Surprisal as a Predictor of Essay Quality

Gaurav Kharkwal
Department of Psychology
Center for Cognitive Science
Rutgers University, New Brunswick
gaurav.kharkwal@gmail.com

Smaranda Muresan
Department of Computer Science
Center for Computational Learning Systems
Columbia University
smara@ccls.columbia.edu

Abstract

Modern automated essay scoring systems rely on identifying linguistically-relevant features to estimate essay quality. This paper attempts to bridge work in psycholinguistics and natural language processing by proposing sentence processing complexity as a feature for automated essay scoring, in the context of English as a Foreign Language (EFL). To quantify processing complexity we used a psycholinguistic model called *surprisal theory*. First, we investigated whether essays' average surprisal values decrease with EFL training. Preliminary results seem to support this idea. Second, we investigated whether surprisal can be effective as a predictor of essay quality. The results indicate an inverse correlation between surprisal and essay scores. Overall, the results are promising and warrant further investigation on the usability of surprisal for essay scoring.

1 Introduction

Standardized testing continues to be an integral part of modern-day education, and an important area of research in educational technologies is the development of tools and methodologies to facilitate automated evaluation of standardized tests. Unlike multiple-choice questions, automated evaluation of essays presents a particular challenge. The specific issue is the identification of a suitable evaluation rubric that can encompass the broad range of responses that may be received.

Unsurprisingly then, much emphasis has been placed on the development of Automated Essay Scoring (henceforth, AES) systems. Notable AES systems include Project Essay Grade (Page, 1966; Ajay et al., 1973), ETS's *e-rater*® (Burstein et al., 1998; Attali and Burstein, 2006), Intelligent Essay Assessor™ (Landauer et al., 2003), BETSY (Rudner and Liang, 2002), and Vantage Learning's IntelliMetric™ (Elliot, 2003). The common thread in most modern AES systems is the identification of various observable linguistic features, and the development of computational models that combine those features for essay evaluation.

One aspect of an essay's quality that almost all AES systems do not yet fully capture is sentence processing complexity. The ability to clearly and concisely convey information without requiring undue effort on the part of the reader is one hallmark of good writing. Decades of behavioral research on language comprehension has suggested that some sentence structures are harder to comprehend than others. For example, passive sentences, such as *the girl was pushed by the boy*, are known to be harder to process than semantically equivalent active sentences, such as *the boy pushed the girl* (Slobin, 1966; Forster and Olbrei, 1972; Davison and Lutz, 1985; Kharkwal and Stromswold, 2013). Thus, it is likely that the overall processing complexity of the sentence structures used in an essay could influence its perceived quality.

One reason why sentence processing complexity has not yet been fully utilized is the lack of a suitable way of quantifying it. This paper proposes the use of a psycholinguistic model of sentence comprehension called *surprisal theory* (Hale, 2001; Levy, 2008) to quantify sentence processing complexity. The rest of the paper is organized as follows. Section 2 describes the surprisal theory, and discusses its applicability in modeling sentence processing complexity. Section 3 details our investigation on whether essays' average surprisal values decrease following English as a Foreign Language training. Section 4 presents a study where we investigated whether surprisal can be effective as a predictor of essay quality. Lastly, Sec-

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 54–60,
Baltimore, Maryland USA, June 26, 2014.

The	**judge**	**who**	**angered**	**the**	**criminal**	**slammed**	**the**	**gavel**	**Mean**
5.64	6.94	6.93	11.60	2.32	9.19	16.92	1.94	4.68	7.35
The	**judge**	**who**	**the**	**criminal**	**angered**	**slammed**	**the**	**gavel**	**Mean**
5.64	6.94	6.93	4.20	9.21	13.73	16.65	2.21	4.69	7.80

Table 1: Surprisal values of two example relative-clause sentences. The values were computed using a top-down parser by Roark et al. (2009) trained on the Wall Street Journal corpus.

tion 5 concludes the paper.

2 Surprisal Theory

The *surprisal theory* (Hale, 2001; Levy, 2008) estimates the word-level processing complexity as the negative log-probability of a word given the preceding context (usually, preceding syntactic context). That is:

$$\text{Complexity}(w_i) \propto -\log P(w_i|w_{1\ldots i-1}, \texttt{CONTEXT})$$

Essentially, the surprisal model measures processing complexity at a word as a function of how unexpected the word is in its context. Surprisal is minimized (i.e. approaches zero) when a word *must* appear in a given context (i.e., when $P(w_i|w_{1\ldots i-1}, \texttt{CONTEXT}) = 1$), and approaches infinity as a word becomes less and less likely. Crucially, the surprisal theory differs from n-gram based approaches by using an underlying language model which includes a lexicon and a syntactic grammar (the language model is usually a Probabilistic Context-Free Grammar, but not restricted to it).

