First-Year Composition as “Big Data”: Examining Student Revisions at Scale

Chris Holcomb
English Language and Literature
University of South Carolina
Columbia, South Carolina
holcombc@mailbox.sc.edu

Duncan Buell
Computer Science and E
University of South Carolina
Columbia, South Carolina
buell@acm.org

ABSTRACT

Approaching First-Year Composition (FYC) as a “big data” phenomenon, we have prototyped software to study revision in a large corpus of student papers and thus to address a question central to Composition and Rhetoric scholarship: “What role does revision play in students’ writing processes?” After running our program on a corpus of student writing, we see that our computational analysis challenges past research on revision and extends the methodological reach of Composition and Rhetoric to include “big data” analytics.

Keywords

first-year composition, revision, text analysis

1. INTRODUCTION

For many First-Year Composition (FYC) programs, revision is the centerpiece of their writing pedagogies. Students draft each major writing assignment, receive feedback from their peers and instructors, revise their papers based on that feedback, and submit all of their drafts and final versions at the end of the semester under a single cover (i.e., the writing portfolio). The assumption here is that students improve their writing through these multiple and guided revisions. However, given the number of papers students produce during a typical semester (it’s 9,000 to 12,000 at our institution), how can we know, at a program level and on a routine basis, what happens between all these first and final drafts? How often and how much do students revise, what specific features do they typically change, and do their revisions match, exceed, or fall short of the learning outcomes and more general expectations of the FYC courses?

In answering these questions, the scholarship on revision has been fairly consistent: students revise infrequently, and, when they do make changes to their papers, they typically focus on minor edits and surface errors [3, 5, 6, 11]. According to Bazerman, “students tend to revise essays shallowly,” focusing primarily on “phrasal adjustments and sentence correctness” [1, p. xii]. Arguing along similar lines, Sommers says that students typically “understand the revision process as a rewording activity”—that is, finding just the right word or eliminating lexical redundancies [11, p. 381]. Harr and Horning claim that while students occasionally “revise extensively,” they are “more likely to stick to surface correction and small changes” [4, p. 4]. All in all, and especially when compared to more experienced writers, students lack a robust approach to revision, one that includes revision strategies that extend beyond word- and phrase-level changes.

As valuable as this research has been in helping us understand and respond to student revision, it is limited in two important respects, limitations that Faigley and Witte acknowledge in their own and prior studies and that still seem applicable today. First, owing to the “complexity of the analysis” involved, researchers have restricted their studies to only a “small number of subjects” [3, p. 411]. Faigley and Witte, for instance, include only 18 subjects in their study while Sommers [11] includes 40, Horning [5] includes 9, and Treglia [13] includes 43. Second, while explaining the causes of revision, researchers focus too narrowly on the “skill of the writer” and thus ignore a range of other “situational variables” that contribute to revision or its absence ([3, p. 410]; see also [7, pp. 258-264]). In other words, “revision cannot be separated from other aspects of composing, especially during that period when writers come to grips with the demands of the particular writing situation.” Research that neglects these “situational variables” is “likely to be skewed” [3, p. 411].

Both these limitations involve problems of scale: too few subjects and too few variables considered. Towards overcoming these limitations as well as answering the question with which this essay begins (“How can we know what happens between all of these first and final drafts?”), we approached revision, and FYC more generally, as a “big data” phenomenon. More specifically, we built a corpus of first and final drafts from our students’ portfolios and developed software to process them. This software allows us to examine revisions in student papers, to explore correlations between these revisions and the situational variables that may influence them, and to perform both of these operations at scale. What we found differs considerably from past research: unlike students in other studies, ours rarely focused on minor edits and surface corrections; instead, when they did revise, their changes primarily involved deleting and, more frequently, inserting complete sentences. What this suggests more generally is that our students see revision not as a "rewording activity," but as a sentence deletion and insertion activity, treating their original drafts as fixed structures into which they plug or unplug not words, but sentences.

In the rest of this paper, we describe our data set, the program we developed to analyze it, and the results it produced. We conclude by outlining future directions for our project and how "big data" analytics informs that work.

2. DATA AND PROGRAMMING

FYC at the University of South Carolina is taught in about 150 (fall semester) and 120 (spring semester) sections, each with about 20-24 students who each write three or four draft and final papers, for a rough total of about 10,000 pairs of papers each semester. These are submitted to a content management system from which we download the papers. Earlier downloads have been manual; we have devised a system for a more automatic script for download. Most of these are submitted as dot doc or dot docx files, which can be turned into ASCII text with a Python program. Scripts and programs convert these to a standard file naming and clean the ASCII files of the various Unicode or nonstandard characters that would complicate later processing (smart quotes, em dashes, en dashes, ellipses, and so forth).

We do lose some data along the way. A small fraction of the papers are submitted in formats other than dot doc or dot docx, and at present we do not process these. Subsequent versions of our code may be able to make use of pdf, or Pages, or odt files, for example. We have not done that yet, though. We are at present drawing rather coarse conclusions from a corpus that is already large, and we would not expect students submitting pdf files, for example, to be statistically different as writers from students submitting doc files. We remark that each paper averages a little less than 10,000 characters, so that 10,000 pairs of papers is only about 200 megabytes of data each semester. This is substantial enough to require some management and organization but is by no means problematic; the quantity of data is less a management problem than is separating the files into class sections, keeping track of which papers come from which standard assignment, etc.

