International Journal for Educational Integrity

Investigating academic plagiarism: A forensic linguistics approach to plagiarism detection

Rui Sousa-Silva
Universidade do Porto, Portugal
r.sousa-silva@lflab.pt

Keywords: plagiarism, plagiarism detection, translingual plagiarism, forensic linguistics, computational forensic linguistics

This paper was assessed by the Editors of the *Conference Proceedings of the Plagiarism Across Europe and Beyond Conference* (Brno, Czech Republic, 12–13 June 2013) as a 'best conference paper'. It was then forwarded to the *IJEI* for consideration. The paper has now undergone additional double-blind peer review and as a result of subsequent revisions is substantially different from the original version presented at the Czech conference.

Abstract

Automatic plagiarism detection tools have evolved considerably in recent years. Owing in part to the recent technological developments, which provided more powerful processing capacities, as well as to the research interest that plagiarism detection attracted among computational linguists, results are nowadays more accurate and reliable. However, most of the plagiarism detection systems freely and commercially available are still based on similarity measures, whose algorithms search for similar or, at most, identical strings of text, within a more or less short search distance. Although these methods tend to perform well in detecting literal, verbatim plagiarism, their performance drops when other strategies are used, such as word substitution or reordering. This paper presents the results of a forensic linguistic analysis of real plagiarism cases among higher education students. Comparing the suspect plagiarised strings against the most likely originals from a legal perspective, it is demonstrated that strategies other than literal borrowing are increasingly used to plagiarise. A forensic linguistic explanation of the strategies used and why they represent instances of plagiarism is then offered, and examples are provided to illustrate why existing software fails to detect them. The paper concludes by arguing that commonly used detection software packages can be effective in identifying matching text, but are not necessarily good plagiarism detection systems. More in-depth research and improvements in computational linguistics and natural language processing are required to increase the accuracy and reliability of the machine-detection procedure.

Plagiarism and forensic linguistics

Plagiarism, which in its most basic form consists of passing off someone else's work as one's own, has attracted considerable media attention in recent years, mostly due to the high profile of people involved. Examples include the case of the German Defence Minister Karl-Theodor zu Guttenberg[1], who, in 2011, (temporarily) renounced his doctorate title and eventually resigned, as a result of accusations that he had plagiarised when writing his doctoral thesis. In Romania, the Prime Minister Victor Ponta[2] was accused, in 2012, of plagiarising substantial portions of his doctoral thesis, and faced pressure to resign. More recently, last year, suspicion was raised by

© International Journal for Educational Integrity Vol. 10 No. 1 June, 2014 pp. 31–41 ISSN 1833-2595

'plagiarist hunters' that the German Education Minister Annette Schavan had plagiarised at least 130 passages of her doctoral thesis; as a result of the suspicion, the University of Düsseldorf, which had awarded her PhD, revoked her title, after conducting an official process to rescind it, and she later resigned[3]. A few years earlier, a journalist of the Portuguese quality newspaper *Público* was accused of plagiarising *Wikipedia* and the *New Scientist*[4], and more recently the journalist of *The Independent* Johann Hari was suspended for plagiarising news articles[5]. Cases of academic plagiarism are also known. In 2010, for example, a Portuguese university lecturer resigned following accusations that she had plagiarised her doctoral thesis[6]. However, in the academy not all cases make it to the news. Rather on the contrary, most of them tend to be resolved internally, by disciplinary boards or the lecturers/ tutors themselves, depending on the respective institution. Academic plagiarism is, nonetheless, considered an unacceptable practice, which brings along severe penalties up to having their titles rescinded, even if the instances of plagiarism are not found until a later date.

In these, as in most cases, plagiarism is seen both as an immoral and an illegal act. Like any other instance of 'theft' or 'misappropriation', plagiarism is morally wrong, as well as academically and socially condemnable. It is this feature of plagiarism that higher education institutions and policies attempt to repair when students plagiarise. However, plagiarism often represents a misappropriation (Jameson, 1993) of personal property, and a violation of both moral and financial rights (Leitão, 2011).

Unsurprisingly, therefore, it is nowadays a serious legal offense in Common Law and Civil Law systems alike. Moreover, as a result of the proliferation of general principles of copyright law, it is now commonly accepted that authors should be granted the right to financially explore their work, as much as the right to the 'paternity' and integrity of that work (Pereira, 2003). Especially in cases of academic plagiarism, it is mostly the guarantee of the moral rights of the original author that needs to be considered.

