Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption.

Maury Courtland[†], **Aida Davani**[‡], **Melissa Reyes**[*], **Leigh Yeh**[‡],
Jun Leung[*], **Brendan Kennedy**[‡], **Morteza Dehghani**[*‡], and **Jason Zevin**[*†]
[†] Department of Linguistics
[‡] Department of Computer Science
[*] Department of Psychology
University of Southern California
{landerpo, mostafaz, reyesmel, leighyeh,
junyenle, btkenned, mdehghan, zevin}@usc.edu

Abstract

Cognitive tests have traditionally resorted to standardizing testing materials in the name of equality and because of the onerous nature of creating test items. This approach ignores participants' diverse language experiences that potentially significantly affect testing outcomes. Here, we seek to explain our prior finding of significant performance differences on two cognitive tests (reading span and SPiN) between clusters of participants based on their media consumption. Here, we model the language contained in these media sources using an LSTM trained on corpora of each cluster's media sources to predict target words. We also model semantic similarity of test items with each cluster's corpus using skip-thought vectors. We find robust, significant correlations between performance on the SPiN test and the LSTMs and skip-thought models we present here, but not the reading span test.

1 Introduction

Generalization of experimental results crucially relies on the validity and representativeness of the experiment to study the phenomenon of interest. Researchers therefore invest considerable resources in experimental design, particularly in controlling for systematic confounds. When experiments rely on language samples for stimuli, this issue is further complicated because participants bring their complex and diverse language histories into the lab. When participants' language experiences differ systematically and the experiment does not control for this, a confound arises that compromises experimental validity and leads to systematic bias. This is the case for many cognitive tests that standardize language materials in the name of equality, whereas a more equitable approach would be to normalize test difficulty for individuals based on their experience.

One of the primary reasons for the traditional standardization approach over a normalization approach is that creating stimuli that are natural and free from confounds is a difficult laborious undertaking (e.g. as attested by Cutler (1981); Kalikow et al. (1977); Calandruccio and Smiljanic (2012)). The time required to create language stimuli is made worse by the fact that experiments can typically only use each target word or phrase once over the course of the experiment, meaning each stimulus must be uniquely created. In addition to the effort required, experimenter bias and error possibly significantly affect results (Forster, 2000).

While previous automation attempts have reduced experimenter bias, error, and workload (e.g. Lahl and Pietrowsky (2006); van Casteren and Davis (2007), vs. Hauk and Pulvermller (2004)'s manual selection) the process still relies on language statistics calculated from corpora unrepresentative of many participants' language experiences (e.g. Coltheart (1981); Linguistic Data Consortium (1996); Kucera and Francis (1967); Thorndike (1944), etc.). This mismatch between the language statistics used to generate test items and participants' actual language experiences represents a persistent confound detracting from experimental validity and perpetuating testing bias.

Our method allows participants to report for themselves the language they are comfortable with and regularly consume. Allowing participants to define their own language experiences ensures stimulus representativeness, increases fairness, and captures individual variability. This moves away from a model that gives researchers the power to define which language materials are representative across all participants (e.g. *Black Beauty* and *Little Women*: Thorndike (1944)) and moves towards a model that empowers participants to define their own language variety. To this end, we develop a method for evaluating lan-

guage experience's effect on cognitive test performance. In this work, we examine the relationship between the language that participants report consuming in media and their performance on two language-based cognitive tasks. We predict that participants' greater familiarity with the particular language variety of test items (as measured by semantic similarity and statistical predictability) will decrease test difficulty, resulting in higher scores.

Our previous results showed that participants cluster into distinct populations based on media consumption habits (Courtland et al., 2019). We determined media consumption habits by administering a self-report survey, asking participants what media content they currently consume in a variety of categories (Movies, Books, TV, etc.) as well as what they consumed in their formative years. K-means clustering identified two main clusters of participants based on the media sources they share in common. These clusters differ significantly in their performance on a test of verbal working memory (Daneman and Carpenter, 1980) and test of functional hearing (Kalikow et al., 1977). This is especially noteworthy considering we found the clusters to be orthogonal to (i.e. evenly distributed across) the traditionally used demographic variables we elicited at the end of the survey (e.g. Race, Socioeconomic Status, etc.). Here, we pursue a linguistic explanation for this performance difference by modeling the language comprising the sources participants reported consuming and examining its relationship to their performance on the behavioral tests.

