The creation of Base Rate Knowledge of linguistic variables and the implementation of likelihood ratios to authorship attribution in forensic text comparison

Sheila Queralt

SQ – Lingüistas Forenses, Spain

Abstract. *This article contributes to the research challenges that Forensic Linguistics faces in the 21st century – to compare texts of unknown authorship with the same reliability as other disciplines that consider forensic evidence. This research implements advanced statistical techniques within the field of forensic text comparison that improve the reliability of linguistic evidence furnished in Court and assess its significance. The first part of the analysis creates a Base Rate Knowledge for some of the most relevant linguistic variables in Peninsular Spanish texts. The second part applies statistical tests to variables with discriminatory potential to identify the samples of the authors and also assesses the reliability of the results in a posteriori classification. The implementation of the likelihood-ratio framework in the third part improves the reliability of linguistic evidence provided in court and offers probabilistic results to assist not only the judge and jury but also the linguistic expert in order to carry out more rigorous testing and extensive performance analysis of the data.*

Keywords: *Forensic text comparison, Authorship Analysis, Idiolect, Multivariate methods, Likelihood ratios.*

Resumo. *Este artigo contribui para os desafios da investigação enfrentados pela Linguística Forense no século XXI, de modo a comparar textos de autoria desconhecida com a mesma fiabilidade que outras disciplinas que consideram a prova forense. Este estudo implementa técnicas estatísticas avançadas na área da comparação de textos forenses para aumentar a fiabilidade da prova linguística fornecida em Tribunal e para avaliar a sua significância. A primeira parte da análise cria uma base de referência para algumas das variáveis linguísticas mais relevantes em textos de espanhol Peninsular. A segunda parte aplica testes estatísticos a variáveis com capacidade discriminatória para identificar as amostras dos autores, bem como avaliar a fiabilidade dos resultados em classificação a posteriori. A implementação de um quadro de razão de verosimilhança na terceira parte aumenta a fiabilidade da prova linguística fornedida em tribunal e oferece resultados probabilísticos para apoiar, não só o juiz e o júri, mas também o perito* *linguístico, de modo a realizar testes mais rigorosos e uma vasta análise do desempenho dos dados.*

Palavras-chave: *Comparação de texto forense, Análise de autoria, Idioleto, Métodos multivariados, Razão de verosimilhança.*

Introduction

Over the last decades courts from several countries such as the United States, the United Kingdom or Spain have increasingly called on the expertise of linguists. The cases in which expert linguists give evidence can be diverse, from disputes about plagiarism, to trademarks, voice identification, linguistic profiling or authorship attribution cases. But the most frequent cases in forensic linguistics involve the comparison of an unknown sample (anonymous text) and a set of known texts from a suspect or several suspects. The estimation of the similarity between those two or more sources was traditionally approached by linguists using a verbal scale which may be based on estimations of probabilities or on opinion thresholds set by the expert (see for example Broeders 1999, Champod and Evett 2007 or Sjerps and Biesheuvel 2007). This traditional approach can be conceived to an extent as quite subjective considering that it is based on the linguistic expert's experience and may vary from expert to expert.

In the past, this traditional approach has been consigned to other forensic sciences that consider evidence such as DNA, fingerprints or handwriting. In parallel to the guidelines established, among other institutions, by the Committee on Identifying the Needs of the Forensic Sciences Community, which, for instance, states in its report that "a strong and reliable forensic science community is needed to maintain homeland security" (2009), therefore pointing towards the need of consolidating forensic techniques, the volume of forensic evidence and sophisticated forensic methods have increased over the last two decades. Consequently, multivariate and probabilistic methods have been developed in an attempt to evaluate the strength of the comparison of the quantifiable properties of known and unknown samples.

The most renowned probabilistic methodology across a broad spectrum of forensic sciences is the Likelihood-Ratio (henceforth LR) framework. In the last decade, research has proved the validity of LR models for assisting experts in forensic sciences to interpret evidence (Aitken and Taroni, 2004; Evett, 1998) and in the words of Fenton and Neil (2012: 2) expressing the "proper use of probabilistic reasoning has the potential to improve dramatically the efficiency and quality of the entire criminal justice system". Furthermore, the LR methodology fulfils the new needs of forensic individualization, applying transparent and testable procedures.

In the light of the aforesaid considerations, this article proposes the implementation of multivariate statistical methods and the LR framework for forensic text comparison through the analysis of linguistic variables. This methodology is implemented in threat texts written in Peninsular Spanish.

Methodological and theoretical framework

The concept of 'idiolect' has been the centre of some sociolinguistic variation studies such as Abercrombie (1969), Biber (1988), Biber (1995), Biber *et al.* (1998), Guy (1980) and also forensic linguistics studies, for instance, Queralt and Turell (2012), Cicres Bosch (2007), Gavaldà Ferré (2011), Spassova and Grant (2008), Spassova (2009) and Turell

(2010). The hypothesis of the existence of an idiolect makes it possible to establish a measure of idiolectal similitude to be able to state the probability of whether two linguistic samples have been produced by the same writer or not. This approach is widely accepted by the forensic linguistics community around the world as the approach to deal with the problem of questioned authorship. Nevertheless, the theory of idiolect is one of the long-standing and ongoing debates in the discipline. A number of scholars have identified practical issues that prevent this axiom from being demonstrated (e.g. Coulthard 2004: 432, Turell 2010: 217, Wright 2013: 46-47). And some have relied on alternative concepts to explain why forensic text comparison is possible, such as idiolectal style, consistency or pair-wise distinctiveness between authors (see, for instance, Turell 2010 and Grant 2010).

However, in this study the author wants to highlight that it is possible that every single person has a unique idiolect, but whether or not that is the case, it is surely true that people do develop a style and that each person's style is distinguishable from the styles of most other writers. As such, the more successful a method is in measuring the distance between the styles of different authors (even those of people with similar linguistic backgrounds), the more it should be viewed as a useful method.

In forensic text comparison, as in forensic voice comparison, the analysis of linguistic evidence does not consist only in describing the linguistic features that the unknown text contains. It also implies determining the degree of similarity between the writer's dependent features obtained from the unknown sample, and the writer's dependent features obtained from the known sample by the suspect (Gonzalez-Rodriguez *et al.*, 2006: 332).

