Legitimated Plagiarism: An investigation of textual borrowing in official documents

Rui Sousa-Silva

Centre for Forensic Linguistics

School of Languages and Social Sciences

Aston Triangle

Birmingham B4 7ET, UK

E-mail: r.sousa-silva@linguisticaforense.pt

Abstract

Plagiarism, which in its most basic form consists of using someone else's text without acknowledgement, is generally considered a fraudulent activity, in both academic and non-academic contexts, with ethical and legal implications. On the ethical side, plagiarists can be demanded public apology; more severe, legal implications can involve student-plagiarists being failed, academic degrees being rescinded, or plagiarists being pushed to resign. However, some official documents are apparently immune to charges of plagiarism, and traditionally remain unchallenged.

This paper presents the results of an analysis of four sets of texts, only two of which were considered to be plagiarism. All texts were analysed against three criteria (quantitative and qualitative) used in forensic linguistic analysis of plagiarism: directionality, volume of textual overlap and linguistic markers. The comparison of the features present, the quantification of overlap and the determination of the direction of the borrowing demonstrate that all texts share the same type of criteria, and that in some instances these criteria are even more marked in the unchallenged texts – indicating that all texts should be considered plagiarism. Hypothesising that the legitimacy of some texts and the illegitimacy of the others are apparently dependent on the respective textual genre, as well as on the common practice, this paper challenges the legal and ethical principles behind this legitimacy, and concludes by arguing that all texts are equally fraudulent, both from a legal and from an ethical perspective.

Keywords: Fraud; Plagiarism; Forensic Linguistics; Linguistic Analysis.

1. Introduction

Plagiarism has attracted the interest of different disciplines over time, with cases of textual borrowing labelled differently as fraudulent behaviour or, conversely, as legitimate text reuse. In the academy, plagiarism is an unacceptable practice; students are required to properly acknowledge their sources, lecturers and tutors are ever more demanding with their students, academic journals are increasingly intolerant to 'unacknowledged' or 'improper' text reuse, university administrations (although not always for the right reasons) encourage and cooperate with investigations of plagiarism and, maybe as a consequence, the general public seems to cooperate. Recent cases of plagiarism involved a lecturer of a Portuguese university, who resigned in 2010 as a result of accusations that she had plagiarised her doctoral thesis. Later, in 2011, the German Defence Minister Karl-Theodor zu Guttenberg (temporarily) renounced his doctorate title and eventually resigned following accusations of plagiarism. In 2012, the Romanian Prime Minister Victor Ponta was accused of having plagiarised large portions of his doctoral thesis, and faced pressure to resign. These cases demonstrate that even when produced in academic contexts, plagiarism can and does have implications out of the academy, but the consequences of plagiarism are not limited to academic work. A few years ago a journalist of the Portuguese quality newspaper *Público* faced accusations that she had plagiarised, and more recently the journalist of *The Independent* Johann Hari was suspended after he was found to have plagiarised news articles. The 'public judgement' of cases such as those of Guttenberg, Hari or Porta suggests, if anything, that plagiarism is believed to be immoral and an act of deception, but ultimately it is the illegal nature of the borrowing that accounts for legal action being adopted. By this token, plagiarism is not only immoral, but also illegal; and consisting of knowingly concealing a fact or misrepresenting aspects of the truth, it is a fraudulent activity.

One of the problems is finding evidence that a certain text has been plagiarised. In cases of barefaced, straightforward borrowing, where the text is copied and pasted word-for-word, a simple comparison of the suspect text against the potential source to determine the amount of overlap, together with the identification of the date of publication and considerations of access to each text by the other author, may be sufficient to determine whether a suspect text is a result of plagiarism or on the contrary, an original product. Most instances are not, however, as simple as *verbatim* borrowing, involving sophisticated strategies of text manipulation. Owing to the technological developments of the last decades, access to more information made it easier to plagiarise any text in any part of the world; luckily, those same developments also facilitated the detection of those instances. Maybe as a consequence, plagiarists had to invent new ways of plagiarising without being easily detected, and this meant deleting or adding to the original, replacing words or phrases, changing the word order, paraphrasing the original or translating from another source. In this case, more advanced linguistic methods are necessary to compare the suspect text against the potential original to then conclude whether the text has been plagiarised. The type of linguistic analysis developed and used by forensic

linguistics has demonstrated good results in plagiarism detection. Previous and ongoing research in forensic linguistics, which approaches linguistic analysis in the scope of the interaction between language and the law, has demonstrated that the likelihood that two texts may have been produced independently can often be determined accurately, and this data can be used as evidence or, at least, as an investigative tool. An additional problem is that some cases of textual borrowing are not even challenged – even when they reuse significant portions of text from other sources – on the grounds that they are texts of a particular genre, often formulaic, to which the requirements of original authorship do not apply, as is the case of some documents of legal or technical nature that circulate around the world, unacknowledged and unattributed.

The purpose of this study is to challenge the assumption that plagiarism only applies to academic contexts, and that texts of a particular genre are not bound to respect the same principles as texts of other genres or disciplines. Using a combination of three quantitative and qualitative criteria (*directionality*, *volume of overlap* and *linguistic markers* and *strategies*), this study determines the extent of borrowing in two sets of texts that were found to have been plagiarised (academic texts and news articles), identifying which linguistic markers and strategies – and to what extent – are present or absent from the borrowed texts. Subsequently, this method is applied to two other sets of texts that are traditionally beyond suspicion: two applications for public funding and three Memoranda of Understanding (MoUs) signed by the three bailed out Eurozone countries (Greece, Ireland and Portugal) and the European Union (EU) and the International Monetary Fund (IMF). The comparison of the four sets of documents suggests that, if the criteria above (which have been found to be valid and reliable in real forensic cases of plagiarism) are applied, applications for public funding and MoUs can also be challenged as fraudulent behaviour, as much as academic and news plagiarism.

This paper is structured as follows. Section 2 reviews the literature on plagiarism across different disciplines. Section 3 explains how the research is operationalized; it describes the corpus of texts analysed in this study and the analytical method employed. The findings of this analysis are presented in section 4. Section 5 presents a summary of the findings and discusses future implications of this work.

