Journal of English for Academic Purposes 12 (2013) 264–272

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Do journal authors plagiarize? Using plagiarism detection software to uncover matching text across disciplines

Yu-Chih Sun*

Institute of Teaching English to Speakers of Other Languages, National Chiao Tung University, 1001 University Road, HsinChu City 300, Taiwan, ROC

Keywords:
Plagiarism
Matching text
Plagiarism detection software
Turnitin
Paraphrasing
Self-plagiarism

A B S T R A C T

The current study aims to explore the extent of matching text in published journal articles and how the number of authors and their various official languages influence the extent to which matching text appears. Six hundred journal articles in Science, Technology, Engineering, and Mathematics (STEM) and Social Science were randomly selected and screened by both plagiarism detection software (Turnitin) and human raters(s). The results indicate that disciplinary differences do exist in terms of the degree of matching text incidences. Journal articles in STEM tend to contain significantly more consecutive matching text from other sources than Social Science journal articles. However, it is not clear if this is discipline dependent. In addition, authors tended to have more consecutive text copied from their own previously published works than that of others' publications. Furthermore, the greater the number of authors an article has the more consecutive text-matching can be observed in their published works. Additionally, authors located in contexts wherein English is an official language do not differ significantly from those in contexts wherein English is not an official language on their Turnitin scores and the number of 30-word or longer strings of consecutive matching text from self-published articles and self-and-others' publications combined.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Plagiarism is defined as the reproduction of ideas and/or language from source materials without sufficient attribution to the source (Abasi, Akbari, & Graves, 2006; Pecorari, 2003). Thus, writers are required to paraphrase and acknowledge their sources through appropriate citation. In other words, one should integrate and interpret different sources in one's own words (Ballard & Clanchy, 1991). This process involves not only reproducing and extending ideas, but also reflecting upon and reiterating their meanings. Plagiarism can also involve copying others' text and copying one's own publication. In the latter case, it is defined as self-plagiarism or the duplication of one's publication (Bretag & Carapiet, 2007).

Plagiarism in higher education is a widespread and growing problem (Sun, 2009; Bennett, 2005; Bull, Collins, Coughlin, & Sharp, 2001; Butakov & Scherbinin, 2009; Selwyn, 2008). Research findings show plagiarism rates are now of epidemic proportions and a majority of students report having plagiarized in some form or another during their studies (Emerson, Rees, & McKay, 2005; Franklyn-Stokes & Newstead, 1995; Graham, Monday, O'Brien, & Steffan, 1994; Handa & Power, 2005). Factors such as time constraints may lead to deliberate cheating (Ashworth, Bannister, & Throne, 1997). Yet the trend may also have arisen inadvertently, due to a lack of understanding what plagiarism is (Yeo, 2007), poor training in citation skills (Ellery,

* Tel.: +886 3 5712121x52755; fax: +886 3 5739033.
E-mail address: sunyc@mail.nctu.edu.tw.

1475-1585/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jeap.2013.07.002

2008), poor academic skills, or an inability to paraphrase appropriately (Keck, 2006). Furthermore, students using English as a second/foreign language face additional challenges, often having to sort out the ambiguities of acceptability in recycling of language while at the same time still struggling to develop competency in grammar and vocabulary (Sun, 2009, 2012).

Another driving force for plagiarism is the profusion of easily retrievable resources from the Internet. The changing nature of online sharing and the ubiquity of digital content tempt students to cut and paste (Underwood & Szabo, 2003; Warn 2006). This has led to a cultural shift in the digital age and changes in attitudes towards information ownership (Flowerdew & Li, 2007; Wood, 2004). As Weiler (2004) states, the learning patterns of the so-called "generation Y" have developed around seeking information rather than reading and critiquing content on their own (Weiler 2004); elsewhere it has been noted that generation Y consider cutting and pasting a legitimate research method (Chaky & Diekhoff, 2002). The re-emphasis of the values of academic integrity and development of strategies to cope with this widespread phenomenon have thus become indispensable.

1.1. Ambiguity of the concept of plagiarism

Despite years of discussion and investigation, plagiarism is still a fairly ambiguous concept, and its exact definition remains vague in many areas (Liddell, 2003). As Mulcahy & Goodacre (2004) indicate, issues of academic integrity are usually not black and white. Some researchers have tried to employ operational measures to define plagiarism and legitimate paraphrasing. For example, Shi (2004) defines legitimate paraphrasing as "no trace of direct text-matching of two or three consecutive words from source texts" (p. 178–179). Benos, Fabres, & Farmer (2005) proposes a similar definition, arguing that the "duplication of words and phrases, however brief, may be indicative of plagiarism" (p. 62). Oshima and Hogue (1999) define inappropriate sourcing as when a paraphrase "contains the same vocabulary and sentence structure as the original" (p. 90).

Other studies adopted varying scales of percentages of matching text as a measure of the acceptability of paraphrasing. For example, Keck (2006) cites 50% or more similarity with original sources as the threshold of plagiarism, 20–49% for minimal revision, and less than 19% for an acceptable revision. Pecorari (2003) points out that a 40% text-match with a source "puts the writers at risk of a plagiarism accusation" (p. 325).

