March 2, 2012

Source Text Re-use in Engineering Thesis and Dissertation Literature Reviews: A Comparison [Transcript]

Edward J. Eckel, Mr., Western Michigan University

Available at: https://works.bepress.com/edward_eckel/6/
Source Text Re-use in Engineering Thesis and Dissertation Literature Reviews: A Comparison


Introduction

Good afternoon. My name is Ed Eckel and I am the engineering and applied sciences librarian at Western Michigan University Libraries, where I have been since 2006. The title of my talk today is "Source Text Re-use in Engineering Thesis and Dissertation Literature Reviews: A Comparison." What I mean by the term “source text re-use" is what is basically called copying, plagiarism, textual appropriation (Shi 2006, 2010), and textual borrowing (Chandrasoma 2004) in academic writing. I will use all four terms interchangeably today. I want to talk about how I became interested in this topic and what led me to conduct this study. I will give you an overview of my project methodology and then basically dive into the preliminary results. Finally, I will briefly discuss the implications of my study, while also offering some suggestions for engineering graduate programs.

Around the time that I became the engineering librarian at Western, I became aware of a big scandal taking place at the Russ College of Engineering at Ohio University. A master’s engineering student at Ohio, searching through theses in the college library for inspiration for his master project, discovered that a number of engineering master's theses and dissertations contained material copied from other theses and dissertations at Ohio University, as well as from books and scholarly articles (Tomsho 2006). This was written about in the Wall Street Journal (Tomsho 2006), The Chronicle of Higher Education (Luegenbiehl et al. 2006; Wasley 2006, 2007), ASEE Prism Magazine (Grose 2006), and even talked about on NPR.

About six months after this incident, I was in the midst of another research project, a citation analysis of engineering master’s theses and dissertations at my university, when I inadvertently stumbled upon a completely plagiarized thesis. In the course of the ensuing investigation (in which other instances of apparent plagiarism were discovered), I started to wonder whether the incidents at Ohio and WMU were unusual clusters, or if this kind of
thing was common in engineering graduate writing. I subsequently decided to use a modified version of the Google search strategy that I had used in the plagiarism investigation at my own institution to do a research study on source text re-use or copying in engineering theses and dissertations from the ProQuest Dissertations and Theses database.

In the course of the literature review that I conducted in preparation for this project, I discovered a paper by McCullough and Holmberg (2005) in which they used the Google search engine to search strings of text from 210 master's theses in various disciplines from the Web. They found what they called "suspected plagiarism" in 27%. What was most interesting to me about their study was that they found the highest incidence of plagiarism in engineering theses - 43% in computer/science engineering and 38% in mechanical/aerospace engineering. These results were intriguing, although methodological problems in their study, as well as a lack of details regarding exactly how much and what type of material was copied, meant that it was hard to know whether these results were really valid. However, their study did seem to suggest that plagiarism in engineering graduate writing could be a larger pattern, not just a local issue.

Pilot Study

I initially conducted a pilot study in which I searched strings of text from the literature review sections of a random sample of 100 English language engineering master's theses from 2007 in the Google search engine. The results from that pilot study were published in the journal Science and Engineering Ethics last year (Eckel 2011). That study appeared to indicate that, while not at crisis proportions, enough textual appropriation occurs in master's level engineering writing to cause some concern. I will briefly touch upon the results of that study in a little while in the context of the results of my current study.

Doing that study helped me develop a more rigorous methodology for this study.

Methodology

I have two basic research questions for this study:

- Do engineering master's theses have longer verbatim text matches than doctoral dissertations? In other words, are they copying longer sections from their sources?
- Do engineering master's theses have more verbatim text matches than doctoral dissertations?

Based on my experience with the copying in engineering theses and dissertations at my institution, my basic hypotheses are yes for both questions. Master's students seem to have more difficulty using source text appropriately in their scholarly writing.
I used the *ProQuest Dissertations and Theses* database for easy access to full text PDFs of the documents. I did a random sample of 125 English language engineering theses and dissertations completed in 2009. My ultimate sample will include 150 of each.

I focused on the literature review sections, Googling 7 word strings of text (basically the first seven words from every fifth line) for a total of thirty searches per thesis/dissertation. For each source I found, I recorded the source title and type (journal, book, etc.), URL, # of words copied, date of publication, and whether or not a source for that string was cited. Often if a source was cited, it was a different source than the one I found on Google. But I was more concerned with seeing if any sources were cited for copied material than if I was actually finding the ORIGINAL source of the copied textual material. Long textual matches were counted for all searches that they covered. (In other words, a long text match usually would cover textual material from several adjacent sections, and would therefore count as multiple verbatim text hits.)

Based upon my two research questions, I looked at two measures from that data: the longest verbatim copied text string per thesis and dissertation, and the percentage of searches out of the 30 total per thesis and dissertation that contained verbatim hits.