2. Literature review

"Un simple imitateur est un estomac ruiné qui rend l'aliment comme il le reçoit: un plagiaire est un faussaire". This quote, attributed to Voltaire, reflects some of the basic principles underlying plagiarism, in particular that it consists of reusing the product of someone else's work without acknowledgment, and hence make it pass as one's own. Portraying the imitator both as an *estomac ruiné*, who is unable to properly address and manage their sources, and as a *faussaire*, whose acts are counterfeit, Voltaire describes plagiarism as an immoral act, with social and legal implications. Additionally, if we take Garner's definition of fraud as knowingly misrepresent the truth or conceal "a

material fact to induce another to act to his or her detriment" (Garner, 2009: 731), then we have to aver that plagiarism is a fraudulent behaviour. But perspectives of how plagiarism cases should be handled, and whether and how they should be subject to punishment have not been unanimous over time. Although it has been argued that plagiarism is either immoral (Garner, 2009) or unethical (Goldstein, 2003), rather than illegal, this has been refuted in the literature (e.g. Finnis, 1991; Eiras and Fortes, 2010), and practice has demonstrated that plagiarism cases can actually be subject to court trial. In fact, in most legal systems (including the Portuguese and the English) the concept of 'illegal' is based on the principle of that which is 'immoral', so that what is immoral often becomes illegal (Finnis, 1991; Eiras and Fortes, 2010); and consequently plagiarism is subject to trial in common law courts, as has been demonstrated by Turell (2008). In the case of the Portuguese jurisdiction, plagiarism is equated with theft and contrafaction – both representing criminal conduct. However, the following four criteria should be met for that conduct to be considered an infringement: (i) be deceitful; (ii) use other people's work and pass it off as one's own; (iii) be a mere reproduction of someone else's work; and (iv) reproduce someone else's work in such an identical way that it has no individual value. These reflect, to a great extent, the provisions of international agreements on copyright and intellectual property, and those of the Berne and Paris Conventions in particular, especially in terms of originality and requirements of being fixed in material form, as those Conventions influenced intellectual property law in general, and copyright law in particular in several countries (Bently and Sherman, 2009).

Unsurprisingly, then, plagiarism has frequently been associated with metaphors of crime, such as 'theft' (Angèlil-Carter, 2000) and 'misappropriation' (Jameson, 1993). These authors concur that this association is not independent of the attribution of property rights to intangible goods, in the late eighteenth century and mid-nineteenth century, which granted the right to individual property and required a legal framework to cope with the infringement of those rights. However, plagiarism is not always considered on the grounds of the oversimplified dichotomist relation between the moral – or ethical – and legal grounds. On the contrary, its multidisciplinary nature contributed to the development of different outlooks, which made plagiarism more than a purely moral and/or legal debate. Education and intercultural studies, in particular, have addressed plagiarism critically. Firstly, as Howard (1995) argued and Pecorari (2008) later demonstrated, differing degrees of penalties are required in academic contexts where instances of plagiarism can indicate a lack of academic writing skills, rather than an intention to deceive, so that borrowing in the academy should be punished when the student attempts to deceive the reader, but not when the instances of plagiarism resemble a 'patchwork' resulting from an unsuccessful attempt of the student at writing academically – especially considering that such 'patchwriting' is part of the educational experience of the students (Howard, 1995). The principles that plagiarism can be punished on equal terms in academic and non-academic contexts (Howard, 1995; Howard and Robillard, 2008; Angèlil-Carter, 2000; Pecorari, 2008) and that, even within academic contexts, plagiarizing students should be treated equally are therefore

challenged. Secondly, claiming that (academic) policies against plagiarism – and the corresponding penalties – represent the hegemony of the West, by transferring to other contexts (such as the academic context) provisions that have been established by copyright agreements comprising mostly western countries, intercultural studies (Scollon, 1994; 1995) have challenged the principle that plagiarism is a universal concept, equally understandable by everyone – with disregard for other mitigating circumstances, such as the 'cultural respect' for the Masters materialised on quoting them or the unsuccessful – yet unintentional – attempt to build upon prior authoritative authors (Howard, 1995).

Finding evidence of plagiarism and demonstrating that it represents a fraudulent activity, whose deceptive nature results e.g. from lying (Eiras and Fortes, 2010), presupposes the ability to detect the instance(s) of plagiarism and fraud, which can both be daunting tasks. Linguistic evidence is necessary to demonstrate the borrowing, whether the suspicions 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) result from the reader feeling that they have read the material elsewhere, from a systematic, manual or machine-assisted analysis of the texts to find similar/identical/matching strings, or from an intrinsic, stylistic analysis of texts to identify different, often clashing writing styles that may indicate multiple authorship. This type of evidence, provided via a forensic linguistic analysis of the texts, has been increasingly used, both in academic and non-academic contexts.

Forensic linguistics, which consists of using linguistic methods and analyses in forensic contexts, has been used effectively to investigate and provide evidence of fraud, e.g. in the detection and investigation of text reuse, including plagiarism and collusion. As reported by Coulthard and Johnson (2007), in academic contexts linguists have been increasingly asked by colleagues to detect, investigate and/or confirm – or refute – instances of student plagiarism, but the potential of linguistic analysis to proving textual borrowing in non-academic contexts is also demonstrated. Citing the example of the document 'Iraq: Its Infrastructure of Concealment, Deception and Intimidation' – more commonly known as the 'Dodgy Dossier' – which the British government presented to the United Nations in 2003 to justify their invasion of Iraq, the authors discuss the extraordinary textual identity between this document and a prior academic article. A linguistic analysis of the two texts demonstrated that the official governmental document had been substantially plagiarised from this article, with changes in spelling only (from American English to British English). Similarly, discussing a case that was taken some years ago to the Spanish courts involving copyright disputes, Turell (2008) provided linguistic evidence that a translation of Shakespeare's *Julius Caesar* into Spanish derived from a previously published translation of the same book, rather than having been produced independently.