It is this view of plagiarism as both an immoral and an illegal action that legitimates punitive actions in academic and non-academic contexts including, among others, rescinding titles. But these considerations of plagiarism both as an immoral act and an illegal action bring additional problems, the most challenging of which consists of determining the degree of intentionality underlying the instance of plagiarism. As Howard (1995) claimed, Angèlil-Carter (2000) later argued and Pecorari (2008) subsequently reiterated, academic plagiarism is more often a problem of academic writing skills (or their lack thereof), than an intentional attempt at passing off someone else's work as one's own. Likewise, Scollon (1994, 1995) and Thompson (2002) sustained that non-compliance with academic writing procedures and conventions was often more a result of clashing intercultural aspects, than an intention to deceive. Additionally, if text re-use is taken to represent a form of authorship, as Robillard (2008) argues, then a clear distinction has to be made between improper, unintentional borrowing, and intentional plagiarism (Howard, 1995). In their research, both Angèlil-Carter and Pecorari attempted to determine the plagiarists' intention by identifying the instances of textual borrowing and then interviewing the suspect plagiarists. They concluded that, in some cases, the students misattributed their sources inadvertently, whereas other cases suggested that the plagiarists acted with the intention to deceive.

Determining the suspect's intention by interviewing them, however, may not always be a possible investigative method. Firstly, due to the reported increasing number of plagiarism cases[7], most universities will rarely have sufficient human and technical resources to investigate all cases thoroughly and properly. Secondly, if we consider that some instances of deception pass unnoticed even in courts of law, depending on the 'expertise' of the deceivers, then lecturers/tutors and educational institutions can hardly be expected to properly identify all instances of deceptive plagiarism. In addition the plagiarist may himself/herself misjudge their case, either by wrongly

admitting the truthfulness of false positives, or by denying the truthfulness of true positives. Not to mention the need for proper evidence that proves the claims for – or against – plagiarism. Finding evidence raises specific challenges, whether it is a case of plagiarism (where a text borrows from (an)other source(s) without acknowledgement) or collusion (where two or more people work collaboratively on the same text and pass off each individual document as an original), especially when the plagiarist has practiced a deceptive act whose nature results from lying (Eiras & Fortes, 2010). Firstly, as Eggington (2008) concluded, deception can hardly be detected linguistically. Secondly, as Coulthard and Johnson (2007) argued, it is not the linguist's task to detect the plagiarist's intention; on the contrary, they sustained that it is the linguist's task to establish whether two texts have been produced independently or otherwise. Analyses of this type, which are based on the comparison of suspect texts and potential originals, have been used in academic, as well as non-academic contexts, and are the basis of most plagiarism detection software packages.

However, as a result of the technological developments of the last decades – especially the internet – more information is now more readily available, including to students, which makes it easier to pass off someone else's work as one's own, by copying and pasting the original text 'as is', or by making minor or more substantial alterations to it. At the same time, due to the massive volume of information available, it is now more difficult for any reader to intuitively identify a text or text passage as an instance of plagiarism. But as Coulthard and Johnson (2007) argue, the technological developments that made it easier to plagiarise also made it easier to detect instances of plagiarism. The need to detect instances of plagiarism that are missed by intuition, together with the recent technological developments and the growing interest of computer scientists and computational linguists, led to the development of a plethora of plagiarism detection software packages.

Existing plagiarism detection software can operate based on two different approaches: *external plagiarism detection* and *intrinsic plagiarism detection* (Potthast et al., 2009). The latter aims to detect instances of plagiarism in cases where the reader is intuitively led to the suspicion that the text has been borrowed from other sources, but does not know any original texts against which the text can be compared. The detection procedure is, in this case, based on an intrinsic, stylistic analysis of the suspect text, in order to identify stylistic inconsistencies that can be used to challenge the authorship. Although this procedure may represent a valuable contribution, from an investigative perspective, by not contemplating the original source from which the text was lifted, it lacks the evidential value required to demonstrate the instance of plagiarism. Most common plagiarism detection software packages currently available operate via an external analysis, by establishing a comparison between the suspect text(s) and the known originals, in order to determine the degree of similarity or identity between the texts. This procedure, which is used (even if with minor or major adjustments) by most detection systems – including *Turnitin* and *SafeAssign* – works by scanning the texts and applying computational string-matching techniques to identify words, phrases, sentences or paragraphs that, having been copied and pasted from another source 'as is', or subsequently altered, are identical or similar to the original text. Systems that use this approach perform well in detecting identical texts, based on *verbatim*, word-for-word borrowing, but less well when changes are introduced to the original text. In this case, the detection gradually becomes more difficult to handle computationally, up to a point where it becomes impossible.

