The relevance of context in plagiarism detection: The case of a professional legal genre

Victoria Guillén Nieto
Universidad de Alicante (Spain)
victoria.guillen@ua.es

Abstract

This paper focuses on plagiarism detection in a professional legal genre: the lawsuit (Alcaraz Varó & Hughes, 2002; Cabré, 2003: 163-199; Ruiz-Garrido, Palmer-Silveira & Fortanet-Gómez, 2010; Nagy, 2014: 261-273). Its main aim is to analyse the major approaches to extrinsic plagiarism detection, namely computer-based approaches and language-based approaches, either form-only-based or integrated. Our discussion attempts to answer five questions: (a) What type of language evidence can be obtained through an extrinsic plagiarism detection tool? (b) What type of language evidence can a text-only based approach to plagiarism provide? (c) What type of language evidence can an integrated approach to plagiarism provide? (d) What are the strengths and weaknesses of each of these approaches? And (d) how relevant can the language evidence provided by each approach be for the legal decision? The analysis is grounded in an exemplary case study on plagiarism between lawyers that was tried in a high court of justice in Spain (Judgment 107/2017, 2nd March 2017). Findings from this paper confirm that plagiarism detection is a complex, multilayered task going beyond and above the discovery of copied text of an earlier original work into another, and show the relevance of context in discerning between real cases of plagiarism and those that are not.

Keywords: copyright infringement, extrinsic plagiarism detection, professional legal genres.

Resumen

La relevancia del contexto en la detección de plagio: el caso de un género profesional jurídico

Este artículo aborda la detección de plagio en un género profesional jurídico: la demanda (Alcaraz Varó & Hughes, 2002; Cabré, 2003: 163-199; Ruiz-Garrido, Palmer-Silveira & Fortanet-Gómez, 2010; Nagy, 2014: 261-273). Su objetivo es analizar los principales métodos para la detección de plagio externo, a saber, métodos asistidos por herramientas informáticas y métodos basados en el análisis lingüístico, bien de naturaleza formalista o bien de naturaleza integrada, de los textos sospechosos. Las cinco preguntas a las que se intenta dar respuesta son: (a) ¿Qué tipo de evidencia se puede obtener con una herramienta diseñada para la detección de plagio externo? (b) ¿Qué tipo de evidencia se puede obtener con el análisis lingüístico de los textos sospechosos? (c) ¿Qué tipo de evidencia se puede obtener con el análisis integrado de los textos sospechosos? (d) ¿Cuáles son las fortalezas y las debilidades de cada uno de estos métodos? Y (e) ¿qué importancia puede tener la evidencia obtenida con cada método en la decisión judicial? El estudio se basa en un caso ejemplar sobre plagio que fue juzgado en un tribunal superior de justicia en España (Sentencia 107/2017, 2 de marzo de 2017). Los resultados de esta investigación confirman que la detección de plagio es una tarea compleja y de múltiples niveles que va mucho más lejos de la mera identificación de texto copiado de una obra original anterior en otra nueva, y demuestran la relevancia del contexto para discernir entre los casos de plagio que son reales y los que no lo son.

Palabras clave: infracción de derechos de autor, detección de plagio externo, géneros profesionales jurídicos.

1. Introduction

“Intertextuality” (Chaski, 2013: 333) is a specialised area of forensic linguistics in which the expert linguist analyses suspicious textual similarity in trade mark conflicts (Guillén-Nieto, 2011: 63-83; Shuy, 2012: 449-462) and in cases involving copyright plagiarism (Woolls, 2003: 102-112; 2012: 517-529; Turrell, 2004: 1-26; 2005: 275-298; 2008: 265-299; Shuy, 2008; 2012: 449-462; Coulthard et al., 2010: 523-538; Butters, 2012: 463-477). In the past, plagiarism detection was the exclusive domain of well-informed human readers. Their job was to discover if a questioned document had unacknowledged borrowings from one or more reference texts, and to assemble a case from the passages found in common between the works compared. Since the 1990s, plagiarism detection has been mainly software-assisted with the result that the task can be performed automatically and much faster than if it were performed manually (Lukashenko, Graudina & Grundspenkis, 2007: 1-6; Woolls, 2010: 576-590; Hage, Rademaker & van Vugt, 2010: 1-26). This paper focuses on plagiarism detection in a professional legal genre: the lawsuit

(Alcaraz Varó & Hughes, 2002; Cabré, 2003: 163-199; Ruiz-Garrido, Palmer-Silveira & Fortanet-Gómez, 2010; Nagy, 2014: 261-273). It aims at analysing the major approaches to plagiarism detection, namely computer-based approaches and language-based approaches, either text-only based or integrated.[1] We attempt to answer five questions: (a) What type of language evidence can be obtained through an extrinsic plagiarism detection tool? (b) What type of language evidence can a text-only based approach to plagiarism provide? (c) What type of language evidence can an integrated approach to plagiarism provide? (d) What are the strengths and weaknesses of each of these approaches? And (e) how relevant can the language evidence provided by each approach be for the legal decision?

The analysis is grounded in an exemplary case study on plagiarism between lawyers that was tried in a high court of justice in Spain (Judgment 107/2017, 2nd March 2017). For purposes of analysis, different tools were employed in the investigation, namely an extrinsic plagiarism detection system, specifically *CopyCatch Gold* v2 (Woolls, 2002), and language-based approaches, either form-only-based or integrated.

2. Plagiarism and copyright infringement

The *Merriam-Webster online dictionary* defines plagiarism in the following terms:

> (a) To copy and pass off (the expression of ideas or words of another) as one's own: use (another's work) without crediting the source. And (b) to present as new and original an idea or work derived from an existing source.

From this definition, it can be inferred that plagiarism is, basically, the act of copying another author's original work without attribution along with the subsequent deception involved in the same act. Green (2002: 205-206) refers to plagiarism as a violation of the European Doctrine of Moral Rights. This consists of three basic parts: (a) the right of integrity, (b) the right of disclosure, and (c) the right of attribution. Of these, the right of attribution is probably the most relevant in the context of plagiarism: "An author or artist has the right both to be identified as the author of any work that she has created and to prevent the use of her name as the author of a work she did not create" (Green, 2002: 206). By contrast, in the US, the doctrine of moral rights is much more limited to the degree that it has no application to literary works.

