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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>β-ms semantic distance</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Stories</td>
<td>1.25</td>
<td>2.766</td>
<td>0.006</td>
</tr>
<tr>
<td>Dundee</td>
<td>5.73</td>
<td>4.759</td>
<td>5.59e-4</td>
</tr>
<tr>
<td>UCL</td>
<td>16.36</td>
<td>7.853</td>
<td>2.76e-10</td>
</tr>
</tbody>
</table>

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

Random intercepts: slopes for all of the above by subject, by-word random intercepts.

Incremental semantic distance:

you would come to a valley that is surrounded by moors as high as mountains.

Incremental 5-gram surprisal:

you would come to a valley that is surrounded by moors as high as mountains.

Evidence of semantic processing difficulty in naturalistic reading
Cory Shain¹, Richard Futrell², Marten van Schijndel³, Edward Gibson², William Schuler¹, and Evelina Fedorenko²
¹Ohio State, ²MIT, ³Johns 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]
  - Word length, position in sentence, 5-gram surprisal, and Natural Stories [8, 21]
  - Random effects: slopes for all of the above by subject, by-word random intercepts.

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.
- Random effects: Effects of all 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 situs 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.
- Natural Stories and Dundee are written in the English language and Natural Stories is a low-frequency word in corpora, has a high surprisal value but relatively low semantic distance from surrounding words.