The problems imposed on the computational detection of plagiarism are due mainly to search space restrictions. Since any two texts are expected to share a high number of words, most of which are grammatical (and consequently used less 'uniquely'), flagging all individual items that are shared between the two will lead to the wrong identification of an instance as plagiarism. Therefore, search space restrictions have

been introduced to 'teach' the system that not all overlapping words should be flagged as plagiarism; on the contrary, the system is instructed to flag as plagiarism only overlapping strings of co-occurring words of a certain length in the original and suspect texts. By determining the minimum number of words that must co-occur, as well as the maximum number of new words that are altered, introduced or deleted from the string before a text can be considered an instance of plagiarism, the system avoids flagging *false positives* (i.e. misidentified strings of plagiarism). Consequently, if a string of overlapping text is below a certain number of words, or if the number of words that are altered, deleted from or introduced to the original text is above a certain threshold, the system traditionally identifies it as original text. This raises some problems. On the one hand, as Woolls (2010) explained, and as is commonly advertised by plagiarism detection software packages, the volume of overlapping text that is calculated usually requires a manual, human analysis, in order to confirm or otherwise reject a certain flagged instance as plagiarism. On the other hand, if we consider that, the more an unattributed text is manipulated, the higher the plagiarist's intention to plagiarise (Sousa-Silva, Grant, & Maia, 2010), then the more a text is altered, the more severe the instance of plagiarism, and the lower the likelihood that it will be identified by the machine.

Linguistically grounded approaches are therefore required, not only to raise suspicion, but also to investigate a text and provide evidence that it has been lifted from another source. This is required to explain the linguistic strategies adopted, and additionally to assist lecturers/tutors and disciplinary board members, among others, in determining the plagiarist's intention. The type of linguistic analysis conducted by forensic linguists has shown good results in this respect.

Although it is often considered that the impact of academic plagiarism is limited to the academy, the cases discussed above demonstrate otherwise. While, on the one hand, the bad academic practice is reprehended, on the other hand the suspect's ethical and moral principles, and their fitness for the job, are challenged. In such cases, suspicion often suffices to socially impact the suspect plagiarist's life, but solid evidence is required to legally support the decision adopted, especially when this involves definite actions up to rescinding or revoking a title. Research into forensic linguistics, which consists of applying linguistic methods and analyses in forensic contexts, has been used effectively in cases of fraud where linguistic evidence is vital, and has demonstrated that the likelihood that a text – or set of texts – has not been produced independently can be determined accurately. Moreover, as has been demonstrated (Turell, 2008), such data can be used not only as an investigative tool, but also as evidence.

The purpose of this study is to challenge the assumption that plagiarism detection software can effectively identify the most serious instances of plagiarism, where the plagiarist has heavily and intentionally manipulated the text to deceive his/her readers. Using a combination of descriptive linguistic analyses of instances of academic plagiarism, this study presents and discusses some cases that, owing to their nature, can be missed, in whole or in part, by plagiarism detection systems. This study indicates that *word substitution* and *reordering*, as well as *translation*, are some of the strategies used by plagiarists to mislead the detection systems.

This paper is structured as follows. The following section explains how the research is operationalised; it describes the corpus of texts analysed in this study and the analytical method employed. The findings of this analysis are presented in the subsequent section, which is followed by a discussion of the findings. The paper concludes with a summary of the findings, and points towards future research directions.