To better understand surprisal, consider the following two example sentences:

(1) *The judge who angered the criminal slammed the gavel.*
(2) The judge who the criminal angered slammed the gavel.

Both sentences are center-embedded relative clause sentences that differ in whether the subject or the object is extracted from the relative clause. Critically, they both share the same words differing only in their relative order. Behavioral studies have found that object-extracted relative clause sentences (2) are harder to process than subject-extracted relative clause sentences (1) (King and Just, 1991; Gordon et al., 2001; Grodner and Gibson, 2005; Staub, 2010; Traxler et al., 2002; Stromswold et al., 1996). The surprisal values at each word position of the two example sentences are shown in Table 1.

As we can see from Table 1, the mean surprisal value is greater for the object-extracted relative clause sentence. Hence, the surprisal theory correctly predicts greater processing cost for that sentence. Furthermore, it allows for a finer-grained analysis of where the processing cost might occur, specifically at the onset of the relative clause (*the*) and the end (*angered*). Other differences, such as greatest difficulty at the main verb are shared with the subject-extracted relative clause, and are plausible because both sentences are center-embedded. These predictions are consistent with patterns observed in behavioral studies (Staub, 2010).

In addition to relative clauses, the surprisal theory has been used to model various other behavioral findings (Levy, 2008; Levy and Keller, 2012). Moreover, corpora analyses examining surprisal's effectiveness revealed a high correlation between word-level suprisal values and the corresponding reading times, which act as a proxy for processing difficulties (Demberg and Keller, 2008; Boston et al., 2008; Frank, 2009; Roark et al., 2009).

Thus, the surprisal theory presents itself as an effective means of quantifying processing complexity of sentences, and words within them. Next, we discuss a series of evaluations that we performed to determine whether surprisal values reflect quality of written essays.

3 Experiment 1

In the first experiment, we investigate whether an essay's mean surprisal value decreases after suitable English as a Foreign Language (EFL) educational training. Here, we make the assumption that EFL training improves a person's overall writing quality, and that surprisal value acts as a proxy for writing quality.

		Total		Syntactic		Lexical	
Topic	**Term**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Analysis	Term 1	6.34	3.32	2.37	1.86	3.97	3.24
	Term 2	6.28	3.30	2.34	1.85	3.94	3.23
Arg.	Term 1	6.24	3.29	2.34	1.85	3.90	3.23
	Term 2	6.15	3.36	2.28	1.85	3.87	3.24

Table 2: Means and standard deviations of total surprisal, syntactic surprisal, and lexical surprisal for *Analysis* and *Argumentation* essays

3.1 Corpus

We used the Uppsala Student English corpus provided by the Department of English at Uppsala University (Axelsson, 2000). The corpus contained 1,489 essays written by 440 Swedish university students of English at three different levels. The total number of words was 1,221,265, and the average length of an essay was 820 words. The essays were written on a broad range of topics, and their lengths were limited to be between 700-800 words. The topics were divided based on student education level, with 5 essay topics written by first-term students, 8 by second-term students, and 1 by third-term students.

To facilitate comparison, we chose similar topics from the first and second-term sets. We thus had two sets of essays. The first set consisted of *Analysis* essays which are written as a causal analysis of some topic, such as "television and its impact on people." The second set consisted of *Argumentation* essays where students argue for or against a topic or viewpoint. We further imposed the restriction that only essays written by the same student in both terms were selected. That is, if a student wrote an essay on a chosen topic in the first term, but not the second, or vice-versa, their essay was not considered. This selection resulted in 38 pairs of *Analysis* essays and 20 pairs of *Argumentation* essays across the two terms, for a total of 116 essays.

3.2 Computing Surprisal

We computed the surprisal value of each word in an essay by using a broad-coverage top-down parser developed by Roark et al. (2009). The parser was trained on sections 02-24 of the Wall Street Journal corpus of the Penn Treebank (Marcus et al., 1993). Essentially, the parser computes a word's surprisal value as the negative log-probability of the word given the preceding words using prefix probabilities. Thus, the surprisal value of the i^{th} word is calculated as:

$$\textsc{Surprisal}(w_i) = -\log \frac{\text{PrefixProb}(w_{1\ldots i})}{\text{PrefixProb}(w_{1\ldots i-1})}$$

Moreover, it decomposes each word's surprisal value into two components: syntactic surprisal and lexical surprisal. Syntactic surprisal measures the degree of unexpectedness of the part-of-speech category of a word given the word's sentential context. On the other hand, lexical surprisal measures the degree of unexpectedness of the word itself given its sentential context and a part-of-speech category.

For every essay, we measured the syntactic, lexical, and total (i.e., summed) surprisal values for each word. Subsequently, the averages of the three surprisal values were computed for every essay, and those means were used for further analyses. Henceforth, surprisal values for an essay refers to their mean surprisal values.