Similarly, we do admit that our "cleaning" process could introduce corruptions in ways that might make some detailed analysis difficult or impossible. Again, however, we do not imagine that a few such character changes, if done consistently to draft and final, would change the overall analysis currently being done.

To analyze the data set, we used Python programs (only about 2500 total lines of code) together with the Natural Language ToolKit (NLTK) [10] and limited use of the Stanford NLP package [12] for processing the data. The NLTK routines were used primarily for breaking the documents into sentences and paragraphs. Having broken both draft and final versions into sentences, we used edit distance, which is a standard measure of similarity [8, 9, 14], to compute the "similarity" between sentences in draft and final versions. Using this measure, we were able to quantify the "distance" between draft and final sentences by looping through those sentences and aligning pairs of sentences whose distance falls within a gradually increasing threshold. On its first pass, our program aligns sentences with an edit distance of zero (no difference between the sentences). On its next pass, it looks in between aligned sentences and aligns in the intervening space the pair of sentences with the smallest pairwise distance. And then it does so again, and then again until it reaches a point where the smallest pairwise distance exceeds 50% of the worst-case distance (the worst case is the distance achieved by deleting each word from the draft sentence and then inserting each word from the final sentence). We chose this as the program's stopping point after visually inspecting scores of sentences and determining that, beyond 50% of the worst-case distance, the program would likely be aligning two different sentences.

3. RESULTS – STUDENT REVISIONS

When we ran our program on a test corpus, the results surprised us because they differed from what the scholarship on revisions says we should be seeing. In other words, unlike other studies of revision which found that students typically focus on minor changes in diction, punctuation, and grammar, we found that when our students revised, the majority of their changes involved deleting and, especially, adding sentences. Consider Figure 1. This stacked bar chart shows the percentages of unchanged sentences (light blue), lightly edited (red) sentences, sentences deleted from the draft or inserted into the final (green), and heavily edited (purple) sentences. By far, the largest portion of sentences fall into the unchanged category. That is, the bulk of student writing survives unaltered from first to final draft. When we consider text that students actually changed, the bulk of those changes involve deleted and inserted sentences, followed by heavily edited and then lightly edited sentences. So while students do edit their text to some extent, their primary revision strategy involves treating their drafts as relatively fixed structures into which the plug or unplug, not words, but complete sentences.

We remark that almost no great shifting of text occurs in our student papers. Our alignment algorithm is somewhat naïve in that it anchors the initial alignment to unchanged sentences and then continues with that alignment. Clearly, if entire paragraphs were moved, our algorithm would work poorly and we would see anomalous results for those papers. In fact, we see this happening in only a very small fraction of the papers.

4. FUTURE DIRECTIONS

Our next steps involve explaining the revision practices we are observing. In other words, having gained a better sense of what happens between all those first and final drafts, we now plan to explore why it happens. Toward that end, our work will continue to be informed by big-data analytics. What do we mean by this? The phrase "big data" refers to a large data set and to a collection of computational techniques for analyzing it. Both meanings apply to our project. Our corpus will eventually consist of tens of thousands of papers, a size much too large for humans to analyze in detail, so we will use natural language processing to capture and quantify features that, taken together, offer a linguistic pro-

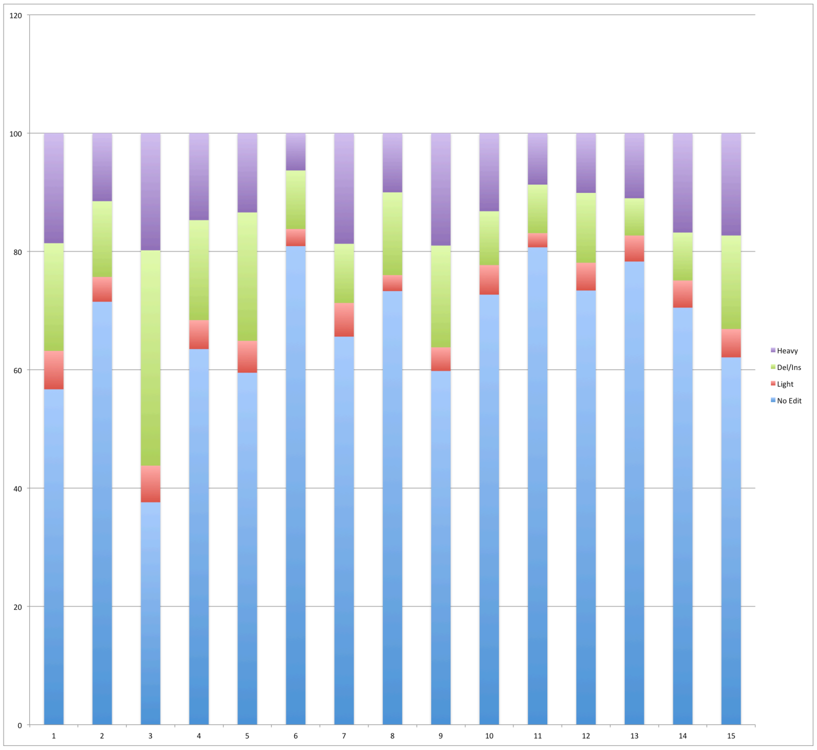

Figure 1: Fractions of sentences unedited, edited, or inserted/deleted in a sample of 15 sections.

