A semi-automatic authorship attribution technique applied to real forensic cases involving Judgments in Spanish

Sheila Queralt Estevez and M. Teresa Turell Julià *

Abstract. *Recent studies in forensic authorship attribution report on newly developed techniques, although it seems that, in this quest for valid and reliable identification markers, syntactic structure has shown to be less appealing, which is easily explained by the fact that syntactic variables are clearly more complex, more difficult to process and less frequent than variables at other linguistic levels. This paper presents a series of experiments in forensic authorship, whose aim is to evaluate the discriminatory potential of sequences of linguistic categories (POS n-grams) by using real forensic legal texts written in Ecuador Spanish. The hypotheses tested in these experiments are that a) the most frequent tag sequences will discriminate effectively between authors and b) both bigrams and trigrams will both show this discriminatory capacity. Experiments were carried out by making use of two morpho-syntactically annotated corpora from two real forensic cases, consisting of disputed Judgments (the N for case A = 1; the N for case B = 1) and non-disputed texts (Judgments and other legal texts) from two author candidates in each respective case, with five non-disputed texts for each candidate author in each case. In both cases, a control corpus of non-disputed texts was used, with three authors and five texts per author, and a total of fifteen non-disputed texts. All texts were analysed both qualitatively and quantitavely, in the latter case by running Linear Discriminant Analysis. Preliminary results confirm the hypotheses for both bigrams and trigrams in each case.*

Keywords: N-grams, forensic written text comparison, authorship attribution, Spanish, forensic real cases.

Introduction

Language reflects a series of linguistic traits that can be used in authorship attribution contexts. So far there is not a single method or technique that can be used in forensic

*IULA, Universitat Pompeu Fabra

analyses or in expert witness consulting. In order to understand the status of authorship attribution it is very important to bear in mind the complementary nature of forensic linguistic evidence very much concerned with methodological reliability.

The cumulative evidence considered in forensic authorship attribution has involved the use of several methods and techniques such as the use of reference corpora, type-token ratios, hapax legomena (de Vel *et al.*, 2001), vocabulary analysis (Hoey, 2005; Turell, 2004a,b; Woolls and Coulthard, 2007), sequences of linguistic categories, also called Part-Of-Speech n-grams, in forensic analyses (Bel *et al.*, 2012; Queralt *et al.*, 2011; Queralt and Turell, 2012; Spassova and Turell, 2007; Spassova and Grant, 2008; Spassova, 2009; Turell, 2010), among others.

The forensic linguist's role in forensic text comparison is to observe those linguistic variables and data which might be decisive for determining, among several candidates, who the author of a particular spoken or written text is, which is the research question addressed in the cases presented in this paper. Forensic linguists should base their analyses on valid and reliable methods and techniques by a) undertaking experimental research on real world texts, outside case work, b) applying the same techniques to real forensic case texts, c) using statistical analyses to establish the significance of results, and d) making use of corpus linguistics, and many other approaches, both qualitative and quantitative. Finally, forensic linguists should make this information much more comprehensible to the judge and court.

Recent studies in forensic authorship attribution report on newly developed techniques, although it seems that, in this quest for valid and reliable identification markers, syntactic structure has shown to be less appealing, which is easily explained by the fact that syntactic variables are clearly more complex, more difficult to process and less frequent than variables at other linguistic levels.

This paper presents a series of experiments in authorship attribution, whose aim is to evaluate the discriminatory potential of sequences of linguistic categories (POS n-grams) by using real forensic legal texts written in Ecuador Spanish.

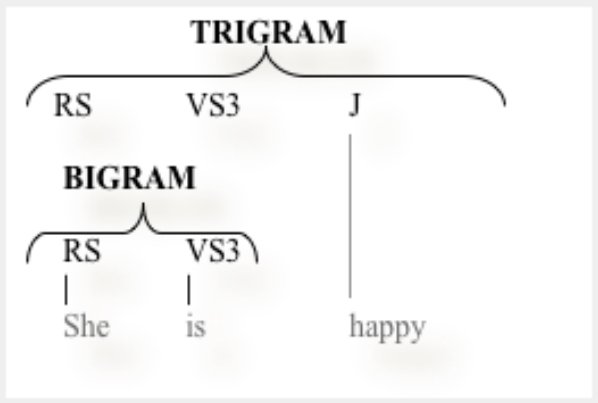
Aim

The theoretical aim of this paper is to show the usefulness of the concept "idiolectal style" (Turell, 2010) in forensic text comparison in its application to the study of non-discrete variables such as sequences of linguistic categories. The methodological aim behind the analysis of two sets of real forensic texts considered in this paper is to establish among several candidates who the author of a written disputed text is in order to help the Court in their decision. The protocol established by our lab, once determined that there is enough linguistic evidence to proceed with the analysis, involves a qualitative and a quantitative approach. In this paper we only focus on the quantitative approach by using the sequences of linguistic categories and its subsequent analysis to establish the statistical significance of the results and to ensure research validity and reliability.