3.3 Results and Discussion

Table 2 reports the means and standard deviations of the three surprisal measures of the essays.[1] As can be seen, there seems to be a reduction in all three surprisal values across terms, and second term essays tend to have a lower mean surprisal than first term essays. To analyze these differences, we computed linear mixed-effect regression models (Baayen, 2008; Baayen et al., 2008) for the two essay categories. Each model included Term as a fixed factor and Student as a random intercept.

While our analysis shows that essays in the second term have an overall mean surprisal values less than than essays in the first term, these differences were not statistically significant. There are a number of factors that could have influenced these results. We made an assumption that only a single term of EFL training could significantly improve

[1] It is important to note here that these means and standard deviations are computed on mean surprisal values per essays and not surprisal values at individual words.

	Total		Syntactic		Lexical	
Score	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Low	6.22	0.39	2.46	0.22	3.76	0.29
Medium	6.10	0.34	2.35	0.17	3.75	0.26
High	6.09	0.28	2.27	0.14	3.82	0.24

Table 3: Means and standard deviations of total surprisal, syntactic surprisal, and lexical surprisal for the three different essay score levels

essay quality, and hence decrease overall surprisal values of essays. However, it is likely that a single term of training is insufficient, and perhaps the lack of a significant difference between surprisal values reflects no improvement in essay quality across the two terms. Unfortunately, these essays were not previously scored, and thus we were unable to assess whether essay quality improved over terms.

4 Experiment 2

In the second experiment, we directly examined whether surprisal values are related to essay quality by using a dataset of pre-scored essays.

4.1 Corpus

For this experiment, we used a corpus of essays written by non-native English speakers. These essays are a part of the Educational Testing Service's corpus which was used in the first shared task in Native Language Identification (Blanchard et al., 2013)[2] .

The corpus consisted of 12,100 essays, with a total number of 4,142,162 words, and the average length of an essay was 342 words. The essays were on 8 separate topics, which broadly asked students to argue for or against a topic or a viewpoint. Each essay was labeled with an English language proficiency level (*High*, *Medium*, or *Low*) based on the judgments of human assessment specialists. The distribution of the essays per score-category was: *Low* = 1,325; *Medium* = 6,533; and *High* = 4,172. In order to ensure an equitable comparison, and to balance each group, we decided to choose 1,325 essays per score-category, for a total of 3,975 essays.

4.2 Computing Surprisal

As in Experiment 1, for every essay we measured the syntactic, lexical, and total surprisal values for each word. We computed the averages of the three surprisal values, and used those means for further analysis.

4.3 Results and Discussion

Table 3 reports the means and standard deviations of the three surprisal values for every essay per score-category. We analyzed the differences between the means using linear mixed-effects regression models (Baayen, 2008; Baayen et al., 2008). Essay Score was treated as a fixed effect and Essay Topic was included as a random intercept. The results indicate that *Low*-scoring essays had a significantly greater mean total surprisal value than *Medium* or *High*-scoring essays. However, the difference in mean total surprisal values for *Medium* and *High*-scoring essays was not significant. On the other hand, for syntactic and lexical surprisal, the means for all three essay score levels were significantly different from one another.

We further evaluated the three surprisal values by performing a correlation test between them and the essay scores. Table 4 reports the output of the correlation tests. All three surprisal values were found to be significantly inversely correlated with essay scores. However, only syntactic surprisal obtained a correlation coefficient of a sufficiently large magnitude of 0.39.

A similar evaluation was performed by Attali and Burstein (2006) in their evaluation of the features used in ETS's *e-rater* system. Interestingly, the magnitude of the correlation coefficient for syntactic surprisal reported here is within the range of coefficients corresponding to *e-rater*'s features when they were correlated with TOEFL essay scores (see Attali and Burstein, 2006, Table 2). Granted, a direct comparison between coefficients is not recommended as the datasets used were different, such a finding is still promising. Overall, the results shed a positive light on the use of surprisal, specifically syntactic surprisal, as a feature for automated essay scoring.

Despite the promising pattern of our results,

[2]Copyright © 2014 ETS. www.ets.org

Dep Var	ρ	t-value	p-value
Total	-.15	-9.87	$< .001$
Syntactic	-.39	-26.53	$< .001$
Lexical	.08	5.35	$< .001$

Table 4: Pearson's R coefficients between the three surprisal values and the essay scores

they must be taken with a grain of salt. The dataset that we used did not contain the actual scores of the essays, and we had to work with broad classifications of essay scores into *Low*, *Medium*, and *High* score levels. A possible avenue of future work is to test whether these results hold when using finer-grain essays scores.

5 Conclusions and Future Work

We proposed the use of the *surprisal theory* to quantify sentence processing complexity for use as a feature in essay scoring. The results are encouraging, and warrant further evaluation of surprisal's effectiveness in determining essay quality.