Investigations of instances of plagiarism are usually based on a comparison of the suspect text(s) against possible originals to find linguistic data to conclude that the suspect text is actually a derivative text, or otherwise an original. In its simplest form, this analysis could imply matching the suspect text against the source and highlighting the identical strings, usually sequences of several words that are copied *verbatim* from another source and used without acknowledgement. In cases of *verbatim* plagiarism, where the original text is borrowed word-for-word, 'as is', a simple comparison of the suspect and the original is sufficient to identify the overlapping, identical phrases, sentences or even paragraphs. However, the investigation is considerably more complex when the derivative text is frequently edited, whether to disguise the authorship and make it pass as their own, or in a failed attempt to write properly – as is often the case in academic writing. Such changes can consist, for example, in altering the word order or reformulating the sentence structures, in paraphrasing the original text or changing the cohesion and the coherence of the original. These strategies, which involve simple to sophisticated alterations in grammar, punctuation, syntax, and semantics or even in vocabulary and discourse, make the detection procedure more difficult. Changes in grammar, punctuation, syntax and word order usually imply text re-ordering in a sense that the sequences of identical words are interrupted, building an apparently distinct original – albeit containing the same, non-original ideas, and possibly reusing some vocabulary. As a result of the reuse of a number of identical words in a different order, the detection procedure fails to identify sequences of identical words of a length that is significant to be deemed plagiarism. In order to overcome problems of this type, more sophisticated methods are required, such as the one used by Johnson (1997): discarding strings and sequences or chains of words (the methods traditionally used to detect plagiarism) and grammatical items, which are close sets of words (hence smaller in number, and likely to be shared anyway), she concentrated on the analysis of shared lexical items. After calculating the percentage of lexical types in the set of three suspect documents against that of a set of three non-suspect documents, she found the percentage of overlapping types (i.e. the number of types of lexical vocabulary occurring in the text) in the latter set to amount to 20%, compared to 72% in the former, and concluded that an analysis of lexical overlap is robust even to changes in syntax and word reordering, allowing the detection of instances of plagiarism that are usually missed by search of identical strings. As justified by Coulthard and Johnson (2007), this can be theoretically grounded on the principle of linguistic uniqueness, i.e. that even the same person writing on the same topic in different occasions would be expected to word the text differently; in the case of two or more texts each authored by different people, lexical overlap would therefore indicate either that one is derivative from the other(s) or that they have been produced collaboratively.

This type of lexical analysis may, however, be of limited usefulness in analysing the coherence and cohesion of instances of plagiarism, since edits introduced to those plagiarised texts usually involve changing the words of the original to reflect a textual or extra-textual, 'real-world' reality that does not necessarily match that of the source text. In other words, the plagiarist may or may not retain the

strings and sequences/chains of words of the original, as well as lexical items related to the topic of the text, but elements of coherence and cohesion can be adapted to convey a coherent link to the plagiarist's reality. Conversely, such instances might be detected more effectively by identifying inconsistencies usually revealed in "referential style" (e.g. inconsistent use of imperative or infinitive verb tenses in forms of address), "decontextualisation" (e.g. by omitting parts of text that otherwise contribute to contextualising the text reused), and "inversion of structural elements" that result in conceptual inconsistencies (Turell, 2008). However, in her analysis Turell considered flaws in the text identified by these linguistic markers to detect plagiarism; in cases where alterations are successfully made to retain the coherence and the cohesion of the plagiarised text, the linguistic analysis of the text needs to concentrate on the analysis of differences surrounding identical textual elements, more than on the analysis of similarities and inconsistencies.

Prior authorship and volume of borrowing are two other criteria used to identify instances of plagiarism. Turell (2008) makes a good case for both of them. Prior authorship, which is usually determined by the date of publication, in most cases help resolve issues of directionality, i.e. by determining which text is the original and which one is the derivative, except in cases where the dates of publication are very close, or when the two texts have a contemporary production (Turell, 2008: 282). On the other hand, considerations of volume, which are based on the assumption that the higher the percentage of overlapping text, the more likely it is that two (or more) texts have not been produced independently, are relevant both in academic and non-academic contexts. In academic contexts, universities seem to often base their definitions of plagiarism on the assumption of "substantial" borrowing (Coulthard and Johnson, 2007). In non-academic contexts, volume of borrowing can be addressed according to different levels (or degrees) of plagiarism: (i) "uncredited verbatim copying of a full paper"; (ii) "uncredited verbatim copying of a large portion (up to 50%) from a paper"; (iii) "uncredited verbatim copying of individual elements (paragraph(s), sentence(s), illustration(s), etc.)"; (iv) "uncredited improper paraphrasing of pages or paragraphs"; and (v) "credited Verbatim Copying of a Major Portion of a Paper without Clear Delineation"[1]. Empirical evidence demonstrates that using quantitative measures such as similarity in overlapping vocabulary, shared once-only words, unique vocabulary and shared once-only phrases (Johnson, 1997; Woolls and Coulthard, 1998; Woolls, 2003; Turell, 2004) can effectively contribute to the start of the analysis, but it is also admitted that "[t]aken in isolation, it is possible that all these measurements do not discriminate sufficiently" (Turell, 2008: 288). Alternatively, biased judgements of plagiarism based solely on quantitative criteria (plagiarism thresholds) can be avoided by using a combination of quantitative and qualitative analyses to demonstrate, based on the principle of idiolect and linguistic uniqueness (Coulthard, 2004; Coulthard and Johnson, 2007), that being very unlikely that two different people at different occasions produce identical text, the amount of identical text across the

[1] See IEEE "Guidelines for Adjudicating Different Levels of Plagiarism" (2006: 63-65).

documents from the same sets could indicate that they were either (a) produced by the same person(s), (b) produced by different person(s), with or without the knowledge of the other(s) or (c) be both based on a third text.

3. Methodological considerations

The purpose of this paper is to investigate whether textual overlap across different documents is indicative of plagiarism, and whether those instances of borrowing are considered a fraudulent activity. This research is operationalized using the three criteria discussed in the previous section, and whose validity has been demonstrated, not only in detecting different types of plagiarism, but also in discarding false positives as a means to improve the detection method: (i) prior authorship; (ii) volume (amount) of overlap; and (iii) linguistic markers.

The criterion of prior authorship is the one used most frequently to detect the directionality of the borrowing, and hence determine which text is the original and which one is the derivative. This criterion ensures unequivocal results when the two texts have been produced at two chronologically different points in time and their date of production is known, but tends to break when the date of production of the texts is unknown, when the texts are produced roughly at the same point in time, or in cases of collusion, where both texts are the result of a collaborative production.

