Evidence of semantic processing difficulty in naturalistic reading

Cory Shain[1], Richard Futrell[2], Marten van Schijndel[3], Edward Gibson[2], William Schuler[1], and Evelina Fedorenko[2]

shain.3@osu.edu

[1]Ohio State, [2]MIT, [3]Johns Hopkins

Background

- Although language is used to convey and infer meaning, existing work on naturalistic sentence processing focuses on lexical and/or structural determinants of comprehension difficulty.
  - Lexical frequency [6, 20]
  - Parse probability [5, 23]
  - Dependency locality [3, 19]
- Incremental semantic decisions are harder to estimate from corpora.
- Vectorial word representations have been shown to contain semantic information [14, 12] and predict human responses [16, 1, 18, 4].
- To the extent that word vectors map to human semantic space, they can help us study the incremental cost of moving around that space.

Methods

- Embed all content words using 300d GloVe vectors [17] pretrained on the 840B word Common Crawl dataset.
- Compute mean vector distance between current word and all content words preceding it in the sentence.
- Transform reading times with Box-Cox [2].
- **Testing procedure:** Ablative likelihood ratio testing of linear mixed effects models
- **Fixed effects:** Word length, position in sentence, 5-gram surprisal (KenLM [10] trained on Gigaword 3 [9]), and PCFG surprisal ([22] parser trained on WSJ [13] re-annotated into Generalized Categorial Grammar [15]), plus (eye-tracking only) saccade length and accumulated surprisal [21]
- **Random effects:** Slopes for all of the above by subject, by-subject and by-word random intercepts
- **Spillover optimization**: Spillover position optimized on exploratory data using fixed effects models. All predictors remained *in situ* except: Dundee (5-gram surprisal spillover-1), UCL (saccade length spillover-1), and Natural Stories (PCFG surprisal spillover-1). Main effects were spillover-1.

Data

- Three reading time corpora:
  - **Natural Stories** [8]
    - Constructed narratives, self-paced reading, 181 subjects, 485 sentences, 10,245 tokens, 848,768 fixation events
    - Post-processing: Removed sentence boundaries, events for which subjects missed 4+ comprehension questions and fixations < 100 ms or > 3000 ms.
  - **Dundee** [11]
    - Newspaper editorials, eye-tracking, 10 subjects, 2,368 sentences, 51,502 tokens, 260,065 fixation events
    - Post-processing: Removed document, screen, sentence, and line boundaries
  - **UCL** [7]
    - Sentences from novels presented in isolation, eye-tracking, 42 subjects, 205 sentences, 1,931 tokens, 53,070 fixation events
    - Post-processing: Removed sentence boundaries
- Data split: 1/3 exploratory, 2/3 confirmatory

Question

Does semantic distance of a word from its context cause processing difficulty during naturalistic reading?

Corpus	$\hat{\beta}$-*ms* semantic distance	t	p
Natural Stories	1.25	2.766	0.006
Dundee	5.73	4.759	5.59e-4
UCL	16.36	7.853	2.76e-10

Table 1: Likelihood ratio testing results for mean semantic cosine distance on Natural Stories, Dundee, and UCL. Reading times were transformed using [2] and $\hat{\beta}$-ms was computed by backtransformation, and is therefore only valid at the backtransformed mean, holding all other effects at their means.

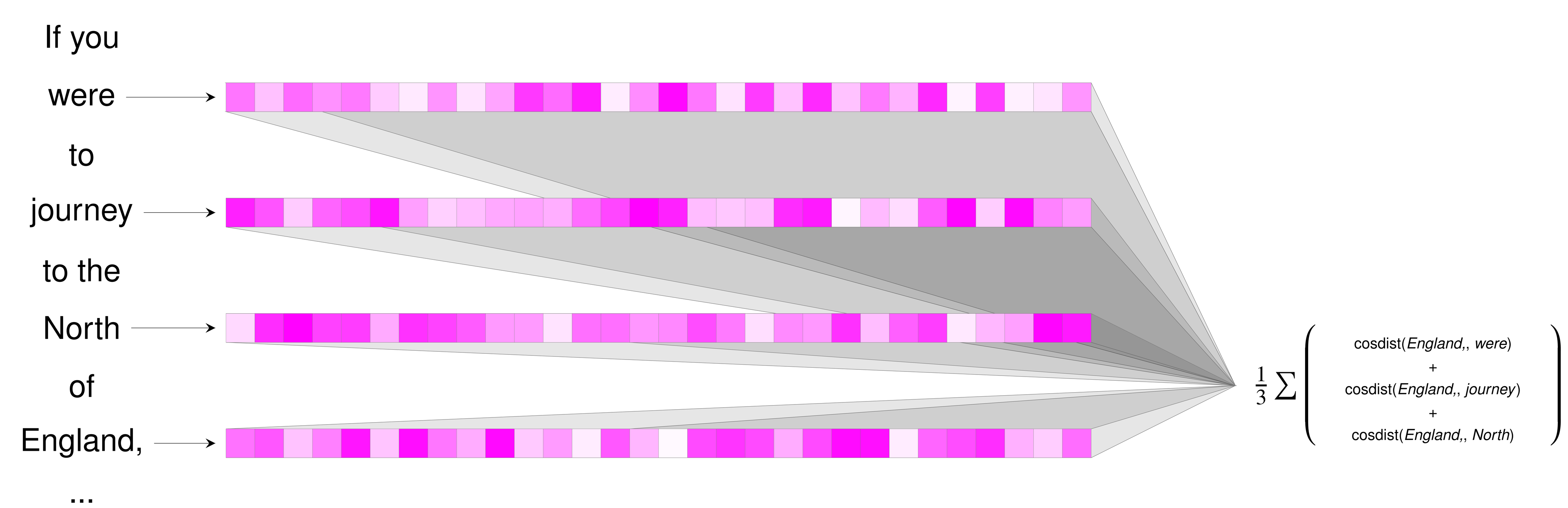

Figure 1: Visual illustration of computation of semantic distance. Content words (left) are cast into real-valued vectors (center) using GloVe. The semantic distance of the word *England* in this example is computed as the mean cosine distance of the embedding of *England* from the embeddings of its preceding content words (*were*, *journey*, and *North*). Non-content words are treated as having distance 0.