To accomplish this, we use neural network language models to learn the joint probability function of word appearances in a corpus. Learning the probability of a word appearing at a certain position in a sentence can be difficult due to sparse representation in the training corpus. However, we choose these models based on their ability to capture long-distance statistical dependencies within a sentence: an advantage they enjoy over n-grams (Bengio et al., 2003). We examine a vanilla long short-term memory (LSTM) model and an attention-based model (Bahdanau et al., 2014). Both are based on recurrent neural networks and are designed to exploit semantic information distributed throughout a sentence to model the probability distribution of vocabulary words appearing as the sentence-final word (Sundermeyer et al., 2012). In addition to modeling the predictability of sentence-final words, we also use a recurrent neural network based encoder to capture sentence-level semantics (Kiros et al., 2015). We use this model to examine whether semantic familiarity affects participants' performances. We model semantics by embedding test items and corpus sentences in a high dimensional vector space and observing the distances between each item and its neighbors from the corpus. We predict that greater semantic similarity and greater sentence-final word predictability as captured by these models will correlate with participants' performance on our cognitive tasks.

2 Methods

2.1 Corpora and Behavioral Data

Participants were recruited from the USC undergraduate population (N=70) and on a local community college campus (L.A. Trade-Tech, N=25). To test language ability, participants complete the reading span task developed to assess verbal working memory (Daneman and Carpenter, 1980) and the speech perception in noise task (SPiN) developed to assess functional hearing (Kalikow et al., 1977). In the reading span task, participants read sets of sentences aloud while remembering the last word of each sentence. At the end of a set, they report the full sequence of sentence-final words in the set (with no partial credit). Set size increases (from 2 to 7) every three sets until participants cannot correctly recall any set at that length, at which point the task is terminated. The SPiN task presents spoken sentences over headphones masked with 12 talker babble (a combination of 6 male and 6 female voices speaking continuously). At the end of the sentence, participants are asked to report the final word of the sentence. We present the SPiN at +6dB SNR based on pilot results. We chose these tests for the important, yet often unacknowledged, role language processing is likely to play in both.

To capture participants' diverse language experiences, we use a proxy measure: the language materials they choose to consume regularly. Participants report these sources by completing an online survey of their current and formative media consumption habits. Using their responses, we aggregate the language data contained in these sources into corpora. We collect the sources for the corpora from *Springfield! Springfield!* and *YIFY Subtitles*, online repositories of television scripts and

movie subtitles. In total, we collect 1027 scripts of complete series (e.g. all episodes of *Futurama*) and 194 movie subtitles. We then clean the sources by removing information that does not reach viewers (e.g. stage directions, parenthetical notes, etc.). Each corpus is then tokenized into sentences for model training.

2.2 Neural Cloze Model

Cloze probability refers to the probability of encountering the last word of a sentence given the sequence of words that precede it (i.e. all non-final words of that sentence). That is, given a sentence of words w_1 through w_n, the cloze probability is expressed by: $P(w_n|w_1...w_{n-1})$. This conditional probability is a particularly important metric for our purposes because of the privileged position sentence-final words enjoy in scoring both of our behavioral tasks (cf. Duffy and Giolas (1974)'s effect of predictability on task performance). Both our behavioral tasks place participants in a condition of increased cognitive burden (either using adverse listening conditions or simultaneous verbal storage and processing demands) and then ask them to identify or remember the last word of a sentence (Daneman and Carpenter, 1980; Kalikow et al., 1977). If these words are predictable for a given participant, top-down processing can alleviate the cognitive burden of online language processing, making the task easier (Winn, 2016). If participants systematically differ in their ability to predict these sentence-final words, as might be caused by different language experiences, the task would effectively be easier for one group of participants, leading to higher scores.