As we have already mentioned, apart from being considered academic dishonesty, violation of the workplace honour code, and a breach of professional ethics, plagiarism can also become a legal issue, if it involves "copyright infringement". In both common law and civil law systems, copyright is a set of exclusive rights granted to the creator of an original work. The rights covered include: (a) to copy the work, (b) to issue copies of the work to the public, (c) to rent or lend the work to the public, (d) to perform, show or play the work in public, (e) to communicate the work to the public, and (f) to make an adaptation of the work. These exclusive rights are not absolute, but limited in time, e.g. 70 years after the death of the author is the set time for literary works, and by restrictions of copyright law. Furthermore, the notion of "fair use" or "fair dealing" is an important limitation upon what constitutes copyright violation. It refers to acts that are permissible without infringing copyrighted material, such as for example private research, copies for educational purposes, news reporting, caricature, parody, and pastiche. Plagiarism violates the above-mentioned exclusive rights granted to the creators of original works by copying the work without permission, and/or by distributing it. However, it is important to note that not all plagiarisms are copyright infringements. For instance, one can plagiarise from sources that are in the public domain and consequently out of copyright, or one can plagiarise short, or even long, uncreative passages without attribution, but this is unlikely to involve a copyright infringement. Green summarises the essential difference between plagiarism and copyright infringement in these words: "Copyright Law protects a primarily economic interest that a copyright owner has in her work […] whereas the rule against plagiarism protects a personal, or moral, interest" (Green, 2002: 202). In both common law and civil law jurisdictions, copyright infringement is actionable by the copyright owner, and can be punished in a court of justice for prejudices caused by copyright infringement and violation of moral rights. It should be noted that copyright infringement is only subject to criminal prosecution in extreme cases, specifically if the plagiarism is intended for purposes of commercial advantage or private financial gain, namely "piracy" and "counterfeiting".

3. Criteria regarding the types of use of another author's work

Lawyer Nettel Díaz analyses the criteria provided by Maurel-Indart (as cited in Nettel Díaz, 2013: 135-152) regarding the types of use of another author's work. These are:

1) The borrowing is either direct (verbatim) or indirect (modified).
2) The borrowing involves either a part or the whole of a reference text.
3) The borrowing is either intended or unintended.
4) The borrowing is either evident or hidden.

Combined in various ways, these four criteria may give rise to different types of borrowings that do not necessarily have to involve plagiarism, as depicted in Table 1.

	Borrowing				
	Total		***Partial***		
	Evident	Hidden	Hidden		Evident
	Intended	*Intended*	*Intended*	*Unintended*	*Intended*
DIRECT	Topic Analogy		Collage	Coincidental match	
			Quotation	Common ground	
		Topic analogy	Similarity of ideas	Similarity of ideas	Similarity of ideas
INDIRECT	Adaptation	Adaptation	Literary School Literary movement	Coincidental match	Literary School Literary movement
	Translation		Pastiche		Pastiche
	Summary	Summary	Reminiscence	Reminiscence	Parody
	Analysis				Analysis
			Allusion		Allusion

Table 1. Criteria regarding the types of use of another author's work (translated from Nettel Díaz 2013: 145).

The creation of new academic, scientific, professional, literary and artistic knowledge draws, to a greater or lesser extent, on the content of previous works. Therefore, one must always presume a certain amount of similarity between texts that are of the same type and topic related. In Table 1 above, the cases that are likely to involve real plagiarism, either relating to copyright infringement or not, are shaded. These cases concern two criteria: (1) the borrowing is intended and (2) the borrowing is hidden. Consequently, any borrowing used without permission, not giving credit to the source, intended, and hidden may be seen as pieces of language evidence pointing to plagiarism. Lawyer Nettel Díaz (2013: 150-151) argues that it is also essential to discern between the "fair use" or "fair deal" and the "unfair use" of the borrowing. Furthermore, lawyer Aznar Auzmendi has attracted attention to what seems to be a concluding legal criterion for Spanish courts of justice in relation to copyright infringement plagiarism (Supreme Court, Judgment of February 20th, 1992 (RJ 1992/1329)):

Although in common language we often feel tempted to call plagiarism any coincidence between two works that strikes us, the truth is that there is no such plagiarism when the works compared are sufficiently different and clearly distinguishable, even if they have points of exposure in common.[2]

From this excerpt, it can be inferred that, according to Spanish legal practitioners, a conclusive criterion for determining that a plagiarism does not imply copyright infringement should be the fact that the works under comparison are "sufficiently different" and "clearly distinguishable", especially in terms of their genres, their communicative purposes, their superstructure, and their macrostructure. For example, it could be argued that a scientific paper and a teaching project are "sufficiently different" and "clearly distinguishable"; or that an informative article and a Massive Open Online Course (MOOC) are also "sufficiently different" and "clearly distinguishable". In conclusion, it has been shown that the task of plagiarism detection, whether including copyright infringement or not, is a complex, multilayered task going beyond and above the discovery of copied text of an earlier original work (the reference text) into a new one (the questioned text). For this reason, it is vital for expert linguists to be able to answer a number of essential questions when reporting on plagiarism. The following list of ten questions is proposed as useful guidance for accomplishing such a task:

1) Is the reference text a copyrighted work?
2) Does the reference text contain original ideas?
3) Has a substantial amount of original text been copied?
4) Could the borrowing fit in the category of "fair use" or "fair deal"?
5) Did the suspect have permission from the author of the reference text to copy original ideas or a substantial amount of text into the questioned text?
6) Does the borrowing in the questioned text embrace the whole or only a part of the reference text?
7) Is the borrowing direct (verbatim) or indirect (modified)?
8) Is the borrowing evident or hidden?
9) Is the borrowing intended or unintended?
10) Are the works under comparison sufficiently different and distinguishable?

