SIMILARITY THRESHOLD TO DETECT PLAGIARISM IN SPANISH

MONTSE MARQUINA
SHEILA QUERALT
UNIVERSITAT POMPEU FABRA

Abstract. Plagiarism is, unfortunately, a very common practice in today's society. This practice occurs in several areas: journalistic, educational, literary and scientific world, among others. Previous studies on the detection of plagiarism establish that it is unlikely that two authors write independently two identical sequences of more than seven words in English. This article examines whether this established similarity threshold in English can be applied into the Spanish language. In addition, the possible variability in the threshold according to the text genre in which the sequence has occurred been also taken into account in this study. For this reason, a selection of utterances in Spanish, from different genres (journalism, literary and scientific) has been analysed. Results show that the similarity threshold for the Spanish language is lower than for the English language regardless of genre. Findings of this study will contribute to furnish more reliable results in Court, in cases of plagiarism detection.

Key words: *Uniqueness threshold, idiolectal style; plagiarism detection, Spanish language.*

Resumen. El plagio, desafortunadamente, es una práctica muy común en la sociedad actual. Dicha práctica se produce en diversos ámbitos: el mundo periodístico, educativo, literario y científico, entre otros. Estudios previos sobre la detección de plagio establecen la improbabilidad de que dos autores escriban independientemente dos secuencias idénticas de más de siete palabras en inglés. Este artículo estudia si ese umbral de singularidad establecido en inglés puede aplicarse al español. Además en este estudio también se tiene en cuenta la posible variabilidad en el umbral según el género textual en el que se ha producido la secuencia. Por este motivo, se ha analizado una selección de oraciones, en español, pertenecientes a distintos géneros: periodístico, literario y científico. Los resultados muestran que el umbral de singularidad del español es menor que para el inglés independientemente del género textual analizado. Los resultados de este estudio permitirán ofrecer resultados más fiables en casos de detección de plagio ante los Tribunales.

Palabras clave: *Umbral de singularidad, estilo idiolectal, detección del plagio, español.* .

1. Introduction

Plagiarism cases increased exponentially from the nineties concurring with the technological boom which facilitated the access to information. Hence it became easier to reuse the words and ideas of others and reproduce them as your own.

Most of the Intellectual Property Laws around the world characterise plagiarism as an offence in their judicial system, but the perception of being plagiarised or to plagiarise somebody is very different depending on the community’s culture. In Common Law (also referred as Anglo-Saxon Law) countries such as United States, Great Britain or Australia,

REVISTA ELECTRÓNICA DE LINGÜÍSTICA APLICADA (ISSN 1885-9089)
2014, Número 1, páginas 79-95
Recibido: 21-05-2014
Aceptación comunicada:10-07-2014

plagiarism is simply considered as an offence. However, In Civil Law (also referred to as Continental) countries such as Spain, Mexico or Argentina, to steal somebody else’s ideas or words is still accepted or is more prone to be accepted. This is the reason why expert linguists are less frequently summoned to give evidence in plagiarism cases in Spanish Courts than in any other Common Law country.

Plagiarism is unfortunately becoming a very common practice in many areas such as the world of journalism, literature and also in the academic scene. Therefore, it is not surprising that studies on plagiarism taking into account different areas of knowledge or settings have increased leaps and bounds in recent year. Nonetheless, the participation of expert linguists in judicial procedures and research studies varies from country to country. It is fundamental that forensic linguists give expert opinions in Court with internal and external validity through widely accepted theories and methodologies among the scientific community. Hence, the most important aim in plagiarism detection is “the establishment of the threshold level of textual similarity between texts which is going to be decisive to determine if that similarity is suspicious” (Turell 2008: 274).

This paper derives from the question which arises from previous studies on forensic linguistics such as Coulthard (2004), Olsson (2004) and Coulthard and Johnson (2007). In those works utterances from real forensic cases were selected and reached the conclusion that a string of seven words (or 40 characters) was enough to state that it was formulated by a unique author. Furthermore, studies such as Culwin and Child (2010) analysing academic works achieved very similar results. Therefore, findings lead to the premise that it is very unlikely that two people produce the same utterance when is formed by 7 or more words. This study aims to validate whether this premise may be also applied to the Spanish language and, therefore, to establish the similarity threshold level between two texts to determine whether both text are written by the same author.