Even with abundant attempts to quantify what constitutes plagiarism and legitimate paraphrasing, there is an apparent absence of consensus among researchers and teachers (Campbell, 1990; Pecorari, 2003; Shi, 2004; Sutherland-Smith and Carr, 2005; Yamada, 2003). Some studies consider surface-level paraphrasing strategies as inappropriate, such as only substituting a word with its synonym (Angélil-Carter, 2000; Shi, 2004), inserting or deleting some words from the source (Shi, 2004), changing the syntactic structure of the source (Shi, 2004), or reordering words or phrases from the source (Keck, 2006).

However, Howard (1995) proposes the term "patchwriting" to refer to the text-matching of source texts that involve only surface-level changes, such as modifying the words using synonyms and changing the verbs. Howard (1995) views patchwriting as "paraphrasing the source's language too closely" (p. 799). He (1995) points out that patchwriting is a crucial and indispensible stage for novice writers to learn and develop writing skills. This sort of dependence on source texts allows novice writers to eventually become mature ones. He argues that patchwriting should be viewed as a positive educational opportunity, rather than as intentional academic misconduct or plagiarism. That is, patchwriting deserves a pedagogical rather than punitive response. Pecorari coins this process "source-dependent imitation learning" and points out that learning is rarely a straight line from novice to mastery, "The novice academic writer must crawl before being able to walk" (Pecorari, 2003, p. 320).

In addition, research findings have also shown that discipline-specific standards and expectations are not yet clear (Crisp, 2004; Sutherland-Smith & Carr, 2005). As Dahl (2007) puts it, "The finer details about what constitutes plagiarism and what not are contested, and in fact are often different from one subject area to another even within the same school or university" (p. 173–174). For example, in Barrett, Barrett and Malcolm (2005)'s study, Computer Science students were more likely to copy others' words than those from Automotive Engineering or Electronics. In addition, according to Dahl's (2007) study, business students are more prone to academic plagiarism than students from other disciplines. Given the already ambiguous definition of plagiarism and legitimate paraphrasing, plus diverse standards and expectations from different academic disciplines, the issue of plagiarism becomes even more complex for academia.

1.2. Plagiarism detection software

Turnitin is an online plagiarism detection (text-matching) tool (www.turnitin.com). It compares a submitted document against a vast electronic database of billions of websites, millions of periodicals, books, and archives of assignments and documents previously submitted to Turnitin from all over the world (Turnitin, 2006). Turnitin produces an "originality report" for each submitted document that highlights matched texts using color coding and links to what it deems to be the original source of the match in its database (Turnitin, 2006). The text-matching system gives an overall plagiarism score (similarity score) from 0% (being completely original) to 100% (being completely matched) (Stetter, 2008). Low percentage reports can be excluded through a filtering system on Turnitin. By using the "side-by-side" feature, with the submitted document on the left and the similar text on the right, the process of comparison is made easy.

The implementation of plagiarism detection software such as Turnitin as part of higher-learning institutions' assessment strategies has been shown to reduce instances of matching texts (Ercegovac and Richardson, 2004). It has also been established as a formative learning tool, giving students instant feedback (Rolfe, 2011). Turnitin can be used as a diagnostic tool to

identify students who may need help in paraphrasing and referencing other scholars' works (Barrett et al., 2003). It can also help students examine the originality of their own work, heighten awareness of plagiarism, and prevent them from acquiring text from Internet sources without appropriate paraphrasing and referencing (Stetter, 2008; Sutherland-Smith & Carr, 2005).

However, plagiarism detection services like Turnitin are not without their shortcomings and should not be considered a panacea for plagiarism (Sutherland-Smith & Carr, 2005). First, the originality report produced by Turnitin is considered to be most useful as a deterrent rather than a solution to plagiarism. Thus, it takes more active strategies to reduce the problem of plagiarism (Savage, 2004), as "there are many anecdotal stories of how students have attempted to 'cheat' the Turnitin system" (Jones & Moore, 2010, 426). It is suggested that modeling appropriate paraphrasing behavior, regular practice, and rewarding successful performance are more effective than punishing inappropriate behavior (Martin, 2004). Second, not all sources are included in the databases, e.g. older textbooks or non-electronic journals (Allan et al., 2005). Therefore, finding no matches in the originality report cannot guarantee there have been no incidences of plagiarism.

Third, there is no means of checking for the plagiarism of ideas (i.e. ideas taken from others but expressed in one's own words) or false authorship (e.g. a paper written by someone else). Fourth, the system relies purely on technical solutions and sometimes generates unreliable judgments (Carbone, 2001). Such text-matching services can only provide a measure of detecting possible plagiarism; human scrutiny and academic judgment are still required to verify each matching passage (Barrett & Malcolm, 2005; Jones, 2008; Sutherland-Smith & Carr, 2005), as for example, Turnitin cannot distinguish accurately cited texts from other copied materials (Purdy, 2009). Fifth, Turnitin cannot identify "common" phrases (Jones, 2008). Thus, Turnitin should be viewed as a matching text tool, rather that a plagiarism detection tool (Jones, 2008). Allan et al.'s (2005) study indicates that even those cases in which there is less than 25% of matching text, there still exists certain unacceptable instances of plagiarism. Thus, judgments based on the level of text-matching as per the originality report is not sufficient; further investigation is required to determine the level of plagiarism (Allan et al., 2005).

