Electronic Plagiarism Checkers: Barriers to Developing an Academic Voice

KATHLEEN GILLIS, SUSAN LANG, MONICA NORRIS, AND LAURA PALMER
TEXAS TECH UNIVERSITY

RECENTLY, WE EMBARKED UPON A LARGE SCALE EXAMINATION of two popular electronic plagiarism checkers—Turnitin.com (Tii) and SafeAssignment (SA). Two specific events encouraged this effort. The first was an invitation from our assistant vice provost to participate in an upcoming university roundtable discussion that sought to answer the question "Should our campus purchase a site license for plagiarism detection service and, if so, which product would best meet our needs?" Second, our university was revising its writing intensive criteria, and faculty who taught these courses were interested in finding ways to enhance students' use of writing as a tool for learning while not increasing the amount of time they had to spend assessing that writing. Admittedly, none of us were fans of plagiarism detection applications; as is the case with many faculty members, our attitudes toward these applications had been formed after only limited contact with them. To combat this bias, we chose to examine the reports generated by each application after submitting a total of 400 freshmen essays to the two applications under consideration.

Why freshmen essays? First, decisions about writing programs, whether they be first-year or full WAC/WID programs, must reflect local conditions. In this case, the First-Year Writing Program at Texas Tech University was in the process of moving toward a WAC/WID emphasis with hopes that this would constitute the first step of instituting a four-year writing program in the College of Arts and Sciences. It was important to learn how each of these plagiarism-detection systems interprets writing from students who are currently engaged in a WAC/WID-like First-Year Writing Program. Second, we felt we had to move beyond anecdotes and conduct a more robust study, one that actually involved the submission of a large number of documents. While many studies have tested the accuracy of the programs and their ability to deter or prevent plagiarism, the samples were small in number, ranging anywhere from two to 150 drafts

DOI: 10.37514/WAC-J.2009.20.1.04

(Braummoeller and Gaines 2001, Purdy 2005, Marsh 2004). Thus, the results could not assist us in our effort. We believed that the results obtained from testing on a larger scale would provide us with important insight as to how these applications may or may not impact WAC/WID-based pedagogy. The results suggest that plagiarism detection applications are not productive tools for WAC instructors as the applications' approach to writing is inconsistent with WAC pedagogy. That is, in lieu of good pedagogy, the applications often penalize students for doing exactly what we want them to do: learn the basic language structures used by people who are writing about a common topic in a given discipline.

We think it imperative, then, that both WAC administrators and faculty teaching in WAC programs know what these applications do with actual texts and in what contexts the use of such applications may or may not be beneficial to student learning. To that end, we'll start by providing a brief description of what each program does and how each represents its findings to instructors and students. We will then discuss our methods for testing both applications and analyzing the data before moving into a discussion of our qualitative and quantitative analyses. We will conclude by examining how basic tenets of WAC pedagogy and these applications conflict, and by considering in what scenarios, if any, these applications should be used.

A Quick Description of the Applications and the Process

As of fall 2007, both the Tii and SA systems allow assignments to be submitted in multiple electronic formats. Within approximately 10 minutes, the programs return a score—as a percentage—that reflects the amount of material in a text that each system has determined matches a source on the Internet or in its databases. Tii and SA have a default mode for evaluating the originality of texts; the default mode produces the originality score seen on the main course screen of the program. As an example, a text with a score of 12% means that 88% is original content, while 12% of the text could be derived from other sources.

This numerical score, called an "Overall Similarity Index" by Tii and an "Overall Matching Index" by SA, is also color-coded to provide instructors with additional meaning about the results; papers may score in the green, yellow, orange and red range and this color will appear next to the numerical score. But the percentages and the scores don't always seem to be telling the same tale. Green, for example, which generally has a positive connotation in the U.S, indicates a low threat; therefore, papers scoring in the green range may be seen as innocuous. However, green-coded reports could indicate a score that is anywhere from 0% to 24% for potentially unoriginal work in both Tii and

SA. Having nearly 500 words of a 2,000-word student assignment matched to at least one other source in the application's database seems far more problematic than a 5% score, or 100 words, of potentially unoriginal material.

The fact remains that both the percentage and color-coded results must be read and interpreted by an instructor or administrator to determine whether the score is actually cause for concern or further action.

Methods and Results

To prepare for the roundtable, we completed an expert review of the functionality of these two systems by inputting 200 texts from Texas Tech University's First-Year Composition database into Tii and SA. The results from that initial round of testing were used in our roundtable in November. Shortly thereafter, we replicated our study with another 200 texts from the same database. The following section describes the data and our sampling methodology as well as our quantitative and qualitative results.

