Digital Methodologies in Forensic Linguistic Authorship Analysis: Social Media Data and Computational Approaches in Geolinguistic Profiling

Dana Roemling
University of Birmingham, UK | University of Helsinki, Finland
danaroemling@gmail.com

Paper Abstract

Research and case work in forensic authorship profiling focuses on inferring social characteristics of unknown authors from their texts, such as age, gender or first language influence, while drawing on foundational work laid in sociolinguistics (Nini, 2018). However, inferring the regional background of an author has received limited attention, although one of the most prominent cases in forensic authorship profiling was resolved recognising the regionalism "devil strip" in a ransom note (see Shuy, 2001).

With computational methods and large corpora of natural language data being available, this study moves away from the traditional approach to geolinguistic profiling by spotting regionalisms and using dictionaries or dialect atlases in the hopes of placing the word in question. For this the study employs a corpus of 21 million German social media posts from the platform Jodel (Hovy & Purschke, 2018) and provides an evaluation of the regionally distributed data in the corpus. Given that geolocated social media data is often sparse and centred on cities, the study uses ordinary kriging (see Wackernagel, 2003), i.e. geospatial statistics, to interpolate the data for unobserved locations, thus enhancing the resolution for location prediction while visualising the results to make them more accessible. Further, the study presents an algorithm to predict the dialect region of an author in question and discusses both the explainability of the prediction in the forensic context and the accuracies reached. The study finds that apart from being a reference tool for qualitative analyses in forensic investigations, this corpus also allows us to extract linguistic features relevant for forensic analyses that are not based on previous knowledge.

Not only does this research advance the field of forensic authorship profiling by reducing the reliance on an analyst's expertise to spot regionalisms, but it also illustrates how interdisciplinary research in linguistics, NLP, digital technologies and forensic science can improve the delivery of justice.

Keywords: authorship analysis, corpus linguistics, dialect classification, forensic linguistics, geospatial statistics

REFERENCES

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on EMNLP*, 4383–4394. https://doi.org/10.18653/v1/D18-1469

Nini, A. (2018). Developing forensic authorship profiling. *Language and Law / Linguagem e Direito*, *5*(2), 38–58.

Shuy, R. W. (2001). DARE's role in linguistic profiling. *DARE Newsletter, 4*(3), 1–5.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer.

Acknowledgements

Dana Roemling was supported by the UKRI ESRC Midlands Graduate School Doctoral Training Partnership ES/P000711/1.

They thank Dirk Hovy and Christoph Purschke for sharing their data with them.

They wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.