Method of analysis

The analyses of instances of plagiarism commonly consist of comparing suspect texts against the putative originals, and highlighting the textual identities and similarities, or alternatively the differences between the texts. One can hypothesise that, the higher the identity between the derivative and the original texts, the easier it is to detect the instance of plagiarism, and the easier the machine detection. Conversely, the higher the number of edits introduced to the derivative text, the more difficult the detection procedure becomes, especially when using detection software. To test this hypothesis, a corpus of academic assignments that were considered instances of plagiarism were used to conduct an extrinsic analysis. The assignments were written in Portuguese by design (S1 and S2) and media and communication (S3, S4 and S5) postgraduate students of two Portuguese universities. A corpus of texts written in Portuguese offers an additional advantage, when compared to English: since Portuguese is morphologically and syntactically more diverse and flexible than English, it offers a greater range of word combinations and inflections, and consequently raises new challenges to the detection procedure. The original sources were also provided by the lecturers/tutors, for comparison.

As shown in Table 1, these assignments are of considerable length (an average of 3,000 words per essay):

Table 1:
Assignments included in the corpus

Student	Number of Words
S1	3,638
S2	1,370
S3	3,333
S4	4,629
S5	2,033

However, for the purposes of the analysis of the linguistic features used to plagiarise, or to assess the impact of these strategies on the detection procedure, this quantification is irrelevant. This is because, on the one hand, there is no correlation between the number of words and the amount of borrowing, and, on the other hand, between the text size and the linguistic strategies used.

The linguistic analysis focused on the nature of the instances that showed changes, in terms of *word substitution*, *word reordering* and *translation*. Since the aim of this study is to identify the nature of the changes operated, no detection software was used at this stage. The potential impact of these alterations on the manual and software detection is explored in the descriptive analysis of the data.

A manual, side-by-side comparison between the original and the derivative text was made, highlighting alterations in grammar, punctuation, syntax, semantics, lexis and discourse. Since the derivative texts were, for the most part, borrowed *verbatim* from the original, the differences, rather than the similarities, were highlighted to signal the alterations introduced, and the identical strings, showing exact matches, were discarded. The next step consisted of the descriptive linguistic analysis of the strings that had been altered by replacing or reordering the words of the original.

Subsequently, the strings that had been translated from the original source were analysed more closely. Finally, those alterations, and specifically their relevance to determining the impact on the machine detection procedure, were investigated, in order to determine whether they are to be expected or, on the contrary, whether they are illicit.

Results of the analysis

The first stage of the analysis consisted of identifying the strings of text containing *word substitution*, *word reordering* and *translation*. Although some of these linguistic strategies are often used to paraphrase, reference to paraphrasing is avoided in this study. This is because paraphrasing involves a deeper rephrasing that goes beyond the three types of alterations discussed, in order to retain the meaning, while using a new form.

Word substitution

Word substitution consists of replacing a word or combination of words with words with identical or similar meaning. Although these replacement words usually retain some sort of semantic relationship with the original text (such as synonymy, hyponymy or hypernymy), they can also be from a different semantic field, especially when they aim to retain the coherence with the extra-textual world. The assignment of S1 presents several instances of the latter. The word 'escola' (*school*) in the original is replaced with 'cultura' (*culture*) in the derivative text; 'um cantor ou uma actriz' (*a singer or an actress*) is replaced with 'um designer ou um artista plástico' (*a designer or a plastic artist*); 'os professores e os pais' (*the teachers and the parents*) is replaced with 'os profissionais e o público em geral' (*the professionals and the general public*); 'educativa' (*educational*) is replaced with 'cultural' (*cultural*); 'um jogo de futebol' (*a football match*) is replaced with 'a performance de um artista' (*an artist's performance*), and 'jogo' (*match*) and 'partida de futebol' (*football match*) are replaced with '*performance*' in both instances. Word substitution is not, however, used exclusively in this assignment. S5, for example, replaces the word 'mesclado' (*entangled*) with the synonym 'embaralhado', 'fotodocumentalismo' (*photo documentary*) with the semantic equivalent 'fotodocumentário', and 'exigem' (*demand*) with the semantically related 'necessitam' (*require*).

In this same assignment, the adjective 'sustentada por' (*argued by*), followed by the author's name, is replaced with the semantic equivalent preposition 'To', followed by the author's name. S4's assignment also shows many cases of word substitution. But, interestingly, most of these are minor, as they result either from the correction of Brazilian Portuguese spelling to the European Portuguese variant, or reflect the different use of prepositions in the two variants. However, there are also substantial lexical substitutions, whose nature involves more than simple spell checks. For example, 'enxergar' (*see*) is replaced with the synonym 'olhar', 'superposição' (*overlapping*) is replaced with its equivalent 'sobreposição', and 'plasmar' (*exhibit*) is replaced with the synonym 'passar'. A more substantial change is operated by the substitution of 'vermelhos' (*reds*) with 'cores vermelhas' (*red colours*) in the derivative text.

