Plagiarism detection systems and international students: detecting plagiarism, copying or learning?

Lucas Introna and Niall Hayes

Centre for the Study of Technology and Organisation,
Lancaster University Management School,
Lancaster University, Lancaster, LA1 4YX, UK.
l.introna@lnacaster.ac.uk

Abstract

This paper explores the question of plagiarism by international students (non-native speakers). It argues that the inappropriate use of electronic plagiarism detection systems (such as Turnitin) could lead to the unfair and unjust construction of international students as plagiarists. We argue that the use of detection systems should take into account the writing practices used by those who write as novices in a non-native language as well as the way 'plagiarism' or plagiaristic forms of writing are valued in other cultures. It calls for a move away from a punitive legalistic approach to plagiarism that equates copying to plagiarism and move to a progressive and formative approach. If taken up such an approach will have very important implications for the way universities in the west deal with plagiarism in their learning and teaching practice as well as their disciplinary procedures.

Introduction

The issue of academic integrity within higher education has received considerable attention in the literature over recent years (Harris, 2001; Carroll & Appleton, 2001; Lathrop, 2000; Dryden, 1999; Myers, 1998; Pennycook, 1996; Scollon, 1995; Howard, 1995, 1993; Deckert, 1993; Sherman, 1992; Kolich, 1983). Much of this literature, coupled with the considerable anecdotal evidence amongst colleagues within our own and other universities, suggests that plagiarism is on the increase. O'Connor (2003) describes one recent Australian study that spanned twenty subjects and six universities. This saw 1925 essays being submitted into Turnitin, an electronic detection service that compares electronic work submitted with the 2.6 billion publicly available pages on the internet, and to all the essays previously submitted to Turnitin for checking. This study found that 14% of essays "contained unacceptable levels of unattributed materials." Further, unacceptable levels of plagiarism were found to be present in all six universities and in over 70% of the subjects.

The report also highlighted that what was detected electronically is just the tip of the iceberg, as Turnitin did not cover most books, journals and paper mills etc (O'Connor,2003).

In relation to the literature that has considered why students plagiarise, Carroll (2002) has suggested that most students are unsure what plagiarism is. She argues that this lack of understanding of what is and what is not plagiarism contributes to students plagiarising unintentionally. Furthermore, Angelil-Carter (2000) claim that there is also a lack of clarity across a university about what constitutes plagiarism and a discrepancy in the way plagiarism is detected and enforced (Biggs, 1994; Ryan, 2000; Scollon, 1995). Others have highlighted the growing staff student ratio as being implicated in the rise in the number of cases of plagiarism. They suggest this results in staff having less time to deal with students as individuals and hence less opportunity to talk through issues regarding writing practices (Angelova & Riazantseva, 1999; O'Donoghue, 1996). Carroll (2002) also argues that the move from examination to coursework and project based assessment has resulted in not just over assessment, but students experiencing continual pressure to attain high marks (Carroll, 2000). Others suggest that poor time management by students, or the institutions setting simultaneous deadlines is a major contributing factor (Errey, 2002).

The purpose of this paper is not to revisit these arguments about the increase (or not) of plagiarism or why students find themselves plagiarising. It is our view that many of these papers and arguments deal with a rather oversimplified view of plagiarism, especially with regard to international students[1] (i.e. non-native English speakers). The purpose of this paper is rather to explore the complex interaction between cultural values, writing practices and electronic plagiarism detection systems as depicted in Figure 1.

The central argument of this paper is that the inappropriate use of electronic plagiarism detection systems (such as Turnitin) could lead to the unfair and unjust construction of international students as plagiarists, with obvious devastating consequences.

[1] We use the term 'international student' to refer to those students that come from a culture in which copying is valued differently—as compared to the UK—and who are non-native English speakers, i.e. have English as a second or third language. Currently these students represent 26% of the UK postgraduate populations (UKCOSA).

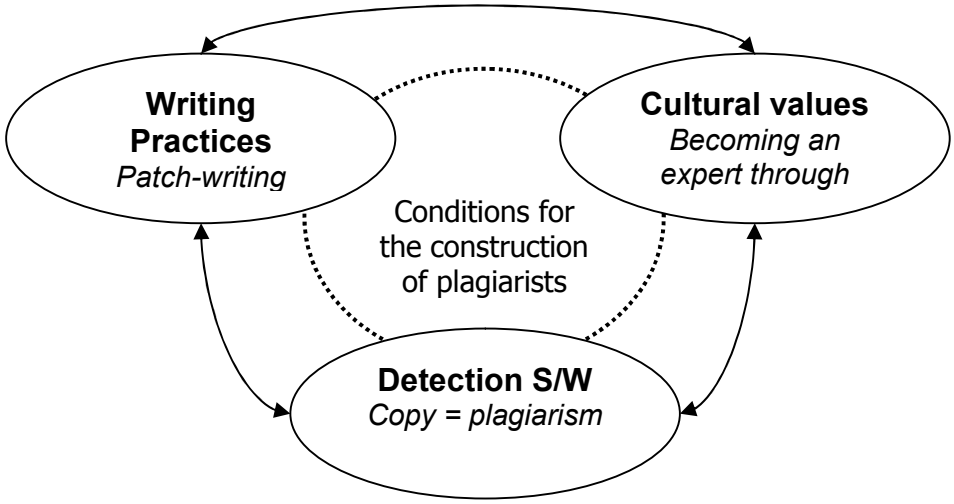

Figure 1: Conditions that mediate the construction of international students as plagiarists