To test whether performance differences on our tasks were due to cloze probability differences, we trained a vanilla LSTM and LSTM with attention on each cluster's corpus to predict the last word of a sentence given all the previous words. The attention-based LSTM model is composed of a layer of LSTM cells that capture the hidden representation of the sequence of words from the beginning of the sentence up to the last word. The final representation for sentence i is shown by H_i (eq. 3, below) and is generated by applying attention weights (α_{ij}, eq. 2) to the LSTM's hidden states, h_{ij}, corresponding to each word j in sentence i of length n. W_s, W_t, u_s, b_s and b_t are learned simultaneously during back propagation (Wang et al., 2016).

$$u_{ij} = tanh(W_s h_{ij} + b_s) \tag{1}$$

$$\alpha_{ij} = \frac{exp(u_s u_{ij})}{\sum_{k=0}^{n-1} exp(u_s u_{ik})} \tag{2}$$

$$H_i = \sum_{j=0}^{n-1} (\alpha_{ij} * h_{ij}) \tag{3}$$

Using a fully connected and a softmax layer, we then calculate the probability of each word w in the vocabulary appearing immediately after the sequence as p_w(i.e. at the end of that sentence).

$$v_{iw} = W_t H_i + b_t \tag{4}$$

$$p_w = \frac{exp(v_{iw})}{\sum_{k=0}^{|vocabulary|} exp(v_{ik})} \tag{5}$$

For the experiment, we use a vocabulary consisting of the 10k most frequent words in the corpus. The hidden size of the LSTM and attention vectors are set to 100. We use 300-dimensional GloVe word embeddings as the semantic representation of the words (Pennington et al., 2014).

2.3 Skip-thought Vectors

To obtain a quantitative measure of semantic similarity, we embed test items and sentences from each cluster's corpus in a high dimensional vector space and measure the distance of each test item to neighboring items from the corpus. To encode target and corpus items into vectors, we use combine-skip-thought vectors as detailed in Kiros et al. (2015). These encode sentences using RNNs with GRU into a 4800-dimensional vector which is the concatenation of a 2400-dimensional uni-directional encoder and a 2400-dimensional bi-directional encoder (1200 dimensions for backwards and forwards each). Results from the original paper show that these vectors capture a high degree of sentence-level semantics, particularly as it relates to encoding similarity as vector-space distance: the closer two sentences are in the embedded vector space, the more semantically related they are. We therefore take the distances in this embedded vector space to be indicative of how typical a test item's semantics are given the corpus of a participant's cluster.

We measure each test item's mean distance from all corpora items using the Taxicab distance (L^1 norm, eq. 6) and standardized Euclidean distance (eq. 7):

$$\sum_{i=1}^{n} |u_i - v_i| \tag{6}$$

$$\sqrt{\sum_{i=1}^{n} (u_i - v_i)^2 / V[x_i]} \tag{7}$$

where $V[x_i]$ is the variance vector over the components of all vectors.

We also measure the mean distance to the closest 100 corpus neighbors in the event that similarity to all corpus items proves less informative than similarity to the closest matches from the corpus.

3 Results

3.1 Neural Cloze Model

For each test item, we correlate each cluster's LSTM activation of the sentence-final word with that cluster's mean behavioral performance (i.e. the percent of the cluster's participants who answered that item correctly). We use rank correlation as we are uncertain of how linear the mapping between predictability and performance benefit will be.

We observe significant rank correlations between the activation of both clusters' vanilla LSTMs and their respective mean performances on the SPiN items ($\rho(48) = .39, p < .01$ for cluster 1, $\rho(48) = .46, p < .005$ for cluster 2). We observe weaker but still significant correlations between the attention-based LSTM activations and mean performances on SPiN items ($\rho(48) = .31, p < .05$ for cluster 1, $\rho(48) = .29, p = .05$ for cluster 2). This poorer performance of the more complex model is noteworthy. We observe no significant rank correlations between any model's activations and performance on the corresponding span task item (see Table 1).

3.2 Skip-thought Vectors

For each cluster, we test for a correlation between the distance from all its corpus items to a given test item and the mean performance of its participants on that item. Given uncertainty of whether the distance-performance relationship will be linear, we use rank correlation. Using the distance metrics in eqs. (6) and (7), we observe significant rank correlations between vector-space distances and performances on the SPiN task (see Table 1 for test statistics, all $\rho(48), p < .005$) but not the span task. In addition to the mean distance of all items, we calculated the distance to the closest 100 neighboring corpus items and obtained similar results.