The criterion of volume is based on the assumption that, the higher the amount of textual overlap between two or more texts, the higher the probability that those texts have not been produced independently, and the more severe is the judgement of plagiarism and fraudulent behaviour. Obviously, this applies only to instances of plagiarism where the derivative text has borrowed *verbatim*, word-for-word, from another original, to such an extent that a comparison of the texts shows a significant identity in phasing and lexical choice. Conversely, the volume of overlap tends to drop in cases where the original texts is edited, to a lesser or greater extent, for example when the phrasing and the sequences of words are altered in such a way that the word order is changed, the grammatical words are replaced with alternative equivalents to match the new word order, and mostly lexical words are retained. The volume of textual identity can be expected to drop even more in cases where only works or ideas are borrowed, rather than the exact words. This is the case where the plagiarist paraphrases the original text, or where the borrowed instances are translated from an original in another language. In this latter case, relying solely on the criterion of volume to detect, investigate and judge plagiarism can be ineffective.

On the contrary, an investigation that is based on the analysis of linguistic markers is effective both in detecting plagiarism and, by describing the process behind the borrowing and justifying it, in investigating the direction of the borrowing. A deeper linguistic analysis has the potential to reveal more complex and sophisticated strategies used by plagiarists. Firstly, an analysis of lexical overlap

that discards plagiarism detection based on sequences of words is effective in identifying text edits such as changes in punctuation, grammar, and syntax. Following from this, an analysis of the meaning relations established between words, considering possible replacements – either syntactic or paradigmatic – is effective in detecting cases of paraphrasing. Thirdly, an analysis of the referential style, contextualisation and order of structural elements is able to identify issues of coherence and cohesion, being effective in detecting plagiarism (when inconsistencies are found in referential style, contextualisation and order of structural elements), or conversely in explaining how and why a certain instance, albeit consistent, has been altered to disguise the derivative text. Finally, plagiarism of ideas operated via translation of text originally written in another language can be detected only by 'guessing' the language of the original and translating it into that language for textual comparison.

This study is based on the linguistic analysis of a *corpus* of four sets of real, naturally occurring texts, each set of a different genre, including academic and non-academic documents: university student essays (set 1), news (set 2), applications for public funding (set 3) and official institutional documents (set 4). The linguistic analysis is performed to check the texts for the existence or absence of the three criteria described – i.e. prior authorship, volume and linguistic markers – so as to investigate whether instances of plagiarism in these texts are judged as fraudulent behaviour or legitimated by their genre.

Set 1: Academic Texts

This set includes two texts submitted by two groups of university students for assessment; the authors of text A were design students, native speakers of Portuguese; the authors of text B were students of communication sciences, non-native speakers of Portuguese, spending a year abroad in Portugal. In both cases, the suspicions of plagiarism were raised by the lecturers, who intuitively identified clues in the texts suggesting that they might have been authored by a third party. Interestingly, those clues were not related to unexpected writing standards: text A was well written, as would be expected from students at this level of education; text B contained several instances of the students' native language in their writing in Portuguese, and which were not found unusual, given their origin. A deeper linguistic analysis was necessary to investigate whether the two texts originated elsewhere and, in that case to determine the amount and type of borrowing. To allow for a textual comparison of the two texts and consequently determine the amount of overlap, the suspect text (written in Portuguese) was translated into the expected language of the original (Spanish).

Set 2: News Articles

This set of texts includes two cases of news plagiarism. Text 1 was published in a Portuguese quality newspaper. The suspicions of plagiarism were raised by the Portuguese news agency, Lusa, who produced the copyright-protected news piece. They found that a significant amount of textual material

had been lifted, which represented plagiarism. Text 2 was published in the Sunday supplement of the Portuguese quality newspaper *Público*. The accusations of plagiarism were triggered by a reader, who realised upon reading a piece on sunscreens that he had already read it in English elsewhere. He complained to the newspaper, which started an investigation on the case to determine whether there was reason to believe that the article was plagiarised. The newspaper also enquired the journalist, who until a very late stage in the investigation rejected having adopted a fraudulent behaviour. *Público* concluded that the journalist had plagiarised from the *Wikipedia* and *The New Scientist*, and demanded an apology from the journalist. Interestingly, this and subsequent cases represented a turn on the assessment of news plagiarism, the judgements of which have traditionally been more relaxed, as Coulthard and Johnson (2007) pointed out. Again, to allow for a textual comparison of the suspect text against the originals and an analysis of overlap, as with text 2 of set 1, the newspaper article was translated into English (the language of both the *Wikipedia* entry and the *New Scientist* article).

Set 3: Applications for Public Funding

This set includes two texts allegedly submitted by two different partnerships (albeit sharing some partners) to a call for applications for public funding in the amount of about 200,000 Euro each. Upon assessing the two proposals, the assessment experts found that a few applications contained some identical contents, but two in particular shared an apparently significant amount of identical strings. The outcome of the quality assessment of the proposals by the experts resulted in one of them being rejected, and the other being approved. As a result of the concerns that (a) the two proposals had not been produced independently and (b) the approved proposal would be entitled to funding of hundreds of Euros (despite being identical to the one rejected), an investigation of the textual overlap was conducted; the two texts were compared to determine the amount and type of shared text, and whether that amount of overlap was to be expected in this particular genre (applications for funding), especially considering that all applications are based on a set form provided by the funding agency.

Set 4: Memoranda of Understanding (MoUs)

The fourth set of documents includes the MoUs signed between the European Commission (EC), the European Central Bank (ECB) and the International Monetary Fund (IMF), on the one hand, and Greece, Ireland and Portugal, on the other hand, and which preceded the loan recently granted to the three countries. Each MoU is composed of a Letter of Intent, a Memorandum of Economic and Financial Policies and a Technical Memorandum of Understanding. The three memoranda were signed at different points in time; Greece signed first (3 May 2010), followed by Ireland (3 December 2010) and finally Portugal (17 May 2011), over one year after the first agreement was signed with Greece. Some degree of overlap, as well as some differences in wording, would be expected from the three

memoranda. In particular, given that the genre is that of an agreement of a legalese nature, the documents would be expected to share an identical outline, with similar headings, and obviously identical wording in the contact details of the parties that remain constant across the three documents. A first reading of the set of documents revealed an apparently unusual volume of overlap, which could be explained by the fact that the three countries were affected by the same crisis, in similar contexts and under similar constraints. Nonetheless, given the cultural and infrastructural diversity of the three countries, an agreement addressing the individual specificities of each country in a different wording would be expected.