Sample Texts

The texts for both phases of the study were extracted from Texas Tech University's First-Year Composition database, which contains all assignments submitted by every first-year writing student since 2002. For the initial sample, texts were selected randomly across all sections of ENGL 1302, Advanced College Rhetoric, Fall 2006. The assignment description—a 2,000-word research paper with 8-10 sources—was standardized for all sections as part of the curriculum.

Because all texts in the First-Year Composition database are tagged with a serial number, we were able to generate a list of random serial numbers via a SQL query. In total, 200 texts were extracted and then two of us took one-half of the texts and entered them into both Tii and SA to compare how each system evaluated the same content. During this first phase, we experienced some difficulties learning each of the applications being tested; consequently, 44 texts of the original 200 were eliminated from the study because of technical difficulties, leaving us with 156 texts.

Our second sampling of texts was extracted from the Spring 2007 sections of ENGL 1302. Although some details of the assignment description were different, the core requirements to produce a 2,000-word submission with 8-10 sources were the same. Again, the database was queried, 200 papers were randomly sampled, and 100 each were provided to two of us, who submitted them to both Tii and SA for evaluation. In this phase, all 200 texts were included in the results. We'll first discuss results for the initial sample of 156 texts (Phase One) and then of the entire set of 356 texts (Phase Two).

Phase One

Using the initial sample of 156 texts, we compared the numerical and color-coded scores produced by Tii and SA on those texts. Initially, our working hypothesis was that each program, when given the same text, would produce a similar score. Our hypothesis was based on the corporate literature from Tii and SA that indicated web pages, student papers, scholarly sources, proprietary databases, as well as commercially-available newspapers and books would be used as the sources for comparison. While we knew the sources used by each program would not be identical, it made sense that there would be overlap in areas such as websites and news media. We expected to see some small variations in the scores—around 2% to 3% in most cases.

Our results on the first data set of 156 texts immediately refuted our initial hypothesis that Tii and SA would produce relatively similar originality scores. A preliminary glance at the originality scores indicated that Tii and SA were not, in most cases, producing a similar score for the same paper. Variations in the originality scores between the two applications commonly ranged from lows of 4% up to differences of 15%; some scores varied more than 20%. Of the 156 texts, the average difference in originality scores between the two programs was 9%; this turned out to be statistically significant where $p < 0.001$.[1]

We found that the originality scores clustered most heavily in the 0%-25% range but that, as per above, the scoring variations were perplexing. SA indicated that 61 texts of the 156 received a zero; this meant all of the content in these texts was original. However, the results from Tii were quite different—only 2 texts received a score of zero. An originality score of 0% in SA could result in a score of 7% in Tii. It quickly became obvious that we would need to know why the variation occurred and whether or not one of these applications was actually more accurate than the other in its detection of potentially unoriginal material.

Each program's overlay of a color-coded scale for the originality scores also proved to be enigmatic for us. Most of the originality scores in both programs were color-coded as green; however, in SA, scores under 10% were coded as white. Should instructors view papers scoring in the green or white range, which implies that the texts had little or no unoriginal material, as automatically acceptable? Perhaps. Yet, even 5% of a 2,000-word paper is 100 words. We wondered how many administrators or instructors would see this as appropriate, and how much time faculty would spend confirming or rejecting the results produced by Tii or SA. Additionally, we wondered if there was any

[1] A non-parametric test, the Wilcoxon test, was used as the data was not representative of the normal curve.

educational value to these colors and numbers for either students or faculty, or if they served as a ploy to divert attention away from the text itself and back to the application results—in short, it seems plausible that the applications become more important than the writing.

Phase One: Qualitative Results

To determine why the variation occurred in the results, we randomly selected twenty of the first 156 sample texts for further examination. In these texts, we looked for patterns in what was marked as "unoriginal material." In Tii, we used the function to exclude both quoted and bibliographic material (recall that the 2007 version of SA did not have a comparable function for excluding the bibliography). To compensate, we manually excluded all properly quoted material and bibliographic information when examining texts in SA. Both systems highlight portions of the text that are deemed "unoriginal." The marked text can be anywhere from a short phrase (e.g. 3 – 4 words that may be separated by an article or preposition) to a paragraph or more in length. (SA is more likely to mark complete sentences than Tii.) We also noted if the material was identified as matching a

Student source—another student's paper submitted at either the host institution or another institution.