	Cluster 1		Cluster 2	
	SPiN	span	SPiN	span
Vanilla LSTM	.39	-.03	.46	-.15
Attn. LSTM	.31	.02	.29	-.03
Taxicab	.486	.075	.519	-.022
Std. Euclid.	.408	-.048	.440	.092

Table 1: Mean behavioral performance on SPiN target items is significantly rank correlated to both LSTM activations and skip-thought distances for both clusters. We find no significant correlations with the span test for either cluster.

4 Discussion

Language models tailored to the media consumption of different "clusters" of English speakers predict performance at the item level on a test of functional hearing (SPiN). In particular, LSTM models, which are perhaps the most natural way to model a task in which the predictability of the final word in a sentence has a strong influence on performance, correctly predict accuracy for each cluster. For the reading span task, in contrast, neither type of model correctly predicted performance. It is possible that the models are not capturing the relevant linguistic information for reading span or that reading span simply depends less on language (and language experience) overall than SPiN. An alternative explanation, however, comes from the difficulty in handling span performance data and its scoring. In the span task, items are presented in a fixed order, and difficulty increases from trial to trial as participants are required to maintain more items in working memory. This makes scoring at the item level difficult to interpret. Given these complications with the scoring procedure, it is possible that item-level analysis of the reading span is uninformative and invalid compared to the straight-forward scoring procedure of the SPiN.

Regarding the SPiN task, the robustness of the correlation between skip-thought vector mean-neighbor distances and participant performance is curious, however. The interesting aspect of this relationship is the direction of the correlation: that as the distance from corpus neighbors increases, performance on the item *increases*. This implies that unusual items are scored better on than famil-

iar ones. This finding is not necessarily at odds with the finding of the neural cloze models: that increased predictability of the last word positively correlates with performance on that sentence. The two models differ in several key aspects which may explain their differences. Firstly, skipthought distances do not capture statistical predictability but rather semantic similarity, so while the last word (or in fact the sentence as a whole) may be semantically odd, it also may be relatively easy to predict the last word from the rest of the sentence. Secondly, skipthoughts operate at the level of the entire sentence rather than at the level of just the last word, which means that all of the words contributing to their embedding but the sentence-final one do not directly factor into the scoring of behavioral performance. This means that the majority of the linguistic information they encode is uninformative for capturing predictability of the last word, which is a direct correlate to how the task is scored. Lastly, skipthoughts are capturing the semantic novelty of a sentence. It is possible that the increased attentional resources these items demand above overly typical items actually causes participants to perform better on these items rather than worse. This must be tested further before concrete conclusions can be drawn, but it represents an interesting future direction for study.

We believe the results obtained here are an initial step toward taking participants' self-reported language experience into account in interpreting their performance on cognitive tests. In light of the evidence that a connection likely exists, we support the approach of normalizing, rather than standardizing, the language of cognitive tests. We predict normalization will produce tests that are simultaneously more fair and more valid. Regarding increased validity, the use of dynamically generated corpora would afford a significant benefit over static corpora by reducing sampling error. Every corpus necessarily contains idiosyncratic sampling error affecting results (Clark, 1973). The repeated use of norms generated from a single corpus (e.g. as was traditionally taken from Kucera and Francis (1967) or Thorndike (1944)) amplifies this noise and its role in experimental results. The construction of dynamic corpora we are planning will mitigate this effect by providing multiple samples across which real statistical regularities are likely to replicate, while sample noise is not (like bootstrapping: Efron (1979)).

While the eventual goal of this work is to generate valid and fair stimuli *ex nihilo* given people's language models, the evaluation of existing stimuli materials represents a necessary first step taken here. The development of models capturing linguistic features that predict behavioral performance provides the possibility for using these models to identify or synthesize fair test items. Modeling the relationship between language experience and task performance allows rapid prototyping and evaluation of stimuli sets with previously unfeasible speed. This allows a much larger set of candidate stimuli to be evaluated affording new levels of rigor to the test creation process. This speed also opens the door for individual personalization of test items, a task far too labor-intensive to perform manually. Our future work will test our models' ability to create test stimuli equitable across diverse language communities.