A sophisticated substitution is operated by S1 in the phrase 'opção que condicionará' (*an option that will condition*). Originally used as a subordinate clause, this phrase is reused in the derivative text as part of a new sentence: 'Esta opção irá condicionar'. A new demonstrative is introduced ('Esta'), which had been omitted in the original, the relative pronoun 'que' (*that*) is deleted from the derivative, and the future tense of the verb *to condition*, 'condicionará', is replaced with the infinitive form of the verb, 'condicionar', preceded by the future tense of the auxiliary verb 'ir'. These alterations result in a new wording that, although semantically identical to the original, is morpho-syntactically different, and sufficient to trick machine detection systems.

In some cases, words are used to replace punctuation. S4, for example, replaces the semi-colon with the adversative 'mas antes' (*on the contrary*). Likewise, in S5's assignment, 'Mais:' (*additionally*), whose specific meaning in this context is marked by the use of the colon, is replaced with the lexical equivalent 'Depois'.

Word reordering

Word reordering is used to describe the linguistic operations whereby the original words are reused, but in a different order. Although this linguistic strategy is not as common as word substitution, the corpus includes several examples of this. S4 uses this linguistic device as a plagiarism strategy several times: 'se deixar envolver' (*let themselves involve*) is replaced with the more European Portuguese standard 'deixar envolver-se'; 'que foram assim chamados por' (*that were thus called by*) is replaced with 'assim foram chamados pelos'; the Brazilian Portuguese syntax 'ele se separaria' (*he would depart*) is replaced with the European Portuguese 'separar-se-ia'. S5 also uses this strategy: 'No sentido lato, entendemos por fotojornalismo' (*In the general sense, by photojournalism we mean*) is reordered 'por fotojornalismo no sentido lato, entendemos'.

Even more complex is the sophisticated case of reordering operated in the following Example 1: the original *Adaptando ao fotojornalismo uma sistematização das funções da linguagem no discurso informativo sustentada por Jesús González Requena(41), poderíamos (…)* is reordered *Para Jesús Requena, adaptando ao fotojornalismo uma sistematização das funções da linguagem no discurso informativo poderíamos (…)*. In this case, the name of the author is edited (*González* is deleted), the comma (as well as the note number) is deleted, and the reporting phrase ('sustentada por Jesús González Requena') is altered ('Para Jesús Requena'). As a consequence, a maximum of 11 running words are retained in the derivative text, five of which are grammatical items, of a chain of 25 running words.

Translation

Finally, translation is used to refer to instances of 'translingual plagiarism' (Sousa-Silva, 2013) where a writer has an original translated from another language, via human or machine translation, and uses it as his/her own original, while omitting the source. An extensive example of this strategy is provided by S3's assignment. This assignment includes a literal translation of an original in Spanish into Portuguese that retains, for the most part, the original punctuation, lexis and syntax. Besides some spelling mistakes ('aerosois' instead of 'aerosóis' for aerosols), this assignment shows several other mismatches. In terms of lexis, some words are wrongly used, either because they do not exist in Portuguese (e.g. 'decoraciones', 'carácter espontáneo'), or because, when existing, they have a different meaning (e.g. 'pintada', 'mural', 'rótulos'). Moreover, as a simple search using a common internet search engine demonstrates, the phrase 'escritor de graffiti' is common in Spanish, but not in European Portuguese. Likewise, some phrases like 'Pois bem', 'Assim mesmo', 'Agora bem', and 'à hora' as a translation of 'Pues bien', 'Así mismo', 'Ahora bien', and 'a la hora', respectively, indicate a wrong literal translation. Syntactically, a wrong transfer is noted in phrases like 'ia alguém a se arriscar' as a translation of 'iba alguien a arriesgarse'.

In terms of grammar, this assignment consistently shows a wrong use of uppercase after a colon; although this may be common is Spanish, in Portuguese lowercase is to be expected after a colon. Additionally, there are some problems with the concordances; for example, the phrase 'Numerosos foram as tentativas' is grammatically wrong in Portuguese because, as this is a gender-sensitive language, 'Numerosas' (instead of 'Numerosos') is expected in order to retain the consistency with 'tentativas'. In Spanish, however, since 'intentos' is a masculine noun, 'Numerosos' is used.

