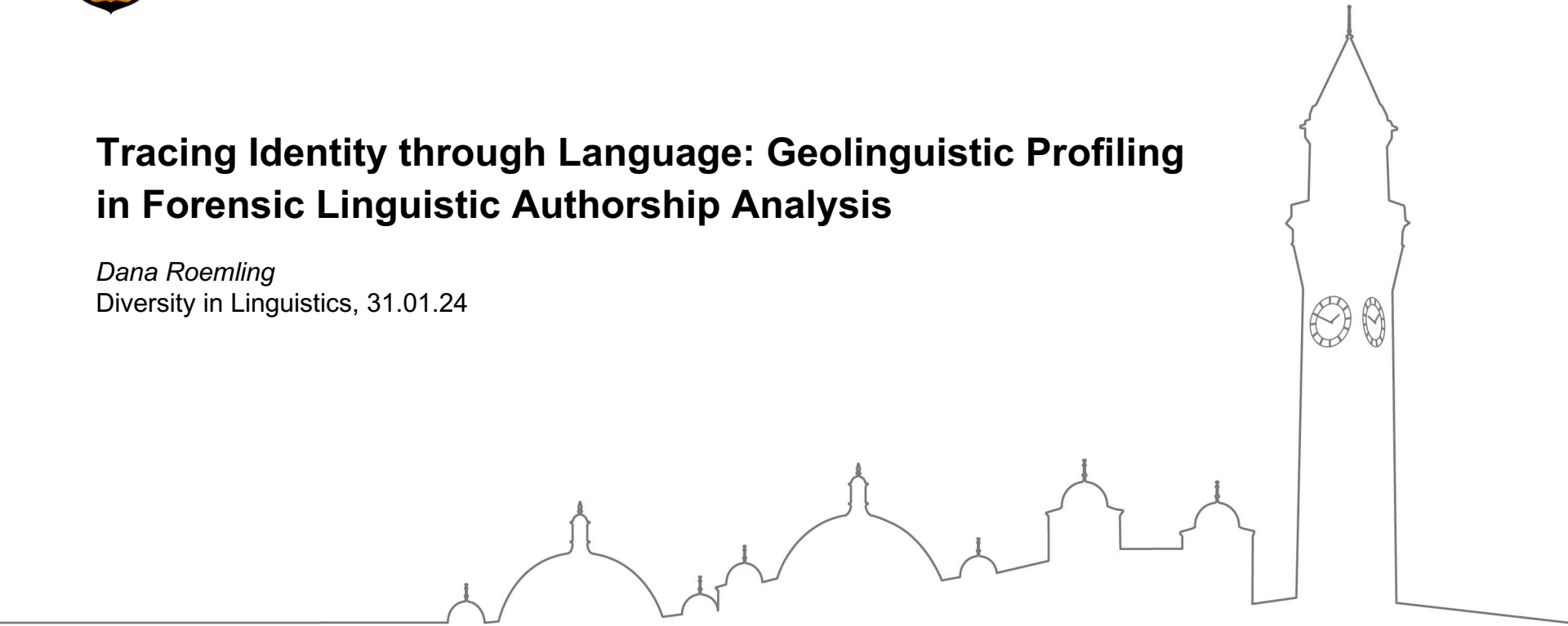
UNIVERSITY OF BIRMINGHAM

Tracing Identity through Language: Geolinguistic Profiling in Forensic Linguistic Authorship Analysis

Dana Roemling

Diversity in Linguistics, 31.01.24



Today

- Quick Intro: Forensic Authorship Analysis & Profiling
- My PhD Project: Dialect / Geolinguistic Profiling
  - Corpus
  - Analysis & (preliminary) Results & (preliminary) Conclusion
  - Forensic Application & Outlook