Finally, there is some criticism about the integrity of Turnitin due to the design of its digital archives, which do not require authorial permission and protection (Purdy, 2009). For every submitted document, Turnitin's default setting is to add that document to its archive and then make it open for public access. As Purdy argues, mandatory archiving in Turnitin contrasts sharply with other media such as wikis in which archival control is distributed among users. This violation of students' intellectual property rights deserves much attention in education (Howard, 2007).

Previous studies also indicate that Turnitin serves as an opportunity to broaden discussions on the precise standards and expectations to which students are held in terms of avoiding plagiarism and, concomitantly, on how students are taught to interpret, synthesize, paraphrase, or quote rather than an oversimplification of the percentages offered in the originality reports provided by the plagiarism detection service (Howard, 2007; Stapleton, 2012). In other words, a response to concerns about plagiarism should have a thoughtful approach to the numerical output generated by Turnitin, rather than as viewing it as a time saver or exclusively as the efficient tool the plagiarism detection service promises (Purdy, 2009).

1.3. Purpose of the study

Much of the research on plagiarism dictation tools has focused on their effectiveness in deterring plagiarism, students' and teachers' attitudes towards plagiarism and paraphrasing, or comparing students' plagiarism and paraphrasing behavior after using the tool. Previous studies have indicated that the definitions of plagiarism and legitimate paraphrasing are still obscure and differences do exist across fields and disciplines (Crisp, 2004; Sutherland-Smith & Carr, 2005). It is of interest to learn how teachers and researchers practice intertexuality in their writing for publication and to learn if there are any discrepancies in terms of what they teach and what they do.

Furthermore, as previous studies have revealed that multi-authorship might lead to dilution of responsibility (Bennett & Taylor, 2003; Rennie, Yank, & Emanuel, 1997), it would be of interest to learn how authorship might impact the intertexuality of journal articles. Finally, as previous studies have indicated that students learning English as a first or second/foreign language differ in their perceptions, attitudes, and behavior on plagiarism and paraphrasing (Currie, 1998; Shi, 2004), it would be of interest to understand the extent to which published authors in contexts wherein English is an official language differ from their counterparts, and students in general, in intertexuality.

To date, little, if any, research has aimed to examine the extent of paraphrasing or consecutive matching text practiced by published authors. Nor has research aimed to examine the relationship between the number of authors of a single article, the language background of these authors, or the extent to which these authors engage in text matching. As researchers and professors usually serve as learning models for students in best practice for paraphrasing and source referencing, in-depth investigations of published works in terms of matching text could shed light on their common practice of the issue. For these reasons, the research questions guiding this study are as follows:

1. Do journal articles from Sciences, Technology, Engineering, and Mathematics and Social Sciences differ in their Turnitin scores?
2. Do journal articles differ in the extent of consecutive matching text across disciplines?
3. Does the number of authors influence Turnitin scores or the number of incidences of consecutive matching text?
4. Are there regional differences in terms of the official language used for Turnitin scores and incidences of consecutive matching text?

2. Methodology

2.1. Data collection

To examine the extent of matching text in published journal articles across fields and disciplines, Thomson Reuters's Web of Knowledge categorization was chosen for its wide range of source material and then narrowed to two broad fields, Science, Technology, Engineering, and Mathematics (STEM) and Social Sciences. The selection of the sub-disciplines was based on the variety and coverage of scopes. The sub-disciplines for STEM include Biology, Computer Sciences, Engineering, Math, and Physics, whereas the sub-disciplines for Social Sciences include Business, Education, Linguistics, Management, and Sociology. A random selection of six journals from each selected discipline was employed. Each journal under the selected disciplines in Thomson Reuter's Web of Knowledge was designated a number and a random sample chart was used to locate the journals for the study.

Next, ten articles were selected from each journal. The articles included in the study were all from the first issue of the selected journal published in 2010. If the first issue contained less than ten articles, the subsequent issues of the journal in that year were used until there were ten articles selected from the journal.

A total of 600 selected journal articles were uploaded to the text-matching detection system Turnitin to check the originality of the articles. The originality reports include indicating the extent of similarities between the uploaded papers and those found in the Turnitin database. Each article was given a score ranging from 0 (no indications of text-matching) to 100; the higher the score the more matches found by the text-matching detection system.

The data collected from the Turnitin text-matching records and additional human examination by researchers included (1) the originality score for each journal article under investigation provided by Turnitin, (2) the frequency of different strings of consecutive matching text (five or more words from the original source material found *in sequence*), (3) the country of affiliation of the corresponding author (or first author if not identified), (4) the number of authors of each article, (5) the source of matching text (from the author's own publications or from others' publications), (6) discipline (STEM or Social Sciences), and (7) field of study. Item 1 was collected from Turnitin output. Item 2 was based on Turnitin output with human scrutiny to tally and collect the data. Items 3 to 7 were collected by the researcher based on the information presented for each selected article recorded in Thomson Reuter's electronic online database.

