Pilot study for the evaluation of linguistic evidence in forensic text comparison by the creation of a Base Rate Knowledge and the implementation of Likelihood Ratios

Sheila Queralt

Universitat Pompeu Fabra, Spain

Researcher
ForensicLab
Universitat Pompeu Fabra
Spain

Awarding Institution:
Universitat Pompeu Fabra, Spain

Date of award: 2015

Keywords: Forensic text comparison, authorship attribution, idiolectal style, Base Rate Knowledge, Likelihood Ratio.

Over the last 20 years courts from several countries 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 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 or difference between those two or more sources was traditionally approached by the knowledge and experience of the linguist. This traditional approach has been conceived subjective to a certain extent considering that it is based on the expert linguist's experience and may vary from expert to expert.

In other forensic sciences that consider evidence such as DNA, fingerprints or handwriting this traditional approach has been consigned to the past. Over the last two decades, the volume of forensic evidence and sophisticated forensic methods has increased dramatically. 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 main goal of this PhD dissertation was to propose the implementation of a methodology protocol within the field of forensic text comparison that improved the reliability of linguistic evidence furnished in Court since it enabled to assess the significance of the findings.

This purpose was achieved by creating a Base Rate Knowledge (BRK) for some of the most pertinent linguistic variables in Peninsular Spanish texts. The creation of the BRK was essential to implement the likelihood ratio framework in forensic text comparison since one must assess the similarity and the typicality of the variables from the known and unknown samples in contrast with a potential population of offenders.

The second step was to select a subset of variables with a high classification potential to carry out the contrast against the population. Thus, an implementation of the likelihood-ratio framework for forensic text comparison was performed, which could improve the reliability of linguistic evidence provided in court and which will offer probabilistic results that could be assessed not only by the judge, but also by the linguistic expert. All in order to conduct more rigorous testing and extensive performance analysis of the data.

The design of the corpus took into consideration the importance of the availability of all the relevant sociolinguistic information of the individuals and its relation to the forensic casuistry. Thus, the corpus collected for the study is a simulation of the forensic reality: letters with threatening content and a relatively short amount of authors and samples per author. Two different corpora were compiled: one for the BRK and another one for the LR. The corpus for the LR was made of 22 man and 25 women, each providing two samples per individual. And the corpus to obtain likelihood ratios comprised 100% of women and 6 letter per each author.

A broad range of linguistic variables were analysed and they can be divided into four main groups: complexity, lexical, syntax and pragmatic. The methodology protocol implemented in this PhD dissertation achieved a correct classification of 75%.

The most important contributions of this proposal were associated with its innovative, original and transferable character and with its reliable results, which will be useful for the field of forensic text comparison:

- The compilation of unified databases of real-world texts in Peninsular Spanish in order to achieve a population distribution (BRK) of linguistic variables.
- A common statistical method based on advanced multivariate statistical methods and the LR framework.
- A first approach to the establishment of a code of good practice in forensic text

> comparison where control factors are considered during the collection of data, there are sampling procedures and quantitative methods are implemented. A new code of good practice can help to provide more reliable and conclusive results in authorship attribution.

This proposal represents a step forward for the needs and research challenges that Forensic Linguistics has faced in the 21st century – for reaching forensic linguistic tests with a degree of reliability as close as possible to other disciplines that consider forensic evidence. Indeed, it opens up a new research direction in forensic text comparison, not yet considered.

Acknowledgements

The proposed PhD dissertation was part of a wider research project entitled 'Towards the consolidation of a similarity index/idiolectal distance (SI/ID) in forensic idiolectometry' (FFI2012-34601) funded by the Spanish Ministry of Economy and Competitiveness and conducted in the context of the Forensic Linguistics Laboratory (ForensicLab) at the Institut Universitari de Lingüística Aplicada (IULA), a research centre at Universitat Pompeu Fabra (UPF).