Plagiarism prevention or detection? The contribution of text-matching software to education about academic integrity

Dominic Keuskamp and Regina Sliuzas
Student Learning Centre, Flinders University, Adelaide SA 5001, Australia
Email: dominic.keuskamp@flinders.edu and regina.sliuzas@flinders.edu.au
(Published online 1 December 2007)

Developing an understanding of academic integrity within students is one of the core objectives of many Academic Language and Learning (ALL) advisers, and the perceived rise of plagiarism suggests that this will continue to demand our attention. A recently available tool to assist advisers in this role is text-matching software (TMS). Routinely promoted on the basis of its capabilities for “plagiarism detection”, TMS also offers students educative opportunities which appropriately are web-based, given the increasing “web-dependency” of students. This paper examines how TMS can contribute to the role ALL advisers play in developing students’ understanding of academic integrity. Students from across one university were invited to submit their assignments to a TMS program called SafeAssignment™, offered as part of the university’s academic integrity policy. Text-matching reports generated from 21 students were analysed to identify the extent and nature of identifiable plagiarism, and how the software communicated this to students. Overall percentages of text-matching were low, with many students’ texts matching purely on information that was bibliographical, appropriately quoted, generic or technical. However, the quantitative information reported to the students by the software offered less assistance in determining if plagiarism had occurred than the more detailed information to be found in careful interpretation of the text-matching reports. A guide is presented for ALL advisers involved with interpreting reports with students.

Key Words: plagiarism, text-matching software, academic integrity.

1. Introduction
1.1. Background to the issues
The last two decades have witnessed prolific and diverse public and academic discussion regarding plagiarism within higher education (Larkham & Manns, 2002; McCabe, Trevino & Butterfield 2001; Park, 2003). Debate has frequently been heated, with Robin (as cited in Clegg & Flint 2006, p. 374) describing situations in the USA as having “achieved the status of public scandal”, although Clegg and Flint (2006) in the UK have suggested that there is “more heat than light” (p. 373). Similar heights of media/public interest have been reached in recent years regarding plagiarism in the Australian tertiary context. The questions raised by the discussion reflect the complexity of the subject, and include: what are the reasons for the focus of attention on plagiarism and the apparent increase in its occurrence, and how is it best defined, avoided, and dealt with? (Larkham & Manns, 2002).

The attention accorded to plagiarism in part reflects that to many it threatens the very basis of academia; the university community is expected to behave ethically in its pursuit of knowledge,
and trust in ethical behaviour is considered to be “the cultural glue that enables academia to function successfully” (Rosamond, 2002, p. 168). If, as Standler (as cited in Rosamond, 2002, p. 168) argues, academic reputations are built upon “creating new knowledge: discoveries of new facts, new ways of looking at previously known facts… [and] original analyses of old ideas”, then in an increasingly competitive and research-driven tertiary environment, plagiarism must be addressed to protect the very foundation of the tertiary sector. When viewed in this light, it is very possible that attitudes towards borrowing of people’s ideas have tightened as academic and professional environments become more competitive and performance-driven.

Likewise, debate has centred on whether plagiarism has always occurred but has simply become more visible (Marshall & Garry, 2006; McCabe, 2004) as awareness of the issue has increased, or whether the incidence of plagiarism has increased with the explosion of online resources (Hart & Graham, 2004) and the ease with which students can “cut and paste” (Evans & Youmans, 2000; Goddard & Rudzki, 2004; Larkham & Manns, 2002). Others talk of changes in students’ attitudes to information due to its ready availability on the Internet (Szabo & Underwood, 2004). Barberio suggests that there now exists a “one-armed bandit syndrome” (2004, p. 307) and the feeling of hitting “data jackpots” (2004, p. 308), in which ease of retrieval reduces students’ willingness to spend time critically evaluating sources for the most useful and relevant information for their topic. Gorman (2007) also refers to a changing way in which information and texts are viewed:

> in the realm of information – a direct consequence of digitally-enabled information saturation – I see a growing willingness by people to think of the search engines as an ever-available knowledge prosthesis that will provide what we need when we need it. What is too easily forgotten is that education is not about knowing facts but about acquiring contexts and perspectives so that we know what we need to look for and how we might go about looking. Information is always a function of context.