2. Theoretical Framework

The theoretical framework of this study is, in general, forensic linguistics which is the interface between language and uses knowledge of new technologies and statistics; and, in particular, plagiarism detection which implies both the appropriation of an idea as to copy text to express this idea.

In the field of plagiarism detection, there is a distinction between copying of ideas and linguistic plagiarism, whilst there may be copy of ideas without linguistic plagiarism, there can be no linguistic plagiarism without copying of ideas.

Plagiarism of ideas would include:

- The use of structural elements that form the unity of a literary piece of work: plot, characters, place, time, stream of consciousness, and others.
- The use of all or almost all original rhetorical figures from some other literary author, without specific acknowledgement, even if the words used to express those figures are different.
- The copying of a translated version, if the translated version itself makes an explicit contribution, by changing this version from prose to verse, of by dehistoricising a classical work, or historicing a contemporary work.
- In scientific contexts, the use of the same topics in the description of a historical period or ina contribution to a field of specialisation.
- In scientific text books, the reproduction of important structural components of this type of work, such as Activities, Questions and Laboratory Techniques.
- In scientific text books, the reproduction of creative methodology devised to teach a particular discipline. (Turell 2008: 275-279)

Nevertheless, linguistic plagiarism is defined, according to Turell (2008: 281) –based on the opinions of Menasche (1977) and Roid (2008)-, by the following characteristics:

- When exactly the same words and/or sentences are used in order to write about one's own or other people's ideas.
- When there exists paraphrase, that is, when someone uses other people's ideas with his or her own words but makes use of the main bulk of the original words, phrases and sentences.
- When one uses several words and sentences without quotations but changes others.
- When the original syntax is maintained and only words are replaced by synonyms.
- When there is acknowledgement of the original author, but the changes only involve one or two words, word order (WO), voice (active v. passive) and/or the verbal tense and aspect of the sentences or the whole text. (Turell 2008: 281)

The main goal of the forensic linguist in a plagiarism detection case is to find the unique unrepeatable idiosyncratic linguistic features of an individual in order to know whether, two text samples may have been produced independently. This set of idiosyncratic linguistic features has been called **idiolect** which Coulthard (2006) defines as quoted below:

> Every speaker has a very large active vocabulary built up over many years, which will differ from the vocabularies others have similarly built up, not only in terms of actual items but also in preferences for selecting certain items rather than others. Thus, whereas in principle any speaker/writer can use any word at any time, in fact they tend to make typical and individuating co-selections of preferred words. (Coulthard 2006:1)

However, Turell (2010) asserts that to describe idiolects, one should analyse large amounts of linguistic data, oral and written, of each individual, which would be an impossible task in real situations. For this reason Turell (2010) considers that the term idiolect is not suitable in the field of forensic linguistics and proposes the term **idiolectal style**, defined as the particular way in which an individual uses a linguistic system shared by many people, considering that each person uses his/her language in a distinctive way, and it is this personal style that is relevant to forensic linguistics.

Thus, the idiolect can be defined as the selection of linguistic elements that a speaker makes among a set of linguistic elements available in his/her language. Coulthard (2006) adds the notion of **uniqueness of encoding** to the concept of idiolect based on the Sinclair's (1991) principles to create phrases which are described as follows:

> Sinclair's principles are the following:
>
> - **open choice principle:** "language text as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness."
>
> - **idiom principle:** "language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair 1991:109-110)

According to Coulthard (2006), these two principles act when we speak or write. Therefore, it is unknown whether an individual's language production is explained by the first principle, ie, by the speaker's or writer idiosyncratic selection word by word, or by frequent collocations and colligations in the language that are linked by the individual, as suggested by the second principle. However, this author ensures that the greater the length of a sentence, the more likely it is that the uttered sentence may be the result of the open choice principle and therefore, it is very unlikely that two speakers of a language may produce the same sequence by chance.