Content Warning

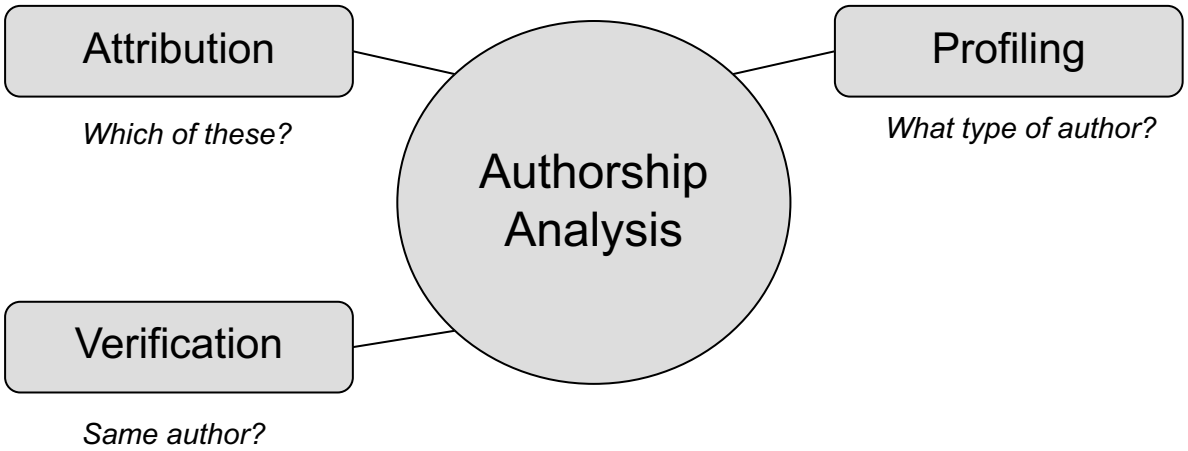
Forensic Linguistics (+ Language & Law)

1) Language as evidence & language in the investigative process
   for example: ransom note, messaging texts in murder investigation

2) Language of the law & language in the legal process
   for example: statutory interpretation, the influence of power & hierarchy in court

Authorship Analysis

- Analysis of texts to infer characteristics about an author
  - e.g. age or gender



Adapted from Wright, 2020

“Do you ever want to see your precious little girl again? Put $10,000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner of 18th and Carlson. Don’t bring anybody along. No kops!! Come alone! I’ll be watching you all the time. Anyone with you, deal is off and dautter is dead!!!”

(Shuy, 2001, p.1)

“Do you ever want to see your precious little girl again? Put $10,000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner of 18th and Carlson. Don’t bring anybody along. No kops!! Come alone! I’ll be watching you all the time. Anyone with you, deal is off and dautter is dead!!!”

(Shuy, 2001, p.1)

“Do you ever want to see your precious little girl again? Put $10,000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner of 18th and Carlson. Don’t bring anybody along. […]”

(Shuy, 2001, p.1)

Devil Strip

devil's strip n Also *devil strip* [Prob from its being a sort of no-man's-land between public and private property; cf **devil's lane**] **chiefly neOH**

The strip of grass and trees between sidewalk and curb.

1957 *AmSp* 32.239 **neOH,** It [=a car] went out of control and jumped the curb, traveling partly on the road and partly on the devil strip. . . [The term] is known throughout the Youngstown, Ohio, area. **1964** *AmSp* 39.293 **neOH,** The Akron term [for the strip of grass or weeds between the sidewalk and the curb] is *Devil strip* or *Devil's strip.* There are a few, however, who think it vulgar or profane (although they recognize it), and to them it is the *berm.* **1966** *DARE* (Qu. N44) Inf SC2, Devil strip. [FW: She [=the Inf] never used it; heard it in Hartsville about 30 miles away. It's supposed to keep the devil out of your house.] **1966** *DARE* File **neOH,** The "parking" or the "boulevard" is known as the "devil's strip" from Cleveland to Youngstown. **1968** *DARE* FW Addit

Dictionary of American Regional English, "devil's strip", 1985

Authorship Profiling

- “[D]etermining the characteristics of an anonymous author, such as their demographic details, from the way they use language”
- The idea is that this can narrow down the pool of suspects
- This is usually done through either:

  The analysis of salient linguistic features *or*

  The analysis of writing style

  (Nini, 2018)

The Analysis of Writing Style

- “[O]ften involves the study of the frequency with which certain features are used, like the study of register variation and takes as the unit of analysis the text itself”
- “A style is a variety of language associated with a particular author or social group […] which is constituted by linguistic features that are pervasive and frequent”
- Problem: “Computational authorship profiling is not necessarily interested in understanding the inner (linguistic) mechanisms of the machine, as long as the accuracy rates are outperforming previous models.”

