div-ling
Diversität in der Linguistik e.V.
Association for Diversity in Linguistics

Dr. Dominic Schmitz
Janina Esser

div-ling.org
contact@div-ling.org

Talk Series: Diversity in Linguistics

31 January 2024
Dana Roemling
Tracing Identity through Language: Geolinguistic Profiling in Forensic Linguistic Authorship Analysis

Within the domain of forensic linguistic authorship analysis, there exist three analytical dimensions: the assessment of shared authorship among distinct texts (verification/clustering), the determination of an author's identity from a pool of potential authors (attribution), and the investigation into the linguistic attributes exhibited by an author (profiling). In this talk I focus on profiling, an area firmly grounded in sociolinguistic theory, which establishes a link between linguistic markers and personal characteristics. Traditionally, research on profiling has focussed attributes such as age, gender (Argamon et al., 2007), and native language (Perkins & Grant, 2018), often relying on qualitative methodologies within the context of the forensic inquiry. However, the exploration of an author's regional origins remains an underdeveloped area within the academic discourse, particularly within the purview of forensic analysis (Chambers, 1990).

With the prevalence of computational methodologies and large corpora of natural language data, this projects moves away from the traditional approach to dialect profiling by spotting regionalisms and using dictionaries or dialect atlases in the hopes of placing the word in question (cf. Shuy, 2001). Commencing with a succinct introduction to authorship profiling, this talk also introduces the expansive corpus of 21 million German social media posts employed in this project (Hovy & Purschke, 2018). I show that regional dialectal variation textualised in this social media data corresponds to dialect areas in current sociolinguistic research. Additionally, I demonstrate that the data is sufficient for regional profiling and show how it can be used in geolinguistic profiling cases by on-the-fly mapping of linguistic examples. The prospect of automating geolinguistic profiling is also taken into consideration. This study contributes to the research in forensic geolinguistic profiling by aiding current qualitative analyses. While it advances current research in German dialectology, it, more importantly, offers a tool to map any word of interest and thus proposes a methodology that is not based on an analyst’s intuition.

References

Argamon, S., Koppel, M., Pennebaker, J.W. & Schler, J (2007). Mining the Blogosphere: Age, Gender and the Varieties of Self-Expression. *First Monday*.

Chambers, J. K. (1990) Forensic Dialectology and the Bear Island Land Claim. *Annals of the New York Academy of Sciences*, 606(1), 19–31.

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on EMNLP*, 4383–4394.

Perkins, R., & Grant, T. (2018). Native language influence detection for forensic authorship analysis: Identifying L1 Persian bloggers. *IJSLL*, *25*(1), 1–20.

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