A variety of different approaches have been developed within our discipline in the quest for quantifying the degree of similarity between samples such as relative frequency of functional or grammatical words (e.g. Burrows 1987 and Burrows 2003), word frequency distributions (e.g. Holmes 2003), vocabulary analysis (e.g. Coulthard 2004, Woolls and Coulthard 1998), and Part of Speech n-grams (e.g. Bel *et al.* 2012, Queralt *et al.* 2011, Queralt and Turell 2012, Spassova and Turell 2007, Turell 2004b and Turell 2004a); and also within other disciplines with more computational aspects, such as Juola (2006), Koppel *et al.* (2009) or Stamatatos (2009).

Nevertheless, quantifying the degree of similarity is not enough in forensic text comparison, one must also consider the rarity or the expectancy of those similar features compared to the relevant population. Coulthard and Johnson (2007) wonder "how can one measure the 'rarity' and therefore the evidential value of individual expressions" (p. 6). In order to calculate the degree of similarity and rarity between written samples one must estimate the population distribution – Base Rate Knowledge – of the relevant linguistic variables in a relevant population (Queralt, 2014: 43). These questions can be addressed by the use of these newly developed probabilistic methods, such as the Likelihood Ratio, which carries out rigorous empirical analyses. Unlike other kinds of evidence such as DNA profile data, forensic linguists deal with continuous and variable data and therefore the analysis has to consider two sources of variability: "the variability within the source (e.g., window) from which the measurements were made and the variability between the different possible sources (e.g., windows)." (Aitken and Taroni, 2004: 322).

In forensic linguistics, we use *inter-individual variation* to refer to the variability between writers and *intra-individual variation* for the variability within one writer. Intra-individual variation, variations across texts written by one author, is another intrinsic characteristic of linguistic data (see Labov 1972: 122, 127, 271-72, 319-25, Chambers 2009: 33-37 and Turell 1995: 20-22). Labov (1972: 208) states that "as far as we can see, there are no single-style speakers. Some informants show a much wider range of style shifting than others, but every speaker we have encountered shows a shift of some linguistic variables as the social context and topic change."

Intra-individual variations may occur in word choice, syntactic structures, grammatical patterns or in other linguistic levels and may be due to genre, time, social context, style, register or other external factors. According to the Saussurean view, the expert can handle intra-individual variation in two ways: on the one hand by treating idiosyncrasy as deviance and, on the other hand, by conceiving the linguistic individual as the set of strategic adaptations chosen from a closed set of conventional possibilities (Johnstone, 1996: 14).

Methodology

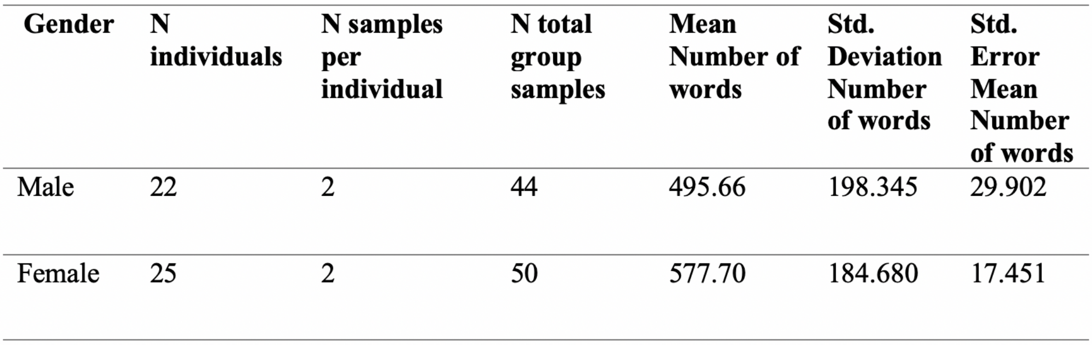
The world of forensic sciences is in continuous change due to the evolution of new technologies and the creation of more rigorous standards. Thus, in order to remain efficient and reliable, forensic sciences – in this particular case, forensic linguistics – need to adapt to these ongoing changes. This research intends to be viewed as a step forward in the direction the field should continue to evolve so as to increase its legitimacy as a forensic science. Specifically, the aim of this study was to implement advanced statistical methods to selected linguistic variables in forensic text comparison. In this respect, the methodology comprised a qualitative analysis and a quantitative analysis grounded on multivariate classical statistics, which can be defined as a simultaneous statistical analysis of a collection of variables and probabilistic methods such as the Likelihood-Ratio framework.

Corpus

One important concern was how to gather a corpus which would be comparable to corpora in the forensic world (typically characterized by a small number of authors, a small number of samples and short texts). The corpus used in this study was designed taking into consideration the importance of the availability of all the relevant sociolinguistic data about the individuals. Therefore, it was possible to avoid the effect of errors in independent variables. Finally, we were able to include texts by 47 informants. All of them were university students. Their native languages are Spanish and Catalan and they qualify as fully balanced bilingual speakers of both languages, since they have equivalent knowledge of both languages at levels corresponding to those of native speakers of each language (Baetens, 1989). All informants were between 18 and 25 years old and came from the Autonomous Community of Catalonia (Spain).

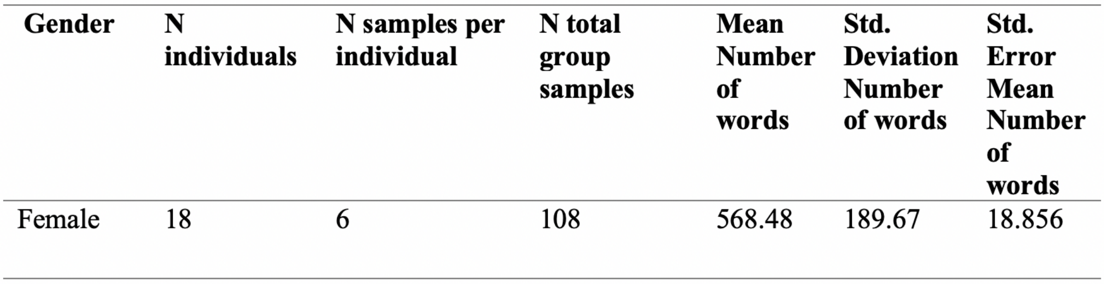
With the aim of gathering a corpus comparable to the forensic reality, participants were given the description of six different situations – one every week – and told to produce a Spanish written threatening message of approximately 600 words with a medium-high level of violence that could be understood as a verbal threat or as actual physiological violence against the recipient of their letter. This procedure resulted in a process of homogenisation of the corpus.

Thus, we compiled two different corpora: one for the BRK (Table 1) and another for the LR (Table 2). The corpus for the LR includes 22 men and 25 women and two samples per individual. The corpus to obtain likelihood ratios comprises 100% of women and 6 letters per each author since informants of this gender displayed the most cooperative attitude and showed willingness to participate in the process all the way through.