The following section discusses these premises and presents the results of the (comparative) analysis of the four sets of texts.

4. Results

This section presents the findings of the analysis of the data contained in each set. The findings reported consider the three criteria identified in the literature: directionality, volume of overlap and linguistic analysis of markers and strategies.

Set 1: Academic Texts

A comparative analysis of suspect text A, written by a group of graphic design students on how to plan a design newspaper, and the expected source, a guide on how to write a school newspaper, shows a very high matching of 91%. This indicates that most of the suspect text, with the exception of a few instances where changes have been introduced, are borrowed *verbatim*, word-for-word from the source – these changes accounting for the remaining 9% of the text.

Most of the edits introduced by the students in the derivative text consist of deleting words or phrases, mainly to make the text more coherent. For instance, all references to the extra-textual world of a school newspaper (e.g. “teachers”, “parents”, “students”, “school activities”…) were omitted. Similarly, some adjectives related to educational activities were replaced with adjectives from the semantic field of cultural activities, references to social roles and social activities of performing arts (“singer”, ”actor”; “football match”) were replaced with social roles of the visual arts (“designer”, “visual artist”; “artist’s performance”). Interestingly, these changes have been retained throughout the derivative text, even where anaphoric references were used, indicating that the student-plagiarists were sufficiently careful to keep the contextual consistency of the text. New words were also added, especially adverbs (“normally”) that interrupt the chain of words borrowed, while retaining the meaning of the text. Additionally, changes were introduced in the word order, influencing the grammar: for example a subordinate clause was transformed into a sentence, and consequently a demonstrative pronoun was introduced; and although the verb was maintained, the tense was changed

by transferring the future tense to the auxiliary verb and using the infinitive tense of the main verb. Although these changes do not represent a paraphrase in the strictest sense, their detection by *verbatim* plagiarism detection systems may be compromised, as happens with paraphrasing. While introducing changes to the text, the students also retained the style of the original, so that the reader can hardly identify multiple "voices" in the text. The directionality of the borrowing could be easily determined by the date of production, as well as by accessibility criteria. On the one hand, the guide on how to create a newspaper from which the students borrowed most of their text was published by the Portuguese newspaper *Público* prior to the students submitting their writing for assessment. On the other hand, while the students had easy access to the guide that was publicly available, it is very unlikely – even if the two texts were contemporary – that the author of the guide had access to the students' academic work.

The second case of student plagiarism involves a group of non-native speakers of Portuguese writing academically in Portuguese. Although the writing standard matched the expectations of the lecturer, considering that the authors of this piece were foreign students writing in another language, suspicions of plagiarism were triggered by the reasoning behind the text. An Internet search of a few strings revealed that the text had been borrowed from another language source and then (poorly) translated into Portuguese. This plagiarism strategy poses several challenges to the plagiarism detection procedure. Firstly, it prevents a straightforward linguistic comparison of the suspect against the source, since the two texts are in different languages. Secondly, even if confirmed, these instances of plagiarism do not consist of the usual 'linguistic plagiarism', but instead of the 'plagiarism of ideas'. The solution to overcoming this challenge and consequently improve the detection procedure consists of converting the two texts into the same language to obtain a rough indication of the volume of borrowing, and thus allowing for a linguistic comparison, as if the two texts were written in the same language.

The very high volume of overlap between the two documents (74%) after conducting this procedure indicates that the student-plagiarists translated the texts literally. A deeper linguistic analysis of the texts also reveals that some other plagiarism strategies, common in same-language 'linguistic' plagiarism, were used. For example, some sentences were edited and words were replaced with semantically-related words, either consciously or due to lack of translation skills. Some differences between the plagiarising and the machine-translated text, especially those operated at the grammatical level (such as verb tenses and demonstrative pronouns), contribute to making the text more grammatically correct. Deletions were also operated in cases where words or phrases were not essential to the meaning of the text, such as specification and examples that do not add to the core argument (e.g. "a halo of mystery" is simply altered to "mystery"). Some alterations span beyond the limits of the sentence, to join clauses that in the original are part of the same sentence. Other edits aim to ensure the consistency of the text with the extra-textual world, thus having implications at the level

of coherence and cohesion. For example, a reference to a railway company in the original is omitted in the derivative text to reflect the extra-textual world of the student-authors. Finally, in some instances, the amount of rewriting (i.e. expressing the same ideas in other words) of the student text *sensu lato* equated with paraphrasing.

Taken together, these results indicate that the suspect text was lifted from another source, in another language, which was available on the Internet. The original was therefore accessible to the student-plagiarists, and not otherwise.

Set 2: News Articles

An analysis of the textual overlap between the news article of the Portuguese news agency Lusa and the online version of the same text that was published in the newspaper website demonstrates a very high identity of 95%, indicating that the main plagiarism strategy used was *verbatim* borrowing. A linguistic analysis of the two texts demonstrates that the remaining 5% are mainly due to minor grammatical and syntactic edits made to the derivative text, and to replacements of lexical items with synonyms. To a greater or lesser degree, all these changes impact the statistics and the performance of the detection systems. For example, in one case two sentences were joined into one sentence via the Portuguese connective "e" (meaning "and"), and a main verb was replaced with another verb followed by the noun form of the original (i.e. "trabalha" was replaced with "desenvolve o seu trabalho"). Some grammatical changes were made to retain the cohesion of the text (e.g. the subject was changed from plural to singular and the verb was adjusted accordingly), and demonstrative pronouns and subordinate clauses that did not impact the meaning of the text were deleted. Conversely, text was also added, such as titles and one sentence stating the number of times that the police had used that method to that date. The latter, in particular, contributes to increasing the credibility of the text by referring to extra-textual facts. Although the two texts were published on the same date (27/1/2009), albeit at different times, the analysis of these linguistic markers corroborates the evidence provided by the time of publication, which demonstrates that the text was first published by the news agency and only then by the newspaper.