**Results**

a) **Longest Verbatim Text String**

![Graph of Length of Longest Verbatim String (per Thesis/Dissertation)](image)

![Bar graph showing the length of longest verbatim strings for 2009 Master's Theses and 2009 Doctoral Dissertations.](image)
Figure 1: Length of Longest Verbatim String (per Thesis/Dissertation)

Figure 1 shows the distribution of the longest verbatim text string for theses and dissertations. The X-axis shows the number of verbatim words copied in a row (in ranges of 20 words) and the Y-axis shows the number of theses and dissertations in each range. I have compressed the range at the extreme right (outliers) to make the chart more readable on the overhead. No values are omitted here.

When you look at this figure, there does not appear to be a big difference between theses and dissertations in terms of the length of strings copied. There are a few more dissertations at the lower end (shorter strings) than theses, but otherwise it is difficult to see much difference between the two groups of writers.

<table>
<thead>
<tr>
<th>MS Summary</th>
<th>Mean</th>
<th>Number (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean:</strong></td>
<td>103</td>
<td>125</td>
</tr>
<tr>
<td><strong>Mean (excluding outliers and zeros):</strong></td>
<td>69</td>
<td>95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PhD Summary</th>
<th>Mean</th>
<th>Number (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean:</strong></td>
<td>52</td>
<td>125</td>
</tr>
<tr>
<td><strong>Mean (excluding outliers and zeros):</strong></td>
<td>60</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 1: Comparison of the Mean Longest Verbatim String

Table 1 compares the mean longest verbatim string for theses and dissertations. At first glance, the mean longest string for the master’s theses looks very different than the mean longest string for the doctoral dissertations, 103 words in a row versus 52 words in a row respectively, an apparent 100% difference. However, if you exclude the extreme values of each group (4 theses and 2 dissertations whose verbatim string lengths were much longer than any others in the data set) and the zero values (people who copied nothing) and recalculate the mean for the remaining theses and dissertations, you find that the mean longest string length for master’s theses drops to 69 words and the mean string length for dissertations rises to 60 words. These values are suddenly very close.

An independent samples t-test for difference of means on the original means for the two groups was significant only to the .100 level. Hence with a 10% chance that the difference
seen here between the means is due to chance variation in the two populations, I have to accept the null hypothesis. There appears to be no significant difference between the two populations in the length of the longest strings that they copy.

Figure 2: Length of Longest Verbatim String per Thesis

2007 and 2009 Master’s Theses

Figure 2 displays the data from my pilot study which analyzed engineering master’s theses completed in 2007 and the current study (which looks at theses from 2009). There was a methodological difference between these two studies in how I gathered the data. For the 2007 master’s theses, I scanned the pages of the literature reviews looking for suspect strings, based on language differences between those strings and the surrounding text, an extremely sophisticated writing style, and other intuitive factors. This was an ad hoc methodology that I dropped for the current study. For the 2009 theses, I simply started with the first line of the literature review sections, and searched the first seven words of every fifth succeeding line in Google, for a total of 30 searches.

However, despite the methodological differences between the pilot study and the current study, the data on the length of the longest verbatim string per thesis appears strikingly similar between the 2007 and 2009 engineering master’s theses, indicating that the two groups of students exhibited similar patterns in the length of the strings they tended to copy.
b) Percentage of Verbatim Hits (/30)

A different picture emerges when one looks at the percentage of verbatim hits obtained out of 30 searches per thesis/dissertation (Figure 3).

![Percentage of Verbatim Hits (Out of 30 Searches)](image)

**Figure 3: Percentage of Verbatim Hits (per Thesis/Dissertation)**

The X-axis displays the percentage of verbatim hits found out of the 30 searches done per thesis and dissertation in Google. The Y-axis shows the number of theses and dissertations in each percentage category. Looking at the data in this figure, one can see that there are more doctoral dissertations (the red lines) at the lower percentages and fewer at the higher percentages. The highest percentage of verbatim hits found for dissertations was 60%, whereas the highest percentage for theses was 100%, a thesis in which every single Google search found a verbatim source. In fact, above 60%, all the data values are for master's theses.

An independent samples t-test for difference of means on the values for master's theses and doctoral dissertations on this measure was significant to the .001 level. So master’s students copy significantly more strings in their literature reviews than the doctoral students.

From these very preliminary inferential statistics, it appears that, while the percentage of theses and dissertations for which I found no copied strings (26 versus 33 respectively) and the length of the strings copied is very similar between the two groups, the master’s
students definitely are copying more strings than the doctoral students, indicating that the master’s students have a slightly bigger problem with source text appropriation.