2.2. Data analysis

Several studies have indicated that the originality report generated by Turnitin can only serve as a reference point and it takes human scrutiny to examine each incidence of text-matching (Barrett & Malcolm, 2005; Jones, 2008; Sutherland-Smith & Carr, 2005). For example, the article under investigation might already be in the database, thus, matching it to the same article already in the database is unnecessary and cannot be considered text-matching. Likewise, text-matching such as embedded quotations or displayed quotations should be excluded from analysis because these follow standardized citation practices.

Formulae (e.g. in mathematics or chemistry) and terminology were excluded in the data analysis because the current study focuses on linguistic text-matching. In addition, certain terminology that is commonly used in academic writing was not considered to be text-matching in the study. Finally, the Turnitin system will detect matches of titles of an article if that article was cited by others. Also, if the Turnitin database includes a particular author's other publications and the author's bio information happens to be same, Turnitin will also count it as matched text. These matches should, of course, be excluded.

Thus, a manual check was employed by the researcher of this study to make qualitative judgments on the appropriateness of textual re-use. In the current study, human screening of each match was conducted across 4247 matches and the following types of matches were excluded from the results: (1) the exact article found in the Turnitin database, (2) text in quotation marks or displayed quotations, (3) formulae and terminology, and (4) article titles and author information. These exclusions speak to the major limitations of the Turnitin service, which requires careful human scrutiny to avoid any misinterpretation of Turnitin results.

After the abovementioned false text-matching incidences were identified and deleted from Turnitin output data, the originality score presented by Turnitin was modified by the Turnitin system automatically. The modified Turnitin scores were then used for data analysis in the study. Since there is no consensus so far regarding the operational definition of plagiarism and the extent of legitimate paraphrasing (how many consecutive copy words from the source is acceptable), the current study employs an arbitrary cutting point of 30-word strings of consecutively copied material for data analysis. It is believed that this is a safe cutting point, as instances of 30 consecutive words taken verbatim from a source is likely to indicate questionable copying (or textual plagiarism). Data analysis was performed using SPSS, version 20.0 for Windows.

3. Results

3.1. Overview of the selected journal articles

A descriptive analysis of the 600 selected journal articles showed that 166 (28%) of them were single-author articles, whereas 434 (72%) of them had multiple authors. A cross-tabulation analysis of the authors' disciplines further indicated that

256 out of 300 articles (85%) in articles in STEM were multi-author articles, whereas 178 out of 300 articles (59%) in Social Sciences were multi-author. Regarding the official language of the corresponding authors' country of affiliation, 343 out of 600 articles (57%) were published by authors in countries where English is used as an official language, leaving 257 (43%) from regions where English is not the official language. A cross-tabulation analysis of the disciplines further indicated that 60% of the STEM articles ($n = 180$) were published by authors in an English-as-an-official-language region, whereas 74% of the Social Sciences articles ($n = 123$) were published by authors in English-as-an-official-language regions.

3.2. Turnitin scores

Table 1 presents the distribution of the Turnitin scores of the 600 selected journal articles in the study. The lower the score the fewer matches found by the Turnitin text-matching detection software. Sixty-four percent of the journal articles ($n = 385$) received a score lower than 10, 86% ($n = 514$) received a score lower than 20, and 98.2% ($n = 289$) of the articles received a score lower than 41. Only a few papers had a high degree of text matching. The mean Turnitin score is 10.065; standard deviation is 10.068, with the range between 0 and 48.

Table 1
Distribution of Turnitin score on selected 600 journal articles.

Turnitin score	Frequency	Valid percent	Cumulative percent
0–10	385	64	64
11–20	129	22	86
21–30	51	9	95
31–40	24	4	99
41–50	11	1	100

To investigate the differences between STEM and Social Sciences fields, articles in STEM had higher Turnitin scores (Mean = 12.60, SD = 11.57) than articles in Social Sciences (Mean = 7.53, SD = 7.50). Table 2 presents the average mean score, standard deviations, and range of Turnitin originality scores in different fields. The results of the analysis indicate that the Turnitin mean score for single-author articles (Mean = 6.92, SD = 8.17) is lower than for multi-author articles (Mean = 11.27, SD = 10.47). In addition, the mean score for authors in contexts wherein English is an official language had a lower Turnitin mean score (Mean = 9.08, SD = 9.53) than those in contexts wherein English is not (Mean = 11.38, SD = 10.62). As the mean score was negatively skewed (Error of Skewedness = .100), simple log transformation and a negative binomial model and Poisson regression model were used for the statistical analysis. In order to normalize the skewed distribution of text-matching data, negative binomial regression, Poisson models designed to deal with count data, and simple log transformation were used in order to address the over-dispersion of the data.

Table 3 presents the results of negative binomial regression with coefficients and standard errors of the combined effects of the variables on Turnitin scores. The four models reported in Table 3 refer to the results of backward stepwise regression, starting with a full model that includes all theoretically relevant predictive variables and then drops each of the variables to see how the model fits. The results reveal that discipline and authorship are significantly related to Turnitin scores. That is, the articles from STEM tend to have a significantly higher mean score on Turnitin (more matching text) than those from Social Sciences ($p = .001$). Also, multi-author articles tend to have significantly lower Turnitin scores than single-author articles ($p < .001$). Nevertheless, whether or not an author was living in a context in which English is an official language was not significantly associated with their Turnitin score ($p = .539$).