The main objective of this research is to study which is the maximum number of identical words in Spanish that can appear in two sentences uttered by different authors.

Thus, which is the maximum length of an utterance in Spanish that determining that they have been produced independently by two different authors? Furthermore, does genre influence the similarity threshold?

Forensic linguistics, particularly in the analysis or plagiarism detection, is based on two fundamental principles. On the one hand, as suggested by Turell (2004), when an author produces a message, whether oral or written, creates a unique and idiosyncratic text in which one can observe a number of authorship marks or linguistic resources that makes it unique. In this sense, Coulthard (2005) states that it is expected that two authors who write about the same topic share a set of lexical and grammatical elements but in no case will be identical. On the other hand, in spontaneous contexts, authors and their recipients are not aware of those authorship marks or linguistic resources and therefore they are unnoticed for the plagiarist who will intend to copy or imitate them (Turell 2004).

The task of the forensic linguist expert is to analyse these linguistic markers and provide the widest possible linguistic data to help confirm or deny the existence of plagiarism. According to Turell (2007), the reliability of the data provided by specialists in the analysis and detection of plagiarism depends on the combination of two types of analysis:

- Qualitative analysis, based on linguistic criteria such as sequential integrity, coherence and cohesion of the texts being compared.
- Quantitative analysis, based on a series of measures and quantitative analytical methods such as the level of shared vocabulary or only once shared words, uses specialized tools and software to detect plagiarism.

3. Objectives

According to Coulthard (2006), the greater the length of a sentence, the more likely that the sentence may be the result of the performance of the **open choice** principle and, therefore, it is very unlikely that two speakers produce the same sequence by chance. Coulthard (2004) proves the above statement with the exercise described below:

1. He selects two sentences from the recording of a police interview used in a real case:
    a. I asked her if I could carry her bags

b. I picked something up like an ornament

2. To demonstrate that although these two sentences are formed by common elements of the language and which can appear separately, it is unlikely that all appear in the same sentence and in the same order; He performs searches on Google and analyzes the number of occurrences provided by the search engine according to the length of the sentence.

Results are displayed in Figure 1.

String	**Instances**
I picked	1,060,000
I picked something	780
I picked something up	362
I picked something up like	1
I picked something up like an	0
if I could	2,370,000
I asked	2,170,000
I asked her	284,000
I asked her if	86,000
I asked her if I	10,400
I asked her if I could	7,770
I asked her if I could carry	7
I asked her if I could carry her	4
I asked her if I could carry her bags	0

Figure 1: Number of occurrences in Google by Coulthard (2004)

As shown in figure 1, from a certain number of words the numbers of cases found on the Internet lower drastically. Following this work, and based on these data, it was considered that if two utterances in English consisting of 7 identical words or more, are most likely to have been produced by the same author or an author has plagiarized other.

On the basis of the foregoing, the research questions of this paper are:

1) whether this threshold –in terms of number of words– may also be applied to other languages such as Spanish, since the corpus used in those studies contained only English utterances;

2) whether genre plays an important role in the establishment of the similarity threshold of utterances.

4. Hypotheses

According to the above objective, the main hypotheses formulated for this study are the following:

a) It is expected that the maximum length of an utterance in Spanish by which we can say that the same statement may have been produced by two different authors independently, may be lower than the threshold found in English. This hypothesis is formulated because Spanish has a more complex grammar than English in terms of the possibility of possessing more alternatives with respect to word order.

b) The establishment of the threshold may be affected by the variable genre, namely journalism, literary and scientific. Each genre manifests different characteristics at their phonetic, morphosyntactic and lexico-semantic linguistic level because their nature is different. For instance, a journalistic text is subject to its content (what, when, where, who) which normally has been provided by the same source and is characterised by simple vocabulary, short sentences and the use of fixed expressions. On the other hand, a scientific text is characterised by the use of technical terms, precise vocabulary and a fixed structure. Furthermore, a literary piece is content free (it can explain a real or an imaginary story) and is characterised by rich vocabulary, long sentences and the use of metaphors. Therefore, it is possible that the similarity threshold might be affected by genre.

