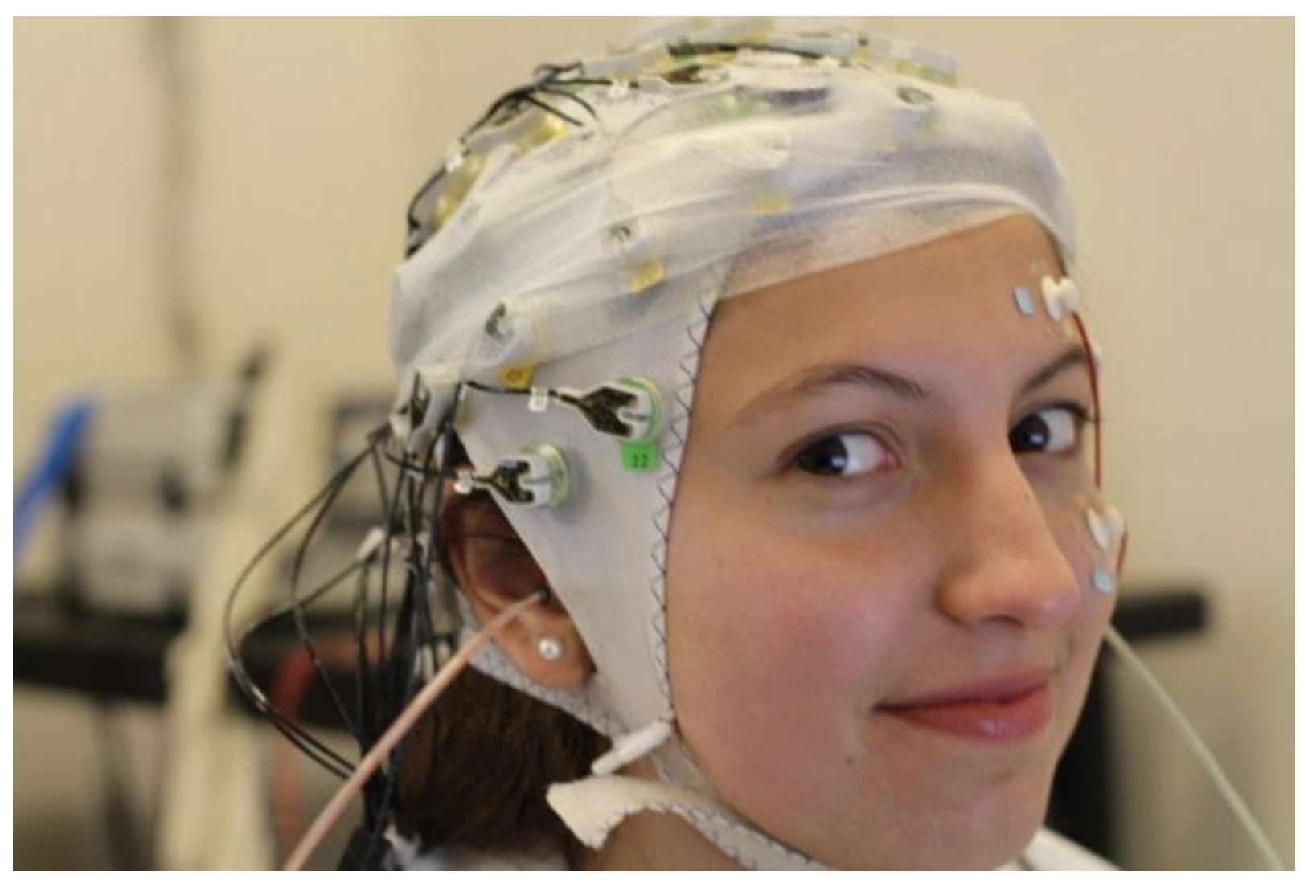
DeepMind

Text Genre & Training Data Size in Human-Like Parsing

John T. Hale (DeepMind & University of Georgia) , Adhiguna Kuncoro (DeepMind & Oxford)
Keith B. Hall (Google NYC), Chris Dyer (DeepMind) & Jonathan Brennan (University of Michigan)

Goal: explain human EEG signal by reference to NLP system

It is possible to account for some aspects of human electrophysiology during language comprehension by reference to the internal states of a deep-learning phrase-structure parsing system. (Hale et al 2018).



Question: does it matter what these systems are trained on?

To find out, we compared NEWSPAPER TEXT to ALICE-LIKE BOOKS.

These text genres were first annotated with phrase structures by a Berkeley-like parser. We then used these trees to train an incremental parser based on Recurrent Neural Network Grammars (Dyer et al 2016, Kuncoro et al 2017). The total probability of all analyses in this incremental parser's beam is the basis for a surprisal prediction. This in turn becomes a predictor in a regression model of human EEG.

ALICE-LIKE according to CosineTop50

This metric, from McClosky et al 2006, is purely lexical in nature. It compares candidate training materials to the attestation counts of the top 50 most well-attested words in a reference corpus.

