Students’ Cyber-Plagiarism

Tuomo Kakkonen

University of Eastern Finland, Finland

Maxim Mozgovoy

University of Aizu, Japan

ABSTRACT

Educational institutions fight against students’ plagiarism, because plagiarizing contradicts the most basic learning principles. Among ways of countering student plagiarism are understanding the reasons behind it and developing measures to detect and prevent it. Wide availability of text and other resources on the Internet makes plagiarizing easy and, thus, commonplace among students. Students’ cyber-plagiarism sets new challenges to educational institutions. Computers and technology, however, also provide solutions to the problem; Plagiarism detection and prevention can be supported by automatic detection technologies that help to reveal instances of plagiarism. Ongoing research in student cyber-plagiarism is also concerned with the various ethical questions that plagiarism and its detection arise.

INTRODUCTION

_Cyber-plagiarism_ is a type of academic dishonesty that consists of reusing whole electronic documents, or parts of them, composed by another author without proper acknowledgment of the original source. _Web plagiarism_ is a specific type of cyber-plagiarism that consists of copying texts
from the Internet. The term student plagiarism is often used to refer to the incidents of plagiarism committed by students who attend educational institutions. Teachers and academics abhor plagiarism because it is inconsistent with pedagogical aims. The mere copying of texts has no educational value. Moreover, it involves students in moral compromise and deception. Letting plagiarists incorporate parts of external texts into their works is, of course, not fair on the honest students who do not plagiarize.

Although the easy availability of information on the Internet has undoubtedly increased the incidence of plagiarism (see, for example, Lathrop & Foss, 2000), the web has also provided educators with means of countering it. For instance, Internet search engines can be used in the detection of plagiarism. A manual detection process is, by any standards, both tedious and labor-intensive, however, because it requires an assessor to insert parenthesized extracts from a suspected text into the web search engine, and to examine the results manually. Fortunately, numerous systems that automate plagiarism detection are available at the present time.

Plagiarism is a widely researched topic. Typing the search keyword “plagiarism” in Google Scholar, as of September 2010, produced close to seventy thousand links to scientific articles. Hence, it is difficult to name just a few representative researchers and works from this massive body of research literature. In order to provide an overview of the most influential work in the area, the list below provides selected works of some of the most frequently cited authors and titles in plagiarism research as well as examples of recent research work:

General references

- Alastair Pennycook, University of Technology Sydney, Australia: Borrowing Others’ Words: Text, Ownership, Memory, and Plagiarism (Pennycook, 1996).
Plagiarism and its prevention in higher education

- Peter Ashworth et al. Sheffield Hallam University, UK: Guilty in Whose Eyes? University Students’ Perceptions of Cheating and Plagiarism in Academic Work and Assessment (Ashworth, Bannister & Thorne, 1997).


Detection


- Xin Chen et al., University of California, USA: Shared Information and Program Plagiarism Detection (Xin Chen et al., 2005).

The most active development of automatic plagiarism detection systems takes place in software companies as part of their business activities. The best known examples of such companies are iParadigms and Blackboard, the developers of TurnitIn (iParadigms, 2010a) and SafeAssign (Blackboard, 2010) plagiarism detection systems, respectively. The technical details of the work done in these companies remain, for obvious reasons, largely unpublished.

OVERVIEW OF THE INTELLECTUAL HISTORY

According to Lynch (2006), the word “plagiary” did not, despite its Latin root, appear in the English language until the beginning of the seventeenth century when it featured in Ben Jonson’s satire The Poetaster (Jonson, 1602). The concept of plagiarism appears to have been unrecognized before
then. In the late nineteenth century, plagiarism among students was a widespread phenomenon and was recognized as a problem by educators (Simmons, 1999). At that time, *essay mills*, libraries of ready-made essays and homework available to students for free or for a fee, were already operating.

The rapid development of computer technologies and the Internet in the 1990s have made plagiarizing much easier; this is especially true for blatant verbatim copying that consists only of copying and pasting text from an online source. The vast amount of text available on the Internet, moreover, makes it tempting for some to resort to dishonest practices when authoring a text. Essay mills also migrated online and their number has grown enormously.

During the 1990s, a number of computerized plagiarism detection systems were developed by various software companies. Academic interest in computer-aided plagiarism detection in software source code dates back to the 1970s (Ottenstein, 1976), but early systems were not reliable enough to be used in practical learning environments, and electronic documents were not widespread. Such systems have become an integral part of the assessment process in many educational institutions. For example, according to iParadigms, their product Turnitin “is used in approximately 9,500 educational institutions (including middle schools, high schools, colleges and universities) across 110 countries” (iParadigms, 2010b).