4. Case study: Plagiarism between lawyers

The case study under discussion concerns plagiarism between lawyers and was tried in a supreme court of justice in Spain (Judgment 107/2017, 2nd March 2017), consequently the dispute must be understood and interpreted within the legal framework of a Roman civil law jurisdiction. Briefly, the case can be summarised as follows: The plaintiff, a junior lawyer, sued a senior lawyer in a legal firm seeking declaration of copyright infringement of an appeal for reversal that she had single authored, and requesting compensation for moral damage. The pair of suspicious texts in the case, therefore, were: (1) the reference text: an appeal for reversal (4,079 words) dated 26th May 2009, and (2) the questioned text: a lawsuit (18,768 words) dated 18th September 2009. In the next two sections, we will look at the type of language evidence three major approaches to plagiarism detection can provide, analysing their strengths and weaknesses, in addition to discussing their relevance for the legal decision.

4.1. Computer-based approach

The computer-based approach can be applied to two major frameworks: (a) "intrinsic plagiarism detection", and (b) "extrinsic plagiarism detection" which we will briefly discuss in the next paragraphs, before moving on to analyse the pair of suspicious texts in the case under discussion.

"Intrinsic plagiarism detection" and "authorship verification" are, according to Stein, Lipka and Prettenhofer (2011: 63-82), "two sides of the same coin". In both cases one is given a single document (there is no reference corpus) and one must deal with the problem of finding the suspicious sections by identifying irregularities or inconsistencies in the author's writing style within the same document, namely stylometric features and average word frequencies (Meyer zu Eissen & Stein, 2006: 565-569) or character-gram profiles (Stamatatos, 2009; Dam, 2013).

By contrast, in "extrinsic plagiarism detection", a reference text or a corpus is given and the task is to identify, using algorithms that compare the questioned document against potential source documents, the presence of identical words, sequences of words, sentences, and paragraphs in common between the suspicious texts (Lukashenko, Graudina & Grundspenkis, 2007: 1-6; Potthast et al., 2010). Some well-known examples of extrinsic plagiarism detection tools are *Turnitin* (the questioned text is compared against potential reference documents available in databases and on the Internet) and *CopyCatch Gold* v2 (a questioned text is compared against a reference text) (Woolls, 2002). Although extrinsic plagiarism detection performs well in identifying copied, or even slightly modified, material, it assumes a closed world:

> […] a reference collection must be given against which a plagiarised document can be compared. This raises the question whether plagiarised passages within a document can be detected automatically if no reference is given, e.g. if the plagiarised passages stem from a book that is not available in digital form (Meyer zu Eissen & Stein, 2006: 565).

From the above quote, some important ideas emerge. Firstly, computer-based extrinsic plagiarism detection may only serve to identify in a questioned text evident, verbatim copy, or slightly modified copy of the whole or of a part of a reference text available in a database or on the Internet. Secondly, computer-based plagiarism detection is difficult once there is a departure from full copy-pasted text of an earlier work, due to the technical difficulties involved in creating suitable software modelling language use. For this reason, extrinsic plagiarism detection software may well not be able to detect lexical and grammatical transformation of a text, namely lexical substitution, paraphrasing and nominalisation. In this context, Vector Space Models (VSMs) seem to provide interesting alternatives because they perform well on tasks that involve measuring the similarity of meaning between words, phrases, and documents, as shown in the works by Turney and Pantel (2010: 141-188), and Mikolov et al. (2013). Furthermore, Latent Semantic Analysis (LSA) and Topic Modelling (LDA) have proved to be successful VSMs in the analysis of lexical similarity and lexical collocations (Osman et al., 2012: 1493-1502).

As mentioned before, the extrinsic plagiarism detection tool employed in the case under discussion is *CopyCatch Gold* v2 (Woolls, 2002). This tool has been selected because, apart from having been originally designed to detect likely plagiarism and collusion between students, it is seen as software for forensic text analysis (Guillén-Nieto et al., 2008: 9-12), as shown in the works by Johnson (1997: 210-225), Turell (2004: 1-26; 2008: 278-280), and Coulthard et al. (2010: 523-538). Before analysing the suspicious pair of texts, the linguist using *CopyCatch Gold* v2 must perform certain tasks. These are:

1) Limit the number of pairs on show by setting the score for similarity. This is typically set up at 50% for works that are topic-related, and at 70% for derivative works e.g. translations.
2) Add to the "Stop list", which contains a list of functional words in the language of the suspicious pair of texts, specific content words that are likely to be shared between the texts under examination.
3) Load the "Stop list" by click-ing on the "Language" button.

4) Select the comparison files, namely the appeal for reversal (the reference text) and the lawsuit (the questioned text).
5) Search for matches between the reference text and the questioned text at both a word level and a sentence level of linguistic analysis by clicking on the "CopyCatch" button.

When searching for similarities between the words compared, the following tasks can be done automatically:

1) Calculate the similarity threshold level between the reference text and the questioned text, namely the score above which one can state that a substantial amount of text has been borrowed.
2) Identify and measure the vocabulary and sentences shared more than once between the reference text and the questioned text.
3) Detect and measure the vocabulary and sentences that are present only once in each separate text and shared only once between them (*hapax legomena*).
4) Discover and measure the vocabulary that is only in one text and not in the other.
5) Get lists of content word and function word frequencies.
6) Obtain a percentage analysis.

We will now move on to present the results yielded by *CopyCatch Gold* v2 in the case under examination: Firstly, the "Similarity Threshold level" yielded a score of 64%, which points to a substantial amount of copied text from the appeal for reversal (the reference text) (4,079 words) into the lawsuit (the questioned text) (18,768 words). Secondly, 62% of the words of the reference text are present in the questioned text. Thirdly, 81% of the sentences in the reference text are also present in the questioned text. Finally, a very high number of *hapax legomena* (328 content words) are found to be only once in the reference text and suspiciously only once in the questioned text too.

Although *Copycatch Gold* v2 does not report on the "originality" of the words or expressions used in the appeal for reversal, the list of content word frequencies the tool provides enables the analyst to seek for "original" words or expressions that can be further analysed with the aid of databases and linguistic tools. *CopyCatch Gold* v2 also assists in the task of finding the excerpts in common between the works compared automatically, preparing the ground for further linguistic qualitative analysis. Additionally, *Copycatch Gold* v2 provides quantitative data, at both a word level and a sentence level of linguistic analysis. These data, which result, from a computer-based approach to plagiarism detection, may serve to answer four of the ten relevant questions earlier proposed as guidelines for expert linguists: a substantial amount (>50%) of text seems to have been borrowed from the reference text into the questioned text; the borrowing embraces, in effect, the whole of the reference text; it is mostly a direct (verbatim) copy, and, for this reason, evident.