5. Methodology

The corpus compiled for this work consists of 60 statements between 10 and 18 words long in peninsular Spanish as you can see in table 1. These 60 statements belong to three different genres: 20 statements belonging to the journalism genre, 20 to the literary genre and 20 to the scientific or academic genre. Phrases were extracted from Spanish real newspapers and literary and academic articles published in Spain.

Genre	**Number**
Journalism	20
Literary	20
Scientific	20

Table 1: Distribution of the corpus by genre

Utterances were selected randomly and were rejected if they contained a number expressed in digits, idioms, collocations, marked structures, acronyms or any other special character. These utterances were used as quoted search terms within a general search engine (Google). The statement “El aprendizaje es significativo cuando se basa en la práctica”, in English “Learning is meaningful when is based on practice”, will be used as example as shown in table 2. The procedure started by introducing the word “El” and then the two word string “El aprendizaje”, and proceeded in the same way until the whole utterance was submitted as a quoted string.

String	**Hits**
El	8.530.000.000
El aprendizaje	15.900.000
El aprendizaje es	9.470.000
El aprendizaje es significativo	98.400
El aprendizaje es significativo cuando	44.700
El aprendizaje es significativo cuando se	11.100
El aprendizaje es significativo cuando se basa	1
El aprendizaje es significativo cuando se basa en	1
El aprendizaje es significativo cuando se basa en la	1
El aprendizaje es significativo cuando se basa en la práctica	1

Table 2: Successive Google searches

This study has used Internet as a corpus because as Olsson (2008) suggests the function words ratios are similar to those in general corpora. However, it must be pointed out that Google expands its corpus everyday, hence, results may change. Therefore, in order to obtain consistent results all Google’s enquires were carried out on the same day.

A descriptive statistical analysis based on the frequencies of linguistic variables was carried out in order to obtain descriptive data. In addition, to confirm that the difference between the different frequencies obtained is statistically significant, the chi-square test (χ^2) was implemented to the analysis.

The χ^2 test allows to test whether the difference between the obtained (o) and the expected frequencies is large enough (e) to say that there are truly significant differences among the number of words that can be produced by the same writer in Spanish. The formula for this comparison is the following:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

Equation 1. Chi-square test formula.

Finally, the Two-way ANOVA test was performed to determine whether the number of hits may be affected by the number of words and genre. The ANOVA enables us to compare the means of more than two groups on two independent variables. By using ANOVA, to examine the differences between the means and decide whether those differences are likely to happen by chance or by treatment effect is possible.

6. Results

6.1. Global results

On the basis of the analysis previously explained, results show that from a certain number of words (5-6) the number of instances found on the network drops dramatically. In figure 2 the number of occurrences retrieved for each of these sequences of words has been noted. We can observe that the largest declines have occurred between 3 and 7 words in the case of scientific and literary text and between 3 and 8 words in the case of journalistic texts.

