processing revealing usage-based variation. *Language and Cognition*, *10*(2), 329–373. doi: 10.1017/langcog.2018.4

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *NeuroImage*, *62*(1), 356–366. doi: 10.1016/j.neuroimage.2012.04.054

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.6/ doi: 10.18653/v1/2020.emnlp-demos.6

Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. *Cognitive Science*. Retrieved from https://api.semanticscholar.org/CorpusID:4877161