Publication source—Internet only; these publication sources are varied and can include news websites, organizations and others.

We should note an important point here: The two categories, above, only identify (sometimes incorrectly) where any "unoriginal material" may have come from. At most, the underlying message that the student receives from the originality report generated is "don't take material from other sources." If the material wasn't deliberately taken from other sources, the report provides no actual instruction or guidance to either student or instructor about whether or not to revise the draft.

This would seem to leave us at a dead-end, unless we ask the question, can the report serve an instructional purpose if we examine *what* the marked text rhetorically represents? That is, what is the context of the marked material? After all, as has been well-documented, even the most sophisticated text-mining software cannot read for context. Our examination of the 20 selected texts revealed that much of the material marked by one or both applications could be described by one of the following categories: topic term, topic phrase, commonly used phrase, jargon, and citation error. In fact, our analysis of these 20 texts revealed that approximately 70% of the text marked

by TII and 83% of the text marked by SA fell into one of the first four categories listed below—none of which indicates plagiarized texts.

- **Topic term**: Short phrases which reflect the subject of the essay. Examples: "the top ten percent rule," "global warming," or "date rape."
- **Topic phrase**: Topic term plus a word phrase. These are usually not quite a complete sentence. Examples include "the dangers of date rape," "students in the top ten percent," and "global warming is a serious problem."
- **Commonly used phrase**: Phrase that could be used in multiple contexts. The phrase is not tied specifically to the topic of one paper. Examples: "Children spend the majority of their day;" "music can be used as." In fact, frequently the topic of the source identified by the system does not match the topic of the essay. For instance, a list of symptoms used in a paper on obesity was flagged as matching a Web site for Viagra.
- **Jargon**: Words or phrases that are tied to a specific topic. Examples: the names of organizations such as PETA in discussions of animal testing; specific terminology such as rohypnol when discussing date rape.
- **Citation errors**: Instances of poor paraphrasing, failure to properly punctuate titles, or other errors in citing material. Of the categories identified, this is the only one that could potentially be labeled as plagiarism. However, we specifically did not try to identify intent in this category.

Thus, in reviewing our results from the data set of 156 papers, we identified some key trends in how Tii and SA produce their originality results. We knew that in each program's default mode, the results for the same paper could be quite different, and that often what was marked as unoriginal material did not fit our university's definition of plagiarism. More often than not, the marked material represented an attempt by the student to use the conventions of academic writing in his or her essay—exactly what we want to see our first-year students doing. Armed with these results, we participated in the roundtable, where our results were discussed with great interest by the approximately 75 administrators, faculty, and student participants. The attention generated by our initial results led to Phase Two of the study, described in the following section.

Phase Two

Following the roundtable, we decided to sample another 200 texts from the Texas Tech First-Year Composition database to see if we could replicate our results and extend our understanding of these applications. This section discusses the results of the full data set of 356 papers the 156 original and 200 additional papers.

As with the first sample, a sizeable majority of the texts in our combined data set contain 25% or less of material derived from other sources. If we consider scores of 25% and under as falling in the green range, 85% of the papers assessed by Tii and 93% of the papers assessed by SA appear to be low threats for unoriginal content.

Table 1: Tii and SA Index Scores Distribution

	Percentage of Unoriginal Material				
	≤ 10%	11-25%	26-50%	51-75%	76-100%
Tii	150 papers	152 papers	46 papers	6 papers	2 papers
SA	277 papers	55 papers	18 papers	4 papers	2 papers

Next we examined the average scores produced by each program. The average originality score across all 356 texts in Tii was 16%; this means that across 356 texts, the entire sample fell into the low or green category. In SA, the average originality score across all texts was 8%. The fact that the average of all scores in both applications fell into the seemingly innocuous 'green zone' was one area of interest for us because it suggests that students were using unoriginal material correctly. This could also indicate that students were in the process of becoming more familiar with the ways in which academics represent knowledge.

The final phase of our analysis focused on a subset of the second sample to determine if the qualitative results we found in the first sample were replicated in the second.

Phase Two: Qualitative Results

In order to ensure that our qualitative results from the first data set were reliable, we decided to expand our analysis. We repeated our qualitative examination on an additional 20 texts from the second data set. When looking at the results from all 40 texts, we found that:

- Tii flags, on average, material belonging to 6 other sources per every 2,000 word draft.
- Of those 6 sources, approximately 4 are student sources (same or other institution) and 2 are publication sources.
- SA flags, on average, material from 2 sources in each draft. These are almost always publication sources.
- Tii flagged 245 instances of unoriginal material.
- SA flagged 22 instances of unoriginal material.

