Decoding Your Email Personality

New York Times Posted online: Mon Jul 25 2011

Ben Zimmer

Imagine, if you will, a young Mark Zuckerberg circa 2003, tapping out e-mail messages from his Harvard dorm room. He never would have guessed that eight years later a multibillion-dollar lawsuit might hinge on whether he capitalized the word "Internet," or whether he spelled "cannot" as one word or two.

But that is exactly the kind of stylistic minutiae analyzed in a lawsuit by Paul Ceglia, owner of a wood-pellet fuel company. Ceglia says that a work-for-hire contract he arranged with 18 year-old Zuckerberg entitles him to half the Facebook fortune. He has backed this up with e-mails "from Zuckerberg" that Facebook's lawyers say are fabrications.

When legal teams need to prove or disprove the authorship of key texts, they call in the forensic linguists. Scholars in the field have tackled the disputed origins of works from Shakespearean sonnets to the Federalist Papers. But how reliably can linguistic experts establish that Person A wrote Document X when Document X is an e-mail — or worse, a terse note sent by instant message or Twitter? After all, e-mails and their ilk give us a much more limited purchase on an author's idiosyncrasies.

The law firm representing Zuckerberg called upon Gerald McMenamin, an emeritus professor of linguistics, to study the alleged Zuckerberg e-mails. McMenamin determined that "it is probable that Zuckerberg is not the author of the questioned writings." He reached his conclusion through a cross-textual comparison of 11 different "style markers," including punctuation, spelling and grammar.

But McMenamin's report has raised eyebrows in the forensic linguistics community. Earlier this month, the outgoing president of the International Association of Forensic Linguists, Ronald R. Butters, publicly questioned whether

McMenamin could actually establish that Zuckerberg likely did not write the e-mails based on such slender evidence. For example, the would-be Zuckerberg e-mails had one instance of uncapitalized "internet," while a sample of e-mails known to be sent by Zuckerberg had two capitalized instances of "Internet." "Are we really doing 'scientific' and 'linguistic' analysis at all when we simply note instances or absences of this or that superficial textual feature?" Butters asked.

Some experts are more optimistic. Carole E. Chaski has developed computer software that categorizes grammatical structures as "marked" and "unmarked": an unmarked noun phrase, for instance, has its main noun at the end of a simple phrase ("our marriage," "a divorce"), while a marked one has the noun in the beginning of a phrase ("anything you ask") or in the middle ("the rest of our lives"). These aspects of a writer's syntax are relatively stable across different styles of writing, Ms. Chaski argues. They are also less prone to technological intervention — compared to spelling and punctuation, which can be changed on the fly by spell-check and autocorrect features.

Benjamin C. M. Fung, who specializes in data mining, and Mourad Debbabi, a cyber-forensics expert, collaborated on a program that can look at an e-mail message and predict who wrote it out of a pool of known authors with an accuracy of 80 to 90 percent. The team identifies bundles of linguistic features, from the position of greetings and farewells in e-mails to the preference of a writer for using symbols (say, "$" or "%") or words ("dollars" or "per cent"). Such features determine what they call a person's "write-print."

Many linguists, however, would challenge the notion that the "fingerprint," a supposedly unique identifier, can be metaphorically applied to writing. Surely we all have our own written quirks and mannerisms — I tend to overuse emdashes, for instance. But there is just too much internal variability in any person's body of writing to imagine that we could take just a bit of it — a handful of e-mails — and recognize some sort of linguistic DNA. That is all the more true when it comes to digital genres like text messages, instant messages and tweets, full of unusual spellings and innovative abbreviations, and often sensitive to the type of device we're using.

Still, these new quantitative approaches hold out the hope of at least differentiating one author from another with a reasonable degree of confidence. This can provide the kind of reliable foundation for research that forensic stylistics as traditionally practiced cannot. Hmm, or is that "can not"?