4.2. Language-based approaches

In this section, we analyse the two main language-based approaches to extrinsic plagiarism detection, namely a form-only-based approach, and an integrated approach.

4.2.1. Form-only-based approach

Typically, a form-only-based approach to extrinsic plagiarism detection consists in identifying matches between the reference text and the questioned text at a word, sentence, and discourse level of linguistic analysis. Communication in specific domains is governed by either academic or professional criteria, and it is characterised by both language-based features and text-based-features. Consequently, the task of detecting plagiarism in a professional legal genre is not an easy one, because the works under comparison are likely to share the same specialist grammatical, lexical, and discursive features. In the case open to discussion, the works compared are, as above mentioned, an appeal for reversal and a lawsuit. Both of them can be classified as text-types of the professional legal genre. Whereas the appeal for reversal can be lodged against acts that exhaust the administrative procedure, the lawsuit is the procedural act by which judicial proceedings are initiated. Both the appeal for reversal and the lawsuit share the specialist language features listed below, among others:

(1) Their communicative goal is to demand the reconsideration of a decision that is thought to be unfair and does not meet the interests of the claimant.
(2) They have a conventional superstructure, namely *Encabezado* (Header), *Hechos* (Facts), *Fundamentos de Derecho* (Legal reasons for decision) and *Suplico* (Plea for demand).
(3) They are expressed in formal register and have a solemn tone.
(4) They are encoded in expository discourse.

(5) They are expressed through legal vocabulary, e.g. formulaic expressions, latinisms, semi-technical words, specific terminology, lexical metaphors, and lexical collocations (Alcaraz Varó & Hughes, 2002: 31-78).

(6) Their syntax is convoluted and long winded, e.g. long nominal groups, abundant gerund and past participle clauses, explicative clauses, archaic verb forms, such as the imperfect future of the subjunctive mode (e.g. *no cumpliere, atendiere*), and extremely long sentences due to the extensive use of hypotaxis and parataxis (Alcaraz Varó & Hughes, 2002: 103-136).

If the appeal for reversal and the lawsuit share the specialist language features above referred, then the question is: What type of language evidence may be indicative of plagiarism in a professional legal genre? As Nagy (2014: 266) has explained, specialist texts are subject to language variation, namely individual variation (idiolect) and regional variation (dialect). Whenever the reference text is rich in idiolectal and/or in regional dialectal features that are specific to its author, and these are also present in the questioned text, one may argue that the texts may not have been written independently from each other. Therefore, the linguistic categories of idiolect and dialect can be useful tools for finding pieces of language evidence hinting at plagiarism.

Idiolect is commonly defined as an individual's unique variety and/or use of language from the level of the phoneme to the level of discourse. At a discourse level, the idiolect of an author can be made evident through other linguistic categories such as "voice" and "stance" (Hyland & Sancho Guinda, 2012). "Voice" involves both individual and social dimensions. On the one hand, the "individualised voice" is, according to Tardy, "[…] a writer's unique and recognizable imprint, associated with authenticity, resonance, authoritativeness, and authorial presence within a text" (Tardy, 2012: 37). On the other hand, from a social perspective, "[…] voice relates to self-representation and authorial presence […] but constructed by the social worlds that the author works within" (Tardy, 2012: 39). The second category indicative of individual variation in a text is, as mentioned before, "stance". Gray and Biber (2012: 15-33) have summarised the different conceptions of "stance" in two: "evidentiality markers" and "affect markers" that are linguistically encoded in grammatical and lexical categories. Whereas "evidentiality markers" express evaluations of knowledge contained in the propositions, "affect markers" express personal feelings, emotions and attitudes associated with persons and/or situations. Table 2 illustrates copy-pasted fragments of the section "Facts" of the appeal for reversal into the lawsuit.

Appeal for reversal	Lawsuit
X. Por lo tanto he debido prepararme durante un año completo, pues es imposible aprobar dicha prueba de otra manera. De hecho son muy pocas las personas que logran superarla ya que el examen de conjunto requiere de un manejo profundizado del derecho español en todas las áreas.	XIII. Por lo tanto he debido prepararme durante un año completo, pues es imposible aprobar dicha prueba de otra manera. De hecho son muy pocas las personas que logran superarla ya que el examen de conjunto requiere de un manejo profundizado del derecho español en todas las áreas.
XIII. Homologué mi título de licenciada con la intención de continuar en España mis estudios de doctorado obteniendo una beca de investigación de Ayudas para la formación de Personal Investigador. En las bases de las sucesivas convocatorias nada dice sobre que se considerará la fecha del título de origen […].	XX.. Homologué mi título de licenciada con la intención de continuar en España mis estudios de doctorado obteniendo una beca de investigación de Ayudas para la formación de Personal Investigador. En las bases de las sucesivas convocatorias nada dice sobre que se considerará la fecha del título de origen […].

Table 2. Examples of copy-pasted text including idiolectal features of the author of the appeal for reversal.