**THE CURRENT KNOWLEDGE STATUS**

Current research is mainly concentrated in the following areas: studying plagiarism as a social phenomenon, detecting plagiarism, examining plagiarism-related and plagiarism detection-related ethical and legal issues, and prevention of plagiarism. The following subsections provide an overview of each of these areas of research.
Plagiarism as a Social Phenomenon

Social studies of plagiarism primarily seek answers to the following questions: Which actions should be termed “plagiarism”? Why do people commit plagiarism? What is society’s attitude to plagiarism? Which types of plagiarism can be identified? Since plagiarism is especially common among students and pupils at all levels of study, many research studies are devoted to answering these questions specifically in the context of student plagiarism.

The concept of plagiarism significantly overlaps but should not be confused with the idea of copyright infringement (Posner, 2007): whereas copyright only protects the exact form in which ideas are expressed, the “stealing of ideas” more accurately constitutes plagiarism. Furthermore, a plagiarist claims to be the original author of the given text, whereas copyright violations might not include false authorship attribution. In the case of student plagiarism, other closely related violations include fabrication and cheating.

Park (2003) mentioned the following most common reasons for student plagiarism: ignorance about correct citing and referencing methods, a desire to obtain better grades by including superior material without proper acknowledgment, inadequate time management skills, and the small perceived risk of getting caught.

In the context of student assignments and other academic works, plagiarism is an academic offence and not a legal offence, and is controlled by institutional rules and regulations (Myers, 1998; Larkham & Manns, 2002). Most universities and other educational institutions regard plagiarism as a form of cheating or academic misconduct, but the specific rules and regulations for dealing with suspected cases of plagiarism are different. Consequently, the penalties imposed for cheating vary from institution to institution, and depend on factors such as the severity of the offense and whether the student admits to the offense. The possible actions include giving a zero mark for the plagiarized work, resubmission of the work, and even dismissal from the university in serious cases.
Plagiarists use various techniques in attempts to disguise plagiarism in their work. Table 1 represents five levels of classification of plagiarism types (Kakkonen & Mozgovoy, 2010).
Table 1.

*Five types of plagiarism*

<table>
<thead>
<tr>
<th>Plagiarism type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Verbatim copying</td>
<td>Copy-paste copying from an electronic source. This includes blatant plagiarism or authorship plagiarism, which refers to taking someone else’s text and putting one’s own name to it. Word-for-word transcription of texts from a non-electronic source.</td>
</tr>
<tr>
<td>(2) Paraphrasing</td>
<td>Adding deliberate spelling and grammatical mistakes. Replacing words with synonyms. Reordering sentences and phrases. Effecting changes to grammar and style.</td>
</tr>
<tr>
<td>(3) Technical tricks</td>
<td>Various tricks that exploit weaknesses of plagiarism detection systems: for example, the insertion of (a) similar-looking characters from foreign alphabets or (b) invisible white-colored letters into what seem to the reader to be blank spaces but are interpreted by a detection system as a character.</td>
</tr>
<tr>
<td>(4) Deliberate inaccurate use of references</td>
<td>The improper and inaccurate use of quotation marks: the failure to identify cited text with the necessary accuracy. Fabrication: providing made-up references that do not exist. Falsification: providing false references, i.e. references exist but do not match the text being referenced. The use of “forgotten” or expired links to sources.</td>
</tr>
<tr>
<td>(5) “Tough plagiarism,” i.e. the types of plagiarism that are particularly difficult to detect</td>
<td>The plagiarism of ideas: the use of similar concepts or opinions without due acknowledgment. Translated plagiarism: translations unsupported by acknowledgment of the original work. Ghostwriters: the use of text produced by an independent “ghostwriter.” Artistic plagiarism: the presentation of someone else's work in a different medium (for example, images, voice or video). Plagiarism of the structure of an argument: involves looking up references and following the structure of the secondary source.</td>
</tr>
</tbody>
</table>
“Traditional” paper and pen plagiarists used and use the same ways of plagiarizing and concealing plagiarism as cyber-plagiarists do. In the list above, type 4, based on exploiting the weaknesses of automatic plagiarism detectors, can be considered as the only type of plagiarism that is totally new to cyber-plagiarism or rather to the era of automatic plagiarism detection systems. Verbatim copying and paraphrasing have always existed; copy-pasting text from Internet is just easier than verbatim copying from offline sources.

Wide availability of free translation services (such as Babelfish and Google Translate) has made translated plagiarism much easier to commit, and one could argue that it is a commoner form of plagiarism than it used to be before the era of Internet and machine translation. Though the quality of automatic document translation is far from perfect, a student can polish the translation manually. This type of plagiarism is especially easy for students who speak several languages. In Europe, for instance, the number of multilingual people is as high as 79% (European Commission, 2005).