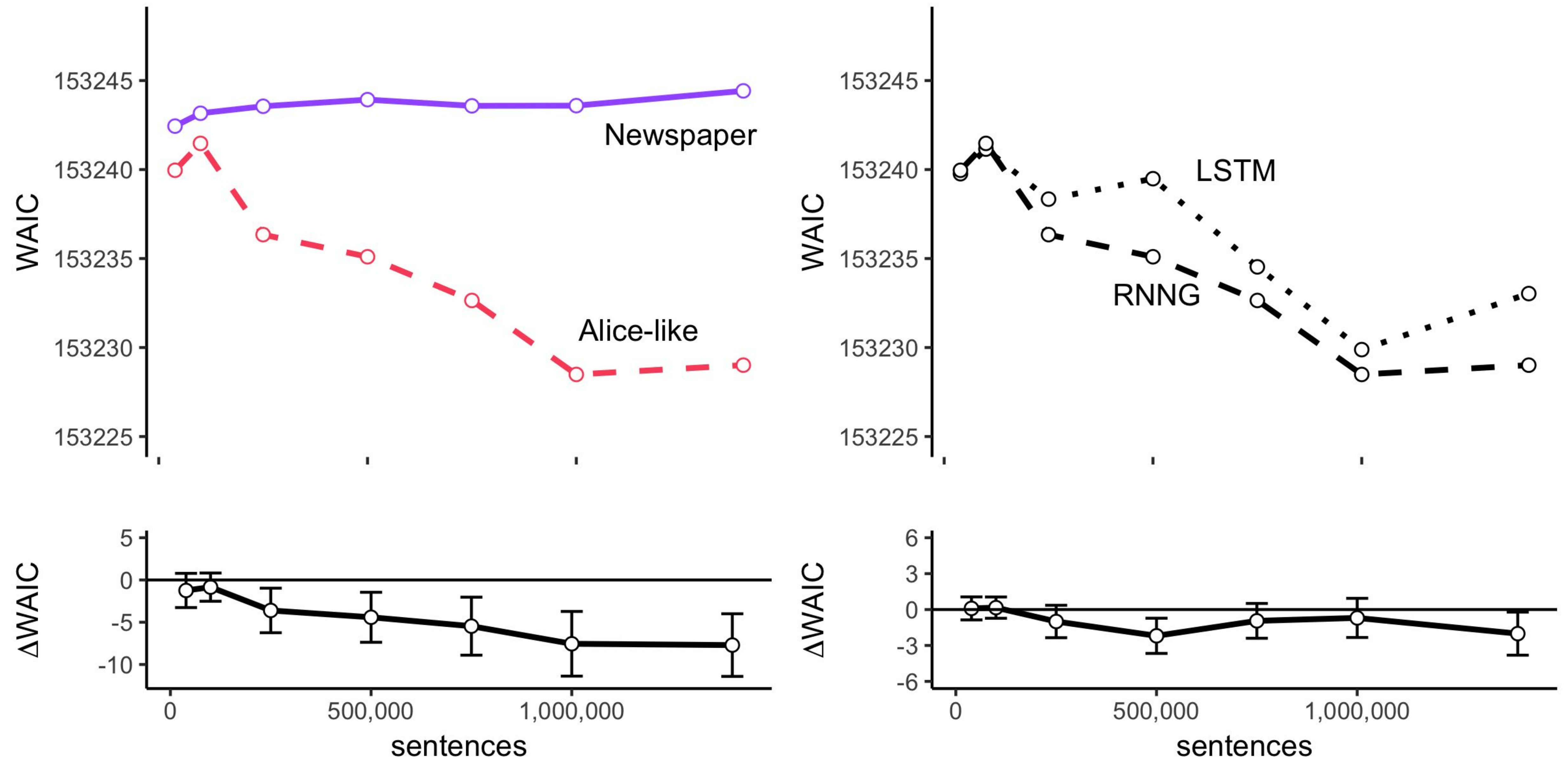
dissimilarity	title	author
0.0584	The Admiral's Caravan	Charles E. Carryl
0.0620	The Secret Garden	Frances Hodgson Burnett
0.0628	The Lodger	Marie Belloc Lowndes
0.0687	The Girls and I: A Veracious History	Mary Louisa Stewart Molesworth
0.0689	What Timmy Did	Marie Adelaide Belloc
0.0724	Little Miss Peggy	Mrs. Molesworth
0.0725	The Girls of St. Olave's	Mabel Mackintosh
0.0741	The Celebrity at Home	Violet Hunt
0.0750	I've Married Marjorie	Margaret Widdemer
0.0752	The Forged Note	Oscar Micheaux
0.0755	Mary Erskine	Jacob Abbott
0.0758	The Bountiful Lady	Thomas Cobb
0.0758	Legacy	James H Schmitz
0.0763	Some Little People	George Kringle
0.0774	In the Wilderness	Robert Hichens

Table 2: Alice-like books from Project Gutenberg

Answer: yes, text genre matters.

Adding more parses of newspaper text to the training set doesn't improve the regression model of human EEG -- but additional parsed examples from Alice-like books do help.

lower WAIC → **better** fit



FAQ

- Did perplexity improve with more training data? *A. yes, in both genres. This dissociates LM perplexity from fit to brain data*
- Is phrase structure crucial? *A. yes. an LSTM performed worse.*
- How did you control the vocabulary? *A. we used a superset vocabulary from the largest training set.*
- What co-regressors went into the EEG modeling? *A. sentence position within the book, word position within each sentence and unigram frequency for prev, current and next word.*
- Really, no N400? *A. that's right; we found no effect in that spatio-temporal region.*

Conclusion:

It is better to train on in-domain data when modeling human language comprehension. Listeners seem to be adapting to the syntactic preferences of a particular genre, as psycholinguists such as Edith Kaan have suggested.