Table 2
Descriptive analysis of Turnitin score on STEM and Social Sciences disciplines.

Field	Discipline	N	Mean	SD	Min	Max
STEM		300	12.60	11.57	0	48
	Biology	60	9.62	10.58	0	45
	Computer Science	60	13.28	12.34	0	45
	Engineering	60	14.57	13.04	0	48
	Math	60	12.73	10.18	1	48
	Physics	60	12.80	11.29	0	48
Social Sciences		300	7.53	7.50	0	42
	Business	60	7.77	7.38	0	34
	Education	60	6.57	7.21	1	34
	Linguistics	60	5.78	8.08	0	42
	Management	60	11.17	7.94	1	36
	Sociology	60	6.37	5.60	0	31

Note: STEM refers to Science, Technology, Engineering, and Mathematics.

Table 3
Negative binomial regression of variables on Turnitin scores.

Variables	Models			
	1	2	3	4
STEM	.392*** (.1143)	.533*** (.1116)		.416*** (.1077)
English	−.068 (.1110)	−.063 (.1128)	−.196 (.1055)	
Author	−.534*** (.1203)		−.648*** (.1055)	−.533*** (.1204)

$N = 600$ $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.
The dummy variables (binary indicators) in the table above have been created in the following way.
STEM = 1 (if a STEM field), STEM = 0 (if not).
English = 1 (if English is an official language), = 0 (if not).
Author = 1 (if a single-author publication), = 0 (if not).

3.3. Consecutive matching text

To investigate the relationship between fields of study, the official language of the first authors' countries of affiliation, and authorship on self-text-matching of 30-word or longer strings of consecutive matching texts, a negative binomial regression was used. The negative binomial regression was chosen because it met the statistical assumptions for a dependent variable, i.e. count data. Due to the diverse lengths of the articles under investigation, log word count of the article length was chosen as an offset variable. Table 4 summarizes the negative binomial regression of variables on self-text-matching. The results reveal that both field (STEM and Social Sciences) and authorship are significantly associated with self-text-matching, with $p < .001$ and $p = .004$ respectively, whereas whether or not English is the official language yields no significant difference on self-text-matching ($p = .077$). All coefficients on English as an official language are negative, but none of them is significant at the 95% level. In other words, no significant difference has been found in terms of whether or not self-text-matching is impacted by an author's country of affiliation wherein English is an official language.

Table 4
Negative binomial regression of variables on self-text-matching (30-word or longer consecutive matching text).

Variables	Models			
	1	2	3	4
STEM	.914*** (.1328)	.844*** (.1263)	1.025*** (.1275)	
English	−.222 (.1259)		.225 (.1257)	−.029 (.1177)
Author	−.439*** (.1512)	−.442** (.1514)		−.735*** (.1429)

$N = 600$ $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.
The dummy variables (binary indicators) in the table above have been created in the following way.
STEM = 1 (if a STEM field), STEM = 0 (if not).
English = 1 (if English is an official language), = 0 (if not).
Author = 1 (if a single-author publication), = 0 (if not).

Table 5 presents the results of negative binomial regression on text-matching of both one's own (self-text-matching) and others' works combined. The results revealed for authors located in contexts wherein English is an official language do not differ significantly from those in contexts wherein English is not an official language. Furthermore, the results also reveal that single-authored articles have consistently significantly fewer incidences of consecutive 30-word or longer strings of matching text whereas STEM has significantly more incidences of consecutive 30-word or longer strings of matching text. In conclusion, the results show that the two most consistently significant independent variables for matching texts in published works are field of study and authorship. In contrast, the official language is not a predictor of matching text.

Table 5
Negative Binomial Regression of variables on text-matching of self-authored and other's works (30 or more consecutive matching texts).

Variables	Models			
	1	2	3	4
STEM	.926*** (.1296)	.888*** (.1236)	1.023*** (.1246)	
English	.123 (.1227)		−.129 (.1226)	−.129 (.1150)
Author	−.387** (.1460)	−.391** (.1461)		−.677*** (.1378)

Footnote.
$N = 600$ $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.
The dummy variables (binary indicators) in the table above have been created in the following way.
STEM = 1 (if a STEM field), STEM = 0 (if not).
English = 1 (if English is an official language), = 0 (if not).
Author = 1 (if a single-author publication), = 0 (if not).

4. Discussion

4.1. Fields and disciplines on text-matching

As previous studies have revealed, standards and expectations on text-matching across disciplines are inconsistent (Crisp, 2004; Sutherland-Smith & Carr, 2005); conventions on presenting the work of others are loosely shared between different disciplines via an informal apprenticeship model (Bretag & Carapiet, 2007). Research question one and two aim to examine if journal articles from STEM and Social Sciences differ in their Turnitin scores and the extent of consecutive matching text. The findings suggest that in practice significant differences exist between journal articles written for STEM and those written for Social Sciences. That is, articles in STEM tend to have more matching text than articles in Social Sciences. Previous research has indicated that multiple authors dilute the shared responsibility of producing original material (Bennett & Taylor, 2003; Rennie et al., 1997). One probable explanation is that of the 600 articles selected for analysis in the current study, 28% of articles were single-author and 72% of articles were multi-author; the higher percentage of single-authorship in Social Sciences could have resulted in fewer incidences of matching text.