The variable

The variable for our study has to do with sequences of linguistic categories. In previous publications (Bel *et al.*, 2012; Queralt *et al.*, 2011; Queralt and Turell, 2012; Spassova and Turell, 2007; Spassova and Grant, 2008; Spassova, 2009; Turell, 2010) the term Morpho-syntactic Annotated Tag Sequences – MATS was used, but in order to be coherent with the current literature the term Part-of-Speech n-grams (POS-ngrams) is adopted.

Figure 1. Sequences of linguistic categories (POS n-gram).



The sentence: "ella es feliz siempre" (She is happy always) is tagged in Figure 1. Two examples of POS n-grams can be observed, the bigram RS-VS3 and the trigram RS-VS3-JQ where RS stands for singular pronoun, VS3 for third person singular verb and JQ, for qualifying adjective.

Hypotheses

Results reported in previous experiments with real world texts and the application of the Linear Discriminant Analysis (LDA) methods showed the efficient discriminatory potential of POS n-grams. Thus, the hypotheses formulated for both cases are that most frequent tag sequences will discriminate effectively between authors and that bigrams and trigrams will both show this discriminatory capacity, more effectively than other sequences.

The method

This method involves several phases: firstly, a pre-processing phase, in which texts are revised for misspellings or any other possible errors. Secondly, a morpho-syntactic tagging phase, during which the text is converted into a row of token types and tags. Thirdly, a disambiguation stage, through which texts are disambiguated and errors are corrected. Fourthly, a tag extraction phase – during which the information obtained, refers to the number of POS n-grams types and tokens and on to POS n-grams frequency values to be used in the subsequent statistical analysis. And finally, once the tags have been extracted, a last stage involves the application of LDA, in order to have the different text sets classified by author, and have the results projected onto graphs. LDA could be defined as a multivariate statistical technique with three main purposes:

a) To describe whether the use of the n-grams under analysis (bigrams or trigrams) is statistically significant.

b) To determine which are the n-grams that exhibit the highest potential to discriminate between different authors.

c) To predict group membership when we have an unknown text. For example we may have an anonymous text and we want to know who is the most probable author to have produced this text with regards to n-grams use. In order to predict group membership, this technique creates a discriminant function that is the result of a combination of the

n-grams weighted to maximize the difference between the idiolectal style of several authors.

Cases and results

The cases reported in this paper implied the consideration of a Judgment whose authorship was being questioned and two possible author candidates in each case. Thus, the aim of the work undertaken in this analysis was to help the court decide whether the written style observed in the disputed Judgment showed linguistic similarities with the non-disputed Judgments of Candidate 1 or Candidate 2 for each respective case.

Case A

Corpus

As Table 1 summarises, the data for this case consist of a disputed judgment that was divided into 5 excerpts and 5 different judgements of two possible male authors (Candidate 1 and Candidate 2), written in Ecuador Spanish. For accountability and reliability purposes a corpus of three sets of five anonymous Judgements from three judges used in another case was considered, with a similar length and textual structure.

Table 1. Corpus Case A.

Writers	Gender	Genre	Text information 1600 words
Candidate 1	M	Judgment	5
Candidate 2	M	Judgment	5
Control 1	M	Judgment	5
Control 2	M	Judgment	5
Control 3	M	Judgment	5
Disputed text	M	Judgment	5

Results

Bigrams

Figure 2 shows the LDA projection for bigrams applied to the corpus of Case A. In this figure it can be observed that the disputed excerpts are located in the same area of Candidate 1, while the centroid of Candidate 2 occupies a different side of the graph.

Table 2 shows that the LDA classification method successfully classified 100% of the texts by authors within their own group, while the cross-validation method confirmed that the analysis was 96% correct. The five excerpts of the disputed text were attributed to Candidate 1.

Trigrams

Figure 3 shows the LDA projection for trigrams. This figure illustrates that the excerpts of the disputed text are close to the centroid of Candidate 1 while Candidate 2 is placed far from the disputed texts set.

Figure 2. Linear Discriminant Function Analysis based on bigrams – Case A.