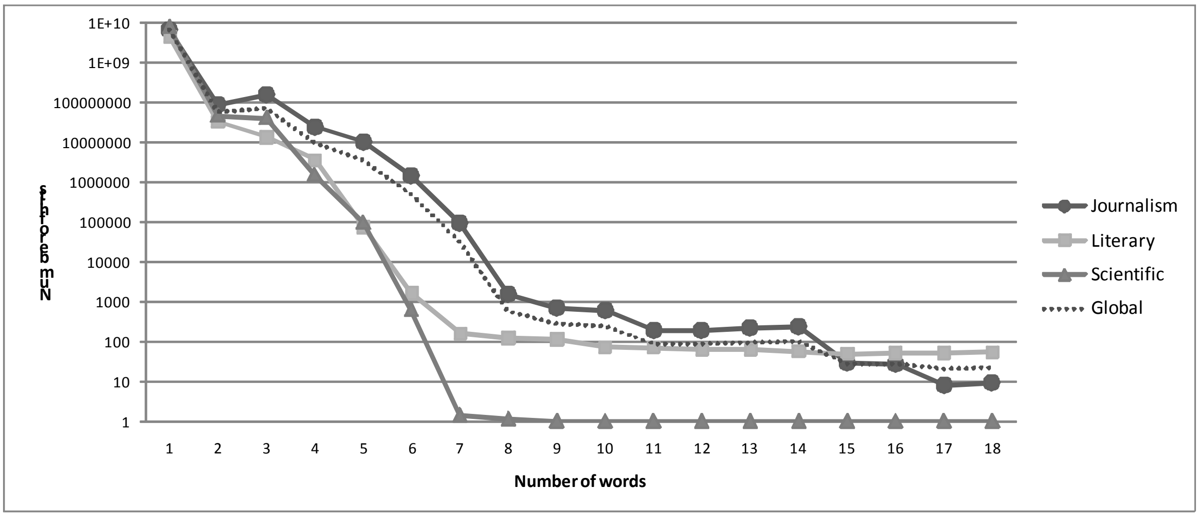

Figure 2: Average hits using progressively longer sequences of words (N)

In order to acknowledge the statistical significance of the findings, a descriptive statistical analysis was carried out for the frequency of occurrence, regardless of genre. Results indicate that, in 51.7% of the cases the threshold is located between 5 and 6 words. Therefore, in Spanish, the number of occurrences from which results drop in Google is considerably lower than in English. More precisely, the statistical mean is 5.65 (SD = 1.774) words and the mode is 6 words. The summary of the descriptive statistics which include mean, mode, standard deviation, range and, minimum and maximum values can be observed in Table 3:

N	Mean	Mode	Std. Deviation	Range	Minimum value	Maximum value
60	5.65	6	1.774	9	2	11

Table 3: Descriptive statistics summary according to the number of words

The distribution of the cases according to the number of words provides the exact percentages of occurrences in each number of words. As shown in figure 3 in a higher percentage of cases (51.67%) sentences between 5 to 6 words were found only written by a single author or a single source. In addition, in less than 25% of the cases, statements of less than 5 words are no longer unique of the author, likewise statements over 6 words.

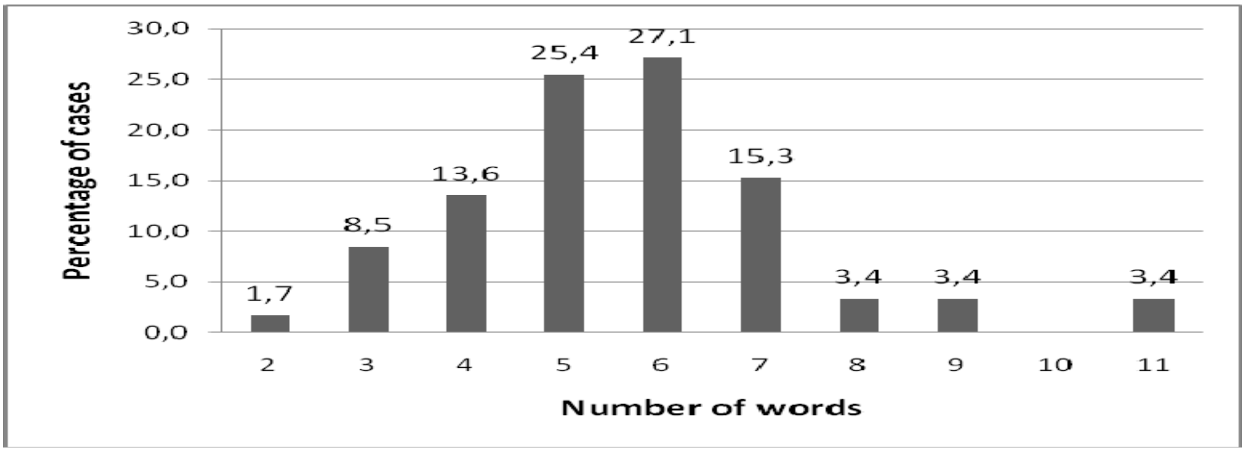

Figure 3: Distribution of cases according to the number of words (%)

In the light of these descriptive findings, it is possible to state that the similarity threshold level in Spanish stands at around 5 or 6 words and therefore if two sentences contain 5, 6 or more words, the expert must be suspicious of plagiarism. In order to determine the exact threshold and its statistical significance a chi-square test was carried out.