We also found that in the 245 instances of allegedly unoriginal material flagged in TII, only 24% of these could be classified as citation errors. The remaining 76% of the material flagged was not the result of intentional or unintentional plagiarism. In fact, 40% of the material flagged in Tii can represent commonly used phrases such as "much more still needs to be done," "the average amount of money spent," "in the state of Texas," and "an epidemic that needs to be taken seriously." Topic phrases accounted for 20% of the material flagged.

In the 22 instances flagged by SA, only 40% of the material could be considered citation errors. As with Tii, commonly used phrases accounted for a significant portion of the flagged material. However, unlike Tii, the commonly used phrases were fewer than the citation errors. Thirty-six percent of the material in SA was commonly used phrases.

Concerns

Our qualitative examination of the 40 drafts raised other issues for discussion. For example, in Tii, the 40 drafts that we examined indicate that 155 of the instances were linked to student sources (63%). In comparing the flagged material, we discovered that the two programs do not flag the same material in the student text, nor do they identify the same potential sources. While Tii does flag material that is improperly cited or poorly paraphrased, it flags so much additional material that finding the possible plagiarism can be difficult. Judging from the patterns observed in flagging material, it appears Tii looks at institutional papers first and then proceeds to examine the Internet. Because of the commonalities in student writing, as noted in our qualitative outcomes, Tii finds more matching content. Tii also tends to flag the most recent source to use a marked phrase. In contrast SA, as discussed earlier, flags an average of only 2 sources per paper and most of this is properly cited material and derived from publication sources. SA also tends to flag entire sentences unlike Tii, which usually flags phrases.

A serious concern for instructors is that neither application has the ability to filter potential sources for context. For example, one paper about obesity contained a flagged phrase, the source of which was identified by Tii as a Web site selling Viagra (http://www.viagra-purchase.com). A report on a paper concerning the deportment of professional football players in Las Vegas contained a tagged phrase which referred us to a site relating to feminist theory. The phrase, "the current policy is not working properly and needs to be changed or amended" was found in a student's summary explaining that NFL players needed strict rules governing their behavior. At the time of our analysis, the most recent occurrence of that phrase was located in the now-defunct

site, http://www.tufffemme.com. However, the site's contextual use of that phrase had no relation to professional sports of any kind.

Another concern is that SA will flag a paper based on students citing the same source. For example, in three different papers about birth control, each student had cited information from Planned Parenthood. The three papers all had a different thesis statement. One paper discussed birth control, in the form of oral contraceptives, in relation to acne and acne treatments. A second paper cited birth control but referenced abstinence as the only viable choice. The third paper discussed birth control as a means to reduce poverty. SA noted that the reference to Planned Parenthood in the Works Cited was also found in another student's paper. The link from http://www.plannedparenthood.org to the material on birth control was included by all three students; thus, SA flagged the entry.

These findings made us want to test these applications in a more thorough manner. First, two of us deliberately "wrote" a draft by compiling text from several different websites and immediately submitted it to Tii. While most of the material was flagged, it was not attributed to the website(s) that it was taken from; instead, the most recent websites posted with the material were flagged. Additionally, a document that contained a substantial amount of material transcribed from several recently published books was submitted to Tii. None of the transcribed material was flagged, and the document received a green rating.

Conflicting Ideologies between WAC Pedagogy and These Software Applications

While these results are troubling enough, perhaps the most direct conflict that emerges between WAC pedagogy and these plagiarism-detection systems occurs when we consider a guiding principle of WAC: "that only by practicing the conventions of an academic discipline will students begin to communicate effectively within that discipline." In our study of Tii and SA, we found that commonly used phrases, such as "In a study from Brown University," or "Researchers have found that X contributes to…", are among the most often flagged as potentially plagiarized material.