Acknowledging the pressures facing many students today, such as those related to time, achievement and the pressure to pass the last module of their topic or program, Szabo and Underwood’s (2004) study of 291 science students at a UK university revealed that 68% used the copy/paste function to embed information into their work without proper acknowledgment, 46% would probably use copy/paste without modification, and a further 20% would definitely use it without modification to prevent failing a topic. Undoubtedly, as Rosamond (2002, p. 171) argues, the emergence of many paper mills has also contributed to the heightened “academic sensitivity” to plagiarism.

Notwithstanding these concerns, while plagiarism superficially appears simple to define, according to Pincus and Schmelkin (as cited in Yeo, 2007, p. 200), it is “not a universally understood or accepted concept in academia”. Most definitions of plagiarism have in common the notion that it involves the presentation by one person of another person’s work as if it were their own, regardless of intent. One particularly problematic area is determining the extent to which original texts must be paraphrased to be acceptable. Should such a decision be based on a maximum number of words in common between the original and the paraphrase? Does this vary among disciplines and the nature of the material? Is one short sentence enough for disciplinary action? These and many more questions are raised, even when original sources have been correctly cited. However, when original sources have not been acknowledged, once similarities between an original and a student’s work have been established, there are often fewer problems in establishing that plagiarism has occurred. This is in part where text-matching software (TMS) comes into play.

### 1.2. Approaches to plagiarism

Academic approaches to plagiarism and transgressions in particular seem to reflect two main perspectives (Martin, 1994). There are those who see plagiarism primarily as theft, will go to considerable effort to detect it, and see punishment for the crime as a natural justice; others argue that much plagiarism occurs unintentionally by students who do not fully understand what
is and what is not acceptable in terms of integrating other people’s ideas into their academic writing, and who take a more educative approach (Anyanwu, 2004; Devlin & Gray, 2007). Of course, perspective is intimately entwined with role with regard to the staff member dealing with both the student and the work itself. For example, faculty-based lecturers are involved with assessing student work and the students’ grasp of the knowledge and concepts in any given discipline. Academic Language and Learning (ALL) advisers, however, are more likely to be involved in assisting students to understand academic conventions and to develop academic literacy, and are more likely to be aware of a number of different reasons that plagiarism has occurred in the writing process (Chanock, 2004). Indeed, the role of ALL advisers in relation to academic integrity and plagiarism is unique in the university environment. ALL advisers are frequently called upon to explain to students at the commencement of their studies (and perhaps later) how academic integrity and plagiarism are defined, and how, without plagiarising, they can integrate information from sources into their written work, using the appropriate referencing conventions. In their role as ALL advisers, they are not involved in grading student work and their approach is educative, rather than punitive. When submitted assignments are identified as containing plagiarism, faculty staff often send students to ALL advisers for detailed explanation and clarification of academic integrity and plagiarism, so that the student will understand how to avoid plagiarism in the future, and in compliance with the university’s obligation to fulfil its educative role.

1.3. Why text-matching software (TMS)?

In the light of the groundswell in literature regarding plagiarism, and to protect themselves from the unwanted media scrutiny received by some high profile cases (e.g., Thompson & Smith, 2004; Watts, 2007), most Australian universities have, in recent years, invested considerable resources in tightening up their policies and procedures for dealing with plagiarism, and have made more explicit their expectations of staff and students in dealing with academic integrity issues. Several Australian universities have incorporated the use of text-matching software (TMS), primarily to help faculty staff detect and provide evidence of electronic plagiarism more efficiently. Dealing with suspected cases of plagiarism is time-consuming for academic staff who may spend hours locating and cross-checking original texts, and then more time in following University policies and processes for dealing with the student appropriately. Moreover, as Anyanwu (2004) points out, plagiarism surfaces most often when academic staff are at their busiest: when marking piles of essays, and when finalising and recording grades. TMS saves staff much time by quickly matching student assignments with electronic sources. Most TMS programs match assignments submitted by students with an intra-institutional archive of other students’ assignments, as well as an electronic archive of articles, web documents and sometimes assignments purchased from paper mills.