Thus, a chi-square test was preformed in order to verify that the differences in the number of words of the sentences searched in Google were not due to chance and that the difference between cases according to the number of words was statistically significant. Table 4 shows observed number utterances which are unique after a concrete number of words and the expected numbers. As can be noted from the residual column, which calculates the ratio of the difference between the observed count and the expected count, the higher residual values (8.3 and 9.3) are obtained for 5 and 6 word utterances. This result indicates that the expert linguist must be suspicious of plagiarism when finding a vocabulary coincidence of 5 or more words.

Number of words	Obseved N	Expected N	Residual
2	1	6.7	-5.7
3	5	6.7	-1.7
4	8	6.7	1.3
5	**15**	**6.7**	**8.3**
6	**16**	**6.7**	**9.3**
7	9	6.7	2.3
8	2	6.7	-4.7
9	2	6.7	-4.7
11	2	6.7	-4.7
Total	60		

Test	
Chi-Square	**39.600**[a]
df	8
Asymp. Sig.	**.000**
a. 0 cells (0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.7	

Table 4: Chi-square results according to the number of words

At the bottom of table 4 the chi-square test can be found. The chi-square test confirms that the difference between the cases is sufficiently high X^2 (8, N = 60) = 39.600, p <0.001], so as to be able to talk about trends or majorities.

In view of the results obtained through the frequentist statistical technique it is possible to state that the exact match of two linguistic sequences formed by 5 or more identical words reveals that such sequences have most likely been produced by the same author, or have not been produced independently.

On the basis of the above results one can validate the first hypothesis of this study which states that the similarity threshold to determine whether to samples have been produced by the same author is lower in Spanish than in English. This is so, because whereas the similarity threshold level had been established in English at 7 words, in Spanish is established at 5 words.

6.2. Results according to genre

The second part of the research, on the other hand, aimed to find out whether genre may affect the similarity threshold of utterances. To that end, this investigation used a corpus from three different textual genres –journalism, literary and scientific articles.

In figure 2 one is able to observe some minor differences in the average number of hits retrieved by Google for each gender. More specifically, journalistic texts seem to place the threshold in a greater number of words (3-8) than scientific and literary text (3-7).

To carry out a closer examination, a distribution of the exact percentage of cases (according to number of words and genre) in which the results were unique from the author was represented in a histogram (figure 4).

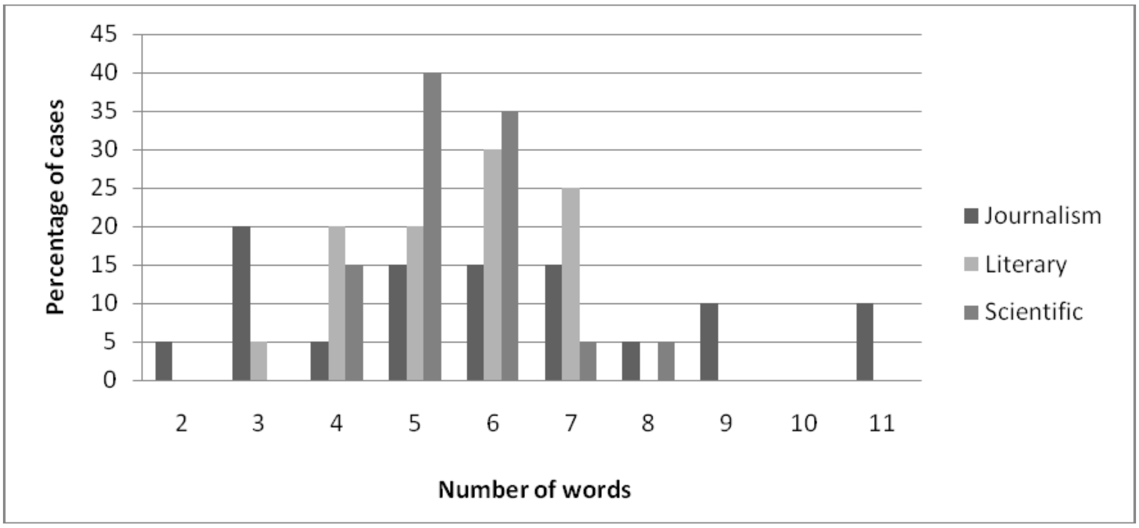

Figure 4: Distribution of cases according to the number of words and genre (%)