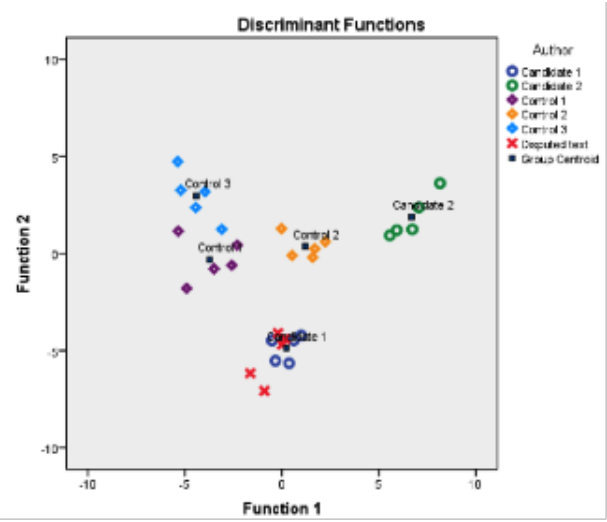


Table 2. Classification and cross-validation results for bigrams – Case A

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	**5**	0	0	0	0	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	1	4	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

96,0% of cross-validated grouped cases correctly classified.

As shown in Table 4, 100% of the texts were classified successfully while the cross-validation confirmed that the analysis was 92.0% correct. Trigrams classified four excerpts of the disputed text to Candidate 1 and one excerpt to one of the control authors.

The interpretation of these statistical results using LDA for POS n-grams revealed that there was a quite high probability that the author of the disputed Judgment was Candidate 1.

Case B

Corpus

As Table 4 summarises, the corpus for this second case consisted of a disputed judgment fragmented into five excerpts from two possible male authors (Candidate 1 and Candidate 2), written also in Ecuador Spanish. In order to optimize the discriminatory potential of the variable under analysis, a set of anonymous Judgements from another case was also used.

Figure 3. Linear Discriminant Function Analysis based on trigrams – Case A.

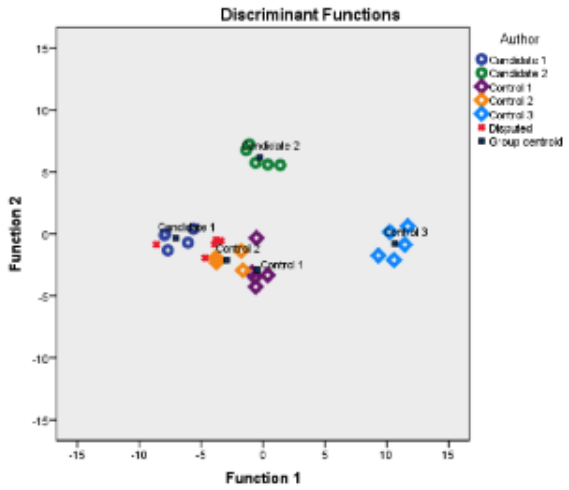


Table 3. Classification and cross-validation results for trigrams – Case A

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	**5**	0	0	0	0	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	5	0	0	0	5
		Control 1	0	0	5	0	0	5
		Control 2	0	0	1	4	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

96,0% of cross-validated grouped cases correctly classified.

Table 4. Corpus Case B.

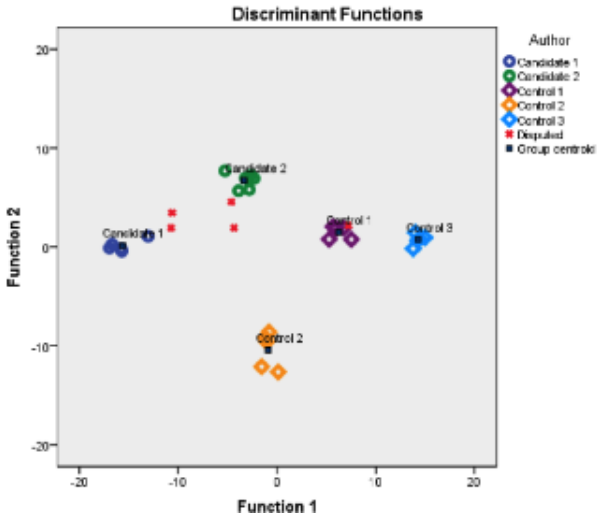
Writers	**Gender**	**Genre**	**Text information 800 words**
Candidate 1	M	Judgment	5
Candidate 2	M	Judgment	6
Control 1	M	Judgment	5
Control 2	M	Judgment	5
Control 3	M	Judgment	5
Disputed text	M	Judgment	5

Results

Bigrams

Figure 4 presents results for bigrams of the LDA applied to Case B text sets. This figure shows that the disputed excerpts are closer to Candidate 2 than to Candidate 1.