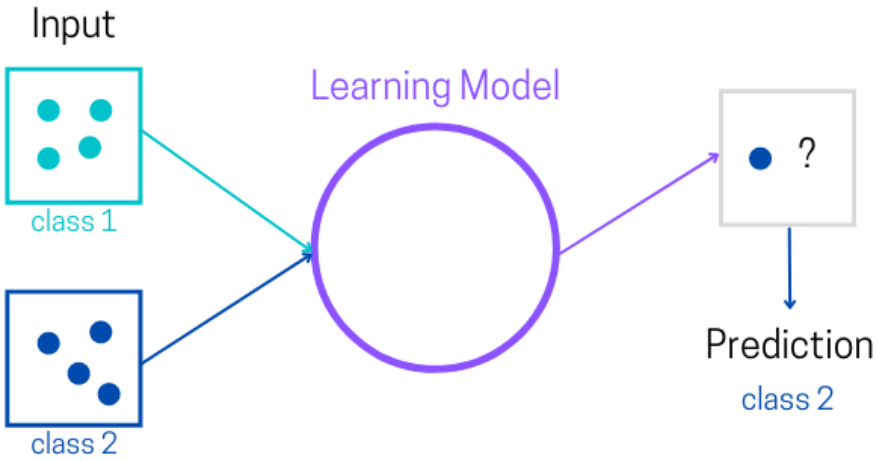
(Nini, 2018)

The Analysis of Salient Linguistic Features

- “[T]he application of sociolinguistic knowledge on a case by case basis to extract *ad hoc* linguistic features that are markers of a certain demographic background”
- “[I]nvolves the linguist’s experience in discovering dialectal or sociolinguistic features that can reveal clues about the background of the author”
- Problem: ”[R]elies almost entirely on the knowledge and intuition of the forensic linguist”

(Nini, 2018)

Best of both worlds: Linguistically informed computational approach

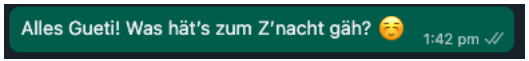


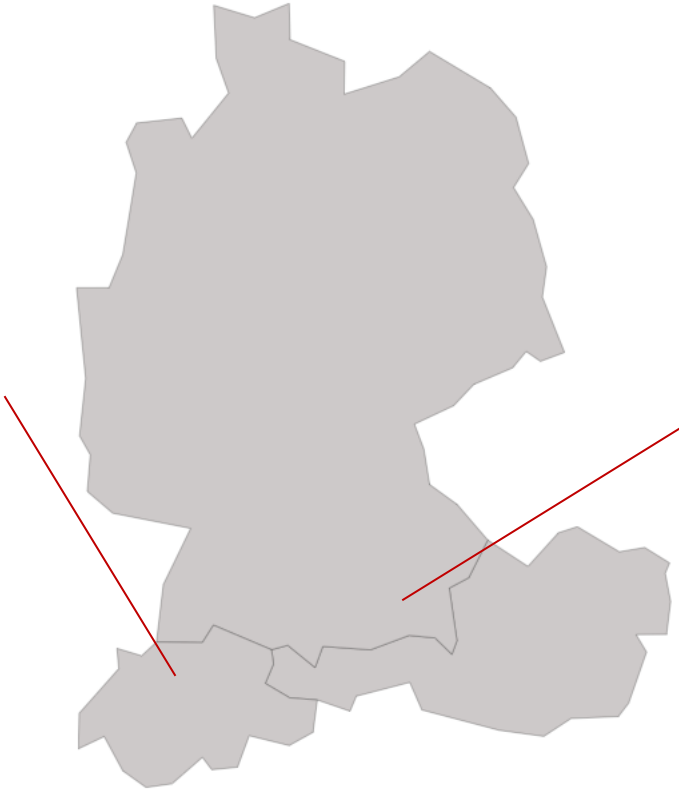
Supervised learning model.
Adapted from Devopedia (2022)

Put $10,000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner of 18th and Carlson

(Shuy, 2001, p.1)

Which regional variety?