4.2. Authorship and text-matching

Research question three examines if the number of authors influences Turnitin scores and incidences of consecutive matching text. The results of the study reveal that single-author articles tend to have significantly lower Turnitin scores (less matching text) and fewer incidences of consecutive-matching text than multi-author articles. One possible explanation is that articles with multiple authors might reduce the responsibility of individual authors and thus make the resultant work more vulnerable for matching texts. Several previous studies have addressed the issue of obscured and diluted author responsibility for the contents of multi-author scholarly papers (Bennett & Taylor, 2003; Rennie et al., 1997; Vesterman, 2002; Vuèkoviæ-Dekiæ, 2003). Multiple authors may maximize credit for a publication but at the same time multi-authorship minimizes responsibility shared among authors. As Vesterman (2002) points out, in a co-authorship writing context the discovery of mistakes can be time-consuming and often unrewarded, and responsibility for uncovering fraud is usually diluted by multi-authorship. Rennie et al. (1997) further states that unless the contribution of each author is disclosed to readers, credit and accountability cannot be assessed. Thus, the system of authorship is flawed in this sense and a conceptual and systematic change is called for to reflect the problems of multi-authorship and to ensure accountability.

4.3. Official language and matching texts

Research question four examines if there are regional differences in terms of the official language used for Turnitin scores and incidences of consecutive matching text. Previous literature suggests that students from different countries hold different perspectives on paraphrasing and matching text (Currie, 1998; Shi, 2004) and those from a "collectivist culture" might see matching text as a positive collaboration with other source authors (Barker, 1997, 115). Furthermore, numerous studies have shown that using English, as the language of international scholarship, puts nonnative-English-speaking scholars at a disadvantage (Canagarajah, 1996; Flowerdew, 1999a, 1999c, 2000; Gosden, 1996; Liu, 2004; Salager-Meyer, 2008; St. John, 1987). This disadvantage lies mainly in the linguistic challenges they face in terms of expression and argumentation, limited vocabulary, and the increased possibility of first language (L1) transfer (Flowerdew, 1999b). However, the quantitative findings of the current study indicate that authors in contexts wherein English is an official language do not differ significantly from their counterparts on their Turnitin scores or the number of 30-word or longer strings of successive matching text from self-published articles and self- and-others' publications combined. One probable explanation regarding the conflicting findings of the current study with previous studies could be that previous studies mostly used a qualitative approach to collect and analyze data through interviews or the textual analysis of writing examples, e.g. comparing the first draft and final version either from students writing (Currie, 1998; Gosden, 1996) or researchers' writing (Flowerdew, 1999b, 2000; St John, 1987). The quantitative approach employed in this area usually employs surveys (Barker, 1997) or the analysis of a limited number of writing examples (Shi, 2004). Few, if any, research has utilized a quantitative approach to analyze massive amounts of published works by researchers. Thus, the findings of the current study can shed new light on the influence of authors (writers) in contexts wherein English is an official language on the actual practice of published authors' intertextuality.

5. Conclusion

In conclusion, the findings of the study can be seen as an opportunity to broaden the discussion on issues of text-matching and paraphrasing across fields, disciplines, authorship, and language context. The numerical data generated by Turnitin is only a starting point for those researchers, instructors, editors, and institutional policy makers involved in scholarly publication to rethink what constitutes text-matching and paraphrasing and why.

5.1. Limitations and recommendations for future research

This study does have some limitations. First, it is a purely quantitative study and does not provide detailed accounts or examples of various types of matching texts. Future research employing a qualitative approach could provide a more in-depth understanding of the patterns of matching texts across disciplines, authorship, and authors' first language. In addition, the current study is based solely on data collected from published works and analyses based on human scrutiny and Turnitin output; future research examining the epistemic knowledge and perceptions of scholars could shed new light on how other variables may influence writers' behavior on paraphrasing and intertexuality. Finally, the current study focuses on large-scale ($n = 600$) journal samples; it did not survey authors' demographic information or identify each author's first language. Future research that examines individual authors' backgrounds more closely could also shed new light on how an author's background influences the intertexuality of their writing for publication.

Acknowledgment

The author is deeply grateful to the anonymous reviewers for their insightful feedback. The project was sponsored by the National Science Council in Taiwan (NSC 101-2410-H-009-034-MY2).

