D56 Comparing Literary, Biometric, and Forensic Approaches to Author Identification

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

After attending this presentation, attendees will understand how to distinguish among the literary, biometric, and forensic approaches to linguistic evidence. Attendees will learn about the four corners of linguistic evidence and how the biometric approach relates to two of these four broad areas. After the presentation, attendees will be able to assess methods and their suitability for use in case investigations and trials.

This presentation will impact the forensic science community by providing a framework for assessing literary, biometric, and forensic approaches to linguistic evidence and how these lessons can be generalized to other forensic disciplines.

Linguistic evidence is evidence in a criminal or civil investigation that hinges on language itself to determine an element of the crime or civil action, such as the authorship of a threatening letter or the speaker of a bomb threat. Generally, linguistic evidence falls into one of four corners: identification, text-typing, intertextuality, and linguistic profiling. Two of these four corners illustrate the intersection of forensic and biometric approaches to linguistic evidence. Identification concerns both speaker (phonetic patterns) and author (lexico-syntactic patterns). Linguistic profiling concerns the sociolinguistic dimensions of language (e.g., age, gender, educational level, dialect, and associates an individual with a group identity).

In some recent high profile cases such as the J.K. Rowling pseudonym, literary author identification has been confused with forensic author identification. Author identification in the forensic setting is difficult and far more difficult than literary author identification for several reasons.[1-3] First, the language of forensic data is less sophisticated and more prone to ungrammaticality than literary language, so that automated parsers make far more mistakes on forensic data than on literary data. Second, the quantity of forensic data is usually much smaller than literary data, as forensic data might consist of tweets and other microtexts or emails while literary data can be tomes and novels. Third, in the forensic setting, the known and questioned documents come from different linguistic registers and genres, while literary data is consistently literary, all novels, all fiction, and all a literary register. These three differences mean that methods developed for the literary authorship question are not transferable to the forensic authorship question; however, the approaches can look similar to the public or to news reporters when the approaches both use computational tools.[1,4] Hence, Chaski makes the distinction between computational stylometry and forensic computational linguistics.[1]

A biometric is a measurement that uniquely identifies or individualizes a person by a biologically based characteristic or behavior. Language is biologically- and neurologically-based; the vocal tract and brain can be measured for unique characteristics. But language is typically seen as a behavioral biometric — it is the behavior of speaking or composing text that is proposed as the identifying and perhaps even individualizing characteristic. Behavioral biometrics are prone to intra-person variability, disguise, and imitation. The biometric and forensic approaches to individuality regarding face, iris, and fingerprint have been contrasted well by Geradts, Ruifrok, and Zoun; Houck and Seigal; and Vorder Bruegge, and these differences apply to author identification.[5-7] The forensic author identification method developed by Chaski is influenced by the biometric approach because it is rooted in a biological/neurological conception of language, especially syntactic theory, and focused on behaviors that are almost impossible to disguise or imitate due to human memory constraints, but it is clearly within standard forensic methodology because it relies on a closed pool of extra-linguistically identified suspects — the line-up model for finding comparable knowns to which to match the questioned item.[1,3,8] Further, like the biometric paradigm, the forensic author identification research conducted for this study is experimentally produced outside of any litigation in a proactive way and returns error rates for specific data quantities/qualities. But the biometric approach raises several research questions that are yet to be addressed adequately: Can author identification select an author in the wild, not in a lineup? Can a Bayesian approach solve this problem? How can error rates be calculated if results in the line-up model are 100% accurate, even under the dampening effect of leave-one-out cross-validation? While forensic author identification must be divorced from literary pursuits, the courtship with the biometric approach deserves a great deal of attention.

* Presenting Author

References:

1. Chaski, C.E. 2013a. "Re Forensic Linguistics: Five data Handling Issues." *Linguistic Society of America Annual Meeting*, Boston, MA.

2. Chaski, C.E. 2013b. "Best Practices and Admissibility in Forensic Author Identification." Journal of Law & Policy, Brooklyn Law School, Brooklyn, New York.

3. Chaski, C.E. 1997. "Who Wrote It? Steps Toward a Science of Authorship Identification." National Institute of Justice Journal. September 1997. Also available through National Criminal Justice Reference Service: http://www.ncjrs.org NCJ 184604.

4. Chaski, C.E. 1998. A Daubert-Inspired Assessment of Current Techniques for Language-Based Author Identification. ILE Technical Report 1098. Also available through National Criminal Justice Reference Service: http://www.ncjrs.org NCJ 172234.

5. Geradts, Z., Ruifrok, A. and Zoun, R 2006. "Forensic Implications of Identity Management Systems. Proceedings of the American Academy of Forensic Sciences, 58th Annual Scientific Meeting, Seattle, WA. 2006.

6. Houck, M. and Seigal, J. 2010. Second Edition. Fundamentals of Forensic Science. Academic Press.

7. Vorder Bruegge, R. 2004. "Biometrics and Forensic Science: What's the Difference?" Biometric Consortium.

8. Chaski, C.E. 2005. "Who's At the Keyboard? Recent Results in Authorship Attribution." International Journal of Digital Evidence. Volume 4:1. Spring 2005. Available at http://www.ijde.org.

Author Identification, Behavioral Biometric, Linguistic Evidence

* Presenting Author