The second group of documents consists of a suspect text on sunscreens that was published in the Portuguese newspaper *Público* and the two sources from which the journalist allegedly plagiarised, the *Wikipedia* and the *New Scientist*. The comparison of the original texts with the translation of the derivative text into English returned a surprisingly low percentage of textual overlap: only 41%. This is below the 50% percentage that Turell (2008) claimed is acceptable in some circumstances, and demonstrates that there is at least some volume of *verbatim* plagiarism, although other strategies might have been used more prominently. A linguistic analysis of the texts, however, suggests that this owes to the fact that, contrary to the academic text 2 in set 1, the translated version of the article was edited heavily, mainly for grammar and syntax, as would be expected from a professional writer. The article

introduced new text, for example by elaborating or explaining the ideas of the original, or adding names (e.g. of the researchers and their research institutions), and rephrasing or paraphrasing other instances of the original. Additionally, it combined several sentences of the original into a single sentence via subordinate clauses and used a degree of embedding that is unusual in English, that translation engines are bound to miss. By doing this, the journalist also changed the word order considerably. As a consequence of all these changes, the 41% volume of overlap consists mostly of lexical items reused from the original that the translation engine translated successfully. This analysis demonstrates that the suspect and original documents share an extraordinary level of identity, so it is very unlikely that the texts have been produced independently. The dates of publication and accessibility indicate that the suspect article borrowed from the two originals, and not otherwise. Taken together, these criteria confirm the judgement of plagiarism passed by the newspaper.

Set 3: Applications for Public Funding

The date of production of the two applications for funding is unknown, since only the date of submission was given. Although none of the two partnerships submitting the application disputed the original authorship, the linguistic analysis suggests that the production of the two texts is contemporary. Text A includes several spelling mistakes that have been corrected in Text B, which might suggest that text B is the derivative and text A is the source; nevertheless, the fact that text A includes a tool whose title is identical to the application B shows otherwise. It is therefore plausible that those mistakes have been corrected as a result of a careful proofreading of the application before submission. Considering that each of the two documents includes edits to the other text, rather than one of them consistently editing the other, the textual evidence collected from the comparative analysis suggests that the two documents have not been produced independently, but instead either collaboratively, or else they both derived from another, third text. Secondly, the investigation of the volume of overlap shows that the percentage of shared text is higher than would be expected, even for texts that are of the same genre and that follow the fixed form of the funding agency: an overlap of 75% in the description of the project characteristics, and an overlap of 92% in the description of the activities planned – with an overall percentage identity in phrasing of 83.5%. These percentages were calculated considering only identical text, therefore not including changes in spelling or other edits. These very high percentages of overlap indicate that most of the text was reused *verbatim*, word-for-word.

Thirdly, an in-depth comparative linguistic analysis reveals some markers that are common in cases of plagiarism, such as textual elements that were adapted to reflect changes introduced in the text; for example in one case the two proposals use different verb tenses: one uses the present, whereas the other uses the present continuous. An analysis of the verb in context demonstrates that this tense change was determined by sentence structure changes. Additionally, some textual elements which are

unexpected in standard English were identified in both texts, including erratic spacing around punctuation marks, omission of punctuation at the end of sentences, non-standard subject-verb agreement and other grammatical errors. Likewise, both applications share uncommon spelling mistakes in exactly the same contexts, as well as some instances where different spellings of the same words are used in identical contexts in the two documents, for instance "per cent" and "percent". Additionally, both documents share inconsistencies at the level of coherence/cohesion, such as inconsistent self-referencing (i.e. quoting the name of other application when referencing to themselves). These indicate that text was copied and pasted reciprocally between the two applications. On balance, this analysis suggests that the two documents resulted either from a fraudulent, unauthorised collaborative work of collusion, or from another document, contemporary or produced previously.

Set 4: Memoranda of Understanding (MoUs)

As reported in the previous section, some degree of overlap and differences in wording would be expected in the three MoUs. An analysis of textual overlap however revealed a significant and unexpected percentage of sharing across the three documents; the agreements signed by Ireland and Portugal show a 75% overlap, the documents signed by Greece and Ireland share 77% of text, and the memoranda signed by Greece and Portugal share 82% of text. *Verbatim* borrowing is therefore the main type of text reuse across the three documents, but a closer pair-wise comparison suggests that some of the differences are due to alterations – often minor, or of a coherent and cohesive nature, reflecting the contextual situation of each country. That is the case, for example, of the loan amounts or the named entities involved; in some cases only the names of organisations (e.g. central banks) are altered: "the Central Bank will direct" in the Irish (IE) MoU is versioned "the BdP [Banco de Portugal] will now direct" in the Portuguese (PT) MoU. In other instances, words are replaced with synonyms: "to achieve the objectives of the economic program" in the Greek (GR) MoU is versioned "to meet the objectives of the economic programme" in the PT MoU. Sometimes, synonymy is applied to named entities: "the Fund" (GR MoU) is replaced with "the IMF" (IE MoU), and other edits reflect the referential context of the documents: "policies contained in the MEFP" (GR MoU) is reworded "policies contained in this letter" (IE MoU). Grammatical, syntactic or lexical edits are also employed; for example "which can be drawn over a period of 36 months" and "under the programme" (IE MoU) are reworded "which could be drawn over a period of 36 months" and "under this programme" (PT MoU); and "We sent a parallel request" (GR MoU) is reworded "We also send a parallel request" (IE MoU) and "We are also sending a parallel request" (PT MoU). Other differences are simply due to spelling or number: "The implementation of our program will" (GR MoU) is replaced with "The implementation of our programme will" (IE MoU), "report on the fulfilment" (GR MoU) is replaced with "report on the fulfillment" (PT MoU) and "quarterly disbursements" (GR MoU) is replaced with

"quarterly disbursement" (IE MoU). An example reflecting syntactic changes involves one case of passiviziation: "a memorandum […] will establish a clear framework" (GR MoU) is rephrased "a framework […] will be established between the government and" (IE MoU). At times, edits reflect at the level of punctuation: "European Financial Stability Mechanism/European Financial Stability Facility" (IE MoU) is reworded "European Financial Stability Mechanism and the European Financial Stability Facility" (PT MoU). Other instances, such as "return to durable growth" (IE MoU), are paraphrased "restoration of sustainable growth" (PT MoU). As in other sets, edits sometimes include additions from previous documents: "to cover the balance of payment needs" (PT MoU) is added to the sentence reused from the IE MoU. The directionality of the three MoUs is not challenged, since the date of publication of the three documents is known to the public: a MoU was first signed by Greece in 2010, followed by Ireland a few months later, and finally Portugal in 2011.