Gender	N individuals	N samples per individual	N total group samples	Mean Number of words	Std. Deviation Number of words	Std. Error Mean Number of words
Male	22	2	44	495.66	198.345	29.902
Female	25	2	50	577.70	184.680	17.451

Table 1. BRK corpus distribution.



Gender	N individuals	N samples per individual	N total group samples	Mean Number of words	Std. Deviation Number of words	Std. Error Mean Number of words
Female	18	6	108	568.48	189.67	18.856

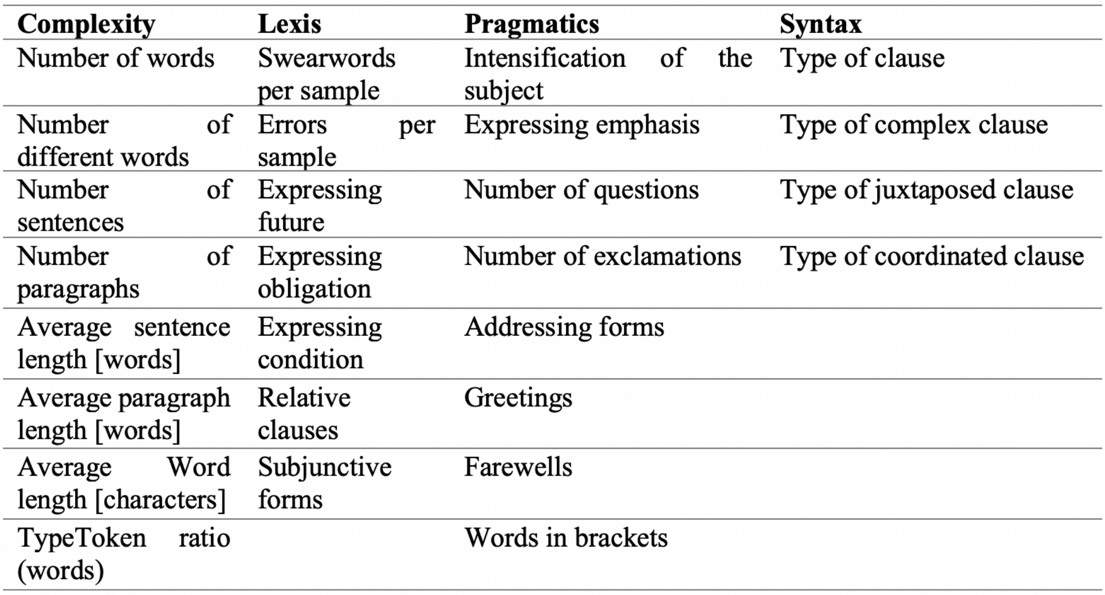
Table 2. LR corpus distribution.

Variables

A linguistic variable is the representation of a linguistic feature that can be expressed in different ways with the same meaning. The linguistic variables in this study took the following fundamental characteristics into account: the variable ought to be highly frequent and stratified (Labov, 1972), show a high inter-individual variability and a low intra-individual variability, and also be relatively easy to extract and calculate (Nolan, 1983: 11), its variants should be interchangeable in some contexts (Tagliamonte, 2006: 73) and, finally, each variable ought to be as independent of other variables as possible (Rose, 2002: 52).

We also considered variables whose discriminatory potential had been evaluated in previous studies like Grant and Baker (2001), Chaski (2001), Wright (2013). And lastly, we considered variables which had been relevant in forensic linguistics casework carried out in the laboratory in which the author has worked.

A broad range of linguistic variables were analyzed and divided into four main groups: complexity, lexis, pragmatics and syntax. Table 3 shows a summary of the analyzed linguistic variables.



Complexity	Lexis	Pragmatics	Syntax
Number of words	Swearwords per sample	Intensification of the subject	Type of clause
Number of different words	Errors per sample	Expressing emphasis	Type of complex clause
Number of sentences	Expressing future	Number of questions	Type of juxtaposed clause
Number of paragraphs	Expressing obligation	Number of exclamations	Type of coordinated clause
Average sentence length [words]	Expressing condition	Addressing forms	
Average paragraph length [words]	Relative clauses	Greetings	
Average Word length [characters]	Subjunctive forms	Farewells	
TypeToken ratio (words)		Words in brackets	

Table 3. Summary of the analyzed variables.

Complexity measures analyzed in this study include the number of words per document, vocabulary richness (number of different words), the number of sentences and paragraphs, average lengths for sentences, paragraphs and words, and type-token ratio. This group was the only one analyzed semi-automatically by a perl code designed *ad hoc* and reviewed manually. The remaining groups were analyzed manually by the researcher.

In the analysis of lexis, frequencies of swearwords and errors per sample were calculated. Other features considered were whether the author used *ir a* + infinitive or the future tense to express future, *deber* + infinitive or *tener que* + infinitive to express obligation and whether the author used *como* or *si* to express condition.

Concerning the field of pragmatics, the distribution – presence or absence – of the first person singular personal pronoun, i.e. *yo*, was calculated in order to identify its intensification when present. The different ways of expressing emphasis such as capitalization, repetition or punctuation were also considered. Other pragmatic variables were the number of exclamations and interrogations used, the formality or informality of addressing pronouns, and the types of greetings and farewells, since they are reported by previous studies as possible authorship markers Wright (2013). Finally, the use of brackets to interject other text was evaluated.

Syntax was analyzed through an observation of the clause types used by the authors, i.e. complex or simple clauses, types of complex clauses – coordinated, juxtaposed or subordinated – and types of juxtaposed or coordinated clauses.

Method

This study proposes a combination of qualitative and quantitative approaches. Schmied (1993) notes that a "qualitative analysis is often a precursor for quantitative analysis, since before linguistic variables can be classified and counted, the categories for classification must first be identified".

During the qualitative analysis, linguistic features were identified in the data but no attempt was made to assign frequencies to those linguistic features. Instead, ambiguities inherent to the Spanish language were recognized. For instance, the word ‘que’ in Spanish (*that* in English) can be used in a corpus as a relative pronoun or as a conjunction. In contrast, features were classified and counted during the quantitative analysis. The measurement of the distribution of and the correlation between features led to the identification of characteristics which are likely to be genuine of the writer and therefore representative of his/her ‘idiolectal style’ and which reflect the author’s behavior.

The statistical analysis was divided into two stages. The first stage consisted of the application of multivariate statistical techniques, which constitute an improvement of univariate analysis because “it incorporates information into the statistical analysis about the relationships between all the variables”, according to Izenman (2008: 1).