SafeAssignment™ for MyDropBox, the TMS program employed in this study, operates as outlined above, and generates a report that: (1) provides the percentage of text in the assignment that matched text from archived works, (2) highlights sentences singly or in groups that have a substantial portion of matched text, indicating the percentage of matching text, and (3) provides links to the possible original sources containing the matched text. However, the reports must then be analysed, as text-matching alone does not necessarily indicate plagiarism. Rather, the matching text may be a string of words which has been quoted and cited correctly, a bibliographical entry in common with another work, or other legitimate matches of text. According to SafeAssignment™’s website (http://www.mydropbox.com):

SafeAssignment™ is an innovative approach to plagiarism prevention, providing educators with an effective solution for checking originality and deterring plagiarism in academic environment [sic]. SafeAssignment™

1 In 2007, all rights to SafeAssignment™ were acquired by Blackboard Inc., which in July 2007 unveiled a new plagiarism prevention service called SafeAssign™.
works with papers students turn in electronically and employs a proprietary technology to identify unoriginal content, including paraphrased or otherwise altered text. After a series of comprehensive plagiarism checks, SafeAssignment™ generates convenient and easy-to-read reports, where all unoriginal material is highlighted and linked to its online or database sources.

SafeAssignment™ is promoted as both a tool for plagiarism prevention and as a solution to “checking originality and deterring plagiarism”, as described above. It does not, however, detail how it helps to prevent plagiarism. Presumably, the sheer fact that lecturers are going to be able to detect plagiarism more easily acts a deterrent in itself. If students are able to submit their work and generate a text-matching report prior to submission, then also presumably, if they analyse any part of their work which has been identified as matching to another work, the students would be able to learn from the report how and why it matches, make some adjustments to their work to reduce or eradicate matching, and improve their paraphrasing and referencing skills in the process. This paper examines the degree to which SafeAssignment™ can contribute to the ALL adviser’s role in communicating with students to develop a better understanding of academic integrity.

2. Methodology

We analysed the text-matching reports of 21 students of a single, medium-sized Australian university who voluntarily submitted their assignments to SafeAssignment™. All reports from submissions in the second semester of 2006 were included in the study, thus the sample was essentially self-selected. Although no explicit data were collected on student characteristics to preserve their anonymity, the student group comprised international and local students, including NESB students, from a range of disciplines. Reports were analysed to identify the extent and nature of identifiable plagiarism, and how SafeAssignment™ communicated this to students. Information provided by reports is predominantly in numerical form, so to parallel this approach a quantitative analysis of the problem was undertaken. Each of the 21 reports was analysed individually and data were collected on a number of variables, as outlined in Table 1. The judgement of plagiarism within each highlighted sentence was based on a subjective assessment of the extent of verbatim copying or inadequate paraphrasing of material which was not deemed generic or technical.

Table 1. Variables for which data were collected from SafeAssignment™ reports.

<table>
<thead>
<tr>
<th>Assignment-level variables</th>
<th>Sentence-level variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of matching text*</td>
<td>Percentage of matching text per sentence*</td>
</tr>
<tr>
<td>No. of highlighted sentences</td>
<td>Longest verbatim-copied string</td>
</tr>
<tr>
<td>Plagiarism adjudged (y/n)</td>
<td>Plagiarism adjudged (y/n, and reason for judgement)</td>
</tr>
</tbody>
</table>

*As provided in SafeAssignment™ report.

All reports were analysed by the first author, and all instances of adjudged plagiarism were verified by the second author. Means per assignment and per sentence were calculated for the variables listed in Table 1, and compared between reports with and without adjudged plagiarism to determine how well each of the variables predicted its instance.

3. Results

Overall, the percentage of matching text per assignment was low, with a mean of 4% reported matching text, and 5 reports without matched text. However, closer inspection at the sentence level revealed that 6 of the 21 reports (29%) exhibited adjudged plagiarism within the sentences highlighted by SafeAssignment™. Neither the percentage of matching text per assignment nor the mean percentage of matching text per sentence differed substantially between the groups of
reports with and without evidence of plagiarism. However, the mean number of highlighted sentences and the mean length of the longest verbatim-copied string differed more markedly. In those reports with adjudged plagiarism, almost 4 times as many highlighted sentences and a verbatim-copied string length over 4 times as long were evident. The verbatim-copied string length from sentences exhibiting plagiarism reached 51 words. Assignments with adjudged plagiarism on average contained 3-4 sentences with plagiarised text.

Fifteen percent of instances of adjudged plagiarism were as a result of inadequate paraphrasing, assessed as either phrase interchange or mere synonym substitution. Of the highlighted sentences that were not adjudged to be plagiarised, 85% contained text of a generic nature, or proper nouns.