Discussion of the findings

The analysis of the four sets of documents above considering directionality, volume of overlap and linguistic markers and strategies demonstrate that text reuse employs identical strategies, independently of the text genre. Firstly, no significant discrepancies were identified in the directionality of the borrowing of the four sets of texts, except for those in set 3; contrary to the other sets, in which the directionality could be determined based on the date of production/publication, the texts in set 3 suggest being contemporary. A deeper linguistic analysis was therefore conducted to identify clues (e.g. completeness, referential style, contextualisation, order of structural elements, …) to the direction of the borrowing. The outcome of this analysis indicated that the borrowing was bidirectional, consequently suggesting that the two texts were either the result of collusion (a result of collaborative writing), or both were borrowed from another text – but not in any case produced independently.

Secondly, the quantitative analysis of textual overlap demonstrates that all documents in the four sets reuse a significant amount of text – the lowest being text 1 of set 2, with 'only' 41% of matching in vocabulary. Despite this percentage being below the 50% baseline described by Turell (2008) as sometimes considered to be acceptable, the newspaper did not hesitate to pass a judgement of plagiarism and act accordingly. If this percentage is considered the baseline in this study, we can then conclude that, quantitatively, all other documents in the remaining sets are plagiarism, including, in ascending order: academic text 2 of set 1 (74%), the MoU of Ireland and Portugal (75%), the MoU of Greece and Ireland (77%), the MoU of Greece and Portugal (82%), the applications for public funding (83.5%), the academic text 1 (91%) and, finally, the news article 1 in set 2 (95%). A quantitative comparison of the documents across the four sets can be found in Table 1. However, it is also relevant to consider a qualitative analysis of the data, especially in order to analyse instances of suspect plagiarism where the text is intentionally manipulated.

Thirdly, the analysis of the linguistic markers used across the four sets, which can be found in Table 2, suggests that not all markers are used in all documents. For example, translation is used only by texts 2 of sets 1 and 2. Interestingly, the Greek MoU, despite being written in English, includes a table whose headings (months) are in French, pointing to an inconsistent decontextualisation often found in student plagiarism. Another marker that was not used consistently across all sets is paraphrasing. *Sensu stricto*, paraphrasing implies expressing the meaning of a certain proposition in other words, so in this analysis a text was considered to be paraphrased when a string of text was rephrased (same meaning conveyed using at least some different words). In this corpus, paraphrasing has been found in all texts, except text 1 of set 2 (which has been mostly borrowed *verbatim* from the original) and the texts in set 3. Conversely, all other markers were employed in all corpus documents; all texts borrowed passages *verbatim* from other sources – or among themselves, in the case of collusion – as the percentages in table 1 demonstrate. Similarly, all documents in all sets make grammatical, syntactic and lexical alterations, as well as in punctuation, including reordering, additions, replacements and deletions. No one particular type of alterations was employed by one particular set of texts, meaning that certain linguistic markers and strategies are not used more prominently in the texts of a particular genre, but on the contrary they tend to be shared by all sets of different genres. Similarly, all texts make alterations in referential style, contextualisation and order of structural elements to ensure their coherence and cohesion.

Table 1: Volume of Textual Overlap in the four sets of texts.

	Sets			
	Set 1	Set 2	Set 3	Set 4
% Textual Overlap	91% / 74%	95% / 41%	83.5%	75% / 77% / 82%

Table 2: Description of Linguistics Strategies across the four sets of texts.

		Sets			
		Set 1	Set 2	Set 3	Set 4
Linguistic Markers	*Verbatim*	✓ / ✓	✓ / ✓	✓	✓ / ✓ / ✓
	Edits	✓ / ✓	✓ / ✓	✓	✓ / ✓ / ✓
	Paraphrasing	✓ / ✓	– / ✓	–	✓ / ✓ / ✓
	Coherence/cohesion	✓ / ✓	✓ / ✓	✓	✓ / ✓ / ✓
	Translation	– / ✓	– / ✓	–	– / – / –

Taken together, these findings indicate that the same criteria were found in all texts in the four sets. It is therefore relevant to question why certain texts are more prone to charges of plagiarism. Coulthard and Johnson (2007) argue that some disciplines are more unwilling to tolerate plagiarism than others, stating that academics, for instance, are more demanding than journalists regarding source citation. However, as the case of text 2 in set 2 demonstrates – and recent events, such as the case of Johann Hari, of *The Independent* confirm – even journalists are nowadays subject to charges of plagiarism. On the other hand, although an investigation of plagiarism was run on the texts of set 3, the panel investigating the case concluded that the borrowing was tolerable, despite the evidence presented above, and that the application was entitled to the funding.

Conversely, despite the irrefutable evidence that the MoUs analysed in this paper have not been produced independently, to my knowledge they – or even other documents with which they may share text – have never been challenged for plagiarism or collusion. One possible explanation for the legitimacy of text reuse in non-academic contexts might be the fact that practitioners often have to replicate text, and it would be impractical to cite their sources in technical documents – in which case refutations of plagiarism are based on concepts of textual genre and formulaic nature of documents, more than on the determination of prior authorship, the calculation of the volume of textual overlap and sharing of linguistic markers and strategies. However, this permissibility raises ethical and legal issues. On the legal side, professional organisations such as the IEEE counter the claims that textual borrowing is justified when employed by practitioners, in professional and technical contexts. This applies both to cases where the text is borrowed from another author, or from texts previously written by oneself. On the ethical side, it should be asked whether well-trained, highly-renowned, well-paid, high-profile professionals can be allowed lower standards than those imposed on untrained, unskilled and inexperienced students. Moreover, it can be asked whether it is ethical to place the future of an entire country in the hands of a 'one-size-fits-all' type of agreement that has been replicated from and for different countries – with disregard for the country's specificities, and whether it is ethical to approve an application for public funding that, for the most part, reproduces the contents of another, different application, with different aims and objectives, based on its "unique quality".