This 'inappropriate' use that we refer to flows from three sets of interrelated assumptions or misunderstandings:

a. A misunderstanding of the writing practices used by those who write as novices in a non-native language
b. Inappropriate assumptions about the way 'plagiarism' or plagiaristic forms of writing (such as copying) are valued in other cultures
c. A dualistic view of plagiarism that does not take into account the practices and values referred to in (a) and (b) above.

We would argue that plagiarism is not a simple phenomenon. It is not a simple choice between cheating and not cheating. There are a number of complex conditions that are shaping the actual writing practices of students (and international students in particular). It is not realistic or fair for us to take a reductionistic approach in dealing with plagiarism by international students. The following sections will consider these issues in more detail.

Becoming an academic 'speaker' (with the help of patches)

Non-native speakers and novices in a discourse often 'speak' or write by means of a practice Howard (1993) calls 'patch-writing'. Howard (1993) defines patch-writing as

"copying from a source text and then deleting some words, altering grammatical structures, or plugging in one-for-one synonym-substitutes" (p. 213). She argues that writers often turn to patch-writing when they are unsure of their understanding of the material or lack confidence in the use of a particular language (such as academic language and phraseology). Patch-writing can be seen as a form of mimicking behaviour. Students normally understand how important it is to 'speak' like the teachers and the people they read to become accepted into the community. Howard argues, following Hull and Rose (1989) that patch-writing is a legitimate attempt to "interact with the text, relate it to your own experiences, derive your own meaning from it" (p. 150). This interaction directly with the text in order to derive one's own meaning from it is something most novice readers/writers do in unfamiliar contexts. This type of engagement is often characterised by copying (mimicking) with ever increasing adoption as confidence grows. Defending patch-writing seems reasonable when we consider that we all learn new skills by mimicking or copying others considered exemplary. Let us consider the example of learning a new language, which seems appropriate in this case. One could even argue that native English speakers also need to learn a new language when they are expected to express their ideas in academic writing. In the case of non-native speakers there is a double hurdle to cross.

Most language courses do not teach you the vocabulary and the grammar separately and then expect you to independently construct meaningful sentences. They tend to follow a very different approach. They normally start by teaching you meaningful phrases in situated contexts, such as how to ask for a glass of water in a restaurant. Only once you have mastered a sufficiently large set of situated phrases and understand when and how to use them appropriately can you begin the next step. In the next step you are expected to selectively and carefully change parts of the phrases in appropriate ways. The way you would converse as a beginner would then be to use these phrases as 'building blocks' changing them ever so slightly in various situations in order to express meaningfully in that particular situation. Only once you have become competent at this level of expression can you begin to build phrases independently in order to convey your intentions more precisely. This account of becoming an expert, moving from the 'standard patterns' to specific situated instances, was used by Hubert Dreyfus (1992) to provide a devastating critique of artificial intelligence research.

If we use this basic model for learning a language to understand the steps that international students might be going through when learning to do academic writing then it is easy to understand how they tend to use patch-writing as a way to deal with their lack of skills. When international students go to study in the UK, USA and Australia[2], they are given the vocabulary (theoretical ideas) and the grammar (academic style of writing, rules of structure, rules of argumentation, conventions for referencing, etc) and then are expected to be expert users of the academic language, and to be able to converse (write an essay) by directly creating independent phrases. However, these ideas, rules and conventions (even if they are individually understood) do not provide non native students with the necessary skill to speak the academic language, as it does not with speaking any other language. What is in fact happening is that as beginners, they are using exemplary sentences and paragraphs as situated phrases (or patches in Howard's terminology) to develop competency in 'speaking' academically. This strategy is evident in this comment by a Greek student in an earlier study (Hayes and Introna, 2005): "*taking a bit here and there helps with getting meaning across. Paraphrasing if you are not a native speaker is difficult."* Even the more competent 'speakers' will tend to use meaningfully modified phrases as a way to sustain a conversation (essay)—as is clear from this comment by an English speaking student during the same piece of research: *"If you take all the sentences / paragraphs from other authors – then you have to do the work to put it together – you have learned and need a certain understanding of the topic, it is not just blatant copying."* (Hayes and Introna, 2005).