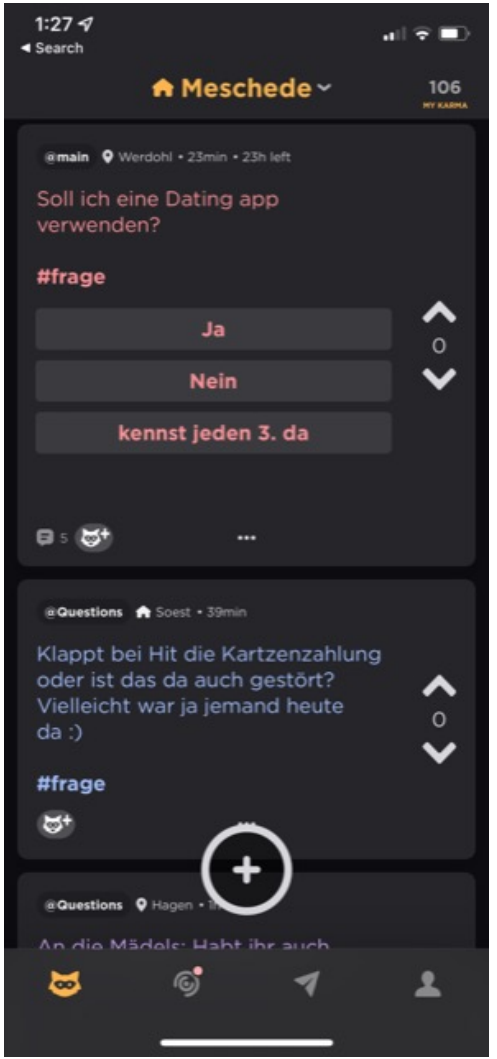




Dialekt, heit oft aa Mundoat, gweanli a 'oatsvaschidnats' Gred oda Gsoad, guit ois a Varietet von ana Sproch, de wo in iran Afkumma, i da raimlichn Asdeanung un im Vaschwinna sozial u historisch o bstimmte Sidlungsraim bundn is. (Wikipedia)

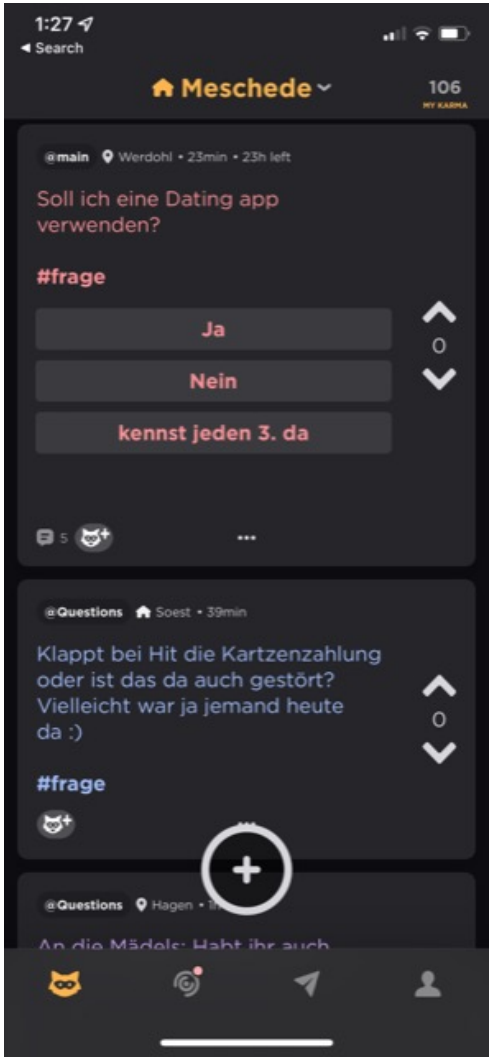
Data

- Jodel: Social Media app
- Interaction in a 10km radius around own location
- Collected 2017 by Hovy & Purschke (2018) to represent German-speaking area (= Austria, Germany, Switzerland)