file of each paper. Once those features are quantified, we will employ other computational techniques (e.g., linear regression and cluster analysis) to search for correlations (and other patterns) among the papers in the corpus.

The program we have already developed supplies us with a relatively finely tuned computational model for revision. Equipped with this model, we have multiple paths forward, and, in the spirit of big data, we will explore as many of them as we can—including, but not limited to, the following:

- Turn each draft-final pair into a four-dimensional vector (i.e., the frequency of unchanged, inserted, deleted, and edited sentences) and use cluster analysis to see if those pairs fall into any groupings. If they do, then look within and across those clusters to see if other written features or situational variables correlate with those groupings.
- Compare the aggregate of deleted sentences (which students presumably thought were bad) with the aggregate of inserted ones (which they presumably thought were better).
- Measure sentence complexity trends in our corpus against those found in other genres, using a distinction between clausal complexity (a characteristic of spoken discourse) and phrasal complexity (a characteristic of academic writing) ([2]). Do students' sentence structures align more with spoken discourse or with academic writing?
- Examine students' use of "evidentials" and compare them against their revision scores. The term "evidentials" refers to linguistic features that signal a writer's source of information and his or her perceptions about its reliability, including reporting verbs (e.g., "say," "think," and "argue"), adverbs (e.g., "actually", "probably", and "certainly"), and modals (e.g., "could," "should," and "must").
- Collaborate with other institutions that have assembled similar corpora of student writing and run their data through our program. By seeing results produced by other institutions, we will gain a better sense of whether the sentence deletion and insertion practice we observed in our corpus is a more general trend or a phenomenon peculiar to our FYC program and its curriculum. Either result would be of interest: if the data from other institutions looks like the USC data, then perhaps we have identified a broad characteristic of student writing. If that data is different, then we will have new questions to ask to determine why one group of students revises differently from the other.

5. CONCLUSION

Thus far, our project addresses one of the limitations Faigley and Witte point out in revision research: that is, rather than restricting our research to "a small number of subjects," we are able to examine revision patterns in tens of thousands of student papers at one go. In doing so, we have unearthed trends in student writing that past studies of revision fail to predict—namely, the prevalence of the sentence deletion and insertion trend. As we move forward with our project, we will address Faigley and Witte's second limitation: too few "situational variables" considered. Having quantified revision, we can now explore correlations between it and dozens of these variables, including grades, student major, teacher feedback, gender, and a host of features in the co-text of student revisions (e.g., sentence complexity, lexical sophistication, metadiscourse, etc.). As we do so, we will continue to enrich our understanding of what happens between all of those first and final drafts.

6. ACKNOWLEDGMENTS

We acknowledge and offer thanks for the financial support for student assistants from the Center for Digital Humanities at the University of South Carolina, and the work of graduate student Gerald Jackson and undergraduate students Brian Flick, Chelsea Reeser, Sam Watson, and Ming Wong at various stages of this research.

7. REFERENCES

[1] C. Bazerman. Preface. In A. Horning and A. Becker, editors, *Revision: History, Theory, and Practice*, West Lafayette, IN, 2006. Parlor Press.

[2] D. Biber, B. Gray, and K. Poonpon. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45.1:5–35, 2011.

[3] L. Faigley and S. Witte. Analyzing revision. *CCC*, 32.4:400–414, 1981.

[4] C. Haar and A. Horning. Introduction and overview. In A. Horning and A. Becker, editors, *Revision: History, Theory, and Practice*, West Lafayette, IN, 2006. Parlor Press.

[5] A. Horning. *Revision Revisited*. Hampton Press, Inc., Cresskill, NJ, 2002.

[6] A. Horning and A. Becker, editors. *Revision: History, Theory, and Practice*. Parlor Press, West Lafayette, IN, 2006.

[7] J. Jones. Patterns of revision in online writing: A study of Wikipedia's featured articles. *Written Communication*, 25.2:262–289, 2008.

[8] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[9] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similaries in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[10] NLTK.org. Natural language toolkit, 2016. `http://www.nltk.org`.

[11] N. Sommers. Revision strategies of student writers and experience adult writers. *CCC*, 31.4:378–388, 1980.

[12] Stanford Natural Language Processing Group. Natural language processing package, 2016. `http://nlp.stanford.edu`.

[13] M. O. Teglia. Teacher-written commentary in college writing composition: How does it impact student revisions? *Composition Studies*, 37.1:67–86, 2009.

[14] R. Wagner and M. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21:168–178, 1974.