Furthermore, it is possible to show that students can use patches from an original document to say something quite different from that which the original author has said. Consider the example below in Table 1. In this example the original text is making the argument that computers in writing can lead to a better quality outcome. The patch-written text uses the original base text as a patch, both in terms of providing meaningful formulations of certain ideas as well as providing an overall narrative structure. Nevertheless, the patch-writer—using the majority of the original text—expresses an *independent* argument to suggest that the research surveyed is only relevant for particular situations and that one cannot make general conclusions from these about the value of computers in writing. Clearly, the patch-written text below would normally be considered in Western universities as being

[2] Australia, the UK and the USA are the three most popular destinations for international students. However, our analysis is also equally relevant to countires whose educations systems resemble these three.

plagiaristic.[3] Nevertheless, we would argue it also indicates independent thought based on informed and sound logical argumentation. In this case the patch-writing is not very sophisticated. More sophisticated forms of patch-writing may use patches from a variety of documents patched together around an independently formulated argument.

The original text	The patch-written text (added text)
In general, the research on word processors and student writing conducted during the 1980's and early 1990's suggests many ways in which writing on computers may help students produce better work. Although much of this research was performed before large numbers of computers were present in schools, formal studies report that when students write on computer they tend to produce more text and make more revisions (Dauite, 1986; Vacc, 1987). Studies that compare student work produced on computer with work produced on paper find that for some groups of students, writing on computer also had a positive effect on the quality of student writing (Hannafin & Dalton, 1987; Owston, 1991;). This positive effect is strongest for students with learning disabilities, early elementary-aged students and college-aged students (Hass & Hayes, 1986; Phoenix & Hannan, 1984; Sitko & Crealock, 1986). Additionally, when applied to meet curricular goals, education technology provides alternative approaches to sustaining student interest, developing student knowledge and skill, and provides supplementary materials that teachers can use to extend student learning. Although earlier research syntheses reveal just modest trends, individual studies of that era have shown that writing with a computer can increase the amount of writing students perform, the extent to which students edit their writing (Dauite, 1986; Etchinson, 1989; Vacc, 1987), which, in turn, leads to higher quality writing (Hannafin & Dalton, 1987; Kerchner & Kistinger, 1984; Williamson & Pence, 1989). **Text from:** Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A metaanalysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Available from http://www.jtla.org.	In general, the research on word processors and student writing conducted during the 1980's and early 1990's suggests that there might be many ways in which writing on computers may not always be helpful for students to produce better work, except in very particular cases or task situations. Although much of this research was performed before large numbers of computers were present in schools, formal studies report that when students write on computer they tend to produce more text and make more revisions (Dauite, 1986; Vacc, 1987). More text and more revisions could lead to better work but not always. One could also argue that it might lead to less disciplined thought processes. Moreover, if students become depended on the possibility of multiple revisions what would that mean for their general writing ability? There are some studies that compare student work produced on computer with work produced on paper find that for only some groups of students, writing on computer also had a positive effect on the quality of student writing (Hannafin & Dalton, 1987; Owston, 1991;). It is however necessary to point out that the notion of 'quality' used in these studies is not uncontroversial. Nevertheless, this positive effect is strongest for students with learning disabilities, early elementary school students and older university level students (Hass & Hayes, 1986; Phoenix & Hannan, 1984; Sitko & Crealock, 1986). This would suggest that the use of computer in writing is not a simple quick fix for all. Nevertheless, when applied to meet curricular boarder educational goals, education technology provides alternative approaches to sustaining student interest, developing student knowledge and skill, and provides supplementary materials that teachers can use to extend student learning. However it should be emphasised that the syntheses of earlier research reveal only modest trends. Thus, although individual studies of that era have shown that writing with a computer may increase the amount of writing students perform, the extent to which students edit their writing (Dauite, 1986; Etchinson, 1989; Vacc, 1987), which, in turn, may lead to higher quality writing (Hannafin & Dalton, 1987; Kerchner & Kistinger, 1984; Williamson & Pence, 1989), there are still significant doubts as to the degree these conclusions can be taken as significant for the larger population.

Table 1: Example of a patch-written text

[3] We use the term Western to refer to universities in Australia, UK and North America. However, it also refers to all countries whose universities are based on the argumentative style that typifies the institutions in the three countries listed above.

This view of an engagement with texts through patch-writing seems to be acceptable practice in many Asian universities as confirmed by a Japanese professor Dryden (1999): "students are supposed to show how well they can understand several books and digest them in a report or a paper. They aren't asked for original ideas or opinions. They are simply asked to show a beautiful patchwork…as long as you mention all the books in your bibliography, you can present the ideas from the books as if they were yours, especially if your patchwork is beautiful" (p. 80). The notion of a 'beautiful patchwork' may seem strange to academics in western universities, but it clearly seems to be quite unproblematic to international students and the institutions they come from.

To conclude: we would claim that patch-writing can and should be seen as a step towards independence in academic writing. Furthermore, we would argue that students can express their own independent arguments through patch-writing that demonstrates *an active and informed engagement with a text*—as indicated by our example above and suggested by Howard (1993) and Pennycook (1996)—rather than mere 'mindless' copying. Thus, we would argue that it is important to move away from a simplistic 'copy = plagiarism' interpretation of patch-writing if effective strategies to deal with plagiarism are to be developed.

