B36 Collecting Ground-Truth, Web-Based Data for Research in Forensic Linguistics

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

After attending this presentation, attendees will: (1) learn about both the utility and the perils of using web-based forensic linguistic research for authorship identification, threat assessment, suicide note assessment, and other language-based issues in criminal and civil investigations; and, (2) be introduced to the Institute for Linguistic Evidence Research (ILER), a web-based platform that addresses the dangers of web-based data collection and provides ground-truth data and automated experimental design for forensic linguistic research.

This presentation will impact the forensic science community by providing document examiners, digital forensic evidence examiners, and other researchers with access to another tool for their work, as ILER provides automated experimental design and text analysis. Because ILER provides ground-truth data, the calculation of accurate error rates is possible.

The world wide web provides an enormous amount of linguistic data, especially through social media sites, blogs, and other fora. It can be extremely tempting for linguists and other forensic examiners to simply "scrape the web" for linguistic data. This procedure has three problems: (1) some ethical issues regarding human subjects;[1] (2) the circularity of using unvetted web-data to solve the problem of anonymous web-data;[2] and, (3) the procedure does not guarantee ground truth data (necessary for error rates to be correctly and accurately calculated). As more and more linguistic data is generated electronically, the need for collecting data from electronic media is obvious, since the medium of generation may influence the message, harking back to Marshall McLuhan's communication dictum. Further, as the forensic science community embraces "normal science" and experimental procedures, the need for examining web-based linguistic data is a valid concern, especially when we consider that an experimental paradigm for validation testing requires data collection in a controlled way.[3] This presentation presents a platform developed to address these issues, a platform that enables an experimental paradigm in forensic linguistics to use web-generated data in a way that ground truth data can be collected. ILER is a web-based platform that enables researchers and practitioners to design experiments, recruit subjects and collect vetted data via the internet. ILER includes human subject protections to solve the ethical issues. ILER includes automated experimental design so that practitioners can create experiments using their own on ILER stimuli for the collection of relevant data. ILER is a closed system so that access is monitored, in the same way that laboratory experiments are monitored, so that the identity of the subjects can be monitored as closely as possible, solving the problem of getting ground truth data from the internet. ILER also includes text analysis procedures so that non-linguists can access automated text analysis of the collected data so that the quantification and pattern identification can be analyzed statistically through statistical routines within ILER and from commercial statistical software. Finally, ILER enables alternate means of data aggregation so that the community can share vetted linguistic data for experimental research and validation testing.

References:

1. McEnery, T. and Hardie, A. 2012. *Corpus Linguistics.* New York: Cambridge University Press.
2. Chaski, C.E. 2013. "Best Practices and Admissibility of Forensic Author Identification." *Journal of Law and Policy.* Volume XX1, No. 2.
3. Mnookin, J. et al. 2011. "The Need for a Research Culture in the Forensic Sciences." *UCLA Law Review,* Volume 8.

Forensic Linguistics, Authorship Identification, Web Data

 ** Presenting Author*