D19 A Validation Test of SynAID in Authorship Attribution Using Legal Transcriptions

Timothy Habick, PhD, Reasoning, Inc., Meadowbrook, PA 19046*

Learning Overview: After attending this presentation, attendees will better understand the feasibility of using one method of Syntax-Based Authorship Identification (SynAID) when part of the evidence involves spoken dialogue.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by providing quantitative evidence regarding the limitations, if any, of using well-vetted transcriptions of spoken dialogue for the purpose of syntax-based authorship attribution.

This presentation reports the results of a research project to evaluate the ability of a syntax-based method for authorship identification. The SynAID functionality in the ALIAS software program was used to identify the author of each of six samples of transcriptions of spoken responses during legal proceedings. Each of three individuals from a single speech community produced responses to a lengthy set of questions during a deposition; they then produced further spoken testimony at trial. All responses were transcribed by professional court reporters.

In order for an authorship attribution software system and protocol to be considered capable of identifying the author of a questioned document, it should be proved capable of correctly identifying the author of known documents.[1-3] A statistical model of the known documents can then be used to classify the questioned documents, as is the standard practice in classificatory statistics.[4-6]

This project considers six sets of statements spoken by three individuals. Every sentence in each set of statements is analyzed syntactically, and the part of speech (or syntactic function) of each word is coded. Once the coding is complete, Statistical Package for the Social Sciences' (SPSS's) Discriminant Function Analysis (DFA), following SynAID protocols, is used to construct a model of the speaker's typical syntactic structures.[2,3] In this case, two models for each speaker will be created, one from the deposition and the other from court testimony.

The validation testing for the authorship attribution of transcribed speech will be considered successful if, based solely on the speaker's syntactic patterns and without regard to the content of the speaker's utterances, the deposition testimony of Speaker A is identified as having the same author as the court testimony of Speaker A, and so on for the other two individuals. Further, the validation testing will be considered successful if Speaker A is differentiated from Speaker B, and so on for the other two individuals. Since the material is dialogic, this study also measured the amount of overlap between the interlocutors as an indicator of mutual understanding as well as potential syntactic interference.[7]

Validation testing has been previously performed on SynAID for written documents in cross-genre documents, such as business letters, love letters, apologies, trauma narratives, and essays.[1-3] Further, validation testing has been previously performed on SynAID for text messaging, where all documents were within the one genre of texting, with the texts bundled into chronologically sequential units of 100 texts.[8] This is the first validation testing that involves spoken language rather than written language. SynAID has not previously been tested using transcribed speech samples. This project will thus provide evidence that SynAID is or is not an appropriate tool for authorship attribution using transcriptions of spoken dialogues.

Reference(s):

1. Chaski, C.E. 2001.Empirical Evaluation of Language-Based Author Identification Techniques. *Forensic Linguistics*: *International Journal of Speech, Language and Law*, Volume 8:1. pp. 1-64. June 2001. Published by University of Birmingham, England.
2. 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.*
3. Chaski, C.E. 2013. Best Practices and Admissibility in Forensic Author Identification. *Journal of Law & Policy,* Brooklyn Law School, Brooklyn, New York.
4. Witten, I.H. and Frank, E. 2005. Second Edition. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.
5. Bruce, P. and Bruce, A. (2017), *Practical Statistics for Data Scientists: 50 Essential Concepts.* Boston: O'Reilly.
6. Norusis, Marija J. (1994) *SPSS Professional Statistics* 6.1. Chicago: SPSS Inc.
7. Vogel, C. 2013. Attribution of Mutual Understanding. *Journal of Law and Policy.* Brooklyn Law School, Brooklyn, New York.
8. Basis Technology LLC. (2018) *Alias*. https://www.basistech.com/case-study/alias/.

Authorship Attribution, Quantitative Evidence, Forensic Linguistics

*Presenting Author