References

Abasi, A. R., Akbari, N., & Graves, B. (2006). Discourse appropriation, construction of identities, and the complex issue of plagiarism: ESL students writing in graduate school. *Journal of Second Language Writing, 15*, 102–117.
Allan, G., Callagher, L., Connors, M., Joyce, D., & Rees, M. (2005). Some Australasian perspectives on academic integrity in the Internet age. In *DUCAUSE Conference [CD-ROM]*. Auckland: University of Auckland.
Angélil-Carter, S. (2000). *Stolen language? Plagiarism in writing*. Harlow: Longman.
Ashworth, P., Bannister, P., & Thorne, P. (1997). Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education, 22*(2), 187–203.
Ballard, B., & Clanchy, J. (1991). Assessment by misconception: cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 19–35). Norwood, NJ: Ablex Publishing.
Barker, J. (1997). The purpose of study, attitudes to study and staff-student relationships. In D. McNamara, & R. Harris. (Eds.), *Overseas students in higher education: Issues in teaching and learning* (pp. 108–123). London: Routledge.
Barrett, R., & Malcolm, J. (2005). Embedding plagiarism education in the assessment process. *International Journal of Educational Integrity, 2*(1), 38–45.
Barrett, R., Malcolm, J., & Lyon, C. (August 2003). Are we ready for large-scale use of plagiarism detection tools?. In *LTSN-ICS Conference Proceedings* (pp. 79–84).
Bennett, R. (2005). Factors associated with student plagiarism in a post-1992 university. *Assessment & Evaluation in Higher Education, 30*(2), 137–162.
Bennett, D. M., & Taylor, D. M. (2003). Unethical practices in authorship of scientific papers. *Emergency Medicine, 15*, 263–270.
Benos, D., Fabres, J., & Farmer, J. (2005). Ethics and scientific publication. *Advances in Physiology Education, 29*, 59–74.
Bretag, T., & Carapiet, S. (2007). A preliminary study to determine the extent of self-plagiarism in Australian academic research. *Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication and Falsification, 2*(5), 1–15.
Bull, J., Collins, C., Coughlin, E., & Sharp, D. (2001). *Technical review of plagiarism detection software*. Prepared for the Joint Information Systems Committee. Available online http://www.jiscpas.ac.uk/documents/resources/Luton_TechnicalReviewofPDS.pdf Accessed 11.12.09.
Butakov, S., & Scherbinin, V. (2009). The toolbox for local and global plagiarism detection. *Computers and Education, 52*(4), 781–788.
Campbell, C. (1990). Writing with others' words: using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 211–230). Cambridge, UK: Cambridge University Press.
Canagarajah, A. S. (1996). "Nondiscursive" requirements in academic publishing, material resources of periphery scholars, and the politics of knowledge production. *Written Communication, 13*(4), 435–472.
Carbone, N. (2001). *Turnitin.com: A pedagogic placebo for plagiarism*. TechNotes. Available online http://bedfordstmartins.com/technotes/techtiparchive/ttip060501.htm Accessed 21.08.06.
Chaky, M., & Diekhoff, M. (2002). A comparison of traditional and Internet cheaters. *Journal of College Student Development, 43*(6), 906–911.
Crisp, G. T. (2004). Plagiarism and the reputation of the university: how to distribute effort between educating students on attribution and rigorous detection of cheating?. In *Proceedings of the Australian Universities Quality Forum 2004* (pp. 52–58) AUQA Occasional Publication.
Currie, P. (1998). Staying out of trouble: apparent plagiarism and academic survival. *Journal of Second Language Writing, 7*, 1–18.
Dahl, S. (2007). Turnitin®: the student perspective on using plagiarism detection software. *Active Learning in Higher Education, 8*(2), 173–191.
Ellery, K. (2008). Undergraduate plagiarism: a pedagogical perspective. *Assessment and Evaluation in Higher Education, 33*(5), 507–516.
Emerson, L., Rees, M., & MacKay, B. (2005). Scaffolding academic integrity: creating a learning context for teaching referencing skills. *Journal of University Learning and Teaching Practice, 2*(3), 12–24.
Ercegovac, Z., & Richardson, J. V. (2004). Academic dishonesty, plagiarism included, in the digital age: a literature review. *College & Research Libraries, 65*(4), 301–318.
Flowerdew, J. (1999a). Writing for scholarly publication in English: the case of Hong Kong. *Journal of Second Language Writing, 8*(2), 123–145.
Flowerdew, J. (1999b). Problems in writing for scholarly publication in English: the case of Hong Kong. *Journal of Second Language Writing, 8*(3), 243–264.
Flowerdew, J. (1999c). *TESOL Quarterly* and non-native speaker-writers: an interview with Sandra McKay. *Asian Journal of English Language Teaching, 9*, 99–103.
Flowerdew, J. (2000). Discourse community, legitimate peripheral participation, and the nonnative-English-speaking scholar. *TESOL Quarterly, 34*, 127–150.
Flowerdew, J., & Li, Y. (2007). Plagiarism and second language writing in an electronic age. *Annual Review of Applied Linguistics, 27*, 161–183.
Franklyn-Stokes, A., & Newstead, S. E. (1995). Undergraduate cheating: who does what and why? *Studies in Higher Education, 20*(2), 159–172.
Gosden, H. (1996). Verbal reports of Japanese novices' research writing practices in English. *Journal of Second Language Writing, 5*(2), 109–128.
Graham, M. A., Monday, J., O'Brien, K., & Steffan, S. (1994). Cheating at small colleges: an examination of student and faculty attitudes and behaviours. *Journal of College Student Development, 35*, 255–260.
Handa, N., & Power, C. (2005). Land and discover! A case study investigating the cultural context of plagiarism. *Journal of University Teaching and Learning Practice, 2*(3b), 64–85.
Howard, R. M. (1995). Plagiarisms, authorship, and the academic death penalty. *College English, 57*(7), 788–806.
Howard, R. M. (2007). Understanding "Internet plagiarism". *Computers and Composition, 24*, 3–15.
Jones, O. K. (2008). *Practical issues for academics using the Tujrnitin plagiarism detection software* (pp. 50). CompSysTech.
Jones, O. K., & Moore, A. T. (2010). Turnitin is not the primary weapon in the campaign against plagiarism. *CompSysTech, 10*, 425–429.
Keck, C. (2006). The use of paraphrase in summary writing: a comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*, 261–278.