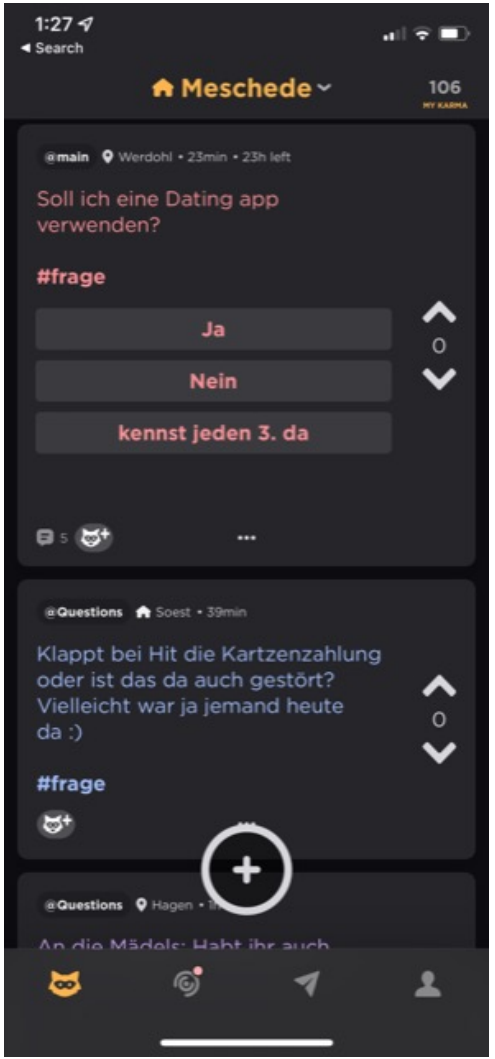
Data

- No accounts / profiles, so “anonymous” interaction
- Demographics:
  - 18-26 → 72%, 27+ → 27%
  - 53% male, 45% female
  - 57% students, 31% employees
- 2.8 million users in the German-speaking area



Data

- The corpus has 239,151,815 tokens at 8147 unique locations in the German-speaking area
- 85% of data in Germany
- Data is split into training (70%), validation (15%) and test (15%), training tokens: 166,538,477

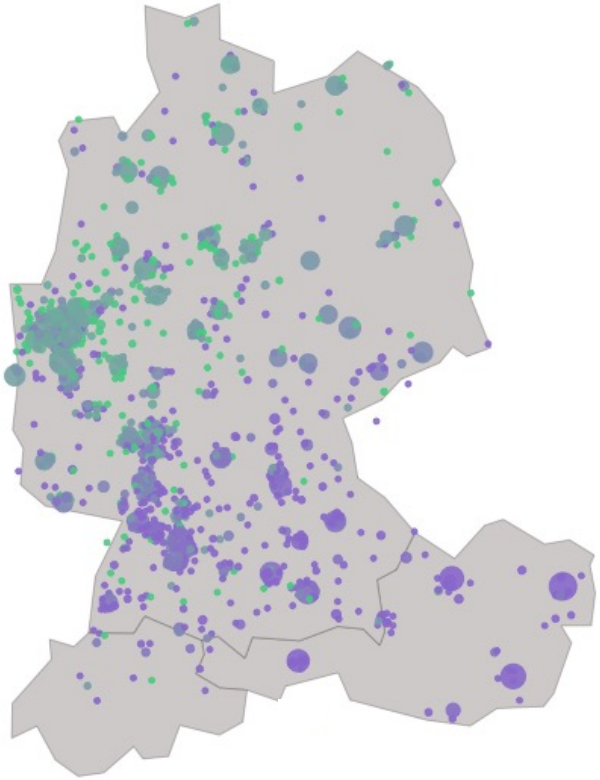


Sample

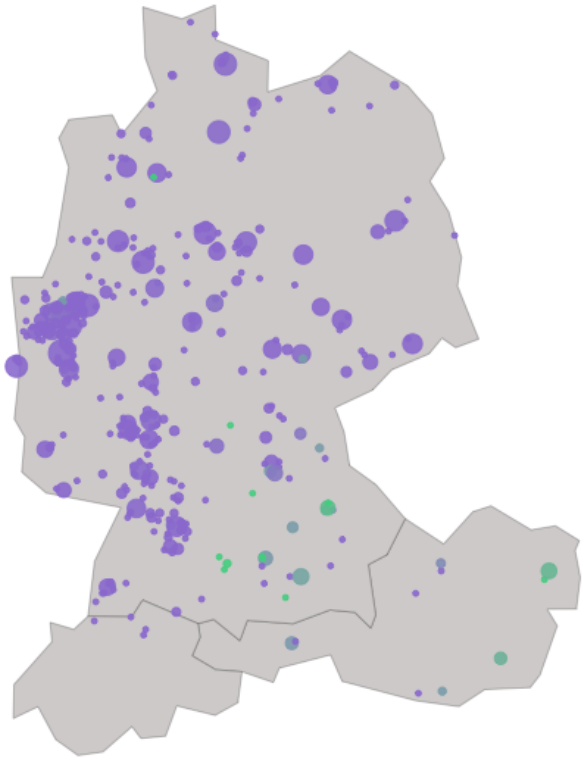
Message	Creation Timestamp	Location	Post ID
Semesterferien: grillen schlafen grillen bar schlafen repeat.. English: *semester break: bbq sleep bbq pub sleep repeat..*	2017-04-04T 22:29:41.814Z	Berlin	58e41e5512e80 a3f0cb6f66b
Ich bin grade leicht verwirrt Werdet ihr Mädels so emotional, kurz bevor ihr eure Tage habt, oder mittendrin? English: *I am slightly confused now Are you girls being emotional just before you're on your periods or in between?*	2017-04-04T 22:04:48.109Z	Hamburg	58e41880a149d 37f12cca9d1