But quantifying the degree of similarity is not enough for our purposes. As stated above, one must also consider the rarity or the expectancy of the distribution and correlation of features found to be similar between corpora in relation to the relevant population. This comparison can be addressed by the use of probabilistic methods such as the Likelihood-Ratio framework which carries out rigorous empirical analyses. Therefore, the second statistical stage consisted on the implementation of the LR framework.

Many researchers and practitioners state that the LR framewok is very well-suited to present evidence in court because it only weighs the impact of the evidence studied by the expert and it does not consider the court’s prior or posterior beliefs. Aitken *et al.* (2011) state:

> To form an evaluative opinion from a set of observations, it is necessary for the forensic scientist to consider those observations in the light of propositions that represent the positions of the different participants in the legal process. The ratio of the probability of the observations given the prosecution proposition to the probability of the observations given the defence proposition, which is known as the likelihood ratio, provides the most appropriate foundation for assisting the court in establishing the weight that should be assigned to those observations. (p.1)

In this particular study, in order to obtain classification and subsequently the LR, we calculated the proximity distances among the author’s samples (inter-variability) and also the distances within the author’s samples (intra-variability). To calculate posterior probabilities for classification four algorithms of calculation were performed by discriminant analysis on the standard deviation of the distances with continuous variables.

The likelihood ratio was calculated considering four different classification tests:

- True positive: number of samples classified as belonging to their real author. 6 possible cases.
- False positive: number of samples classified as belonging to another author. 102 possible cases.
- True negative: Number of samples which are not classified as belonging to an incorrect author. 102 possible cases.
- False negative: Number of samples which are not classified as belonging to their real author. 6 possible cases.

Based on the results of these tests, the validity of the classifications was determined through the use of sensitivity and specificity tests. Sensitivity was defined as the probability of detecting an author's own samples and specificity as the probability of detecting samples that were not produced by that author, that is, the probability of rejecting foreign samples.

The subsequent step was to calculate the LR for each individual and for each of the variables in order to know the probability of the results. In particular, there were two ways to measure the likelihood ratio in this study, positively and negatively:

- Positive likelihood ratio (LR +) is the ratio between sensitivity and difference, that is, the probability that a sample is assigned to its author compared to the probability of a sample not produced by that author also being assigned to him or her.
- Negative likelihood ratio (LR−) is the ratio of the difference and specificity, that is, the probability that a sample is not assigned to its author compared to the probability that the rest of the samples are assigned to the rest of the authors.

$$LR+= \frac{\text{Sensitivity}}{\text{1- Specificity}} = \frac{\frac{TP}{(TP+FN)}}{1-\left(\frac{TN}{(TN+FP)}\right)}$$

$$LR-= \frac{\text{1- Sensitivity}}{\text{Specificity}} = \frac{1-\left(\frac{TP}{(TP+FN)}\right)}{\left(\frac{TN}{(TN+FP)}\right)}$$

Figure 1. Formulas of the Likelihood ratio.

LR+ varies between zero and infinity – the higher its value, the greater the probability of classifying the unknown sample correctly. LR− varies between 0 and 1 – the lower the value, the greater the probability of correctly classifying the unknown sample. In order to assign the unknown sample to its author, these two conditions had to be fulfilled: an LR+ as high as possible and a LR− as low as possible. Thus, an author's samples were classified correctly when they met the following requirements: the group of samples that are classified correctly to their group (true positives) is large; the value of LR+ is very high (> 1000) and the value of LR− is minimal (0).

Summing up, qualitative analysis provides greater richness and precision, whereas quantitative analysis provides statistically reliable and generalizable results (McEnery and Wilson, 2001: 77).

Results

Base Rate Knowledge results

For each of the variables, a population distribution was provided, that is, the most commonly used variant of each variable and the expected frequency of that variant were established. A frequency rate higher or lower than that established by the population distribution may signal a particular characteristic of that author.

For instance, it was observed that the threat letters in the study were not abundant in abbreviations. Nevertheless, the distribution of the abbreviation of 'euros' was considered relevant because of its frequency in extortion letters from real cases. Results showed that the most common way of writing 'euros' in this corpus is in its non-shortened form (64.56%), followed by the sign '€' (33.33%) and, finally, the abbreviation 'EUR' (2.08%).

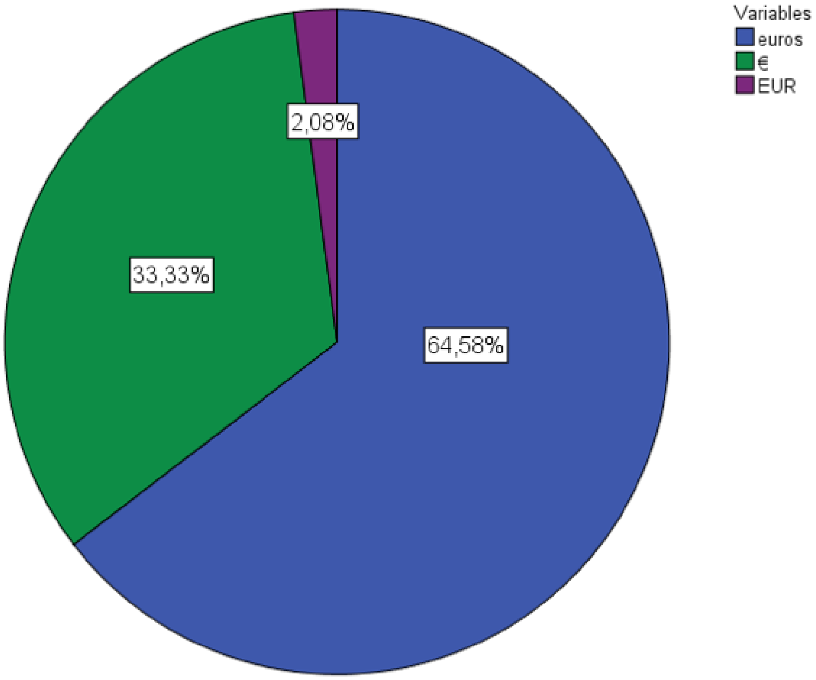


Figure 2. BRK of the substantive euros.

The way a speaker expresses emphasis may also differ in a relevant manner (Figure 3). In this study we analyzed the expression of emphasis by capitalization, the use of punctuation marks and the use of repetition. Results showed that the most common way of expressing emphasis in written texts is capitalization (70.48%), followed by punctuation marks (19.05%) and, finally, by using the repetition of words or expressions (10.48%).

