D20 Challenges and Opportunities for Integrating Epistemic and Evidential Information Into Author Recognition Systems

E. Allyn Smith, PhD, University of Quebec at Montreal, Montreal, PQ H3C 3P8, CANADA*

Learning Overview: The goal of this presentation is to introduce epistemic modality and linguistic evidentiality to the broader forensic science community, presenting empirical results with respect to the potential fitness of these variables for computational linguistic systems. Attendees should be able to understand and evaluate the relevance of these variables for, as an example, authorship analysis systems.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by presenting theoretical and empirical results related to the integration of semantico-pragmatic variables into existing author recognition systems, putting them in the hands of those who are best positioned to apply them to state-of-the-art computational linguistic systems.

Most successful author recognition systems, whether computational linguistic systems (to be preferred in legal applications) or computational stylometric systems, rely more heavily or even exclusively on morphological and syntactic linguistic variables (including here characters and n-grams), partially because these lend themselves more readily to quantification and computational systems.[1-3] Semantics, roughly the study of the literal meaning of utterances, and pragmatics, the study of meaning in context, are less often present (though mentions incorporating discourse functions, and there is recent interest in moving in this direction).[4,5] Systems sometimes mention semantic information as being important, but this often boils down to word choice, such as identifying low-frequency words occurring at a statistically higher-than-normal frequency in a corpus of texts by a given author. In this case, it is the mere presence of a word and not its meaning as such that is of interest. It is not surprising that meaning tends to be neglected, as it is seen as more ephemeral and, indeed, can be hard to pin down despite being related to the syntax. Many aspects of semantics and pragmatics are also known to be genre-specific, which poses an additional challenge for systems that strive for general application.

The two main steps in true computational linguistic authorship systems are the extraction of variables representing the author's style, and the application of various algorithms to the variables in the first step for the purpose of modeling.[1,6] As explained, efficient variables for these systems will have two features: low within-author variability, and high between-author variability.[4,6] This presentation addresses the variable-extraction step, offering an academic linguist's take on semantic and pragmatic variables that might fit the bill of being variable across authors/speakers and less variable within a given author/speaker.

The variables considered here are epistemic modals and markers of evidentiality. Epistemic modals are words and phrases that speakers use to communicate their degree of certainty or commitment to their utterances (or lack thereof).[7] In English, these might include "I'm sure that," "apparently," "I don't know," or even "ummm" or "well," among many others. These are sometimes also called "hedges" in other literatures, especially when expressing uncertainty. Evidential markers give the basis for a speaker's claim.[8] These include expressions such as "I heard through the grapevine," "I saw," etc. Evidentials can implicate epistemic meanings as well. For example, speakers may choose to specify that their statements are hearsay in order to distance themselves from the information.

In this talk, Empirical results from certain works and previous studies that delve into the application of epistemic modals and evidentials in legal settings will be presented.[9,10] Building on earlier work distinguishing legitimate 911 emergency calls from false calls, corpus studies show that legitimate calls contain statistically more epistemic modals than false calls, including those in which a killer calls and feigns innocence, presumably because real witnesses want to be sure that they do not overstate information they are unsure of in a high-stakes situation.[10,11] Speakers often choose to repeat the subtype of epistemic modals they use, which is a further advantage as it distinguishes them from one another. This study tests whether an authorship identification system could define a group of epistemic modal and evidential phrases and search for their relative usage rather than just searching for the presence/absence of one of the many of them to see whether this variable cluster would correctly distinguish authors.

Reference(s):

1. Chaski, Carole E. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4, no. 1 (2005): 1-13.
2. Chaski, Carole E. Best practices and admissibility of forensic author identification. *JL & Pol'y* 21 (2012): 333.
3. Juola, Patrick. The Rowling Protocol, Steven Bannon, and Rogue POTUS Staff: A Study in Computational Authorship Attribution. *Language and Law = Linguagem e Direito* 5, no. 2 (2019): 77-94.
4. Chaski, Carole E. Who wrote it? Steps toward a science of authorship identification. *National Institute of Justice Journal* 233, no. 233 (1997): 15-22.
5. Chaski, Carole E. *Discourse-based authorship identification in the forensic setting.* Talk given at the International Pragmatics Association meeting, Hong Kong (2019).
6. Krsul, Ivan, and Eugene H. Spafford. Authorship analysis: Identifying the author of a program. *Computers & Security* 16, no. 3 (1997): 233-257.
7. Papafragou, Anna. Epistemic modality and truth conditions. *Lingua* 116, no. 10 (2006): 1688-1702.
8. Aikhenvald, A. *Evidentiality*. Oxford University Press, 2004.
9. Grund, Peter J. The nature of knowledge: Evidence and evidentiality in the witness depositions from the Salem witch trials. *American Speech* 87.1 (2012): 7-38.
10. Rioux-Turcotte, Jessica. *Le souci de fiabilité de l'appelant au 9-1-1: description et fonctions interactionnelles du marquage épistémique.* Dissertation, Université du Québec à Trois-Rivières, 2016.
11. Harpster, Tracy, Susan H. Adams, and John P. Jarvis. Analyzing 911 homicide calls for indicators of guilt or innocence: An exploratory analysis. *Homicide Studies* 13.1 (2009): 69-93.

Computational Linguistics, Semantics, Pragmatics