Can we see regional distribution?

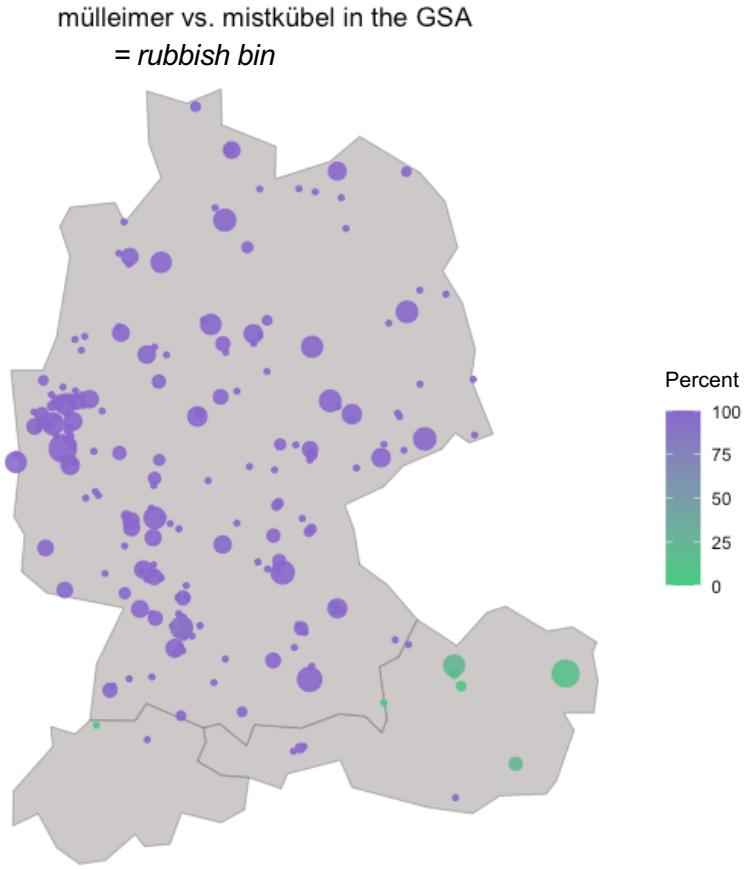
schau vs. guck in the GSA
= *see*

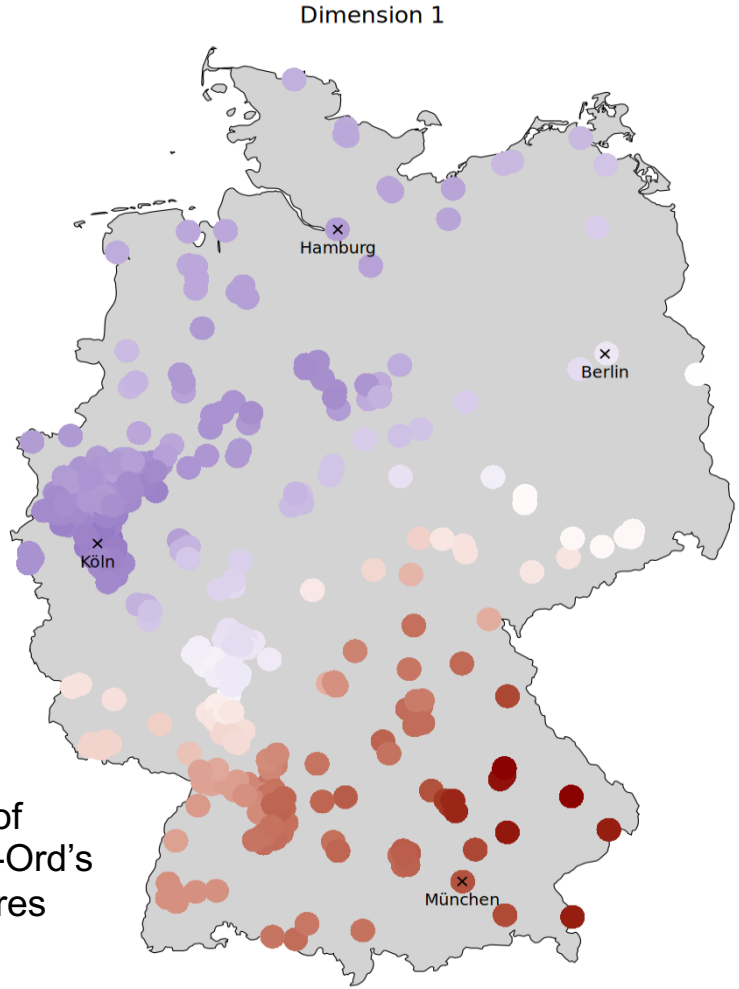