Several recent studies have examined culture-specific attitudes towards plagiarism. For example, Shei (2005) claimed that Chinese students often have more relaxed views of plagiarism compared with the Western “plagiarism as stealing” position. The explanation lies in the traditional Chinese learning model, when imitation of another’s works is considered a natural part of study. Beute, van Aswegen and Winberg (2008), and Alaoutinen et al. (2005) have observed differences in attitudes to plagiarism between students from different African nations and between Finnish and Russian students respectively. It is evident from these findings that special measures have to be taken against plagiarism, as well as other types of academic dishonesty, in multi-cultural learning environments. In addition to strict enforcement of anti-plagiarism codes, proper guidance and an explicit introduction to the concept of plagiarism are necessary.
Detecting plagiarism

The most common method of plagiarism detection relies on the ability of a human assessor to spot it. For example, an author whose writing style is simple might suddenly become uncharacteristically eloquent in some text passages. An assessor can confirm the suspicion of plagiarism by manually matching the suspected plagiarisms to the original sources. Countering plagiarism by using only such manual methods is tedious, time-consuming and prone to errors. In an era of information explosion few assessors, if any, can be familiar with all the relevant sources of a topic to identify the origin of each and every instance of plagiarism.

Typical plagiarism clues useful for manual checking include (Jackson & Kern, 2003; Harris, 2004; Ryan, 1998):

- Peculiar formatting, such as the unnecessary insertion of line breaks or sudden changes in font types.
- Outdated quotations and citations may indicate reuse of an old paper.
- Inconsistent bibliography and quotation style may indicate that the text was composed of several source documents.
- Inconsistent writing style: dissimilar writing style in adjacent paragraphs or compared with the author’s earlier works.
- The use of unexpectedly advanced vocabulary or surprisingly good language constructions relative to the other parts of the document or the author’s earlier works.

Owing to the difficulty of manual plagiarism detection, automated detection systems are widely popular nowadays. From the user’s point of view, the primary differences between the available detection systems include:

- Distribution model. The system can be available as a desktop application (for instance, Sherlock; Joy & Luck, 1999) or as a web service (e.g., TurnitIn; iParadigms, 2010a).
• **Specialization.** The system can be optimized for a particular type of input data, such as natural language texts (TurnitIn) or software source code (e.g., JPlag; Prechelt et al., 2002).

• **Web plagiarism detection capabilities.** The system can be enabled to search the Internet for the potential original sources of the given text (TurnitIn), or just compare locally all the submissions with each other (JPlag).

• **Access and license type.** Typical options are subscription-based access (Turnitin), unlimited license for a fixed price (e.g., EVE2; Canexus, 2000), free access (JPlag), GNU-licensed free software (e.g., Plaggie; Ahtiainen et al., 2006).

Some types of plagiarism are easier to detect than others. Current automatic detection systems, as well as humans, are relatively accurate when it comes to detecting verbatim copying (type 1 in Table 1). Both humans and state-of-the-art plagiarism detection systems are capable of detecting plagiarism that involves paraphrasing (type 2). Heavily paraphrased plagiarism, however, is not easy to spot either by humans or by computers. Plagiarism of type 3 is technically easy to see, but surprisingly most current detection systems do not implement any counter-measures against these simple tricks (Kakkonen & Mozgovoy, 2010). Humans are immune to this type of plagiarism. Type 4 plagiarism (false references) is beyond the capabilities of even the best computerized detectors. Human assessors also find this type of plagiarism quite challenging. As the name suggests, “tough plagiarism” (type 5) is difficult, if not impossible in many cases, for both computers and humans to detect.

**Ethical and Legal Issues**

**Issues common to manual and automatic plagiarism detection.** The most difficult cases of suspected plagiarism to prove are those in which the original source is impossible to find; the lack of evidence makes it hard to prove guilt (Larkham & Manns, 2002; Joy & Luck, 1999). *Ghostwriting* is a serious issue in “traditional” and student cyber-plagiarism alike. If the original text was composed by an external ghostwriter, neither the original nor the fake authors are interested in revealing the truth. An
interesting real-life detective story involving a case of ghostwriter plagiarism is recounted by Zobel (2004). Both the student and the ghostwriter were eventually sentenced in a court of law.

An interesting ethical issue is raised by the open publication of specific plagiarism hiding and detection methods. Such information is helpful to plagiarists, and can be used as a “how not to get caught” guide. An informal book of Quickwit entitled “Plagiarism: How Profs Spot a Cheat” (Quickwit, 2004) is an example of such a guide. It is promoted by the publisher as a valuable resource for both students and professors.