**Table 2. Summary of results from the analysis of 21 SafeAssignment™ reports.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reports with adjudged plagiarism in highlighted sentences</th>
<th>Reports where no evidence of plagiarism in highlighted sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. (percentage) of reports with adjudged plagiarism</td>
<td>6 (29%)</td>
<td>15 (71%)</td>
</tr>
<tr>
<td>Percentage of matching text per assignment*</td>
<td>8%</td>
<td>2%</td>
</tr>
<tr>
<td>Mean percentage of matching text per sentence*</td>
<td>85%</td>
<td>84%</td>
</tr>
<tr>
<td>No. of highlighted sentences</td>
<td>13.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Longest verbatim-copied string</td>
<td>23.2</td>
<td>5.7</td>
</tr>
<tr>
<td>No. of sentences with adjudged plagiarism</td>
<td>3.5</td>
<td>0</td>
</tr>
<tr>
<td>Reason for adjudged plagiarism (% of sentences, as indicated)</td>
<td>Sentences with adjudged plagiarism:</td>
<td>Highlighted sentences:</td>
</tr>
<tr>
<td></td>
<td>Verbatim copy (85%)</td>
<td>Generic (45%)</td>
</tr>
<tr>
<td></td>
<td>Inadequate paraphrase - phrase interchange (10%)</td>
<td>Proper Nouns (40%)</td>
</tr>
<tr>
<td></td>
<td>Inadequate paraphrase - word substitution (5%)</td>
<td>Technical (10%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adequate paraphrase (5%)</td>
</tr>
</tbody>
</table>

*Means from information provided in SafeAssignment™ report.

### 4. Discussion

This study sought to determine the extent to which the reports generated by the TMS program SafeAssignment™ are capable of improving students’ understanding of plagiarism and academic integrity. The results of the research suggest that the program communicates a somewhat clouded message to students regarding whether or not plagiarism has occurred in their assignments. Our results indicate that the statistics provided by the program in its reports – percentages of matching text per assignment and per sentence – are insufficient alone to make a judgement regarding plagiarism. Further interpretation is required, of both a quantitative and qualitative nature. We contend that ALL advisers are in the best position to assist students in the interpretation of reports.

As the first statistic that a student sees once a report has been generated, the percentage of text-matching (per assignment) undoubtedly influences to a large extent the student’s reception of the report. However, the findings of this study have indicated that students are unlikely to gain
much from this figure. It may not be intuitive to students that the percentage is dependent on the word count of the article and thus in itself does not necessarily predict the existence or extent of plagiarism. Likewise, the mean percentage of matching text per sentence differed little between reports with and without adjudged plagiarism. Conversely, a large number of highlighted text strings was a better indicator of plagiarism. Students’ abilities to deal with such quantitative information will depend on their level of academic numeracy. With long-standing evidence of declining numeracy skills in society (Paulos, 1990) and our universities (Chapman, 1988), one questions how fully students are able to interpret the quantitative information provided to them by the reports.

Notwithstanding these results, it should be noted that in this study the percentages of matching text per assignment were low in comparison with other research, for example Barrett and Malcolm (2006). This could be attributed to the success of the university’s academic integrity management, or to those students’ understanding of academic integrity. Alternatively, this may reflect the nature of the student sample, who volunteered to submit their assignments to SafeAssignment™. Our assumption is that this group of students is less likely to plagiarise, and thus percentages may be lower than could be expected in the main student population. It is also possible that the number of instances of plagiarism was underestimated for several reasons. First, students are likely to use sources that are not available electronically and therefore not available for matching. Second, only those sentences highlighted by SafeAssignment™ were investigated in this study, however even a cursory look over the remainder of some of the assignments indicated text that could have been plagiarised, distinct as it was from the surrounding text.

The results of this study will also serve to raise staff awareness of the degree of interpretation they may find necessary for SafeAssignment™ reports. Naturally, staff as well as students are required to exercise their judgement in determining if plagiarism has occurred, and thus time must be devoted for this task, an acknowledged disadvantage of the software. One solution is to set a threshold percentage of text-matching below which it is effectively ignored (Clark & Freeman, 2005). While choosing this option is time effective, the educative benefits of TMS are somewhat diminished, as students are led to believe that a small amount of plagiarism is acceptable. Moreover, this tends to cater for students who look for a simple numerical answer to the question, “what percentage of text-matching is OK?”. For example, the percentage of 11% found in one of the reports analysed in this study would have passed under the 15% threshold for plagiarism used by Barrett and Malcolm (2006). However, the assignment exhibited 9 word strings ranging from 5 to 51 words that appeared to be verbatim copies of the Internet source text.