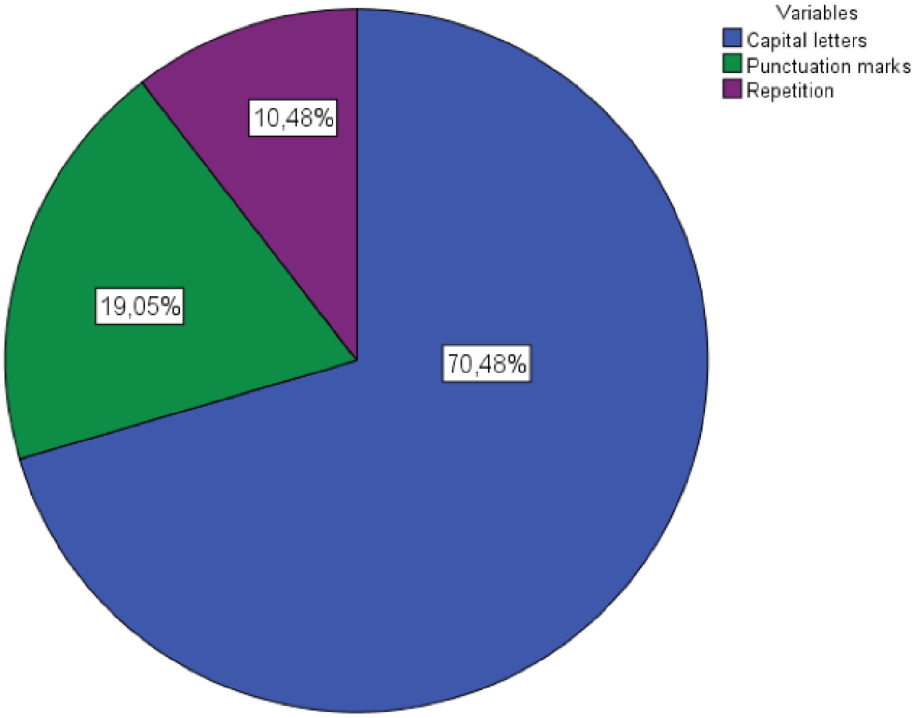


Figure 3. BRK expression of emphasis.

Population distributions of linguistic variables are useful for authorship attribution since they allow the expert to know the mean frequency values for each variant of the variables and, therefore, to what degree a variant may be expected to occur generally. Figure 4 shows the variable of lexical errors and the values for each individual sample. It is worth noting the individual behavior of certain writers. For example, writer 44 often makes significantly more errors than the average population, which is why, in the case of spelling errors, diacritics, and grammatical pleonasms, this author's samples are placed in the extreme values of the graph. Other cases of special interest are those in which the writer often makes a greater number of errors of a single type. For example, writer 35 shows a remarkable number of errors caused by the contact between Spanish and Catalan languages in both samples, and writer 41 shows some difficulties with normative punctuation. Extreme values are indicated with an asterisk and outliers with a circle.

Another example of BRK results is the variable of expression of obligation in Spanish shown in Figure 5. It is relevant to note the cases of authors 32 and 28. Author 32 stands out for using *deber* + infinitive frequently in both samples, while author 28 is the only author who uses *haber de* + infinitive.

Variables with discriminatory potential

Once the Base Rate Knowledge was established, the variables that offered a greater discriminatory potential were selected. Those variables showed low intra-individual variation and high inter-individual variation, thus, it should be possible to distinguish samples among individuals. Table 4 comprises the most discriminating variables.

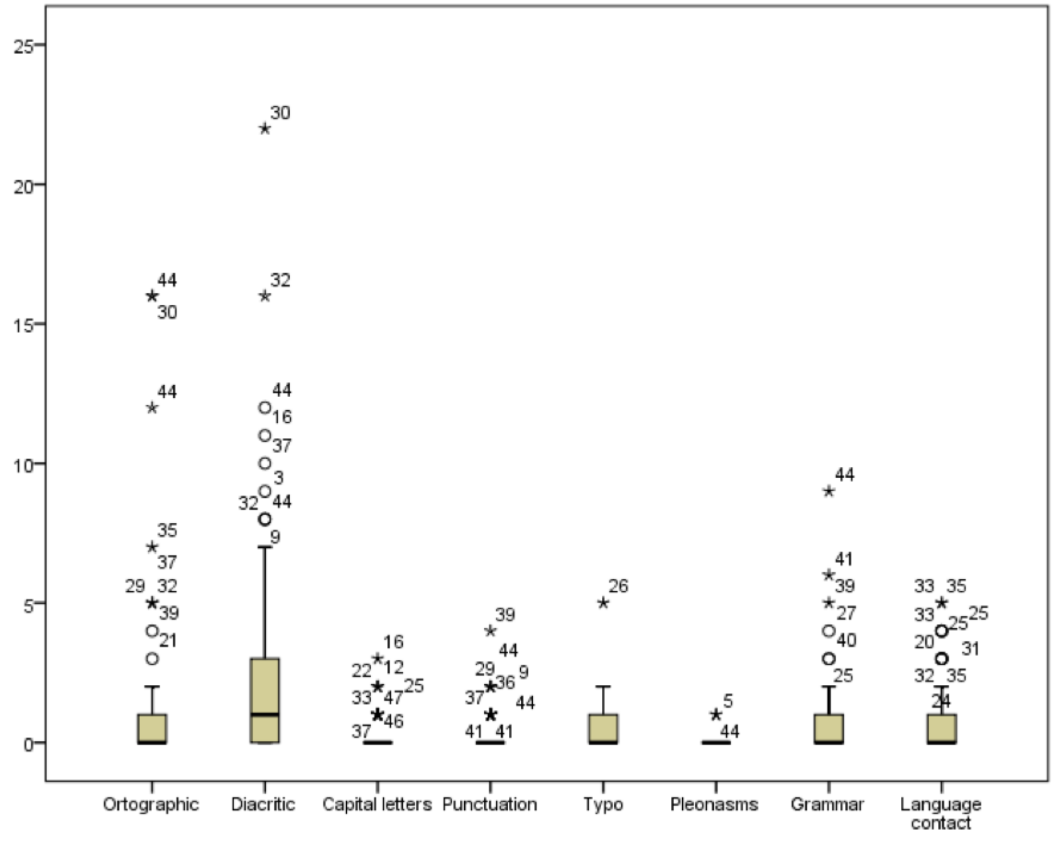


Figure 4. BRK lexical errors.