One can speculate as to why master’s students and doctoral students seem to copy similar lengths of verbatim strings. Perhaps the populations of the two groups are actually quite similar in terms of practical writing experience in the discipline, years of schooling, and facility/fluency with the English language. Remember that many engineering students enter either master’s or doctoral programs right out of their undergraduate years, and hence are in similar age ranges. However, despite the similarities in terms of the length of strings copied, the doctoral students seem to do less copying overall than the master’s students, which may indicate that they either are more sensitive to the issue of plagiarism, get more feedback from advisors about the necessity of avoiding it, or that their dissertations get closer scrutiny (thereby exposing potential problems in advance) than master’s students.

However, despite the differences between the two groups, a still sizable number of both master’s and doctoral students engage in questionable levels of source text copying, both in terms of the length of the strings copied and the number of strings copied (see table 2). 31% of master’s engineering students copied strings of text over 60 words in length. 20% of doctoral students did the same. 21% of master’s engineering students copied more than 25% of their literature review section from their sources (based on the percentage of verbatim hits out of 30 searches), whereas only 7% of the doctoral students did the same.

<table>
<thead>
<tr>
<th></th>
<th>MS</th>
<th>Doctoral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over 60 words in row</td>
<td>39 (31%)</td>
<td>26 (20%)</td>
</tr>
<tr>
<td>Over 25% literature review</td>
<td>26 (21%)</td>
<td>9 (7%)</td>
</tr>
</tbody>
</table>

**Table 2: Overall Patterns of Copying – Master’s vs. Doctoral Engineering Students**

Such patterns of textual appropriation are problematic at best and could get these students in trouble, whether in graduate school or beyond if they continue with this pattern.

**Recommendations**

I would like to make some tentative recommendations regarding what might be done to mitigate or prevent this kind of source text re-use among engineering graduate students. First, I would like to share these results with engineering faculty at my university, and try to start a discussion regarding graduate student writing abilities, particularly in light of
past problems with plagiarism. I think such discussions might focus in particular on master’s students.

Engineering and applied science librarians might collaborate with writing programs to push for better instruction in synthesis skills, preferably at the beginning of graduate programs. Various writing composition researchers have come up with innovative ways to expose students to disciplinary conventions while also teaching them how to write from sources for synthesis. For example, Barks (2001) discusses giving students focused practice with borrowing relatively generic disciplinary source language while not crossing the line into inappropriate text re-use or plagiarism. Finally, I think it would be a very good idea for graduate programs in engineering to develop explicit guidelines on using boilerplate technical descriptions, definitions and mathematical expressions, since, in my experience, these are often used verbatim by engineering graduate students, with and without attribution. Students need help figuring out what is appropriate and what is not. Engineering librarians, working in collaboration with university and college writing centers and graduate advisors, can help them with this.

Whether engineering faculty will take us academic librarians up on such initiatives is another story.

**Future Work**

As mentioned in the methodology section, I hope to finish collecting data on a full sample of 150 engineering master’s theses and 150 doctoral dissertations, and then reanalyze everything, as well as perhaps looking at other measures from the data.

**Conclusion**

The thing that I find interesting about this data is that we sort of assume that graduate students are following the basic conventions of academic writing. But when you look at what they do, a fair number don’t. Based upon this data, can we say that there is a crisis in engineering graduate writing? No I don’t think so. However, while the examples I show here, with few exceptions, are relatively minor, it is obvious that some master’s and some doctoral students still struggle with issues of textual appropriation. I think the bigger point to take away from this study is that this is what we can realistically expect from novice writers in the field of engineering, based upon the level of writing instruction and practice they’ve received up to this point in their educational journey.

So I think getting too judgmental about what graduate students are doing without teaching them what they need to know is kind of pointless. Making reference to the moral and ethical perils of research misconduct and academic dishonesty or recommending the adoption of campus honor codes (McCabe et al. 1999, 2001) misses the main issues at the
core of much of this text re-use. Conventions of source text use and attribution are much more complex than students are usually told. You don’t teach someone to avoid plagiarism by pointing your finger at them, telling them to not to plagiarize and to always cite their sources.

Based upon the literature on novice writers, particularly that which focuses on ESL writers (Barks 2001; Howard & Robillard 2008; Pecorari et al. 2001), and on studies like my own (Eckel 2011), I think that the only way we can expect students to "avoid plagiarism" is by teaching them to write from sources for synthesis. You teach them by showing them (with examples) what’s acceptable in the discipline and what is not. Students need actual answers to questions like: What kinds of text strings can be re-used without consequence? What should be either paraphrased or put in quotation marks? Finally, students must be given practice writing from sources over and over again until they master the skill.

If we want to avoid the patterns of textual copying seen here, this kind of instruction may need to become more integral to graduate programs in engineering, if not other disciplines as well. Otherwise we risk sending these people out into academia and the working world where they will perpetuate this kind of copying.

Thank you.

References