Becoming an expert through reproduction

> *"If you want to write a poem you must first copy three hundred good poems"*
>
> A Chinese proverb according to a Chinese teacher

Patch-writing makes even more sense in other cultures with a different philosophy of language and learning and whose cultural values that do not value individualism, creativity and autonomy. In many Asian cultures copying, especially through large amounts of repetition, is seen as the true route to learning. Yong learners are encouraged to copy good expression and exemplars that they appreciate or are told to be exemplary. This has as much to do with their pedagogical approach as it has to do with their view of language. For example, Pennycook (1996) has argued that the Chinese view of language is quite different to ours: "In this [view of language] primacy is accorded to language and not to the 'real' world, notions such as metaphor, which suggests that some word 'stands for' something else, become quite different because reality is in the language and not in the

world" (p.221). The sinologist Hans-Georg Moeller (2003, p.75) also expresses this view clearly:

> "Chinese theory of 'forms and names' granted an equal ontological status to both the matter *and* the designation of the things. To use a more formal expression, not only the signified but also the signifier was considered to be inherent in the things. *The signifier was not conceived of as an arbitrary 'label' or as being only attached to things a posteriori. Its name belonged to the thing just as much as its form.* The ancient Taoist text Zhuangzi (see Zhuangzi 1947: 72/25/76) says: 'It has a name and it has a shape: this is what establishes a thing'." (Emphasis added)

Obviously, there is an issue with regard to the degree that this ancient view of language is still evident in contemporary everyday practice. Nevertheless, to the degree that it still is, it would suggest that for Chinese students altering the exact expression of something might plausibly be seen as altering the reality of the world itself. Where would the authority to do this come from for a student? Furthermore, capturing the exact expression—through meticulous memorisation—would be seen as capturing the reality as such. Thus, students would be encouraged to express reality by using the words, the exact expression of the expert since the exact expression contain in them the meaning and expertise that they want access to. Several Chinese students mentioned that precise memorising of texts has been the focus of their learning experience throughout all levels of education. Turner (2000) confirms this mode of teaching and learning in her telling account of the Chinese educational context:

> In the classroom, the teacher speaks and the students listen. Asking questions in class is actively discouraged – the teacher/lecturer may ask one or two favoured students questions but may not ask questions at all…The questions are likely to be factual - it is not normal practice to ask students to venture an opinion. Should a student provide an incorrect answer, they tend to receive some kind of rebuke or punishment from the lecturer… [T]he teacher will provide the students with structured notes - usually on the blackboard which the students will copy - and students are not encouraged to take notes independently. … Owing to the competition for places in Chinese higher education…[w]ork is entirely individual - and almost completely examination-based. … [W]riting, in the form of how to style, structure and present a piece of writing is not taught in China… Students, therefore, are unlikely to have encountered essay-writing to any extent… Nor will they have any experience of using references or multiple sources of information to inform their written work or their thinking…. The teaching method emphasises the correct memorisation and

> reproduction of teacher's notes or text book information - referencing is not used, since almost the entire essay [in the exam] may be in the form of *memorised sections of text*. Information is viewed in a unitary way: the teaching of facts. Critical examination of different perspectives on a subject, and the development of an argument is absent within Chinese education. (Emphasis added)

We are not making a judgement about the validity or appropriateness of this pedagogical approach as such. There is some research to suggest that this approach to learning may indeed be effective (Biggs, 1994, 1996). Nevertheless, we are claiming that this approach—which is common in Asia (Marsh & Morris, 1991; Morris & Sweeting, 1995)—would tend to create the conditions under which the exact reproduction of the expert's expression and formulations (as contained in the prescribed textbook) might be seen as necessary to succeed. As one Indian student indicated in our earlier study, the exam questions "will ask us to repeat definitions word for word from the textbook" (Hayes and Introna, 2005).

There could not be a starker contrast to what is expected from international students when they enter the learning and teaching environments in western universities. Instead of relying on the authority of the lecturer and the textbook, international students—especially at postgraduate level— are expected to gain their understanding of a topic from a multitude of sources (journals, books, internet papers, case studies, etc) expressed in the reading list. They are required to be able to read the material and distil from it the important points, arguments and issues, i.e. be able to evaluate the material with regard to authority, content, relevance and appropriateness. International students are expected to be able to give a critical account of the literature and to be able to formulate their own position, pertaining to the material suggested, which they must be able to justify. International students may be expected to present and justify these views openly through discussion and questioning in a group or lecture context. The good students are expected them to move beyond the reading list, which most non native students will already consider to be extensive. Further, they are required to find their own sources, evaluate them critically and incorporate them in an appropriate manner into their arguments. It is clear that completely different sets of skills are called for in these two approaches. In this situation the typical non-native student will often find himself or herself in a situation where they have a huge skills deficit. In such a

situation they will tend to fall back on what has worked in the past, memorisation and the reproduction of 'canonical' phrases as expressions of expertise.

Once western views of language and values in learning are set aside it becomes possible to start appreciating the behaviour of some of the international students in Australian, US and UK universities. In particular the importance that memorisation and the use of exact expressions play in their way of understanding and knowing the world.

