D21 Proper (and Improper) Handling of Data and Analysis in Forensic Linguistics

Carole E. Chaski, PhD, ALIAS Technology, LLC, Institute for Linguistic Evidence, 25100 Trinity Drive, Georgetown, DE 19947*

After attending this presentation, attendees will be able to evaluate proper or improper handling of data in forensic linguistics (forensic stylistics, forensic computational linguistics, or forensic natural language engineering).

This presentation will impact the forensic science community by providing principles for proper data handling for forensic linguistics and other pattern recognition techniques and ways to recognize when data is being improperly handled or analyses are improperly conducted.

Improper handling of linguistic data and analysis impedes the progress of forensic linguistics and its acceptance as a legitimate forensic science. Data management can affect admissibility based on case law and federal/state rules of evidence because proper data handling enables methods to meet legal standards. This presentation covers principles of data management in linguistics, computer science, and forensic science, including ground truth data, human subjects protection, data scarcity, data ill-formedness, data contamination, and statistical analysis. Linguistic (text) data has specific requirements for proper management, but these principles apply to many kinds of forensic techniques.

What kind of data is needed? Ground truth data has known characteristics relevant for a specific task. In authorship identification work, ground truth data would be a set of texts whose authorship is known and verified. It is important to secure ground truth data, but difficult to do. At least one State Superior Court excluded the Federal Bureau of Investigation's (FBI's) Behavioral Analysis Unit (BAU) Communicated Threat Assessment Database (CTAD) due to ground truth problems (*New Jersey v McGuire*). Some advocate using the internet for authorship Identification (ID), but electronic authorial suspicions arise precisely because screen names are pseudonymous.[1-2] Alternative ground truth datasets do exist and are still needed for validation testing.[3-5]

What regulations apply to data sources? Human Subjects Protection (45 Code of Federal Regulations (CFR) 46) defines the standard practice for linguists, but sharing forensic linguistic data must follow both CFR restrictions as well as policing and legal policies. Ethical issues arise in the collection of suicide notes, threats, and predatory chats, including data collection methods, legality, and chain of evidence. Further, discussion of any case while it is still in adjudication is unethical and a potential obstacle to a fair trial. In *Tennessee v Potter*, a forensic stylist was excluded from testifying but presented a talk about the case while the case was in trial. In the JonBenet Ramsey case, a forensic stylist working for the prosecution provided his analysis to *The New York Times* during the grand jury.

What qualities of text data affect method? Three qualities are important: scarcity, ill-formedness, and contaminant-free. Data scarcity is a fact of forensic casework, so methods must select analytical levels to exploit information in minimal amounts of text. Forensic data is measured in the tens and hundreds, not hundred-thousands of words. Analytical levels are thus constrained. In the smallest samples, lexis isn't reliable for authorship ID but grapheme and syntax are.[4, 6-11] Lexis, grapheme, and syntax are standard analytical levels; prescriptive grammar is not.

Ill-formedness is another fact of forensic casework, so analytical procedures must perform on messy input while still preserving it. Spelling, syntax, or punctuation should not be "corrected" by the linguist because this changes

the data, but some have. Standard sociolinguistics preserves data, no matter how ill-formed it may seem to the analyst.[12] In a habeas corpus case, two forensic stylists were engaged by the plaintiff and calculated wildly different sentence lengths; when questioned, one explained that he had "corrected the punctuation" of the data. This is not standard practice in linguistics or forensic science.

Correcting ill-formedness is close to contamination, which occurs when multiple unknowns are assumed to be from the same source and treated as known examples of one unknown source. Data should remain contaminant-free. Samples, whether blood or text, should never be mixed, although stylists regularly mix texts.[13] Alternatively, multiple unknowns can be hypothesized to come from one source, but not assumed to be; they can be tested for internal consistency as a single-source, but only if the expert report makes it clear that such a test has occurred, as in *BWI v John Doe*.

Finally, what statistical practice is required? Statistical analysis should proceed by normal rules for particular statistical procedures. While it is true that a few statistical procedures will still work well even if a requirement is violated, the multiplication rule will not work accurately if its requirement of independence is violated. In an immigration case, a computational linguist applied the multiplication rule on dependent data so that he could get, in his own words, the probability that the attorney requested, below .05. After being questioned about this, the analyst called it, in print, "statistical hand-waving."

Reference(s):

1. Coulthard, Malcolm; Kredens, Krystof. 2012. Corpus linguistics in author identification. In Peter Tiersman and Lawrence M. Solan (eds). *The Oxford Handbook of Language and Law*. New York: Oxford University Press.
2. Schler, J.; Koppel, Moshe; Argamon, Sean; Pennebaker, James (2006). Effects of Age and Gender on Blogging in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.
3. Iqbal, Farkhund; Hadjidj, Rachid; Fung, Benjamin; Mourad, Debbabi. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*. 5(1) 2008.
4. Chaski, Carole E. 2001. Empirical Evaluation of Language-Based Author Identification Techniques. *Forensic Linguistics: International Journal of Speech, Language and Law*. (8)1.1-64.
5. Chaski, Carole E. 2012. Author identification in the forensic setting. In Peter Tiersma and Lawrence M. Solan (eds). *The Oxford Handbook of Language and Law.* New York: Oxford University Press.
6. Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. New York: Cambridge University Press.
7. Brennan, Michael; Greenstadt, Rachel. 2009. Practical attacks against Author Recognition Techniques. Paper at IAAI, Pasadena, CA.
8. Peng F., Schuurmans D., Wang S. 2003. "Language and Task Independent Text categorization with Simple Language Models." In Proceedings of HLT-NAACL, pp.110-117. Edmonton.
9. Keselj V., Peng F., Cercone N., Thomas C. 2003. "N-Gram-Based Author Profiles for Authorship 9. Attribution." In Proceedings of PACLing'03, Halifax, Canada, pp.255-264.
10. Iqbal, Farkhund; Khan, Liaquat Benjamin; Fung; Benjamin; Mourad Debbabi, Mourad. 2010. Email authorship verification for forensic investigation. SAC2010.

11. Chaski, Carole E. 2005. Who's At The Keyboard? IJDE, Spring.
12. Milroy, Lesley. 1987. *Observing & analyzing natural language*. New York: Blackwell.
13. Inman, Keith, and Rudin, Norah. 2000. *Principles and Practice of Criminalistics: The Profession of Forensic Science.* Boca Raton: CRC Press.

Forensic Linguistics, Text Mining, Data