As shown in figure 4, in 45% of the cases of newspaper utterances the text is unique when quoting a sentence between 5 to 7 words. In the case of literary utterances the similarity threshold seem to be situated between 4 and 7 words in 95% of the cases and regards scientific or academic articles the threshold seem to be situated between 5 or 6 words in 75% of the cases. Thereupon, statistical differences in the similarity threshold might exist.

After the descriptive results, a two-way ANOVA analysis was carried out in order to test the statistical significance of the findings which pointed out the possibility that genre may have a statistical impact on the establishment of the similarity threshold. The analysis evaluated the combined effect of genre and number of words and allowed testing whether these variables were statistically significant either separately or in combination (also called interaction).

Since the test for homogeneity of variance within cells (Levene’s test) yielded a significant p = 0.000, result that was expected since neither standard deviations nor the number of samples were equal. Consequently, the p-value adopted for the analysis will be set up at 0.01 instead of 0.05 to ensure reliable results.

In table 5 results we can observe the actual result of the two-way ANOVA –namely, whether either number of words, gender or their interaction (Genre * Number) is statistically significant.

Tests of Between-Subjects Effects						
Dependent Variable: number of hits						
Source	Type III Sum of Squares	gf	Mean Square	F	Sig.	Partial Eta squared
Corrected Model	2.761E21	53	5.210E19	15.929	.000	,504
Intercept	1.080E20	1	1.080E20	33.023	.000	,038
Genre	5.327E18	2	2.664E18	.814	**.443**	,002
Number	2.628E21	17	1.546E20	47.270	**.000**	,492
Genre * Number	1.223E20	34	3.597E18	1.100	**.321**	,043
Error	2.715E21	830	3.271E18			
Total	5.683E21	884				
Corrected Total	5.476E21	883				
a. R Squared= .608 (Adjusted R Squared = .473						

Table 5: Two-way ANOVA results

The focus of the outcome must be in the “Genre”, “Number” and “Genre*Number” rows and the “Sig” column which are highlighted in the table. According to the data genre does not have a statistically significant interaction to the unique number of words to establish the similarity threshold since p = 0.321. We can see from the above table that there is no statistically significant difference between journalistic, scientific and literary texts. However, there are statically (p =0.443) significant differences between the number of words ($p <$ 0.001).

More specifically, results conclude that the interaction between genre and the number of words is not significant, F(34, 830) = 1.100, p = 0.321 and regarding Partial Eta squared results the interaction only accounts for the 4% of the variability on the number of hits. Since there is no significant interaction, main effects are tested. On the one hand, the main effect for genre is not significant, F(2, 830) = 0.814, p = *0.443* and only accounts for the 0.02% of the variability of hits. On the other hand, the main effect for number of words quoted is significant, F(17, 830) = 47.270, p = 0.000 and it accounts for almost 50% of the variability on the number of hits. These values confirm Chi-square results on figure 3 and table 4 that pointed out that there is a statistical significance according to the number of words used as a quoted search and that in 50% of the cases the similarity threshold is situated at 5 or 6 words.

In view of the results obtained, the second hypothesis of this study cannot be validated because genre does not influence the establishment of the similarity threshold level. This outcome might be explained by the proposals of Swales (1990) y Couture (1986) which formulate that while genre imposes restrictions in the structure of the discourse, register does in grammatical and lexical linguistic levels. In this sense, we can postulate that the similarity threshold to be suspicious of plagiarism in Spanish is five matching words and regardless of genre.

7. Conclusions

In this paper we have attempt to contribute to the theoretical and methodological framework on plagiarism detection. Specifically, in terms of establishing the empirical similarity threshold in Spanish language in order to be able to state that a text is suspicious to have plagiarised. This similarity threshold is an imperative to furnish reliable and robust evidence in Court.