Becoming seen as a plagiarist

> Main Entry: **pla·gia·rize**
> Pronunciation: 'plA-j&-"rIz *also* -jE-&-
> Etymology: *plagiary (to kidnap). Transitive senses*: to steal and pass off (the ideas or words of another) as one's own **:** use (another's production) without crediting the source. *Intransitive senses*: to commit literary theft **:** present as new and original an idea or product derived from an existing source
> (from Merriam-Webster Online Dictionary)

How does patch-writing and the copying (mimicking) of the expert manifest itself in the texts of international students? Clearly it is not simply a matter of 'lazy' students 'cutting and pasting' the work of others and presenting this "as their own"—as the definition above suggests. Obviously this does sometimes happen and ought to be taken very seriously. However, it seems rather that there is a complex process of learning and valuing at play in the construction of texts by international students. It is most certainly the intentional use of another's words—as indicated in part of definition above. However, it seems that it is mostly *not* an attempt to present it as original (i.e. as if their own work)—as indicated in the other part of definition above. Thus, the intention is not to deceive but rather to conform to perceived expectations of what it means to learn.

The debate about cases of plagiarism is often characterised by a *dualistic* perspective. Teachers in the UK often argue that if a text contained 'copied' material then it was either intentionally copied (which would be cheating) or it was unintentionally incorporated (which would be sloppy or bad writing practice). This dualistic view of plagiarism does not allow for the account we gave above where a student intentionally uses parts of texts as 'patches' as well as a means to retain the expressions of the expert, yet does not present it, or mean to present it, as their own independent work. Indeed the question of whether it is there own work or not does not come up as relevant at all—as novices it was never expected to be in the first place. In other words we want to argue that there are significant

pedagogical and cultural reasons for using part copies of texts (as patches) that are not simply plagiaristic behaviour.

It is this form of 'plagiarism', which we will call 'grey' plagiarism, that is our concern. This is not and ideal term as 'plagiarism' essentially refers to the intention to deceive, which is mostly not the case. However, we do want to retain it as an acknowledgement that grey plagiarism should only function as *a step towards* independent work and not as an end in itself. Thus, we are proposing a progressive and formative view of plagiarism that sees patch-writing as a step towards independent and critical thought. This is in contrast to the dualistic and punitive view often held in UK higher education institutions.

It is our claim that the implementation of plagiarism detection software (such as Tutnitin) implicitly operates with, or is used with a dualistic punitive approach to plagiarism—i.e. copy = plagiarism = requires discipline. Thus, many international students that engage in patch writing or use parts of texts to retain the expressions of the expert could, and is being, identified as 'plagiarists' by our electronic detection systems. This is both with regard to the design of algorithms as such and the way in which it is implemented. We will now turn to this 'technical' issue.

Algorithms and the detection of plagiarism

Plagiarism detection software detects *copies* not plagiarism. How does it detect copies? A simple approach that could have been adopted by the develops would be to compare a document character by character. However, this approach has a number of problems: (a) it is very time-consuming and resource intensive; (b) it is not sensitive to white spaces, formatting and ordering changes; (c) it cannot detect part copies from multiple sources. To deal with these problems a number of algorithms have been developed. Unfortunately many of these (such ad Turnitin and EVE) are now proprietary software and therefore not available for analysis. However, there is one that we can consider as typical of these types of algorithms: an algorithm called 'winnowing' (Schleimer et al, 2003).

Winnowing, like many other algorithms, makes a digital fingerprint of a document which it then uses to compare documents against each other. The fingerprint is a small and compact representation of the content of the document that can serve as a basis for determining

correspondence between two documents (or parts of it). A fingerprint is created in a number of steps indicated in the Table 2 below.

Sample text submitted to algorithm:

"How to make a cup of tea

1. Get a cup
2. Place a teabag in the cup
3. Fill the kettle and boil
4. Pour boiling water into cup with teabag
5. Wait one minute to brew
6. Add milk and sugar to taste"

Step in the fingerprinting algorithm	**Example execution using sample text**	**Comments**
1. Remove irrelevant information from text	howtomakeacupoftea1getacup2placeateabaginthecup3fillthekettleandboil4pourboilingwaterintocupwithteabag5waitoneminutetobrew6addmilkandsugartotaste	Remove all white space and punctuation to create a continuous (145) character string
2. Create k-grams of the Step 1 text where k is a parameter (here chosen as 5)	howto, owtom, wtoma, tomak, omake, makea, akeac, keacu, eacup, acupo, cupof, upoft, pofte, oftea, fteag, teage, eaget, ageta, getac, etacu, tacup, …	The 5-grams are created as follows: take the first 5 characters together; move one character right; take the next 5 characters together; and continue until the whole document is done. Here we have only done the first 21 characters. There will normally be almost as many k-grams as there are characters in the document (145 in our case)
3. Convert all k-grams into hashes using a hash function	[**77**, 74, 42, 17], [**98**, 50, 17, 98],[**8**, 88, 67, 39],[**77**, 74, 42, 17],[**99**, 29, 80, 52],[**75**, … (these are hypothetical examples)	A hash function[4] is a program that converts a character string into an integer (in the example 'howto' becomes '77'). Note that the conversion does not always produce a unique result.
4. Take a sample of consecutive hashes from the string of all the hashes (at least one from each window) and store this as the digital fingerprint of the document	77, 98, 8, 77, 99, 75	The technique for selecting the sample from the population of hashes (created in step 3) is crucial. If the gap between successive hashed is too big then the 'identity' of chunks of the document can be lost. If it is too small then a large amount of information will be stored as the fingerprint, which will be inefficient (and costly in terms of resources). Winnowing requires at least one hash from a window of hashes indicated by the [] in step 3 above.
5. Store fingerprint for detection purpose		