**Issues specific to automatic detection.** The use of computer-aided plagiarism detection raises a number of ethical and legal issues (see, e.g., Foster, 2002). These issues are caused both by technical imperfection of plagiarism detection algorithms (e.g., a system might incorrectly suspect a student’s work of being plagiarized), and by misunderstanding of the role of plagiarism detection software in the educational process.

Students may complain about the low quality of plagiarism detection systems. In fact, plagiarism detection systems do occasionally make *false detections*. Naturally, a student feels insulted when falsely accused of plagiarism after investing considerable time and efforts in composing their work. False accusations arise partly because of misunderstanding of the role of automatic plagiarism detection. Considering the inaccuracies and deficiencies of automatic detection systems, educators should treat them as auxiliary tools rather than providers of absolute proof of plagiarism. Detection tools alert a teacher to the possibility of plagiarism, but it is the responsibility of the educator to verify the suspicion. It is never enough just to detect plagiarism; the specific instances need to be verified one by one because some suspected cases of plagiarism may turn out to be coincidental similarities between two texts (Joy & Luck, 1999).

Perceived low quality of automated plagiarism detection is not the only objection against plagiarism detection software. Another common argument states that students’ essays should not be
stored in an online database because it violates the authors’ intellectual property rights and taints the students with an unjustified “presumption of guilt.” (Glod, 2006) This issue arises specifically with Turnitin as the system retains an internal database of student essays (See, e.g., Jones, 2007). Responses to this argument are heavily debated. The proponents of storage of student texts argue that the use of a plagiarism detection system is categorically similar to sanctioning the presence of a referee in a football match. Therefore, it does not violate our customary understanding of a person’s presumption of innocence (Foster, 2002). An online database of essays could also be compared with Google cache (a module that automatically collects and stores Internet pages), which is currently considered in most countries as a legal tool.

**Prevention of plagiarism.** At least as important as efficient and accurate detection of plagiarism is its prevention. There is a great variety of techniques that can be used for lowering the instances of plagiarism among students. These include (Lathrop & Foss, 2000; Park, 2003; Posner, 2007):

1. **Better design of assignments.** Narrowly focused assignment topics, specific restrictions on the allowed sources, process approach scheme (different assignments are different stages of the same process). Furthermore, a teacher can require a more personalized task than a usual term paper, such as a problem-solving diary.

2. **Honor codes.** Educational institutions should develop honor codes, and strictly apply them. Each student should be aware that plagiarism is officially considered unacceptable by the institution and that plagiarizing violates the honor code. Breach of the code should, moreover, have consequences for the plagiarist.

3. **Improved student-teacher interaction schemes.** Educators should explain to students how to avoid plagiarism (i.e. what is considered plagiarism, how to cite the sources properly, etc.)
Teachers themselves should set an example by using correct citation and referencing methods in their lecture notes and presentations.

**FUTURE RESEARCH DIRECTIONS**

As computers are nowadays used as the main tool for authoring texts, and online research has become an integral part of the writing process, there is a legitimate concern that plagiarism will become an ever more critical issue for education in general and for writing instruction in particular. Plagiarism prevention will therefore be an increasingly important concern for educational institutions.

Most of the existing plagiarism detection systems are only capable of detecting direct copying and the kind of plagiarism that consists of simple editorial changes such as the addition, removal or substitution of various characters. Since it is unrealistic to assume that all possible sources of plagiarism will be contained in the detection system’s internal collection or in public web access, the attribution of authorship by means of creating stylistic fingerprints of authors is of crucial importance. This capability will help to create more advanced plagiarism detection software for identifying plagiarism of types 1 (verbatim copying), 2 and 3. A field of research referred to as “reference and citation tracking” aims at developing methods of finding citations and references from texts automatically. These techniques will provide a base from which to develop methods of detecting plagiarism type 4 (deliberate inaccurate use of references).

As long as reliable methods for detecting types 1 to 4 of plagiarism do not exist, it is unreasonable to expect successful progress in the detection of tough plagiarism. Hence, tough plagiarism will probably remain as such for the foreseeable future. Another type of plagiarism that goes beyond the capabilities of the state-of-the-art systems is plagiarism that consists of non-textual data such as images, music, or website design. At the present time, one can only speculate as to how widespread these forms of plagiarism might be among students in the field of art and design.
REFERENCES


**ADDITIONAL READINGS**


**KEY TERMS**

Cyber-plagiarism, student plagiarism, plagiarism prevention, plagiarism detection, cheating.