The two copy-pasted excerpts shown in Table 2 belong to the exposition of facts supporting the appeal for reversal. Although the appeal for reversal and the lawsuit, as is the case with other legal text-types, are encoded in template-like writing, which is typically characterised by abundant legal formulas and set phrases, the exposition of facts is the section in which spontaneous writing showing features of the idiolect and regional dialect of the author is more likely to be found. The authorial presence of the junior lawyer is linguistically expressed through different grammatical and lexical categories. For example, verbs are conjugated in the first person singular (*He debido prepararme*/I had to study, *Homologué*/I had my aca-demic certificates recognised); possessive adjectives are in the first person (*mi título*/my academic degree*, mis estudios de doctorado*/my PhD studies, *mi título de licenciada*/my academic certificate); nouns are in their feminine form (*licenciada*/graduate). In addition, the stance of the author of the appeal for reversal is linguistically encoded in epistemic markers. These indicate that the information the author provides is based on her knowledge and direct experience ("I know that…"). The epistemic markers are linguistically realised through grammatical and lexical categories such as the present of indicative. For example, *es imposible aprobar*/it is impossible to pass, *el examen de conjunto requiere de un manejo profundizado del derecho español en todas las áreas*/the examination requires a thorough understanding of Spanish law in all areas; and the phrase *De hecho*/In fact. For example, *De hecho son muy pocas personas las que logran superarla* […]/In fact, very few people manage to pass it [...]. Moreover, the two excerpts of the appeal for reversal copied into the lawsuit are very rich in "boosters" (Hyland, 2000: 179-197) whose function is to demonstrate the confidence of the author as well as to convey a sense of self-assurance and certainty ("I know that…"). Boosters are linguistically realised through nouns, adjectives, verbs and adverbs. Some illustrative examples of boosters shared by the texts under comparison are listed below:

1) *un año completo*/one full year [adjective]
2) *es imposible aprobar dicha prueba*/it is impossible to pass such test [adjective]
3) *son muy pocas las personas*/there are very few people [adverb + adjective]
4) *que logran superarla*/that manage to pass it [verb]
5) *el examen requiere de un manejo profundizado del derecho español*/the examination requires a thorough understanding of Spanish law [adjective]
6) *en todas las áreas*/in all areas [adjective]
7) *nada dice sobre que se considerará la fecha del título de origen*/ nothing is said about what should be considered the date of the original academic title [indefinite pronoun]

Dialect is a regional variety of language that signals where a person comes from; it is distinguished by features of pronunciation, vocabulary, and grammar from other regional varieties of language. In the case under discussion it was found that the appeal for reversal contains words and expressions that are not typical of Peninsular Spanish; these words and expressions are coincidentally present in the lawsuit presumably written in Peninsular Spanish. Some of these instances were looked up in the *Diccionario panhispánico de dudas* (2005) of the Spanish Royal Academy and found to be common anglicisms in American Spanish. For instance, let us take the anglicism: *las becas para las que he aplicado*/the scholarships I have applied for (*las becas que he solicitado* is the correct expression in Peninsular Spanish). Other uncommon words in Peninsular Spanish that are found in the appeal for reversal, which are also present in the lawsuit, are: *Defensoría del pueblo* and *puntaje*. After looking them up in the database *CORPES XXI* of the Spanish Royal Academy, one can confirm that these are dialectal markers of Rioplatense Spanish, namely the Spanish dialect spoken in the River Plate region. More specifically, 1,023 instances of the usage of *Defensoría del pueblo* (*Defensor del pueblo* in Peninsular Spanish) were found in a total of 527 documents. Whereas the highest absolute and relative frequencies relate to the Andean Region (398 documents/19.86%) and the River Plate region (94 documents/2.60%), the lowest absolute and relative frequencies (5 documents/0.05%) point toward Spain. Similarly, 1,493 instances of the usage of *Puntaje* (*Puntuación* in Peninsular Spanish) were found in a total of

703 documents. Once more, the highest absolute and relative frequencies relate to the Andean region (234 documents/11.67%) and the River Plate region (222 documents/6.14%), and the lowest absolute and relative frequencies refer to Spain (10 documents/0.11%).

At a discourse level, apart from the idiolectal and dialectal features that were copy-pasted in the lawsuit, there are other instances hinting at plagiarism. For instance, the thematic sequence in the section *Hechos* (Facts) of the appeal for reversal is organised in chronological order, namely the application for a scholarship is presented as a first fact and the rejection of the application by the local administration as a second fact. Although the lawsuit exhibits a verbatim copy of this section of the appeal for reversal, it can be clearly seen that the facts are presented in a different order. As a result of the act of copy-and pasting, one can see that the thematic sequence of the text of the lawsuit is illogical. In other words, the rejection of the application by the local administration is presented as a first fact (when it should be the second fact), the file of the appeal for reversal is presented as a second fact (when it should be the third fact), and the application for the scholarship is presented as a third fact (when it should be the first fact). Similarly, the formal expression and content in the section *Fundamentos de Derecho* (Legal reasons for decision) of the questioned text are a direct (verbatim) copy of the reference text. Ostensibly, the lawsuit even includes the misprints and grammatical errors found in the appeal for reversal. However, some minor modifications are observed in the questioned text. These include a different thematic sequence in the exposition of the doctrine, addition of information, namely applicable law, legal citations and explanatory clauses. Yet again, the content and formal expression in the section *Suplico* (Plea for demand) of the questioned text is a direct copy of that of the reference text, except for the supplementary information and extra pleas that have been added at the end of the text.

Moreover, the questioned text contains errors resulting from the act of copy-and-pasting from the appeal for reversal into the lawsuit. These include, for instance, misprints, numbering errors, and, as shown in Table 3, the date of the lawsuit mistakenly matches that of the appeal for reversal filed four months earlier. The correct date of the lawsuit should be 18th September 2009, rather than 26th May 2009.

Appeal for reversal	Lawsuit
En Salamanca, a 26 de Mayo de 2008.	Es justicia que pido en Salamanca, a 26 de Mayo de 2008.

Table 3. Error in the date of the lawsuit.

From the above discussion it can be concluded that the language evidence obtained with a text-only-based approach to plagiarism detection is mainly qualitative and may be useful to answer four of the ten relevant questions that were earlier proposed as guidelines for expert linguists: (1) The borrowing is a direct copy of the reference text. (2) The borrowing is evident. (3) There is no language evidence to support the idea that the formal expression and content of the appeal for reversal are "original". However, both idiolectal and regional dialectal features of the author of the appeal for reversal are also present in the lawsuit presumably written in Peninsular Spanish. (4) The suspicious pair of texts are topic-related and share many specialist language features at a word level (e.g. legal lexicon), sentence level (e.g. convoluted syntax and hypotaxis), and discourse level (e.g. predetermined superstructure) of linguistic analysis. Then it cannot be stated that the works compared are "sufficiently different" and "distinguishable".