These commonly used phrases, topic phrases, and jargon are indicative of basic language structures used by most people in writing about a common topic (global warming, date rape, birth control, etc.). Handling citations and executing proper paraphrasing is again a measure of the inexperience of the writers and not necessarily a lack of originality or deliberate attempts to deceive on the part of the students. For example, in our sample drafts, one sentence from one student's draft on global

warming was linked to another student paper as the source. We put the sentence through Google only to discover that ten other sites, many of them maintained by professional or non-profit organizations, had used the exact same sentence in their documents. If our goal, as Susan Peck MacDonald writes, is to move undergraduates from pseudo-academic writing to "expert, insider prose," it seems likely that students will model their writing on examples they are given by instructors or read during coursework or research for projects, especially as they progress through MacDonald's Stages 2 and 3. Students who receive feedback indicating that their attempts might be plagiarized may revert back to such pseudo-academic structures to avoid any possibility of accusation. They might also revert to simple, quick fixes such as "using the thesaurus function in Word" rather than honing their paraphrasing skills, a valid concern voiced by writing center tutors (Brown et al. 24). A more likely consequence, though, is that students will progressively disengage from both formal and informal writing tasks—exactly what WAC/WID programs are designed to combat.

So, What's to Be Done?

We began this study in order to determine whether or not there were any viable reasons to use such applications as Turnitin and SafeAssignment. What we found is that in the context of undergraduate writing, the potential liabilities far outweigh the possible benefits of doing so. In short, the primary benefit of using either application is that instructors may be able to quickly identify material that has been copied from an Internet source or shared by students in multiple sections of a course who submit similar assignments. Additionally, if students use the applications in draft mode, they may be able to identify places in their drafts where they have incorrectly cited or punctuated citations.

However, these benefits pale when we look at the potential problems caused by using these applications. Despite the verbiage on the applications' Web sites to the contrary, nothing about the interfaces suggests an emphasis on teaching or learning about proper citation methods. Consequently, instructors will need to invest a significant amount of time in learning the applications and in preparing students to analyze the results, discard all of the erroneously identified instances of "potentially unoriginal material," and use the remaining data to assist with revision. However, students at MacDonald's Stages 1, 2, or 3 may quickly become discouraged. More significantly, many students may shift from writing to an appropriate human audience to "writing to the software." Susan Schorn notes that students need to move beyond merely knowing who

their readers are to gaining an understanding of them (337). However, the very topic phrases, jargon, or commonly used phrases expected as signals of understanding by human readers may be those very items flagged as unoriginal by the applications. This conflict will not help students become more effective writers in any discipline and may actually promote the type of writing that instructors in many disciplines are trying not to teach—writing that, in its attempt to pass muster with the originality checkers, loses all semblance of a realistic, academic voice.

To ensure that instructors understand the limitations of each application and communicate those limitations clearly to students, WAC/WID coordinators will need to work even more diligently if they are on campuses where site licenses to these applications have already been purchased. Additionally, given today's shrinking budgets and increasing requirement for accountability, campus administrators need to understand that the return on their investments in these systems may not be what they had hoped for. While purchase of these applications might achieve a short term goal of illustrating that the institution is discouraging/cracking down on plagiarism, in the long term such purchases may well co-opt any attempts made to institute the kind of careful pedagogy that enables "students to conduct research, comprehend extended written arguments, evaluate sources, and produce their own persuasive written texts" (Howard 789). It's not a stretch to say that those students using these applications may become disengaged from writing, their coursework may suffer, and, eventually, their performance on such accountability measures as the CLA or other exit exams may be impaired. In the end, monies would be better spent on developing other campus resources for writing instruction than relying on these "quick fixes" that ultimately do not contribute to the educational mission of our institutions.

WORKS CITED

Braumoeller, Bear F., and Brian J. Gaines. "Actions Do Speak Louder than Words: Deterring Plagiarism with the Use of Plagiarism-Detection Software." *PS: Political Science and Politics*. 34.4 (2001): 835-39. Print.

Brown, Renee, et. al. "Taking on Turnitin: Tutors Advocating Change." *The Writing Center Journal.* 27.1 (2007) 7 – 28. Print.

Howard, Rebecca Moore. "Should Educators Use Commercial Services to Combat Plagiarism? No." *The CQ Researcher.* 13.32 (2003) 789. Print.

MacDonald, Susan Peck. *Professional Academic Writing in the Humanities and Social Sciences*. Carbondale: Southern Illinois UP, 1994. Print.

Marsh, Bill. "Turnitin.com and the Scriptural Enterprise of Plagiarism Detection." *Computers and Composition*. 24 (2004) 427-438. Print.

Purdy, James P. "Calling Off the Hounds: Technology and the Visibility of Plagiarism." *Pedagogy*. 5.2 (2005): 275-95. Print.

Schorn, Susan. "A Lot Like Us, but More So: Listening to Writing Faculty Across the Curriculum." *What Is "College-Level" Writing?* Ed. Howard Tinberg and Patrick Sullivan. Urbana, IL: NCTE, (2006):330 – 340. Print.