Discussion

An investigation into plagiarism needs to consider the particular circumstances involved, especially considering the legal implications of plagiarism. Additionally, if as Howard (1995), Angèlil-Carter (2000) and Pecorari (2008) have argued, plagiarism among students is a pedagogical problem that can be, for the most part, resolved by teaching the students how to write academically, then we have to agree that academic and non-academic plagiarism cannot be judged independently of their circumstances. Specifically, in the academy, where instances of plagiarism represent a failed attempt by students at writing academically (Howard, 1995; Pecorari, 2008), plagiarism can result from a legitimate attempt at producing a good piece of writing, or, alternatively, an attempt at obtaining the best possible grade, for the minimum effort. Consequently, if one considers that the principle behind plagiarism is laziness, then only minor alterations are to be expected, as these do not require hard work. This is the case, for example, where only one word or a few words are altered in a long sentence.

Minor alterations of this type do not impact the machine detection procedure, since they are not sufficient to interrupt the minimum chain required from the string matching procedure. Conversely, the machine detection procedure is made more difficult in cases where the alterations introduced break the chain of consecutive words in such a way that the sequence of running, overlapping words is not sufficient to run the detection procedure against an unknown source. Example 1 above illustrates this point well. Taken together, the alterations introduced to the original sentence transform a total of 25 running words into a chain that is broken down into three batches of overlapping words: 11, 1 and 4 words, respectively. Moreover, considering the principle of lexical richness, as discussed by Coulthard and Johnson (2007), even the longest string of these (11 words) loses significance owing to the fact that roughly half of the words are grammatical items, and hence more likely to occur anyway.

Punctuation is another element that impacts the machine detection procedure. In order to avoid the maximum number of false positives, while at the same time attempting to identify true cases of plagiarism, some detection software packages use outer punctuation to divide the text into chunks. Consequently, the detection procedure is affected in cases where words are used to replace punctuation, as is the example where the semi-colon is replaced with the adversative ‘mas antes’.

Machine detection is also hampered in cases where, after dissecting the original into smaller sentences, the plagiarist substitutes at least some of the original words, as in the example ‘opção que condicionará’. In this case, the amount of consecutive words that are shared with the original text is so small that the software can hardly identify the string as plagiarism. Likewise, the detection procedure is also impacted by the addition of new words, together with word substitution or reordering. The phrase ‘que foram assim chamados por’, discussed above, illustrates this point well. The derivative sentence is not only split by the reordering, but also ‘Movimentos como’ (*Movements such as*) is added to the beginning of the sentence, and a sequence of eight running words is interrupted by the determiner ‘o’ (*the*), resulting in five and three running words, respectively; finally, a sequence of nine words is added to the end of the sentence. Taken together, these alterations impact the machine detection procedure, not only by interrupting the chain of consecutive words, but also by increasing the ratio of new words, in relation to the words of the original. As a consequence, the number of reused words, in this particular case, may be lower than the threshold required by the detection system to flag a text as plagiarism, and therefore falsely considered to be an original text.

Translation also represents additional problems to plagiarism detection, starting with the definition of plagiarism. Specifically, translation can be considered a plagiarism

strategy if plagiarism is defined as passing off someone else's works and ideas as one's own, but not if the restrictions imposed on borrowing apply only to words. Since a translation involves a transfer of the meaning of an original in one language using the linguistic signs of another language, these signs are necessarily different from the original ones. Consequently, the text that is lifted from another original is not similar, and much less identical. This represents a problem to computer systems, which need to process texts using comparable patterns to be able to proceed to the string matching. In this case, it is a requirement that the two texts are converted to one common language for comparison. However, this conversion is only possible if the original is known, which requires that the reader either (a) knows the original text, or (b) the text provides linguistic cues that lead the reader into intuitively establishing the language of the original. These cues are usually provided by issues in grammar, punctuation, syntax or lexis, such as the ones discussed in the examples illustrated above. However, these cues can be discounted when the writers are known to be writing in a foreign language, in which case issues with grammar, punctuation, syntax or lexis are to be expected. The challenges imposed by translation on the detection procedure are even bigger when this strategy is used in combination with other strategies, such as word substitution or reordering.