The analysis has determined that in Spanish it is unlikely that two different people independently produce the same linguistic sequence of 5 or more words. Proof of this is the fact that the percentages of frequencies obtained from the descriptive statistical analysis of the corpus of study (see figure 3 and table 3). The difference in the results when compared to the data provided by English studies can be explained by the flexibility and versatility of the Spanish language. In addition, the reliability of the data provided in this work was tested by the chi-square test which confirmed the statistical significance of the results obtained (see table 4). Although the study was carried out with a corpus containing three different genres, ANOVA results (table 5) according to genre indicate that there is no interaction, since the magnitude of the difference between the number of words used as a quoted search does not depend upon genre. Therefore, the similarity threshold which would be able to denote plagiarism is the same regardless of genre (journalism, literary and scientific).

Thus, when an overlapping of two identical sequences of words produced by different authors is observed, it may be explained by the "idiom principle" proposed by Sinclair (1991), based on the existence of common words in a given language. However, if such sequences are formed by 5 or more identical words, the acting principle is the "open choice principle", which is based on the concept of idiolect and on the certainty that an individual’s idiosyncratic style is reflected from a specific length and, therefore, either they have been produced by the same author or one of the sentences has been plagiarised from the other.

Acknowledgements

This research was the brainchild of a conversation with Professor Maria Teresa Turell (Universitat Pompeu Fabra), which is hereby gratefully acknowledged for her priceless guidance.

Bibliographical references

Baldwin, J. 1979. Phonetics and speaker identification. *Medicine, Science and the Law* 19/4: 231-232.

Coulthard, M. 2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics* 25/4: 431-447.

Coulthard, M. 2006. … And then … language description and author attribution. [Available at http://www.aston.ac.uk/downloads/lss/english/Andthen_Coulthard.pdf]

Coulthard, M., and A. Johnson. 2007. *An introduction to forensic linguistics: Language in evidence*. Abingdon: Routledge.

Couture, B. (ed.) 1986. *Functional Approches to Writing: research perspectives*. Norwood, NJ: Ablex.

Culwin, F. and M. Child. 2010. Optimising and automating the choice of search strings when investigating possible plagiarism. *Proceedings of 4th International Plagiarism Conference*. Newcastle.

Ferguson, C. 1979. Phonology as an individual access system: Some data from language acquisition. In C. Fillmore, D. Kempler and W. Want (eds.), *Individual differences in language ability and language behaviour*: 189-201. Nova York: Academic.

Menasche, L. 1977. *Writing a research paper*. Ann Arbor: The University of Michigan Press.

Olsson, J. 2004. *Forensic linguistics: an introduction to language, crime and the law*. Continuum International Publishing Group.

Payne, A. 1980. Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In W. Labov (ed.), *Locating language in time and space*: 143-178. Nueva York: Academic Press.

Roig, M. 2008. *Avoiding those little inadvertent lies when writing papers.* Eye on Psi Chi, Winter.

Sinclair, J. 1991. *Corpus concordance collocation*. Oxford: Oxford University Press.

Swales, J. 1990. *Genre Analysis.* Cambridge: Cambridge University Press.

Turell, M. T. 2004. Textual kidnapping revisited: The case of plagiarism in literary translation. *The International Journal of Speech, Language and the Law. Forensic Linguistics* 11/1: 1-26.

Turell, M. T. 2007. Plagio y traducción literaria. *Vasos Comunicantes* 37/1, 43-54.

Turell, M. T. 2008. Plagiarism. In J. Gibbons and M. T. Turell (ed.), *Dimensions of forensic linguistics*: 265-299. Amsterdam/Philadelphia: John Benjamins.

Turell, M. T. 2011. La tasca del lingüista detectiu en casos de detecció de plagi i determinació d'autoria de textos escrits. *Llengua, Societat i Comunicació: Revista de Sociolingüística de la Universitat de Barcelona 9*: 69-85.

Turell, M. T. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law* 17/2: 211-250.