DANA ROEMLING & JACK GRIEVE

FORENSIC AUTHORSHIP ANALYSIS

Despite the prevalence of written language in the digital age, forensic authorship analysis is an underestimated tool in forensic investigations, which can facilitate profiling authors and identifying authorship.

Imagine law enforcement is faced with a ransom note in a kidnapping case. One of the sentences in the note reads '*Put it in the green trash kan on the devil strip at corner of 18th and Carlson.*' You might notice that the author misspelt *kan* or that they correctly used *18th* and capitalised *Carlson*. This type of evidence could help you infer information about the author, although this can be tricky: It might seem like the author has a low education level, given this misspelling, but they spell other difficult words correctly, and may be trying to disguise their identity. Indeed, this is what was found to have happened in this case, while the feature that ultimately broke the case was the phrase *the devil strip*. This phrase is highly regionally bound and primarily used in the city of Akron, Ohio. This information was then used to narrow down the list of suspects.

This type of linguistic analysis is considered to be an application of forensic linguistics, specifically forensic authorship analysis. In general, authorship analysis is concerned with inferring information about the author of a document of questioned authorship. This could be:

a. to determine whether different texts were authored by the same individual, called *authorship verification*,

b. to assess who is the most likely author of a text given a set of potential authors, called *authorship attribution*, or;

c. to infer characteristics about the author by their language use, called *authorship profiling.*

For example, authorship analysis has been used to assess whether a suspect had actually authored their police statements or to determine whether messages sent from a victim's phone were written by their suspected murderer. Limitations for authorship analysis arise through sparse data, genre constraints or texts being written by multiple authors. But, what features help determine the authorship of a text?

ANALYSING AUTHORSHIP

Even though, theoretically, every individual can use language in any way they please so long as they follow linguistic protocols (e.g., "grey green talk dog" is not a sentence that easily conveys meaning), people have preferences of how they use language. This means there is a degree of linguistic individuality, tendencies of using certain words with certain other words. Based on this assumption authorship analysis can generally assess whether texts were authored by the same individual. For example, in the Starbuck murder case the use of semicolons in a series of questioned emails was pivotal for showing that the emails were written by Jamie Starbuck who had murdered his wife, Debbie Starbuck, and then assumed her identity online.

The linguistic analysis found that he was impersonating her, but the usage of semicolons in the disputed emails was less clear at first. In their undisputed emails, Jamie used relatively few semicolons, while Debbie used them with great frequency. In the disputed emails, semicolons were used far more frequently than had even been observed in Debbie's writing. Further examination, however, revealed that the semicolons in the disputed texts were used grammatically in the same way as Jamie, as opposed to Debbie. It was therefore concluded that Jamie had purposely increased his rate of semicolon usage to impersonate Debbie, but had not appreciated the grammatical pattern that characterised Debbie's usage, thereby revealing himself.

> **Jamie had purposely increased his rate of semicolon usage to impersonate Debbie.**

REGIONAL PROFILING

When there is no comparison material, authorship analysis can still provide important insights into the author of a text. Authorship profiling focuses on the linguistic features that let us predict the social characteristics of an author, for example, age or gender. This type of analysis is rooted in sociolinguistics, the analysis of language and its relationship to society. In dialectology, for example, sociolinguists research the regional distribution of language variation. This research can then be applied to forensic authorship questions and be used for regionally profiling an unknown author, which is an exciting area of current research.

> **Authorship profiling focuses on the linguistic features that let us infer characteristics of an author.**

Profiling the regional background of an author can be done through careful, manual analysis and requires the analyst's knowledge about regional dialect variation, as illustrated in the ransom note example above. This task, which is referred to as geolinguistic profiling, can also be based on statistical or computational methods, for example, comparing the language used in an unknown text to patterns of regional variation observed in large collections of social media data.

This is a topic we are currently working on, developing a method for automatically profiling the regional background of the author of a questioned document through the quantitative analysis of large corpora of English and German social media data. Specifically, our approach involves creating a map for each word in a questioned document showing its regional distribution. These maps can be combined into one map, weighing each word map by its regional strength. An aggregated map like this shows how language used on social media would predict the location of the analysed text and could aid law enforcement in their investigations.

Interested practitioners can find more information on forensic linguistics and contact details of forensic linguists through the global forensic linguistics mailing list (http://bit.ly/mail_fl) and the International Association for Forensic and Legal Linguistics (IAFLL.org).

Dana Roemling is a doctoral researcher at the University of Birmingham. Their PhD research focuses on Geolinguistic Authorship Profiling, and they are interested in Authorship Analysis, Language and Law and Lavender Linguistics.

Jack Grieve is a Professor of Corpus Linguistics at the University of Birmingham. His research focuses on Dialectology, Authorship Analysis, Computational Sociolinguistics and Language Change.

Image credit: © *Donatas Dabravolskas* | stock.adobe.com