Table 2: Steps of a typical plagiarism detection algorithm (based on Schleimer et al, 2003)

[4] A more technical definition of hash function is "A **hash function** is a function that converts an input from a (typically) large domain [input values] into an output in a (typically) smaller range (the *hash value*, often a subset of the integers). (from http://en.wikipedia.org/wiki/Hash_function)

It is in step 4 where most algorithms differ. There are a variety of techniques for determining which hashes to keep as the document fingerprint (see also Brin, S. et al, 1995). The ratio between the total population of hashes and the sample selected for the fingerprint is called the *density* of the fingerprint. Obviously there is a trade-off to consider here. If the fingerprint were not dense enough then it would not be unique and would lead to many false positives (incorrect identification of text as 'copies'). On the other hand if it is too dense then it will be inefficient, as it would require a huge amount of computing resources to process the fingerprints when a comparison is made. What does this mean in practice?

In experiments done by the authors of winnowing using 500000 web pages (in HTML format) it was found that these consisted of 7,182,692,852 bytes of text (approximately 14300 bytes per page). After step 1 these were reduced to 1,940,576,448 bytes of data. This represents an enormous 73% reduction. This means that 73% of the documents consisted of whites spaces and formatting data and only 27% was actual content. This is the redundancy required to make our documents easy to read (spaces between words, lines between paragraphs, headings, formatting, etc.). The results of the first step were hashed in step 3 to create 1,940,576,399 hashes. From these hashes 38,530,846 fingerprints were selected as the fingerprint of the 500000 web pages (approximately 78 bytes of fingerprint per page text). This is a reduction gives a fingerprint density of 1.9855%. This means that the size of the fingerprint (selected hashes) is only 0.536% of the original document size. This implies that we can uniquely identify a document with a fingerprint that is only 0.536% of the size of the original document. Stated differently, it is the same as saying a typical research paper of 8000 words can be uniquely identified by a 'fingerprint' that is the equivalent of a sequence of 43 words selected from the document—obviously the algorithm is more complex than such a comparison would suggest.

With such a reduction will it not be possible that there will be many documents that end up having the same fingerprint? In their experiments with winnowing Schleimer et al (2003) have found that: "82% of the fingerprints selected by winnowing were chosen [occurred] only once; 14% were selected twice; and only 2% occurred three times." Thus, they are fairly confident that they will be able to detect the source document if given a sufficiently large 'chunk' from it. In the case of winnowing it has to be greater than the window (or

chunk) size, which is a user parameter that was set as approximately 100 characters in the reported experiments. This is because the algorithm ensures that it takes at least one unique hash from each window for the fingerprint.[5] Most sentences in this paper vary between 50 and 300 characters. The vast majority is over 100. Thus, any partial copy of a document—greater than 100 characters—will map onto a part of the fingerprint, making it possible to identify the part as belonging to the document identified by the fingerprint. Furthermore, the use of k-grams (k successive sequence of characters) means that the algorithm will be robust against 'noise', i.e. it will not simply match common phrases with copies of those phrases in other documents.

From this discussion it is clear that plagiarism detection algorithms are reasonably robust at linking copies or part copies back to its source document. Let us now consider some of the implications of the implementation of the algorithms for detecting plagiarism.

We need to start by reminding ourselves that plagiarism detection systems (contrary to their name) *do not detect plagiarism*. They only detect copies (or part copies) of documents. This is an important point. Not all copies are plagiarised and not all plagiarism comes in the form of exact text copies. Thus there is not a one-to-one relationship between copies and plagiarism.

To copy is not always to plagiarise

There may be a variety of reasons why a copy does not represent plagiarism. It may be a legitimately referenced quote. It may also be a phrase that coincidently corresponds to a phrase in another document. However, most important for us, it may be a patch in a patch-written text. As argued above, it is very likely that international students, who are novices in academic writing, may present their work through patch-writing. This issue becomes more acute when student essays are batch submitted for checking and a threshold as a percentage of a document copied is set quite low(as one can do in these systems) for cases to be further investigated. One might argue that the international student's patch-writing and use of familiar sources to expressions of the expert will exaggerate the difference between them and the native students even if it is legitimately referenced, thereby 'pushing' the native-speaker down below the line of detection. It is also likely that native

[5] Unless consecutive hashes are the same then it is omitted, which is why there are less hashes than there are sentences in a document.

speakers will be more able to use patches in such a way that they may be identified as paraphrases rather than direct copies. This can easily be done by the careful use of synonyms and slight changes in the structure of sentences. However such more subtle changes require a more sophisticated linguistic ability that may be beyond the level of a non-native speaker.