The combination of strategies is, speculatively, one of the major challenges imposed on software detection systems. Different plagiarism detection software packages have demonstrated different degrees of effectiveness in detecting different plagiarism strategies. Some packages (e.g. *Turnitin*) perform well in detecting identical text, regardless of the nature of the words (lexical or grammatical), whereas others offer the users the possibility of excluding certain strings from the plagiarism report (e.g. *SafeAssign*) or focus on the lexical items to calculate the percentage of plagiarised lexical vocabulary (e.g. *CopyCatch*). However, as the analysis of these corpus texts demonstrates, it is very unlikely that only one strategy is used individually when plagiarising; on the contrary, a combination of plagiarism strategies within the same text is not uncommon. Since, at this stage, it is computationally challenging, if not unrealistic, to combine the possibility of detecting several plagiarism strategies within one same detection system, software packages have until now given priority to one or other strategy. Ascertaining, 'beyond reasonable doubt', that the suspect text is a derivative of the original therefore requires the manual analysis of a human 'detector' (ideally a trained forensic linguist), who is able, as Woolls (2010) argues, to handle the complexity underlying the principle of similarity.

Conclusion

This paper specifically discussed three linguistic strategies used to plagiarise: word substitution, word reordering and translation. It demonstrated, with examples from a corpus of instances of plagiarism, that these strategies are commonly used to plagiarise, and that at least some amount of editing is expected from instances of plagiarism, not the least as a result of proofreading.

A linguistic analysis of these instances, identical to that applied in forensic contexts, was provided, on the one hand to describe how these strategies were operationalised, and, on the other hand, to explain why they represent plagiarism. This analysis illustrated three cases of linguistic operations that existing software packages fail to detect, or misidentify as original text – especially when the text is altered substantially, or when a combination of strategies is used. The latter, especially, has the potential to hamper the machine detection and pass unnoticed, even if it is potentially the most relevant in demonstrating the plagiarist's intention to consciously manipulate the text and pass it off as his/her own. These are some apparently simple, yet relevant issues and complexities that are imposed on the detection procedure. Further software improvements are necessary until systems can efficiently and correctly detect plagiarism, and these may take some time to be implemented.

The future of the machine detection effectiveness is challenging, yet promising. Although existing systems can hardly have sufficient processing power to (a) manage sophisticated dictionaries to identify instances of word substitution as plagiarism, and (b) handle a combination of different plagiarism strategies, the increase in processing power should make this easier in the coming years. Additionally, more research is necessary in the areas of natural language processing and computational forensic linguistics to address the need to be agnostic to the word order when building a word index. Nevertheless, for some areas of plagiarism detection that until recently seemed almost impossible, the future is already here. This is the case of 'translingual plagiarism' (Sousa-Silva, 2013), where translation into or from another language is used to plagiarise.

Moreover, these complexities raise terminological challenges and ethical issues as to whether (most) existing software packages can fairly be called 'plagiarism detection software', or on the contrary whether calling them 'text matching software' (which is what most of them do) is more accurate.

Indeed, until more advances are implemented to address complexities such as the ones identified, the latter is certainly more appropriate. In the meantime, as Woolls (2010, p. 590) argues, "any computer program can only be an approximation of what human readers can recognise and handle". Given the underlying ethical, moral and, more importantly, serious legal implications, special care is advisable to avoid the misclassification of instances of plagiarism. As a first step, the analyses and the reports provided by detection systems can be interpreted with the assistance of a forensic linguistic analysis, so as to discard false positives, on the one hand, while at the same time unveiling hidden true positives that may have been missed by the detection systems.

End notes

[1] See e.g. http://www.theguardian.com/world/2011/mar/01/german-defence-minister-resigns-plagiarism

[2] See e.g. http://www.nature.com/news/romanian-prime-minister-accused-of-plagiarism-1.10845

[3] See e.g. http://www.dw.de/plagiarism-charges-cost-german-minister-phd/a-16544422

[4] See http://static.publico.pt/homepage/provedor/formaDePlagio/

[5] See e.g. http://www.theguardian.com/media/2011/jul/12/johann-hari-suspended-independent

[6] See e.g. http://www.publico.pt/noticia/universidade-do-minho-e-a-primeira-do-pais-a-anular-doutoramento-por-plagio-1472839

[7] See e.g. http://www.plagiarismadvice.org/news/54-growing-problem

References

Angèlil-Carter, S. (2000). *Stolen language? Plagiarism in writing*. Harlow: Longman.