4.2.2. An integrated approach to plagiarism detection

As its name suggests, an integrated approach to plagiarism detection integrates the analysis of linguistic and extralinguistic data. In such a broader pragmatic perspective, the expert linguist examines the way the works compared are affected by (and affect) the social environment in which they were produced. In other words, in an integrated approach to plagiarism detection, the focus of analysis is shifted from the analysis of the formal aspects of the works compared to the analysis of the appropriateness of language use and discourse in the communicative situation in which the case is embedded. In what follows we will refer to the communicative situation framing the case in dispute. van Dijk has explained communicative situations in terms of "context models":

> Such models consist of simple a schema of culturally variable categories used by language users in the interpretation and representation of the communicative situation, such as spatiotemporal Setting, Participants and their different identities and roles, ongoing social Action, Goals and current Knowledge (van Dijk, 2008: 2).

a) Setting

Spatiotemporal information defines language users' ongoing awareness of the space and time in which the communicative situation takes place, and controls the properties of discourse. In the case under discussion, the setting refers to a professional legal context in Spain between 2009 and 2011.

b) Participants

The contextual representation of participants, as well as their social identities, roles and relationships control many text properties. Categories of gender, age, ethnicity and social status are socially construed and control participation structure in discourse. The participants in the case are a junior female lawyer from Argentina and a senior lawyer from Spain. The former had a collaboration contract for professional legal services with the senior lawyer's legal firm. Therefore, it is clear that there is an asymmetric social relation of power between both lawyers in the communicative situation.

c) Action and goals

In his seminal work *How to Do Things with Words* (1975[1962]), Austin laid the pragmatic foundation of speech act theory, that is, the way words can be used not only to present information but also to carry communicative actions of various types, namely social, political, legal, and cultural. These communicative actions are driven by the participants' intentions and can only be defined in terms of the goals they want to reach (van Dijk, 2008: 8). The actions performed by the participants in our case are of a professional legal type. More specifically, these are, on the one hand, to request a higher instance to reverse an administrative decision that is deemed unfair, and, on the other hand, to sue for copyright infringement. These legal actions control, and are controlled by, the specific discourse structures of the professional legal genre. More specifically, the participants in the case use two closely-related text-types of the professional legal genre as instruments of social interaction in order to achieve one communicative goal: to reverse the rejection of an application for a scholarship. These text-types are, as referred above, an appeal for reversal and a lawsuit. While the former puts an end to the administrative proceeding, the latter initiates the legal proceeding in the court of justice. On 26th May 2009, before the junior lawyer had started working as an intern in the legal firm, she had filed an appeal for reversal in her own defence against the local administration's rejection of the scholarship she had applied for. Later, when she was hired by the legal firm, she decided to take legal action against the local administration, and for this reason granted power of attorney to the senior lawyer to act for her in the legal proceeding. It is important to note that the senior lawyer used the appeal for reversal written by the junior lawyer to work on the preparation of what would be the final lawsuit. This was co-signed by the senior lawyer and the junior lawyer. Meanwhile the junior lawyer registered the appeal for reversal as a creation of scientific doctrine in the

Intellectual Property Registry. Subsequently, in 2011, the junior lawyer sued the senior lawyer seeking declaration of copyright infringement of the appeal for reversal and requesting economic compensation for copyright infringement and moral damage.

d) Knowledge

Communicative actions require that participants have or lack specific knowledge, manage knowledge such as presupposition, implications and implicatures, and share knowledge such as the beliefs that are presupposed in public discourse, in our case professional legal discourse. Accordingly, it is important for the expert linguist to have access to such background knowledge that is crucial for the right understanding and interpretation of the case. In the paragraphs that follow, we will refer to some relevant issues that must be clarified before giving an expert opinion.

Firstly, an essential question in the case is whether or not the professional writings of lawyers can be protected by copyright. Article 10.1 of Spanish Intellectual Property Law establishes that all original literary, artistic or scientific creations expressed by any means or support, tangible or intangible, currently known or to be invented in the future, are the object of intellectual property, and includes a list of intellectual creations that can be considered original works among which professional legal genres are not included, but in letter a) reference is made to the writings of forensic reports, in which the professional writings of lawyers can fit. Then, the appeal for reversal can be considered a copyrighted work.

Secondly, a controversial issue was whether or not the senior lawyer had permission from the junior lawyer to use and modify the text of the appeal for reversal which she had single authored. As Love (2002: 40-50) has explained, in forensic linguistic analysis one must discern between different kinds of authorship: a precursory author is the one that provides the source or ideas; an executive author is the one that writes the text; a declarative author is the one that features as author; and the revisionary author is the one that edits and introduces amendments. To the best of our knowledge, the author of the appeal for reversal fulfilled all authorial functions. On the other hand, as mentioned above, the junior lawyer had granted power of attorney to the senior lawyer to act for her in the legal proceeding. This act implies that the senior lawyer was also given permission to borrow text from the appeal for reversal into the lawsuit as well as to modify it, in order to accomplish the communicative goal of the new text. Although the lawsuit was co-signed by the senior lawyer and the junior lawyer, due to the hierarchical relationship existing between them, their authorial functions were different. Whereas the senior lawyer performed the functions of executive author, declarative author featuring in the first place as manager of the legal proceedings, and revisionary author, the junior lawyer performed the functions of precursory author and declarative author featuring in the second place as assistant. Then, rather than considering the lawsuit a collaborative work (Spanish Intellectual Property Law, art. 7), this should be seen as a composite work (Spanish Intellectual Property Law, art. 9). More specifically, a collaborative work is the unitary result of the collaboration of several authors in the writing process, which requires an action at the same level between them, namely without a hierarchical or subordinate relationship. By contrast, a composite work (Spanish Intellectual Property Law, art. 9) is a new work that incorporates a pre-existing work. The composite work, therefore, implies a transformation of a pre-existing work (Spanish Intellectual Property Law, art. 21) giving rise to a derivative work (Spanish Intellectual Property Law, art. 11) resulting from the process of transformation or modification, over which both the author of the transformed work and the author of the work resulting from the transformation into a derivative work will share rights.