To plagiarise is not always to copy

Plagiarism detection systems are based on the principle of character sequence detection, as seen above. This means that it can only identify plagiarism where there is an exact copy made of a string of characters (irrespective of location on the page). This sort of detection will obviously tend to show up those students who tend do retain exact copies of phrases or sentences. It will therefore not detect those that copy structure, arguments or ideas but express these in 'there own words". Thus, plagiarism detection systems operate with the assumption that to plagiarise one needs to use the exact words of another. This is a very 'legalistic' view of plagiarism. It is similar to the legal view of copyright—adopted in most western countries—which suggests that one can only copyright expression and not ideas. Clearly this is a very narrow definition of plagiarism; an assumption that favours the native speaker and disproportionately penalises the non-native speaker. The native speaker has the linguistic skills to eloquently re-express the work of others and remain undetected by the detection algorithm. It is evident that if the task of plagiarism detection is 'delegated' to algorithms then there is a strong possibility that this might create the conditions for constructing international students as plagiarists while also allowing for native speaking plagiarists to remain undetected.

On the (un)construction of plagiarists

There is no doubt that plagiarism is a problem for most UK universities. However, there is also no doubt that this is a complex problem that defies simplistic solutions—as most authors in the field will agree. It is our argument that a reactive punitive response to plagiarism based on an algorithmic detection approach is unfair for the following reasons:

a. It makes inappropriate assumptions about plagiarism, i.e. copy = plagiarism
b. International students are predisposed to the use of exact copies in their writing practice (in patch-writing and keeping the master's voice) and are therefore inappropriately identified as plagiarists. This often leads to further more detailed and meticulous scrutiny. Something other students are not subjected to.

c. Thus, international students tend to be disproportionately identified as plagiarists to the benefit of native speakers who may plagiarise through the unattributed copying of ideas and arguments of others and yet remain undetected.
d. Plagiarism algorithms, or more specifically the assumptions embedded in them, are developed within a western cultural context which makes particular assumptions about the nature of teaching and learning. As such they may unfairly discriminate against those from non-western backgrounds.

Obviously, this argument is still somewhat tentative and needs further evidence for it to become sufficiently robust. Nevertheless, it seems at least plausible. As such we would suggest that there are a number of things that staff in Australian, British and North American institutions could and need to do to address the issues raised by this paper. Let us briefly state them.

a. The issue of plagiarism detection cannot be delegated to an electronic detection system or service. As the quality guru Edwards Deming (1986) said: "you cannot inspect quality into the product". Quality is a systemic outcome of the whole system. Likewise, plagiarism cannot be 'detected' out of the learning process. The elimination of plagiarism requires a systemic approach which involves the whole system
b. An appreciation and understanding of the learning and teaching environment from which our international students come is required in order to create the mechanisms and resources that will make their transition to western systems as easy as possible.
c. Western universities ought to take a formative attitude to plagiarism in which they accept that patch-writing may be a legitimate interim step to the development of independent writing skills. In this regard plagiarism detection systems can act as a mechanism to help students and lecturers to become aware and monitor this transition to independence. This will have important implications for how institutional rules and frameworks for dealing with plagiarism will be formulated and implemented.
d. There is a need to develop a much better understanding of how plagiarism detection systems work. What are the assumptions the make? How do the different parameters interact? How do these favour some forms of plagiarism and not others?

This may be difficult as most systems are based on proprietary systems where the algorithms and code is not available for inspection.

Conclusion

Our paper seeks to provide a further insight into why some students may be identified as plagiarists. We accept that students may plagiarise as a consequence of poor time management and a marks orientation etc. We also accept that due to being disaffected from their studies, students may deliberately plagiarise. Indeed, when there is no or little interest in a subject, plagiarism could be an attractive option. But what about those students who are interested and committed to a subject who are identified as plagiarists? In relation to non-native students, we suggest that there may be an alternate series of explanations for why many non-native students are incorrectly identified as plagiarists. Namely, due to their lack of familiarity with the education context in the west, their limited ability to develop an independent argument, and importantly, them learning to undertake academic writing in a second or third language. Our chapter has pointed to the strong possibility that international students are more likely to be detected as having significant strings of unattributed characters that are copied from another source. This may be due to writing in patches, or due to native students being much better at remaining undetected, and as a consequence rendering non-native writers more visible. We suggest that utilising detection systems too early may take away the opportunities for learning that are required by such students in order to become embedded in a different education context and to develop all the practices that are required to succeed. Detecting out such possibilities to learn while also upholding academic integrity is thus an important challenge when thinking through the use of plagiarism detection systems among overseas students in western universities.