Figure 4. Linear Discriminant Function Analysis based on bigrams – Case B.



As shown in Table 5, the LDA classification method classified 100% of the texts correctly and confirmed that the analysis was 100% correct. However, this table illustrates that three of the five disputed excerpts were attributed to Candidate 2, one to Candidate 1 and another one to Control author 3.

Table 5. Classification and cross-validation results for bigrams – Case B

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	**1**	**3**	0	0	**1**	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

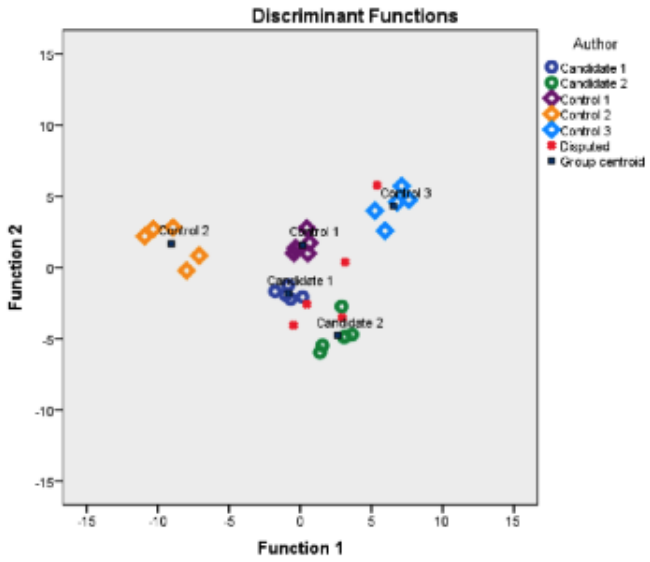
100,0% of cross-validated grouped cases correctly classified.

Trigrams

In Figure 5 the LDA results for trigrams are projected. Most of the disputed excerpts are classified near the centroid of Candidate 2.

The LDA classification method successfully classified 100% of the texts by authors within their own group and the cross-validation method confirmed that the analysis

Figure 5. Linear Discriminant Function Analysis based on trigrams – Case B.



was 100% correct. As Table 6 illustrates, three of the disputed excerpts were attributed to Candidate 2 and two of the excerpts were classified near the control author centroid.

Table 6. Classification and cross-validation results for trigrams – Case B

			Predicted Group Membership					
		Autor	Candidate 1	Candidate 2	Control 1	Control 2	Control 3	Total
Original	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5
		Disputed	0	**3**	**1**	0	**1**	5
Cross-validation	Count	Candidate 1	5	0	0	0	0	5
		Candidate 2	0	6	0	0	0	6
		Control 1	0	0	5	0	0	5
		Control 2	0	0	0	5	0	5
		Control 3	0	0	0	0	5	5

100,0% of original grouped cases correctly classified.

100,0% of cross-validated grouped cases correctly classified.

Results help us conclude that Candidate 1 can be rejected as a possible author of the disputed text and that there exists a moderate probability that the author of the disputed text could be Candidate 2.

Conclusions

We hope that the application of this technique can help forensic linguists to base their analyses on valid and reliable methods and techniques by using, in this case, sequences of linguistic categories in their analyses to be included in their expert witness reports, and to make the information much more comprehensible to the judge and the court, since 80% of the information included in the expert witness's report is usually incorporated

by the judges in their judgments and since the expert witnesses' most important duty is to assist the judge and give reliable forensic linguistic evidence in court.

Our view is that we need to refine our linguistic methods and techniques, and make them as valid and as reliable as possible, so that the unfortunately existing "room for maneuver" is reduce and, in turn, our opinions as forensic linguists are more scientifically grounded.

References

Bel, N., Queralt, S., Spassova, M. S. and Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192–209, Birmingham.

de Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001). Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4), 55–64.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London and New York: Routledge.

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. *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 *The Regional Conference of the International Association of Forensic Linguists*, Kuala Lumpur, Malaysia.

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

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*, University of Zagreb, Dubrovnik (Croatia).

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.

Turell, M. T. (2004a). 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. (2004b). 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. (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. (2007). Tools for the trade. *International Journal of Speech, Language and the Law*, 5(1), 33–57.