The language evidence obtained with an integrated approach to plagiarism detection is qualitative and is essential to adequately interpret the results obtained in the other two approaches to plagiarism detection, namely the computer-based approach and the form-only-based approach. Furthermore, an integrative approach helps us to give an answer to two crucial questions: (1) the earlier work is copyrighted, and (2) the suspect had legal permission from the author of the appeal for reversal to borrow a substantial amount of text and modify it for purposes of elaborating the lawsuit.

5. Conclusions

It was demonstrated that an extrinsic plagiarism detection tool like *Copycatch Gold* v2 (Woolls, 2002) can be suitable for purposes of discovering and measuring substantial formal similarity between two texts. Besides, it was verified that this tool performs well in detecting direct (verbatim), and therefore evident, copy of words, sequences of words, sentences, and longer passages from an author's earlier work into another author's work. However, attention was drawn to the fact that language-based approaches, either form-only-based or integrated, seem to provide more complete language evidence to sustain a real case of plagiarism. Specifically, owing to the findings obtained by means of an integrated approach, in which linguistic elements and extralinguistic data are combined into the task of plagiarism detection, it was possible to give an answer to the three questions that happened to be the most relevant in the case, according to the legal verdict (Judgment 107/2017, of 2nd March 2017): (1) The appeal for reversal is a copyrighted work. (2) The use of the appeal for reversal made by the senior lawyer was appropriate to the context in which the case must be understood and interpreted, because he had been given power of attorney by the author of the appeal for reversal to represent her in the legal proceeding and consequently, he had permission to borrow and modify the text of the earlier work for the elaboration of the lawsuit. (3) According to Spanish Intellectual Property Law, the lawsuit should be seen as a composite work giving rise to a derivative work, rather than as a collaborative work. In spite of the fact that the language evidence yielded by the computer-based approach and the linguistic form-only-based approach clearly hinted at plagiarism, concluding that the author of the questioned text plagiarised the appeal for reversal may seem naive and inappropriate in the eyes of the court of law.

The integrated approach offered the possibility to interpret the quantitative and qualitative data obtained in the two other approaches, bearing in mind the setting, the participants involved, their actions and goals, and knowledge that are relevant to the communicative situation. On these grounds, the high court of justice that tried the dispute finally ruled that there was neither plagiarism nor copyright infringement in the case (Judgment nº107/2017, of 2nd March 2017).

From the analysis of the exemplary case study, it can be concluded that the expert linguist must be cautious when reporting cases involving plagiarism because, on the one hand, not every single instance of copy-pasted text is plagiarism, and because, on the other hand, not every single instance of plagiarism implies copyright infringement. Since plagiarism detection goes far and beyond the task of identifying copied text from an earlier work into another author's work, it would be beneficial for the profession that expert linguists use an integrated approach, analysing the appropriateness of the questioned text in its context of production. The suggested pragmatic turn may have a positive effect on the relevance and accuracy of the linguistic evidence that the expert linguist can provide to the courts of justice.

Acknowledgements

The author would like to thank the editors and reviewers for careful reading, thoughtful comments and constructive suggestions for this article.

Article history:
Received 8 October 2018
Received in revised form 11 March 2020
Accepted 27 April 2020

References

Alcaraz Varó, E. & B. Hughes (2002). *El español jurídico*. Barcelona: Ariel.

Austin, J.L. (1962). *How to Do Things with Words*. Oxford: OUP.

Aznar Auzmendi, J.M. (2013). "Qué es y qué no es un plagio". URL: <https://www.huffingtonpost.es/jose-maria-aznar-auzmendi/que-es-y-que-no-es-un-plagio_b_3042814.html> [28/02/2020].

Butters, R.R. (2012). "Language and copyright" in P.M. Tiersma & L.M. Solan (eds.), *The Oxford Handbook of Language and Law*, 463-477. New York: OUP.

Cabré, T.M. (2003). "Terminology. Theory, methods and applications". *Terminology* 9(2): 163-199.

Chaski, C. (2013). "Best practices and admissibility of forensic author identification". *Journal of Law and Policy* 21(2): 333-376.

Copyright Law of the United Kingdom: Copyright, Designs and Patents Act 1988. URL: <https://www.legislation.gov.uk/ukpga/1988/48/contents> [29/02/2020].

Copyright Law of the United States: Copyright Law. URL: <https://www.copyright.gov/title17/> [29/02/2020].

Coulthard, M., A. Johnson, K. Kredens & D. Woolls (2010). "Plagiarism. Four forensic linguists' responses to suspected plagiarism" in M. Coulthard & A. Johnson (eds.), *The Routledge Handbook of Forensic Linguistics*, 523-538. London, New York: Routledge.

Dam, M. van (2013). "A basic character n-gram approach to authorship verification". *Notebook for PAN at CLEF 2013.* URL: <http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-vanDam2013.pdf> [29/02/2020].

Gray, B. & D. Biber (2012). "Current conceptions of stance" in K. Hyland & C. Sancho Guinda (eds.), *Stance and Voice in Written Academic Genres*, 15-33. Basingstoke: Palgrave Macmillan.

Green, S. P. (2002). "Plagiarism, norms, and the limits of the theft law: Some observations on the use of criminal sanctions in enforcing intellectual property rights". *Hastings Law Journal* 54(1): 167-242.

Guillén-Nieto, V. (2011). "The linguist as expert witness in the community trademark courts". *ITL International Journal of Applied Linguistics* 162: 63-83.

Guillén Nieto, V., Vargas Sierra, C., Pardiño Juan, M., Martínez Barco, P., & Suárez Cueto, A. (2008). Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies, 8*(1), 1-28. Retrieved from https://revistas.um.es/ijes/article/view/49071

Hage, J., P. Rademaker & N. van Vugt (2010). "A comparison of plagiarism detection tools" in *Technical Report UU-CS-2010-2015*, 1-26. Utrecht: Department of Information and Computing Sciences. Utrecht University.

Hyland, K. & C. Sancho Guinda (2012) (eds.). *Stance and Voice in Written Academic Genres*. Basingstoke: Palgrave Macmillan.

Hyland, K. (2000). "Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts". *Language Awareness* 9(4): 179-197.

Johnson, A. (1997). "Textual kidnapping – a case of plagiarism among three student texts". *International Journal of Speech, Language and the Law* 4(2): 210-225.