These methods for promoting equity are likely relevant to education where equality vs. equity is debated as the difference between equal access to educational resources vs. access to resources leading to equal outcomes (e.g. Green (1983); Stromquist (2005); Espinoza (2007)). Language-based cognitive testing and access to education share several features in common. Both are moderated by the complex individual variability of personal experience. Those with the worst outcomes in both are underrepresented among those setting policy and creating tests (National Science Foundation, 2013; Thaler and Jones-Forrester, 2013; Thaler et al., 2015). And most importantly, both also determine relevant real-world outcomes for test takers: the tests we consider here are used clinically to diagnose aphasia (Caspari et al., 1998), Alzheimer's disease (Kempler et al., 1998), schizophrenia (Stone et al., 1998), and age-related cognitive decline (Salthouse and Kersten, 1993). Many cognitive tests use linguistic stimuli to assess other cognitive functions; by identifying specific ways in which individuals' language variety influences their performance, we can start to tease apart potential educationally and clinically meaningful deficits from social and cultural differences between participant groups.

Acknowledgments

This work was supported by the NIH (5R21DC017018-02). Behavioral data used were collected under USC IRB #UP-18-00006.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Lauren Calandruccio and Rajka Smiljanic. 2012. New Sentence Recognition Materials Developed Using a Basic Non-Native English Lexicon. *Journal of Speech, Language, and Hearing Research*, 55(5):1342–1355.

Isabelle Caspari, Stanley R. Parkinson, Leonard L. LaPointe, and Richard C. Katz. 1998. Working Memory and Aphasia. *Brain and Cognition*, 37(2):205–223.

Maarten van Casteren and Matthew H. Davis. 2007. Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, 39(4):973–978.

Herbert H. Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359.

Max Coltheart. 1981. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Maury Courtland, Aida Davani, Melissa Reyes, Leigh Yeh, Jun Leung, Brendan Kennedy, Morteza Dehghani, and Jason Zevin. 2019. Subtle differences in language experience moderate performance on language-based cognitive tests. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, Austin, Texas. Cognitive Science Society.

Anne Cutler. 1981. Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10(1):65–70.

Meredyth Daneman and Patricia A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466.

Joseph R. Duffy and Thomas G. Giolas. 1974. Sentence Intelligibility as a Function of Key Word Selection. *Journal of Speech and Hearing Research*.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Oscar Espinoza. 2007. Solving the equityequality conceptual dilemma: a new model for analysis of the educational process. *Educational Research*, 49(4):343–363.

Kenneth I. Forster. 2000. The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7):1109–1115.

Thomas F. Green. 1983. Excellence, Equity, and Equality.

O Hauk and F Pulvermller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090–1103.

D. N. Kalikow, K. N. Stevens, and L. L. Elliott. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Daniel Kempler, Amit Almor, Lorraine K. Tyler, Elaine S. Andersen, and Maryellen C. MacDonald. 1998. Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. *Brain and Language*, 64(3):297–316.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. *arXiv:1506.06726 [cs]*. ArXiv: 1506.06726.

Henry Kucera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Olaf Lahl and Reinhard Pietrowsky. 2006. EQUIWORD: A software application for the automatic creation of truly equivalent word lists. *Behavior Research Methods*, 38(1):146–152.

Linguistic Data Consortium. 1996. *CELEX2*. OCLC: 1023487640.

National Science Foundation. 2013. Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013: (558442013-001). Technical report, American Psychological Association. Type: dataset.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Timothy A. Salthouse and Alan W. Kersten. 1993. Decomposing adult age differences in symbol arithmetic. *Memory & Cognition*, 21(5):699–710.

Maria Stone, John D. E. Gabrieli, Glenn T. Stebbins, and Edith V. Sullivan. 1998. Working and strategic memory deficits in schizophrenia. *Neuropsychology*, 12(2):278–288.

Nelly Stromquist. 2005. Comparative and International Education: A Journey toward Equality and Equity. *Harvard Educational Review*, 75(1):89–111.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Nicholas S. Thaler and Sharon Jones-Forrester. 2013. IQ Testing and the Hispanic Client. In *Guide to Psychological Assessment with Hispanics*, pages 81–98. Springer, Boston, MA.

Nicholas S. Thaler, April D. Thames, Xavier E. Cagigas, and Marc A. Norman. 2015. IQ Testing and the African American Client. In Lorraine T. Benuto and Brian D. Leany, editors, *Guide to Psychological Assessment with African Americans*, pages 63–77. Springer New York, New York, NY.

Edward L. Thorndike. 1944. *The teacher's word book of 30,000 words*. New York :.

Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Matthew B. Winn. 2016. Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, 20:233121651666972.