Students’ perceptions of using SafeAssignment™ were not addressed in this study. The authors’ own observations suggest that students do not fully understand how SafeAssignment™ “checks” their assignments and what exactly is entailed in the information provided to them. For example, students commonly submit assignments to the program complete with bibliographies and assignment guidelines, bolstering the percentage of text that matches to other students’ assignments and electronic sources. However, in research investigating trials of TMS, students believed that TMS would help their understanding of plagiarism and would also act as a deterrent for potential plagiarists (Clark & Freeman, 2005; Green, Lindemann, Marshall, & Wilkinson, 2005). In both trials, students commented that their use of TMS gave them a sense of security that they would not be accused of inadvertent plagiarism. McGowan (as cited in Green et al., 2005, p. 32) believes that students will become more original and critical if they know that TMS will be used, a contention also borne out in research by Martin (2005). Others believe that the use of TMS will deter both intentional and unintentional plagiarism (Carroll, as cited in Green et al., 2005, p. 32). Further research is needed into student perceptions and software performance to fully elucidate the nature of the relationship between them.

The utility of TMS such as SafeAssignment™ for the detection of plagiarism is not disputed here. In this study the program effectively detected text adjudged as plagiarism among several assignments, including instances of inadequate paraphrase as well as verbatim copying. The
program’s capabilities for detection aside, however, its usefulness for educative purposes remains to be evaluated. Among the emerging literature concerned with TMS, there has been little attempt to document how this can effectively occur. Some even predict that in the case of students who use a “cut and paste” essay writing technique, those students will learn to adapt to the software to avoid detection by finding alternatives for key words (Warn, 2006). However, the research presented here has succeeded in highlighting that where TMS is provided for students to voluntarily “check their assignments”, this should be accompanied by some education in the mechanism of the report generation as well as in the interpretation of the reports. Briggs (as cited in Warn, 2006, p. 204) agrees that “[a]n important component of any anti-plagiarism strategy is the provision of training in correct practice”. Nonetheless, the time needed to educate students in the appropriate use of TMS and interpretation of reports is unlikely to be spent by faculty-based lecturers. It is suggested that the independence from faculty and knowledge of academic integrity held by many ALL advisers places them ideally to take on this role.

ALL advisers can support students’ use of TMS by several means. Apart from explaining the purpose of providing TMS to students, and the logistics of how to generate reports, students will benefit from being familiarised with the concept of academic integrity and with using academic citation before attempting to use TMS such as SafeAssignment™. In addition, students need to have a solid understanding of academic literacy and how arguments are constructed. Unless students understand these concepts, there is the likelihood that in attempting to paraphrase other people’s work, they will rely heavily on the words and sentences of the original authors, rather than focussing on paraphrasing and discussing their ideas. Then, the students may focus on simply reducing their percentage of text-matching until they can “beat” the software, as suggested by Warn (2006). Conversely, others may be alarmed that the percentage of matching text in their report suggests plagiarism, but not understand that their use of particular words and phrases may be quite acceptable. It is further suggested that students engage in training in academic writing conventions, and referencing practices, and be encouraged to seek detailed instruction on how to interpret SafeAssignment™ reports, whether it be through an online resource or in person through either an ALL adviser or their faculty lecturers.

5. Conclusion

In summary, students cannot rely solely on the information provided in reports from TMS such as SafeAssignment™ to determine whether or not plagiarism has occurred, or to educate themselves about academic integrity. In this study, the statistics provided by such reports were insufficient indicators of whether an assignment contained what was judged to be plagiarised text. It is suggested that the capabilities of TMS for educative purposes can be substantially enhanced by explanation of the implications of reports by those well-trained in academic literacy, ideally ALL advisers. As Warn (2006, p. 201) observed, “the quantitative measures derived from the output can be of use, but qualitative judgement is still needed”. To what extent students are literate and numerate enough to do so themselves is yet to be determined, and provides a focus for further research.

Acknowledgements

The authors would like to thank Dr. S. Kutieleh for supporting this research, and staff of the Student Learning Centre, Flinders University, for their assistance with the manuscript.

References