Complexity	Lexical	Syntax	Pragmatic
Number of paragraphs	Number of orthographic errors	Number of subordinate sentences	"Sincerely" (*Cordialmente*, ES)
Number of different tokens	Number of swearwords	Number of simple sentences	Absence of singular first person subject
Number of sentences	Number of language contact errors	Number of complex sentences	
Words per paragrah			

Table 4. Variables with discriminatory potential.

This set of variables was used to calculate the probabilities of success and failure in posterior classifications.

Likelihood Ratio results

Table 5 shows the classification results. Cells shaded in red show four authors who are completely different from the rest because all their samples (6) are classified correctly (meaning true positive) and no samples are attributed to another author (represented in

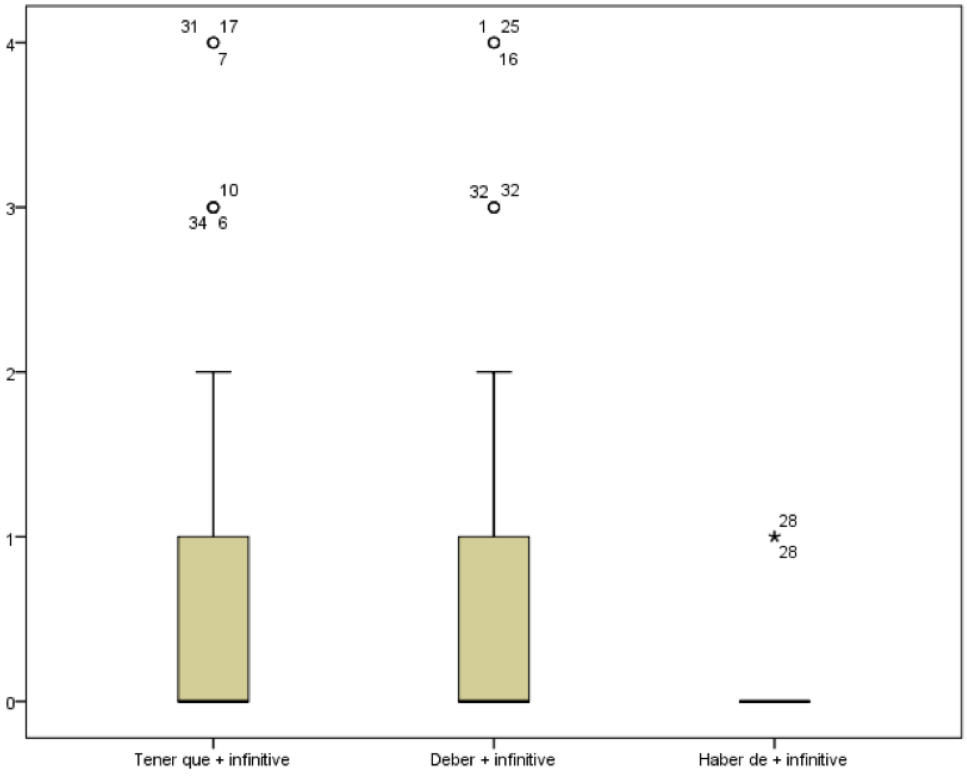


Figure 5. BRK expression of obligation.

the table as false positive). Orange cells indicate 6 authors who are well differentiated from the rest but share more features with other authors and, therefore, some of their samples are attributed to other authors.

Author	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
True positive	6	1	4	6	1	6	6	4	6	3	5	3	6	2	6	6	6	6
False positive	3	2	2	0	0	2	0	7	0	3	9	8	3	2	6	6	4	0
False negative	0	5	2	0	5	0	0	2	0	3	1	3	0	4	0	0	0	0
True negative	99	100	100	102	102	100	102	95	102	99	93	94	99	100	96	96	98	102

Table 5. Classification results.

The validity of the classifications must be determined from these figures, that is, to what extent the classifications obtained would fit more complex and rigorous processes.

Each position on the X axis of Figure 6 represents an individual and on the Y axis the probability of each of the samples. In green we can observe samples which are classified correctly to their author and in red samples which are not classified to the correct author (the number indicates the author which is incorrectly classified), that is, the method's sensitivity. Thus, this graph visually summarizes the probability of detecting an author's own samples.

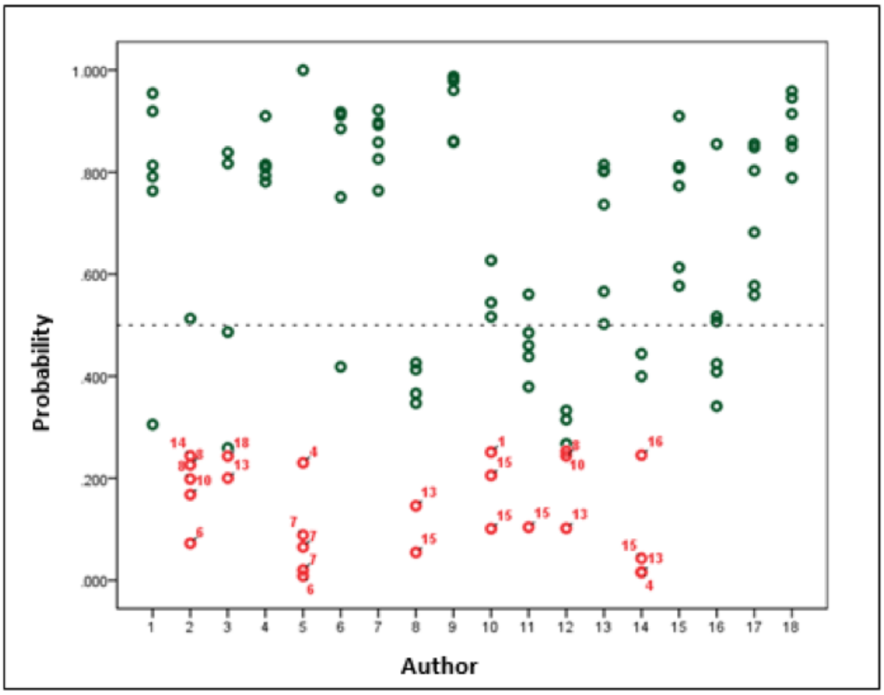


Figure 6. Sensitivity of the method.

According to these results, the classification potential is up to 76.85% and in more than half of the cases the classification probability is greater than 50%. It is important to highlight that all false negatives are below 25% probability and that all true positives are above this probability value.

With regard to the method's specificity, Figure 7 shows samples which are correctly classified in green and samples which are classified as belonging to an incorrect author in orange (notice that the number indicates the real author of the sample).

In this case, true positives are also situated in the higher odds. Furthermore, most of the true positives (77.94%) are above 50% probability and most of the false positives (82.45%) are below that percentage. However, specificity results are not as satisfactory as sensitivity results because false positives are above 25% and even 50%.