On balance of probabilities, it can therefore be argued that, more than being dependent on 'genre' and 'discipline', the legitimacy of institutional plagiarism such as the one detected in the applications for public funding and in the MoUs is one of expectations of truthfulness. As argued by Eggington (2008), in human communication Grice's (1975) maxim of quality (truthfulness) takes precedence over the maxims of quantity, relation and manner, thus creating a "truth bias" (Eggington, 2008: 256). This is underpinned by a default assumption of truth implying that we do not deceive others, and in return are not deceived by them – even more so when the 'other' is a renowned and unchallenged person or organisation. Finally, if we consider, like Garner (2009: 731) that "knowingly misrepresent[ing] the

truth or conceal[ing] a material fact for one's own advantage" is fraud, then the linguistic integrity of these documents should be challenged.

5. Conclusions

Although plagiarism is considered, by force of ethical and legal infringement, a fraudulent behaviour, some types of plagiarism tend to be more permissible and even legitimated than others; recent events suggest that plagiarism in the academy is more readily punished than elsewhere, but it is not only in the academy that plagiarism is subject to accusations of fraud, as cases of news plagiarism demonstrate. However, some non-academic texts are apparently beyond suspicion or even legitimated, especially when produced by renowned, powerful organisations. These texts not only tend not to be investigated, but more importantly they are not even challenged as plagiarism, often on the grounds that their textual genre does not require the same standards of originality as other genres.

However, the analysis of four sets of texts against three criteria commonly employed in the forensic linguistic analysis of instances of plagiarism (directionality, volume of textual overlap and linguistic markers) demonstrates that there is no reason to consider that these criteria can be used as sufficient evidence in some cases (e.g. academic plagiarism and news plagiarism), but not in others (e.g. applications for public funding and institutional documents such as Memoranda of Understanding). On the contrary, evidence based on qualitative linguistic analysis and even quantitative analysis seems to be stronger in some instances of the applications for funding and the MoUs than in instances of academic texts and news articles. On the other hand, the argument that considerations of plagiarism are dependent on the text genre can be undermined by the claims of international organisations such as the IEEE.

Considering therefore that any genre is subject to accusations of plagiarism, and that a forensic linguistic analysis of the documents in sets 3 and 4 employing tried and tested criteria of plagiarism analysis, this study demonstrated that those documents cannot have been produced independently and, subsequently, are subject to investigations of fraudulent behaviour, owing to the improper reuse of textual material for personal gain.

Although further research is necessary to prove or disprove the social implications of this type of fraud, that research needs to go beyond linguistics, and span into discourse analysis. In other words, the linguistic analysis of these sets of texts has demonstrated that the suspect texts have not been produced independently of other texts, on the contrary; they share an uncommon amount of quantitative and qualitative criteria with other texts, and this in itself constitutes a fraudulent act. However, it is the role of discourse analysis to analyse these texts in more detail and determine, considering the social events with which the texts interact, whether there is a hidden agenda, what their implications are, and whether these represent even more fraudulent behaviour.

Acknowledgements

This work was partially supported by Grant SFRH/BD/47890/2008 FCT-Fundação para a Ciência e Tecnologia, Portugal, co-financed by POPH/FSE.

References

Angèlil-Carter, S. 2000. Stolen language?: plagiarism in writing. Real Language Series. Longman, Harlow.

Bently, L. and Sherman, B. 2009. Intellectual Property Law. Oxford University Press, Oxford.

Coulthard, M. 2004. Author Identification, Idiolect and Linguistic Uniqueness. Applied Linguistics 25: 431-447.

Coulthard, M. and Johnson, A. 2007. An Introduction to Forensic Linguistics: Language in Evidence. Routledge, London.

Eggington, W. G. 2008. Deception and fraud, pp. 249-264. In J. Gibbons and M. T. Turell (eds.). Dimensions of Forensic Linguistics. John Benjamins Publishing Company, Amsterdam.

Eiras, H. and Fortes, G. 2010. Dicionário de Direito Penal e Processo Penal. Quid Juris, Lisboa.

Finnis, J. 1991. Intention and side-effects, pp. 32-64. In R. Frey and C. Morris (eds.). Liability and responsibility: Essays in law and morals. Cambridge University Press, Cambridge.

Garner, B. A. 2009. Black's Law Dictionary. MN: West, St. Paul.

Goldstein, P. 2003. Copyright's highway: from Gutenberg to the celestial jukebox. Stanford University Press, Stanford.

Grice, P. 1975. Logic and conversation. Syntax and Semantics 3: 43-58.

Howard, R. 1995. Plagiarisms, authorships, and the academic death penalty. College English 57: 788-806.

Howard, R. and Robillard, A. (eds.) 2008. Pluralizing Plagiarism: Identities, Contexts, Pedagogies. Portsmouth, Boynton/Cook.

Jameson, D. A. 1993. The Ethics of Plagiarism: How Genre Affects Writers' Use of Source Materials. Business Communication 56: 18-28.

Johnson, A. 1997. Textual kidnapping - a case of plagiarism among three student texts?. The International Journal of Speech, Language and the Law 4: 210-225.

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

Robillard, A. E. and Howard, R. M. 2008. Plagiarisms, pp. 1-7. In R. M. Howard and A. E. Robillard (eds.). Pluralizing Plagiarism: Identities, Contexts, Pedagogies. Portsmouth, Boynton/Cook.

Scollon, R. 1994. As a matter of fact: The changing ideology of authorship and responsibility in discourse. World Englishes 13: 33-46.

Scollon, R. 1995. Plagiarism and ideology: Identity in intercultural discourse. Language in Society 24: 1-28.

Turell, M. T. 2004. 'Textual kidnapping revisited: the case of plagiarism in literary translation'. The International Journal of Speech, Language and the Law 11: 1-26.

Turell, M. T. 2008. Plagiarism, pp. 265-299. In J. Gibbons and M. T. Turell (eds.). Dimensions of Forensic Linguistics, John Benjamins Publishing Company, Amsterdam.

Woolls, D. 2003. Better tools for the trade and how to use them. The International Journal of Speech, Language and the Law 10: 102-112.

Woolls, D. and Coulthard, M. 1998. Tools for the Trade .The International Journal of Speech, Language and the Law 5: 33-57.