Liddell, I. J. (2003). A comprehensive definition of plagiarism. *Community & Junior College Libraries, 11*(3), 43–52.
Liu, J. (2004). Co-constructing academic discourse from the periphery: Chinese applied linguists' centripetal participation in scholarly publication. *Asian Journal of English Language Teaching, 14*, 1–22.
Martin, B. (2004). *Plagiarism: Policy against cheating or policy for learning?*. Available online http://www.bmartin.cc/pubs/04plag.pdf Accessed 22.06.12.
Mulcahy, S., & Goodacre, C. (2004). Opening Pandora's box of academic integrity: using plagiarism detection software. In *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference* (pp. 688–696). Available online http://www.ascilite.org.au/conferences/perth04/procs/mulcahy.html Accessed 25.06.12.
Oshima, A., & Hogue, A. (1999). *Writing academic English*. New York: Addison Wesley Longman.
Pecorari, D. (2003). Good and original: plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*, 317–345.
Purdy, J. P. (2009). Anxiety and the archive: understanding plagiarism detection. *Computers and Composition, 26*, 65–77.
Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. A proposal to make contributors accountable. *JAMA, 78*(7), 579–585.
Rolfe, V. (2011). Can Turnitin be used to provide instant formative feedback? *British Journal of Educational Technology, 42*(4), 701–710.
Salager-Meyer, F. (2008). Scientific publishing in developing countries: challenges for the future. *Journal of English for Academic Purposes, 7*, 121–132.
Savage, S. (2004). Staff and student responses to a trial of Turnitin plagiarism detection software. In *Proceedings of the Australian Universities Quality Forum: Quality in a time of change*. AUQA Occasional Publications.
Selwyn, N. (2008). Not necessarily a bad thing: a study of online plagiarism amongst undergraduate students. *Assessment and Evaluation in Higher Education, 33*(5), 465–479.
Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171–200.
St. John, M. J. (1987). Writing processes of Spanish scientists publishing in English. *English for Specific Purposes, 6*, 113–120.
Stapleton, P. (2012). Gauging the effectiveness of anti-plagiarism software: an empirical study of second language graduate writers. *Journal of English for Academic Purpose, 11*(2), 125–133.
Stetter, M. E. (2008). Plagiarism and the use of Blackboard's Turnitin. In J. Luca, & E. Weippl (Eds.), *Proceedings of world Conference on educational multimedia, hypermedia and telecommunications* (pp. 5,083–5,085). Chesapeake: AACE. Available online http://www.editlib.org/p/29077 Accessed 25.06.12.
Sun, Y. C. (2009). Using a two-tier test in examining Taiwan graduate students' perspectives on paraphrasing strategies. *Asia Pacific Education Review, 10*(3), 399–408.
Sun, Y. C. (2012). Does text readability matter? A study of paraphrasing and plagiarism in English as a foreign language writing context. *The Asia-Pacific Education Researcher, 21*(2), 296–306.
Sutherland-Smith, W., & Carr, R. (2005). Turnitin.com: teachers' perspectives of anti-plagiarism software in raising issues of educational integrity. *Journal of University Teaching and Learning Practice, 2*(3b), 94–101.
Turnitin. (2006). Available online http://www.turnitin.com/en_us/ Accessed 11.06.13.
Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: individual propensities in cheating. *British Journal of Educational Technology, 34*(4), 467–477.
Vesterman, W. (2002). The death of the scientific author: multiple authorship in scientific papers. *Common Knowledge, 8*(3), 439–448.
Vuèkoviæ-Dekiæ, L. (2003). Authorship-coauthorship. *Archive of Oncology, 11*(3), 211–212.
Warn, J. (2006). Plagiarism software: no magic bullet! *Higher Education Research and Development, 25*(2), 195–208.
Weiler, A. (2004). Information-seeking behaviour in generation Y students: motivation, critical thinking, and learning theory. *Journal of Academic Librarianship, 31*(1), 46–53.
Wood, G. (2004). Academic original sin: plagiarism, the Internet and librarians. *Journal of Academic Librarianship, 30*(3), 237–242.
Yamada, K. (2003). What prevents ESL/EFL writers from avoiding plagiarism? Analyses of 10 North American college websites. *System, 31*(2), 247–258.
Yeo, S. (2007). First-year university science and engineering students' understanding of plagiarism. *Higher Education Research and Development, 26*(2), 199–216.

Yu-Chih Sun is a professor in the Institute of TESOL at National Chiao Tung University, Taiwan. Her research interests include computer-assisted language learning, academic writing, and second language acquisition.