Coulthard, M., & Johnson, A. (2007). *An introduction to forensic linguistics: Language in evidence*. London and New York: Routledge.

Eggington, W. G. (2008). Deception and fraud. In J. Gibbons & M. T. Turell (Eds.), *Dimensions of forensic linguistics* (pp. 249–264). Amsterdam and Philadelphia: John Benjamins.

Eiras, H., & Fortes, G. (2010). *Dicionário de Direito Penal e Processo Penal*. Lisboa: Quid Juris.

Howard, R. (1995). Plagiarisms, authorships, and the academic death penalty. *College English*, *57*(7), 788–806.

Jameson, D. A. (1993). The ethics of plagiarism: How genre affects writers' use of source materials. *Bulletin of the Association for Business Communication*, *56* (2), 18.

Leitão, L. M. T. M. (2011). *Direito de Autor*. Coimbra: Almedina.

Pecorari, D. (2008). *Academic Writing and Plagiarism: A Linguistic Analysis*. London: Continuum.

Pereira, A. L. D. (2003). Problemas actuais da gestão do direito de autor: gestão individual e gestão colectiva do direito de autor e dos direitos conexos na sociedade da informação. *Estudos em Homenagem ao Professor Doutor Jorge Ribeiro de Faria - Faculdade de Direito da Universidade do Porto*, 17–37. Coimbra: Coimbra Editora.

Potthast, M., Stein, B., Eiselt, A., Bauhaus-universitat Weimar, A. B., & Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In B. Stein et al., eds. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*. CEUR-WS.org, pp. 1–9.

Robillard, A. E. (2008). Situating plagiarism as a form of authorship: The politics of writing in a first-year writing course. In R. Howard & A. Robillard (Eds.), *Pluralizing plagiarism: Identities, contexts, pedagogies (*pp. 27–42). Portsmouth: Boynton/Cook,

Scollon, R. (1994). As a matter of fact: The changing ideology of authorship and responsibility in discourse. *World Englishes*, *13*(1), pp.33–46.

Scollon, R. (1995). Plagiarism and ideology: Identity in intercultural discourse. *Language in Society*, *24*, 1–28.

Sousa-Silva, R., Grant, T. & Maia, B. (2010). “I didn’t mean to steal someone else’s words!”: A forensic linguistic approach to detecting intentional plagiarism. In *4th International Plagiarism Conference: “Towards an Authentic Future” - 2010 Conference Proceedings & Abstracts*. PlagiarismAdvice.org.

Sousa-Silva, R. (2013). *Detecting plagiarism in the forensic linguistics turn*. Unpublished PhD thesis. Birmingham: Aston University.

Thompson, C. (2002). Discourses on plagiarism: To discipline and punish or to teach and learn? In *Australian New Zealand Communication Association (ANZCA) Conference, Queensland, Australia, 10-12 July 2002*. Hosted by Bond University, Queensland.

Turell, M.T. (2008). Plagiarism. *Dimensions of Forensic Linguistics (*pp. 265–299). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Woolls, D. (2010). Computational forensic linguistics: Searching for similarity in large specialised corpora. In M. Coulthard & A. Johnson (Eds.), *The Routledge Handbook of Forensic Linguistics* (pp. 576–590). Milton Park, Abingdon, Oxon; New York, NY: Routledge,

Acknowledgements

This work was partially supported by grant SFRH/BD/47890/2008 FCT-Portugal, co-financed by POPH/FSE. The author would like to thank the reviewers of the *IJEI* for their comments and suggestions.

About the author

Dr Rui Sousa-Silva is author of the PhD thesis ‘Detecting Plagiarism in the Forensic Linguistics Turn’, presented at Aston University, in Birmingham (UK). His research interests lie in within-language and between-language plagiarism, as well as in investigative linguistics and authorship analysis in forensic contexts. He is a member of CLUP, the Linguistics Centre of the University of Porto, and of the Forensic Linguistics Group of UFSC, the Federal University of Santa Catarina (Brazil). He also works as a consultant and has also been involved as a linguistic expert in forensic cases, including court cases.