C1 Forensic Linguistics: Curious and Instructive Parallels Between Voiceprints and Forensic Stylistics

Carole E. Chaski, PhD, Institute for Linguistic Evidence, 25100 Trinity Drive, Georgetown, DE 19947; and John R. Middleton, JD, Lowenstein Sandler, 65 Livingston Avenue, Roseland, NJ 07068-1791*

After attending this presentation, attendees will be able to recognize the difference between current computational research in speaker and author identification and popular but misguided attempts, á la CSI effect, at using language data as forensic evidence.

This presentation will impact the forensic and legal communities by providing lessons learned from emerging technologies in novel aspects of forensic identification, showing that when invalid methods are held to legal requirements for scientific evidence such methods can at least be hampered if not totally excluded from the court system.

Forensic linguistics provides linguistic analysis as evidence. In forensic linguistics, the linguist focuses on answering forensically- significant questions which have arisen in a criminal or civil case. The two most common questions are: Who is speaking on this tape? and who wrote this document? Thus, most research in forensic linguistics has focused on speaker identification and author identification. Both speaker and author identification are classification problems; in each, the linguist is seeking a reliable procedure for classifying some speech or some writing to a sample of known speech or known writing.

The histories of speaker identification and author identification show many curious parallels between the voiceprint for speaker identification and forensic stylistics for author identification. Hollien (2002) and Rose (2002) both discuss many theoretical and empirical shortcomings of the voiceprint, a.k.a. aural-visual method, aural- spectrographic identification method. Although academic phoneticians and acoustic engineers rejected the voiceprint, the technique was unfortunately endorsed and used by a prestigious crime laboratory and associated with at least one university. Voiceprint examiners made unbelievable and inflated claims about how many cases they have been involved in, grandiose claims that each human voice is unique and implied that voices were never confusable by voiceprint examiners. The tenacity with which the voiceprint technique lingers, even in the face of empirical evidence repudiating its accuracy and court rulings against its use as testimony is instructive. Author identification has followed much the same path as speaker identification, with forensic stylistics, which is also called by its proponents: text analysis, discourse analysis, sociolinguistics, or psycholinguistics, as the intellectual equivalent of the voiceprint. The world-renowned linguist David Crystal rejected forensic stylistics as linguistics in a review published in *Language*, the prestigious journal of the Linguistic Society of America. Other linguists have also objected to forensic stylistics being represented as linguistics. Some sections of the Federal Bureau of Investigation have adopted and endorsed forensic stylistics, while other sections have recognized the severe limitations of this method and prevented its use as trial evidence. In independent research projects, Chaski (2001, 2007), St. Vincent and Hamilton (2002), and Koppel and Schler (2003) have provided empirical evidence showing that the forensic stylistics method has an extremely low accuracy. Forensic stylistics practitioners have claimed to work in unbelievable numbers of cases, claimed that the each person has a unique set of vaguely defined stylemarkers, and have never produced any empirical evidence in support of their method. Even court rulings which have prevented forensic stylistics testimony from being allowed in trial, or stipulated that the forensic stylistics expert is not an expert in author identification, or restricted the expert so that he can not state an actual opinion about authorship has not stopped the experts from using or attempting to use the method in case and in court.

Meanwhile, there is exciting current computational forensic linguistics research validating methods in both speaker identification and author identification, such as Hollien (2002); Rodman, McAllister, Bitzer, Cepeda, and Abbitt, (2002); Reynolds, Andrews, Campbell, Navratil, Peskin, Adami, Jin, Klusacek, Abramson, Mihaescu, Godfrey, Jones, Xiang, (2003); Rose (2002); Chaski (2005, 2007); Gamon 2004; Stamatatos, Fakotakis, and Kokkinakis (2000, 2001); Diri and Amasyali (2003); Baayen, van Halteran, Neijt and Tweedie (2002). These techniques are meeting the challenges for scientific evidence under both *Daubert* and *Frye*, but still meet resistance in some quarters. Actual reports, court cases, and depositions are presented to support this historical analysis.

References:

Baayen, H. , van Halteran, H., Neijt , A., and Tweedie, F. 2002. “An experiment in authorship attribution.” Journees internationales d’Analyse statisque des Donnes Textuelles 6.

Chaski, C.E. 2001. “Empirical evaluations of language-based author identification techniques.” Forensic Linguistics (International Journal of Speech, Language and Law), Vol. 8:1.

Chaski, C.E. 2007. “The keyboard dilemma and authorship identification.” In Shinoi, S. and Craiger, P. (eds), Advances in Digital Forensics III. New York: Springer. Pp. 133-146.

Crystal, D. 1995. “Review of Gerald McMenamin, Forensic Linguistics.” Language 71 (2) 381-5.

Diri, B. and Amasyali , M.F. 2003. “Automatic author detection for Turkish texts.” ICANN/ICONIP. Available at

 ** Presenting Author*

www.ce.yildiz.edu.tr/mygetfile.php?id=265.

Gamon, M. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland, August 23 - 27, 2004). Association for Computational Linguistics, Morristown, NJ, 611. DOI= http://dx.doi.org/10.3115/1220355.1220443

Hollien, H. 2002. Forensic voice identification. New York: Academic Press.

Koppel , M. and Schler, J. 2003. “Exploiting stylistic idiosyncrasies for authorship attribution.” in Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D., Xiang. B. 2003. The SuperSID project: Exploiting high- level information for high-accuracy speaker recognition.” In Proc. International Conference on Audio, Speech, and Signal Processing Hong Kong. Available at: www.icsi.berkeley.edu/ftp/global/pub/speech/icassp03-peskin.pdf

Rodman, R. D., McAllister, Bitzer, Cepeda, and Abbitt 2002. “Forensic speaker identification based on spectral moments.” Forensic Linguistics (International Journal of Speech, Languaae and Law), Vol. 9:1.

Rose, P. 2002. Forensic speaker identification. Boca Raton: CRC Press.

St. Vincent , S. and Hamilton, T. 2002. “Author identification with simple statistical methods.” Computer Science Department, Swarthmore College.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. 2000. "Automatic Text Categorization in Terms of Genre and Author." Computational Linguistics 26(4): 471-495.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). "Computer- Based Authorship Attribution Without Lexical Measures."

Computers and the Humanities 35: 193-214.

Forensic Linguistics, Natural Language Engineering, *Daubert*

 ** Presenting Author*