PhD Abstract

Stylistics versus Statistics: A corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails

David Wright

Nottingham Trent University, UK

Lecturer in Linguistics
School of Arts and Humanities
Nottingham Trent University
UK

Awarding Institution:
School of English, University of Leeds, UK

Date of award: 2015

Keywords: Authorship attribution, author profiling, idiolect, corpus linguistics, Enron, email.

This thesis investigates how a corpus linguistics approach can address the main theoretical and methodological challenges facing the field of forensic authorship analysis. This is pursued through three main research aims: to empirically test the linguistic theory of idiolect; to combine stylistic and statistical techniques in authorship attribution; and to augment quantitative evidence in sociolinguistic author profiling with corpus-driven descriptive analysis. The data used to achieve these aims is the Enron Email Corpus, a collection of 60,000 emails and 2.5 million words written by 176 employees of the former American energy company Enron. This unique corpus, used here for the first time

in forensic linguistics, offers a number of advantages for the analysis of authorship. It contains large amounts of naturally-occurring language data for 176 individually identifiable authors and allows for the investigation of the kinds of digital texts which are becoming increasingly common in forensic casework.

Linguists approach the problem of questioned authorship from the theoretical position that each person has their own distinctive idiolect (Coulthard, 2004: 431). However, the notion of idiolect has come under scrutiny in forensic linguistics over recent years for being too abstract to be of practical use (Grant, 2010; Turell, 2010), given that there is little empirical evidence to substantiate the theory (Kredens, 2002). This thesis, therefore, uses a corpus-based methodology to provide evidence of individual linguistic uniqueness and idiolectal variation. Building on research in corpus linguistics and psycholinguistics (e.g. Schmitt *et al.*, 2004; Hoey, 2005; Mollin, 2009) and forensic linguistics (Coulthard, 2004), the analysis investigates the personal and idiolectal nature of collocation patterns and lexical co-selections in authors' writing. Specifically, the analysis develops the notions of 'Base-Rate Knowledge' (Turell, 2010; Turell and Gavaldà, 2013) and population-level distinctiveness (Grant, 2010) by identifying author-distinctive collocation patterns when individual authors' linguistic choices are compared against those of the remaining 175 Enron employees, who can serve as relevant population data. Analyses reveal that even in shared communicative contexts, and when using very common lexical items, individual Enron employees produce distinctive collocation patterns.

The current situation in forensic authorship attribution research is one in which two competing methodologies have developed. On the one hand, there are qualitative stylistic approaches, and on the other there are statistical 'stylometric' techniques. This thesis demonstrates how a corpus linguistic methodology can combine these two divergent approaches. Building on the evidence of idiolectal collocation patterns, this method uses word n-grams between one and six words in length to capture this individual variation in a quantifiable way. An attribution experiment is performed in which word n-grams in combination with Jaccard's similarity coefficient are used to attribute anonymised email samples of between 4 emails (55 tokens) and 459 emails (14,000 tokens) to their correct authors. An average accuracy rate of 92.64% was returned when attributing the largest samples (100% for certain authors), but success decreases to as low as 17.08% with the smallest samples. That said, the method does correctly identify the authors of anonymised email samples as small as 77, 84 and 109 tokens in length. A main advantage of this approach, computed using a specially-designed program Jangle (Woolls, 2013), is that the analyst can identify specifically which word n-grams are responsible for the accurate attribution of emails. This allows us to isolate a set of lexical sequences which are powerful in identifying a particular employee's idiolect and using that to accurately assign authorship. The method developed here draws together the strengths of both stylistic and statistical techniques and produces an approach in which: (i) there is a clear theoretical motivation for the linguistic features being drawn on in the comparison of authors, (ii) that the similarities and differences between authors, and any subsequent attribution of disputed texts, are based on reliable and replicable statistical techniques, and (iii) that the statistical results produced can be explained and described in linguistic terms.

Finally, this synergy between quantitative and qualitative evidence is applied to the problem of author profiling, which seeks to determine social characteristics of a text's author on the basis of linguistic evidence. Current author profiling research is exclusively statistical in nature, and relies on relative frequencies of various linguistic features to discriminate between authors with different social characteristics such as age, gender, ethnicity and native language. Such work has produced very good results. However, given the over-generalisations necessary to 'categorise' different kinds of authors in this way, Coulthard *et al.* (2011: 538) argue that such methods are 'not certain enough to provide evidence to the courts'. The analysis in this thesis seeks to distinguish between male and female Enron employees, and employees with different occupations, on the basis of their email style. Initial analysis follows the trend of previous research (e.g. Argamon *et al.*, 2003) in using the relative frequencies of a wide range of function and content words to identify statistically significant differences in language use across males and females and employees in different occupations in Enron. Only 35/291 of the features utilised identified a difference between genders and 79/291 discriminated between those in different occupations, revealing that the groups of writers are actually more similar to each other than they are different. Furthermore, a closer qualitative analysis of a small selection of these 'discriminatory' features reveals that authors use particular linguistic features in response to different communicative contexts and functions, and to project different aspects of their identity accordingly, rather than because they are male or female, or because they have a particular role in the company. It is argued, therefore, that author profiling work assumes an over-simplified notion of language and identity which, by contrast, is regularly acknowledged in other fields of linguistics (e.g. Johnstone, 1996; Angouri and Marra, 2011). It is also proposed that quantitative results must be augmented by a descriptive analysis of word use in context to more accurately observe the complex relationship between language and authors' social identities.

The methodological and theoretical contributions of this thesis are various, and it is hoped that they serve as a basis for further developing corpus linguistic approaches to forensic authorship analysis.

References

Angouri, J. and Marra, M. (2011). *Constructing Identities at Work.* London: Palgrave.

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23(3), 321–346.

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 24(4), 431–447.

Coulthard, M., Grant, T. and Kredens, K. (2011). Forensic Linguistics. In R. Wodak, B. Johnstone and P. Kerswill, Eds., *The SAGE Handbook of Sociolinguistics*, 531–544. London: Sage.

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis? In M. Coulthard and A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*, 508–522. London: Routledge.

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

Johnstone, B. (1996). *The Linguistic Individual: Self Expression in Language and Linguistics.* Oxford: Oxford University Press.

Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects. In B. Lewandowska-Tomaszczyk, Ed., *PALC'01: Practical Applications in Language Corpora*, 405–437. Frankfurt am Mein: Peter Lang.

Mollin, S. (2009). ‘I entirely understand’ is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics*, 14(3), 367–392.

Schmitt, N., Grandage, S. and Adolphs, S. N. S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In *Formulaic Sequences: Acquisition, Processing and Use*, 127–151. Amsterdam: John Benjamins.

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.

Turell, M. T. and Gavaldà, N. (2013). Towards an index of idiolectal similitude (or distance) in forensic authorship analysis. *Journal of Law and Policy*, 21(2), 495–514.

Woolls, D. (2013). *CFL Jaccard n-gram Lexical Evaluator (Jangle)*. version 2: CFL Software Limited.