Table 6 shows the likelihood ratio results: 5 authors with a maximum positive LR (this value is denoted as > 1000), 10 with minimal negative LR (0.00) and 4 authors with a maximum and a minimum LR+ and LR– respectively.

Thus, the results so far complement recent advances in authorship attribution using LR with the integration of BRK. For example, Ishihara (2017) used word- and character-based features to attribute chatlog messages of different length by 115 authors and estimated the strength of this attributions with LR. The results of his model show a discrimination accuracy of around 76% with the shortest texts (500 words) and of around 94% with the longest (2500 words). On a different study, Ishihara (2014) applied an N-gram language model to a corpus of text messages, again divided into four groups of different sizes.

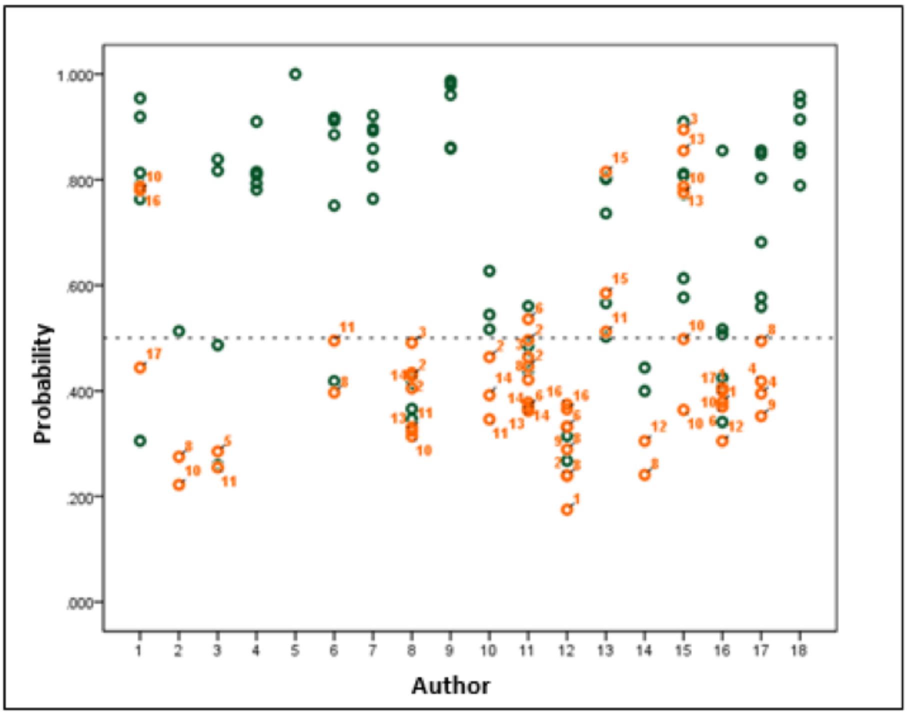


Figure 7. Specificity of the method.

Author	1	2	3	4	5	6	7	8	9
LR+	34.00	8.50	34.00	>1000	>1000	51.00	>1000	9.71	>1000
LR-	0.00	0.85	0.34	0.00	0.83	0.00	0.00	0.36	0.00
Author	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
LR+	17.00	9.44	6.38	34.00	17.00	17.00	17.00	25.50	>1000
LR-	0.52	0.18	0.54	0.00	0.68	0.00	0.00	0.00	0.00

Table 6. LR results.

Conclusions

This technique has correctly classified 75% of the samples, 60% of which with a probability greater than 50%. Finally, it should also be noted that there is a 25% sensitivity threshold since all the texts classified as belonging to their true author are above the 25% threshold and all the texts incorrectly classified are placed below this value.

As mentioned in the Introduction, the Committee on Identifying the Needs of the Forensic Sciences Community at the National Research Council of the United States published a document titled 'Strengthening Forensic Science in the United States: A Path Forward' (2009) which states:

> For decades, forensic sciences have produced valuable evidence that has contributed to the successful prosecution and conviction of criminals as well as to the examination of innocent people. Over the last two decades, advances in some forensic science disciplines, especially the use of DNA technology, have

> demonstrated that some areas of forensic science have great additional potential to help law enforcement identify criminals. Many crimes that may have gone unsolved are now being solved because forensic science is helping to identify the perpetrators. (p.26)

This statement must make the forensic scientific community realize its important role in society and therefore the – positive and negative – implications of its expert evidence. Due to the importance of the forensic expert's task, the community ought to set up a reliable methodology with agreed-upon standards and "should establish a professional body that not only promotes these goals but also certifies experts and, where applicable, accredits training programs and laboratories" (Koehler, 2013: 537).

At a general level, this study can contribute to forensic linguistics and particularly to the field of forensic text comparison, since the proposed methodology can be useful when resolving cases of authorship attribution and the corpus, the variables selected and the methodology may also represent a contribution to Corpus Linguistics, Computational Linguistics and the Likelihood-Ratio framework. Admittedly, however, this corpus has a relatively small number of participants to represent a comparative baseline to establish similar BRK and LR values for another language, which constitutes a significant limitation of the study. Additionally, as is commonly the case with research in forensic linguistics, any conclusions drawn from this study must consider the fact that the samples analyzed were produced in artificial contexts and that texts produced naturally would provide possibly provide more realistic information as to the authors' styles.

At a more detailed level, the most important contributions of this proposal have to do with the compilation of unified database of real-world texts in Peninsular Spanish in order to achieve a population distribution of linguistic variables in threatening letters; a common statistical method based on advanced multivariate statistical methods and the LR framework; a further small step towards the establishment of a code of good practice in forensic text comparison since control factors are considered during the collection of data, there are sampling procedures and qualitative and quantitative methods implemented. The implementation of a code of good practice can help to provide more reliable and conclusive results in authorship attribution.

Notwithstanding these results, there is still much to be done in the field of authorship attribution to reach the precision levels of the results of other forensic sciences taking into account the limits imposed by the nature of the object analyzed. It is necessary to develop and test new approaches to achieve comparable results taking into account the Achilles' heel of each research, for instance, the variability inherent in language.

References

Abercrombie, D. (1969). Voice qualities. In N. N. Markel, Ed., *Psycholinguistics: An introduction to the study of speech and personality.* London: The Dorsey Press.

Aitken, C., Berger, C. E., Buckleton, J. S., Champod, C., Curran, J., Dawid, A. and Jackson, G. (2011). Expressing evaluative opinions: A position statement. *Science and Justice*, 51(1), 1–2.