References

Angelil-Carter S, (2000) *Stolen Language? Plagiarism in Writing*, Pearson Education Limited, UK

Angelova, M., and Riazantseva, A. (1999) "If You Don't Tell Me, How Can I Know?': A Case Study of Four International Students Learning to Write the U.S. Way." *Written Communication,* 16 (4), pp. 491-525.

Biggs, J. and Watkins, D. (1996) *The Chinese Learner*, Comparative Education Research Centre, Hong Kong

Biggs, J. (1994) "Asian learners through Western eyes: an astigmatic paradox", *Australian and New Zealand Journal of Vocational Educational Research,* 2 (2), pp.40-63

Carroll, J. and Appleton, J. (2001) Plagiarism: A Good Practice Guide, JISC, Oxford Brookes University

Carroll, J. (2002) Suggestions for Teaching International Students more effectively, Learning and Teaching Briefing Papers Series, Oxford Brookes University www.brookesac.uk/services/ocsd

Deckert, G. (1993) Perspectives on plagiarism from ESL students in Hong Kong, *Journal of Second Language Writing,* 2 (2), pp.131-148

Deming, W.E. (1986) *Out of the Crisis*. Cambridge, Mass.: Massachusetts Institute of Technology Press.

Dreyfus, H.L. (1992) *What Computers Still Can't Do*. Cambridge, Mass: Massachusetts Institute of Technology Press.

Dryden, L. (1999) A distant mirror or through the looking glass? Plagiarism and intellectual property in Japanese education. In L. Buranen & A. Roy (Eds.), *Perspectives on plagiarism and intellectual property in a postmodern world* Albany, NY: State University of New York Press, pp. 75-85.

Errey, L. (2002) Plagiarism: Something Fishy? …Or Just a Fish out of Water?, OCSLD

Harris, R. A. (2001) *The plagiarism handbook: Strategies for preventing, detecting, and dealing with plagiarism*. Los Angeles, CA: Pyrczak Publishing.

Hayes, N. and Introna. L. (2005) Cultural values, plagiarism, and fairness: when plagiarism gets in the way of learning, *Ethics and Behavior*, vol 15(3), pp 213-231

Howard, R. (1995) Plagiarism, authorship, and the academic death penalty. *College English,* 57 (1), pp. 788-805

Howard, R. (1993) A Plagiarism Pentimento. *Journal of Teaching Writing* 11(2), pp. 233-245.

Hull, G. and Rose, M. (1989) Rethinking Remediation: Toward a Social- Cognitive Understanding of Problematic Reading and Writing *Written Communication,* 6 (2), pp. 139-54.

Kolich, A.M. (1983) Plagiarism: The worm of reason, *College English,* 4, pp.141-148

Lathrop, A. (2000) *Student cheating and plagiarism in the Internet era: A wake-up call*. Englewood, CO: Libraries Unlimited.

Marsh, C. & Morris, P. (eds.) (1991). Curriculum Development in East Asia. London: Falmer Press.

Morris, P.& Sweeting, A.(eds.) (1995). Education and Development in East Asia. New York and London: Garland Publishing Inc.

Moeller, H. (2003) Before and after representation, *Semiotica*, 143, (1/4), 69–77.

Myers, S. (1998) Questioning author(ity): ESL/EFL, science, and teaching about plagiarism. *Teaching English as a Second or Foreign Language*, 3 (2). http://www-writing.berkeley.edu/TESL-EJ/ej10/a2.html

O'Connor, S. (2003) Cheating and Electronic Plagiarism - Scope, Consequences and Detection. Proceedings EDUCAUSE in Australasia. Conference held in Adelaide, May 6-9 (CD-ROM)

O'Donoghue T. (1996) Malaysian Chinese Student's perceptions of what is necessary for their academic success in Australia: a case study at one University, *Journal of Further and Higher Education,* Vol 20 (2), pp.67-80

Pennycook A. (1996) Borrowing others' words: Text, ownership, memory and plagiarism, *TESOL Quarterly,* Vol 30 (2), pp.210-230

Ryan, J. (2000) A Guide to Teaching International Students, Oxford Centre for Staff Development, Oxford Brookes University, Oxford

Schleimer, S., Wilkerson, D. and Aiken, A. (2003) Winnowing: Local Algorithms for Document Fingerprinting. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 2003, pp. 76-85.

Scollon R, (1995) Plagiarism and Ideology: identity in intercultural discourse, *Language in Society,* 24 (1), pp.1-28

Turner, Y. (2000) Chinese Students: Teaching, Learning and Equality in UK Higher Education, *Higher Education Equal Opportunities Network, National Network*

Newsletter for Equal Opportunities Practitioners, Spring 2000, Issue 13, http://www.worc.ac.uk/services/equalopps/HEEON/newsonline.htm#Yvonne%27s)

National Network Newsletter for Equal Opportunities Practitioners Spring 2000 Issue 13

Brin, S., Davis, J., and Garcıa-Molina, H. (1995) Copy detection mechanisms for digital documents. In the *proceedings of the ACM SIGMOD Conference*, pp.398–409.

Sherman, J. (1992) Your own thoughts in your own words, *ELT Journal,* 46 (2), pp.190-198