One point of concern is that the relationship between mean surprisal values and essay scores is likely to vary depending on the general quality of the essays. Here, we used a corpus of essays written by non-native English speakers, and as such, these essays are bound to be of a lower overall quality than essays written by native English speakers. For example, consider the following, somewhat questionable, sentences chosen from the subset of essays having a *High* score:

(3) Some people might think that traveling in a group led by a tour guide is a good way.
(4) This is possible only if person understands ideas and concept.
(5) It is an important decision, how to plan your syllabus.

These examples suggest that even high-scoring essays written by non-native English speakers may not necessarily be flawless, and as such, grammatical acceptability may play a crucial role in determining their overall quality. Therefore, it is possible that for lower-quality essays, high surprisal values reflect the presence of grammatical errors. On the other hand, for better-written essays, moderate-to-high surprisal values may reflect structural variability, which arguably is preferable to monotonous essays with simpler sentence structures. Thus, it is likely that the relation between surprisal values and essay scores depends on the overall quality of the essays in general. For an equitable evaluation, further tests will need to determine surprisal's efficacy over a broader range of essays.

Another critical point is the choice of corpus used to compute surprisal. Whatever choice is made essentially dictates and constrains the grammar of the language under consideration. Here, we used the WSJ corpus and, thus, implicitly made an assumption about the underlying language model. Therefore, in our case, a good essay, i.e. one with a lower surprisal score, would be one which is stylistically closer to the WSJ corpus. Future work will need to investigate the role played by the underlying language model, with special emphasis on evaluating language models that are specific to the task at hand. In other words, it would be interesting to compare a surprisal model that is built using a collection of previous essays with a surprisal model that uses a broader language model.

Lastly, our evaluations were aimed at determining whether surprisal can be an effective predictor of essay quality. Further tests will need to evaluate how well the measure contributes to essay score predictions when compared to related approaches that rely on non-syntactic language models, such as n-grams. Moreover, future work will need to determine whether adding mean surprisal values to an AES system results in a performance improvement.

Acknowledgments

We are indebted to ETS for sharing their data with us, and supporting us through this project. This work would not be possible without their help. We are also thankful to the reviewers for their helpful and encouraging comments. The opinions set forth in this publication are those of the author(s) and not ETS.

References

Helen B. Ajay, P. I. Tillett, and Ellis B. Page. 1973. Analysis of essays by computer (AEC-II). *Final*

Report to the National Center for Educational Research and Development for Project, (8-0101).

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Margareta W. Axelsson. 2000. USE – the Uppsala Student English corpus: An instrument for needs analysis. *ICAME Journal*, 24:155–157.

Harald R. Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Harald R. Baayen. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *Educational Testing Service*.

Marisa Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Proceedings of the Workshop on Discourse Relations and Discourse Marking*, pages 206–210.

Alice Davison and Richard Lutz. 1985. Measuring syntactic complexity relative to discourse context. In David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 26–66. Cambridge: Cambridge University Press.

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

Scott Elliot, 2003. *Automated essay scoring: a cross disciplinary approach*, chapter IntelliMetric: From here to validity, pages 71–86. Lawrence Erlbaum Associates, Mahwah, NJ.

Kenneth Forster and Ilmar Olbrei. 1972. Semantic heuristics and syntactic analysis. *Cognition*, 2(3):319–347.

Stefan L Frank. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 1139–1144. Cognitive Science Society Austin, TX.

Peter C. Gordon, Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27:1411–1423.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 159–166, Pittsburgh, PA.

Gaurav Kharkwal and Karin Stromswold. 2013. Good-enough language processing: Evidence from sentence-video matching. *Journal of psycholinguistic research*, 43(1):1–17.

Jonathan King and Marcel A. Just. 1991. Individual differences in sentence processing: The role of working memory. *Journal of Memory and Language*, 30:580–602.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz, 2003. *Automated essay scoring: a cross disciplinary approach*, chapter Automated scoring and annotation of essays with the Iintelligent Essay Assessor, pages 87–112. Lawrence Erlbaum Associates, Mahwah, NJ.

Roger Levy and Frank Keller. 2012. Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Ellis B. Page. 1966. The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5):238–243.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 324–333. Association for Computational Linguistics.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Dan Slobin. 1966. Grammatical transformations and sentence comprehension in childhood and adulthood. *Journal of Verbal Learning and Verbal Behavior*, 5(3):219–227.

Adrian Staub. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Karin Stromswold, David Caplan, Nathaniel Alpert, and Scott Rauch. 1996. Localization of syntactic comprehension by position emission tomography. *Brain and Language*, 52:452–473.

Matthew J. Traxler, Robin K. Morris, and Rachel E. Seely. 2002. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47:69–90.