Aitken, C. G. G. and Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists.* Chichester, UK: John Wiley & Sons, john wiley ed.

Baetens, H. (1989). *Principis bàsic del bilingüisme.* Barcelona: La Magrana.

Bel, N., Queralt, S., Spassova, M. and Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tomblin, N. MacLeod, M. Coulthard and R. Sousa-Silva, Eds., *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham, UK: Aston University Centre for Forensic Linguistics (Online).

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: a Cross-Linguistic Comparison.* Cambridge; New York: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Broeders, A. P. A. (1999). Some observations on the use of probability scales in forensic identification. *Forensic Linguistics*, 6(2), 228–241.

Burrows, J. (2003). Questions of authorship: Attribution and beyond A lecture delivered on the occasion of the roberto busa award ACH-ALLC 2001, new york. *Computers and the Humanities*, 37(1), 5–32.

Burrows, J. F. (1987). Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, 2, 61–70.

Chambers, J. K. (2009). *Sociolinguistic theory: linguistic variation and its social significance.* Oxford: Wiley-Blackwell.

Champod, C. and Evett, I. W. (2007). Commentary on APA Broeders (1999) 'Some observations on the use of probability scales in forensic identification'. Forensic Linguistics 6 (2): 228–41. *International Journal of Speech Language and the Law*, 7(2), 239–243.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1350–1771.

Cicres Bosch, J. (2007). *Aplicació de l'anàlisi de l'entonació i de l'alienació tonal a la identificació de parlants en fonètica forense.* Phd thesis, Universitat Pompeu Fabra, Spain.

Committee on Identifying the Needs of the Forensic Sciences Community - National Research Council, (2009). *Strengthening forensic science in the united States: A path forward.* Rapport interne.

Coulthard, M. (2004). Author Identification, Idiolect and Linguistic Uniqueness. *Applied Linguistics*, 25(4), 431–447.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence.* London and New York: Routledge.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice*, 38(3), 198–202.

Fenton, N. and Neil, M. (2012). On limiting the use of bayes in presenting forensic evidence. 1–27.

Gavaldà Ferré, N. (2011). Sociolingüística de la variació i lingüística forense. *Llengua, Societat i Comunicació*, 9(49-59).

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M. and Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2–3), 331–355.

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis. In M. Coulthard and A. Johnson, Eds., *Routledge Handbook of Forensic Linguistics.* Routledge.

Grant, T. D. and Baker, K. L. (2001). Identifying reliable, valid markers of authorship: A response to chaski. *International Journal of Speech, Language and the Law*, 8(1), 66–79.

Guy, G. (1980). Variation in the group and the individual. In W. Labov, Ed., *Locating language in time and space.* New York: Academic Press, 1–36.

Holmes, D. I. (2003). Stylometry and the civil war: The case of the pickett letters. *Chance*, 16(2), 18–25.

Ishihara, S. (2014). A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech, Language & the Law*, 21(1).

Ishihara, S. (2017). Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis. *International Journal of Speech, Language & the Law*, 24(1).

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning.* Berlin and Heidelberg: Springer.

Johnstone, B. (1996). *The linguistic individual: Self-expression in language and linguistics.* Oxford: Oxford University Press.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334.

Koehler, J. J. (2013). Linguistic confusion in court: Evidence from the forensic sciences. *Brooklyn Law School's Journal of Law & Policy*, 21(2), 515–540.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9–26.

Labov, W. (1972). *Sociolinguistic patterns.* Oxford: Basil Blackwell.

McEnery, T. and Wilson, A. (2001). *Corpus linguistics.* Edinburgh: Edinburgh University Press.

Nolan, F. (1983). *The phonetic bases of speaker recognition.* Cambridge: Cambridge University Press.

Queralt, S. (2014). Acerca de la prueba lingüística en atribución de autoría hoy. *En Revista de Llengua i Dret*, 62, 35–48.

Queralt, S., Spassova, M. and Turell, M. T. (2011). L'ús de les combinacions de seqüències de categories gramaticals com a nova tècnica de comparació forense de textos escrits. *LSC- Llengua, societat i comunicació*, 9(59-67).

Queralt, S. and Turell, M. T. (2012). Testing the discriminatory potential of sequences of linguistic categories (n-grams) in Spanish, Catalan and English corpora. In *Regional Conference of the International Association of Forensic Linguists 2012*, Kuala Lumpur, Malaysia: University of Malaya.

Rose, P. (2002). *Forensic speaker identification.* London and New York: Taylor and Francis.

Schmied, J. (1993). Qualitative and quantitative research approaches to English relative constructions. In C. Souter and E. Atwell, Eds., *Corpus-based computational linguistics.* Amsterdam: Rodopi, 85–96.

Sjerps, M. and Biesheuvel, D. B. (2007). The interpretation of conventional and 'Bayesian' verbal scales for expressing expert opinion: A small experiment among jurists. *International Journal of Speech Language and the Law*, 6(2), 214–227.

Spassova, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español.* Phd, Universitat Pompeu Fabra, Spain.

Spassova, M. S. and Grant, T. D. (2008). Categorizing spanish written texts by author gender and origin by means of morpho-syntactic trigrams: Some observations on method's feasibility of application for linguistic profiling. In *Curriculum, Language and the Law Inter-University Centre*, Dubrovnik: University of Zagreb.

Spassova, M. S. and Turell, M. T. (2007). The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution. In M. T. Turell, M. S. Spassova and J. Cicres, Eds., *Proceedings of the second european IAFL conference on forensic linguistics, language and the law*, 229–237, Barcelona: Publicacions de l'IULA.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Tagliamonte, S. A. (2006). *Analysing sociolinguistic variation.* Cambridge: Cambridge University Press.

Turell, M. T. (1995). *La sociolingüística de la variació.* Barcelona: Promociones y Publicaciones Universitarias.

Turell, M. T. (2004a). Textual kidnapping revisited: The case of plagarism in literary translation. *International Journal of Speech Language and the Law*, 11(1), 1–26.

Turell, M. T. (2004b). The disputed authorship of electronic mail: Linguistic, stylistic and pragmatic markers in short texts. In *First European IAFL Conference on Forensic Linguistics, Language and Law*, Cardiff: Cardiff University.

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

Woolls, D. and Coulthard, M. (1998). Tools for the Trade. *International Journal of Speech, Language and the Law*, 5(1), 33–57.

Wright, D. (2013). Stylistic variation within genre conventions in the Enron email corpus: developing a textsensitive methodology for authorship research. *International Journal of Speech Language and the Law*, 20(1), 45–75.