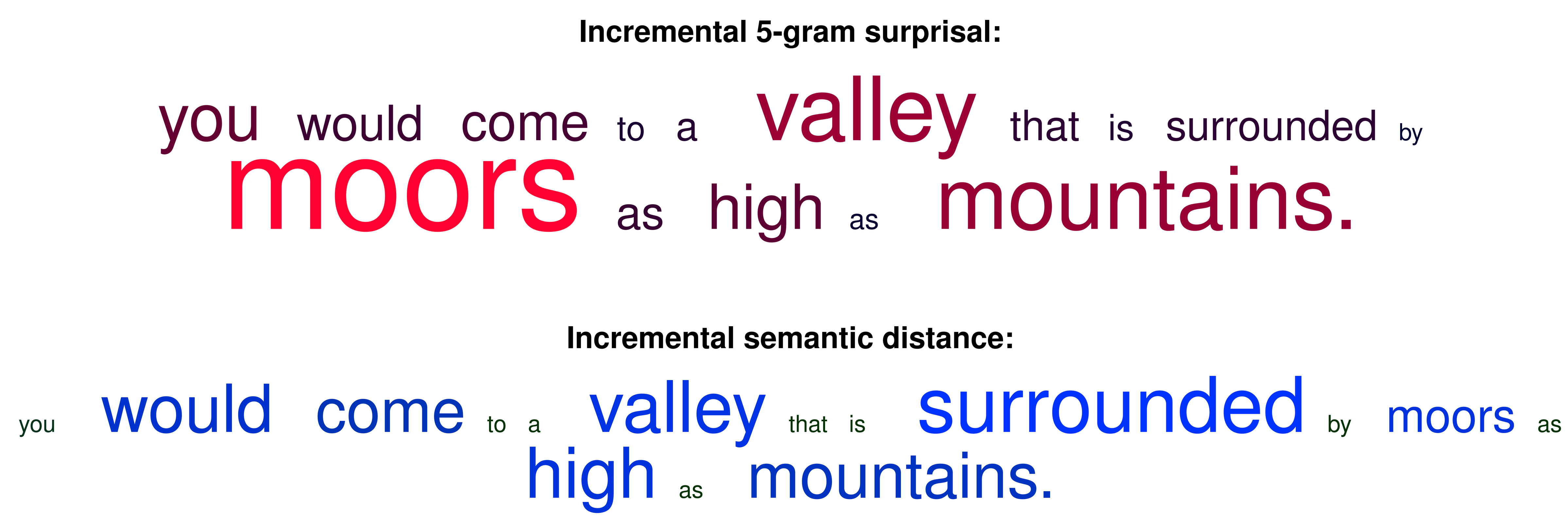

Figure 2: Visual illustration of the actual 5-gram surprisal (top) and semantic distance (bottom) values from the end of the first sentence of the Natural Stories corpus, where large font indicates large value. Although both measures flag content-bearing words, there are important differences. For example, *moors*, a low-frequency word in corpora, has a high surprisal value but relatively low semantic distance from preceding words like *valley*, while *surrounded* has low surprisal but high semantic distance.

Discussion

- We find positive effects and significant contributions to model fit across corpora, suggesting a replicable contribution of semantic distance to processing load.
- Result consistent with at least two (possibly compatible) interpretations
  - Traversing the semantic space might be costly, since semantic targets may not have been primed through spreading activation
  - Semantic distance may partially estimate semantic predictability, and therefore improve on baseline estimates of incremental surprisal.
- Future advances in automatic incremental semantic parsing may help tease apart these possibilities.

References

[1] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014*, pages 238–247, 2014.

[2] George E. P. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[3] Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210, 2008.

[4] Allyson Ettinger, Naomi H. Feldman, Philip Resnik, and Colin Phillips. Modeling n400 amplitude using vector space models of word representation. In *Proceedings of the 38th annual conference of the Cognitive Science Society*, pages 1445–1450, 2016.

[5] Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of CMCL 2012*. Association for Computational Linguistics, 2012.

[6] Stefan Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 2011.

[7] Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45:1182–1190, 2013.

[8] Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. The natural stories corpus. *arXiv*, (1708.05763), 2017.

[9] David Graff and Christopher Cieri. *English Gigaword LDC2003T05*, 2003.

[10] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, 2013.

[11] Alan Kennedy, James Pynte, and Robin Hill. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

[12] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308, 2014.

[13] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyreand Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The Penn TreeBank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, 1994.

[14] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *In Proceedings of NAACL 2013*, 2013.

[15] Luan Nguyen, Marten van Schijndel, and William Schuler. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING 2012*, pages 2125–2140, Mumbai, India, 2012.

[16] Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 38–46, 2011.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, 2014.

[18] Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33:175–190, 2016.

[19] Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 49–58. Association for Computational Linguistics, 2016.

[20] Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.

[21] Marten van Schijndel. *The Influence of Syntactic Frequencies on Human Sentence Processing*. PhD thesis, The Ohio State University, 2016.

[22] Marten van Schijndel, Andy Exley, and William Schuler. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540, 2013.

[23] Marten van Schijndel and William Schuler. An analysis of frequency- and memory-based processing costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics, 2013.