Surprisal as a Predictor of Essay Quality

Gaurav Kharkwal1, Smaranda Muresan2

gaurav.kharkwal@gmail.com, smara@ccls.columbia.edu

1. Department of Psychology and Center for Cognitive Science, Rutgers University, New Brunswick
2. Center for Computational Learning Systems, Columbia University

Summary

• We bring together work from psycholinguistics and NLP.
• Through corpora studies, we examine the relation between sentence processing complexity and essay quality.
• Essays of greater overall complexity tend to have lower scores, and vice versa.

Surprisal Theory

• Surprisal is a psycholinguistic model of sentence processing complexity (Hale, 2001; Levy, 2008).
• Word-level processing cost estimated as negative log-prob of word given preceding context:

\[ \text{Surp}(w_i) \propto - \log P(w_i|w_{1...i-1}, \text{CONTEXT}) \]

Computing Surprisal

• We used a top-down parser trained on WSJ corpus (Roark, 2009), which provided three measures:
  - Syntactic surprisal: unexpectedness of POS cat of word given sentential context.
  - Lexical surprisal: unexpectedness of word given sentential context and POS cat.
  - Total surprisal: sum of Syntactic and Lexical.

Experiment 1 (contd.)

Corpus

• Uppala Student English corpus (Axelsson, 2000).
  - 1,489 essays written by 440 EFL students.
• 116 essays were randomly selected:
  - 38 pairs on topic Analysis
  - 20 pairs on topic Argumentation
  - Each pair written by the same student across 2 terms

Methods

• Computed surprisal values using Roark's parser.
• Evaluated group mean differences across the two terms using linear mixed-effects regression models for the two topics:

\[ \text{Surp} \sim \text{Term} + (1|\text{Subject}) \]

Results and Discussion

<table>
<thead>
<tr>
<th>Topic</th>
<th>Term</th>
<th>Mean (Syn)</th>
<th>SD (Syn)</th>
<th>Mean (Lex)</th>
<th>SD (Lex)</th>
<th>Mean (Total)</th>
<th>SD (Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis</td>
<td>Term 1</td>
<td>2.37</td>
<td>1.86</td>
<td>3.97</td>
<td>3.24</td>
<td>6.34</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td>Term 2</td>
<td>2.34</td>
<td>1.85</td>
<td>3.94</td>
<td>3.23</td>
<td>6.28</td>
<td>3.30</td>
</tr>
<tr>
<td>Argumentation</td>
<td>Term 1</td>
<td>2.34</td>
<td>1.85</td>
<td>3.90</td>
<td>3.23</td>
<td>6.24</td>
<td>3.29</td>
</tr>
<tr>
<td></td>
<td>Term 2</td>
<td>2.28</td>
<td>1.85</td>
<td>3.87</td>
<td>3.24</td>
<td>6.15</td>
<td>3.36</td>
</tr>
</tbody>
</table>

• Despite trends, no consistent indication of an effect of EFL training on essays’ surprisal scores.
• Absence of essay scores prevented direct evaluation of the link between surprisal and essay quality.

Experiment 2 (contd.)

Corpus

• ETS’s corpus used for NLI (Blanchard, et al, 2013).
  - 12,100 essays on 8 topics scored as High, Medium, or Low.
• 3,975 essays were randomly selected:
  - 1,325 per score category.

Methods

• Computed surprisal values as before.
• Performed correlation tests and group mean evaluations using a linear mixed-effects model:

\[ \text{Surp} \sim \text{EssayScore} + (1|\text{Subject2}) \]

Results

<table>
<thead>
<tr>
<th>Score</th>
<th>Mean (Syn)</th>
<th>SD (Syn)</th>
<th>Mean (Lex)</th>
<th>SD (Lex)</th>
<th>Mean (Total)</th>
<th>SD (Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>2.46</td>
<td>.22</td>
<td>3.76</td>
<td>.29</td>
<td>6.22</td>
<td>.39</td>
</tr>
<tr>
<td>Medium</td>
<td>2.35</td>
<td>.17</td>
<td>3.75</td>
<td>.26</td>
<td>6.10</td>
<td>.34</td>
</tr>
<tr>
<td>High</td>
<td>2.27</td>
<td>.14</td>
<td>3.82</td>
<td>.24</td>
<td>6.09</td>
<td>.28</td>
</tr>
</tbody>
</table>

• Although all measures were found to be correlated, only Syntactic Surprisal had a high correlation coeff.

Conclusion

• Inverse relation between surprisal values and essay scores, with Syntactic Surprisal most promising.

Future Work

• How do findings vary across different datasets?
• Does greater processing complexity cause lower essay score?
• How important is training corpus used for computing surprisal?

Experiment 1

Introduction

• Investigated whether EFL training improves essay quality, using essays written by EFL students across various terms.
• Examined whether essays’ surprisal values decrease after training.

Introduction

• Directly investigated link between surprisal and essay quality using a pre-scored set of essays.
• Evaluated whether surprisal values are correlated with essays’ scores.