Juan, M., M. Amengual & J. Salazar (eds.) (2006). *Lingüística aplicada en la sociedad de la información y la comunicación*. Islas Baleares: Servicio de Publicaciones y de Intercambio Científico. Universidad de las Islas Baleares.

Love, H. (2002). *Attributing Authorship*. Cambridge: CUP.

Lukashenko, R., V. Graudina & J. Grundspenkis (2007). “Computer-based plagiarism detection methods and tools: An overview” in *Proceedings of the 2007 International Conference on Computer Systems and Technologies* 18: 1-6.

Merriam-Webster Online Dictionary. URL: <https://www.merriam-webster.com/> [02/03/2020].

Meyer zu Eissen, S. & B. Stein (2006). “Intrinsic plagiarism detection” in *Lecture Notes in Computer Science* 3936: 565-569.

Mikolov, T., K. Chen, G. Corrado & J. Dean (2013). “Efficient estimation of word representations in vector space”. *Computation and Language* 1-12.

Nagy, I.K. (2014). “English for special purposes: Specialized languages and problems of terminology”. *Acta Universitatis Sapientiae, Philologica* 6(2): 261-273.

Nettel Díaz, A.L. (2013). “Derecho de autor y plagio”. *Alegatos* 83: 135-152.

Osman, A.H., N. Salim, S. Binwahlan, R. Alteeb & A. Abuobieda (2012). “An improved plagiarism detection scheme based on semantic role labelling”. *Applied Soft Computing* 12(5): 1493-1502.

Potthast, M., B. Stein, A. Barrón-Cedeño & P. Rosso (2010). “An evaluation framework for plagiarism detection” in *COLING’10: Proceedings of the 23rd International Conference on Computational Linguistics*, 997-1005.

Provincial Court in Salamanca. Judgment of 2nd March 2017 (Nº 107/2017). URL: <https://www.iberley.es/jurisprudencia/sentencia-civil-n-107-2017-ap-salamanca-sec-1-rec-503-2016-02-03-2017-47715175> [29/02/2020].

Real Academia Española (2016). Banco de datos (*CORPESXXI*). *Corpus del español del siglo XXI*. URL: <http://www.rae.es> [02/03/2020].

Real Academia Española (2005). *Diccionario panhispánico de dudas* (*DPD*). URL: <https://www.rae.es/recursos/diccionarios/dpd> [02/03/2020].

Ruiz-Garrido, M.F., J.C. Palmer-Silveira & I. Fortanet-Gómez (2010). *English for Professional and Academic Purposes*. Amsterdam, New York: Rodopi.

Shuy, R.W. (2012). “Using linguistics in trademark cases” in P. Tiersma & L. Solan (eds.), *The Oxford Handbook of Language and Law*, 449-462. New York: OUP.

Shuy, R.W. (2008). *Fighting over Words: Language and Civil Law Cases*. New York: OUP.

Spanish Intellectual Property Law 1 /1996, 12th April. URL: <https://www.boe.es/diario_boe/txt.php?id=BOE-A-1996-8930> [02/03/2020].

Stamatatos, E. (2009). “Intrinsic plagiarism detection. Using character n-gram profiles” in E. Stamatatos et al. (eds.), *Proceedings of the SEPLN’09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, 38-46.

Stein, B., N. Lipka & P. Prettenhofer (2011). “Intrinsic plagiarism analysis”. *Language Resources and Evaluation* 45(1): 63-82.

Tardy, C.M. (2012). “Current conception of voice” in K. Hyland & C. Sancho Guinda (eds.), *Stance and Voice in Written Academic Genres*, 34-48. Basingstoke: Palgrave Macmillan.

Turell, M.T. (2008). “Plagiarism” in J. Gibbons & M.T. Turell (eds.), *Dimensions of Forensic Linguistics*, 265-299. Amsterdam/Philadelphia: John Benjamins.

Turell, M.T. (2005). “El plagio en la traducción literaria” in M.T. Turell (ed.), *Lingüística Forense, Lengua y Derecho. Conceptos, Métodos y Aplicaciones*, 275-298. Barcelona: IULA, Universitat Pompeu Fabra.

Turell, M.T. (2004). “Textual kidnapping revisited: the case of plagiarism in literary translation”. *Speech, Language and the Law* 11(1): 1-26.

Turney, P.D. & P. Pantel (2010). “From frequency to meaning: Vector space models of semantics”. *Journal of Artificial Intelligence Research* 37: 141-188.

Turnitin. URL: <http://turnitin.com/> [02/03/2020].

van Dijk, T.A. (2008). “Context theory and the foundation of pragmatics”, *Studies in Pragmatics* 10: 1-13.

Woolls, D. (2012). “Detecting plagiarism” in P.M. Tiersma & L.M. Solan (eds.), *The Oxford Handbook of Language and Law*, 517-529. New York: OUP.

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*, 576-590. London/New York: Routledge.

Woolls, D. (2003). “Better tools for the trade and how to use them”. *The International Journal of Speech Language and the Law: Forensic Linguistics* 10(1): 102-112.

Woolls, D. (2002). *CopyCatch Gold v2*. CFL Software Development, United Kingdom.

Victoria Guillén is tenured Associate Professor at the University of Alicante, where she directs the Master's in English and Spanish for Specific Purposes and teaches Applied Linguistics and Forensic Linguistics. Since 2009 she has provided professional linguistic service as an expert linguist in Spain, Germany, Sweden, and the US. She is currently doing research on language crimes, especially harassment and hate speech.

NOTES

[1] This approach integrates the study of linguistic and extralinguistic contextual elements into the task of plagiarism detection.

[2] This is an English translation of the source text in Spanish: "De este modo, y aunque en el lenguaje común nos sintamos muchas veces invitados a llamar plagio a cualquier coincidencia entre dos obras que nos resulte llamativa, lo cierto es que no existe tal plagio cuando las obras comparadas resulten suficientemente distintas y diferenciables, aunque tengan *puntos comunes de exposición* (Sentencia de 20 de febrero de 1992)". URL: <https://www.huffingtonpost.es/jose-maria-aznar-auzmendi/que-es-y-que-no-es-un-plagio_b_3042814.html> [28/02/2020].