brötchen vs. semmel in the GSA
= *bread roll*

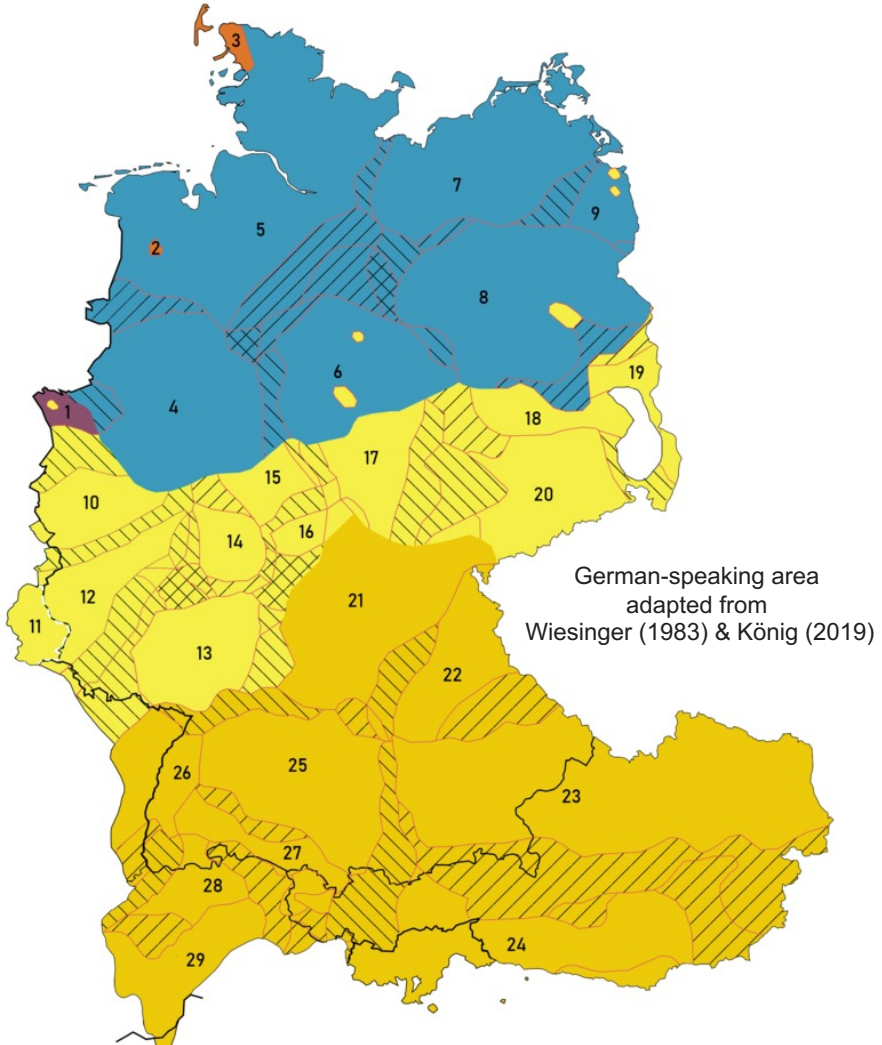


Can we see regional distribution?





PCA of Getis-Ord's z-scores



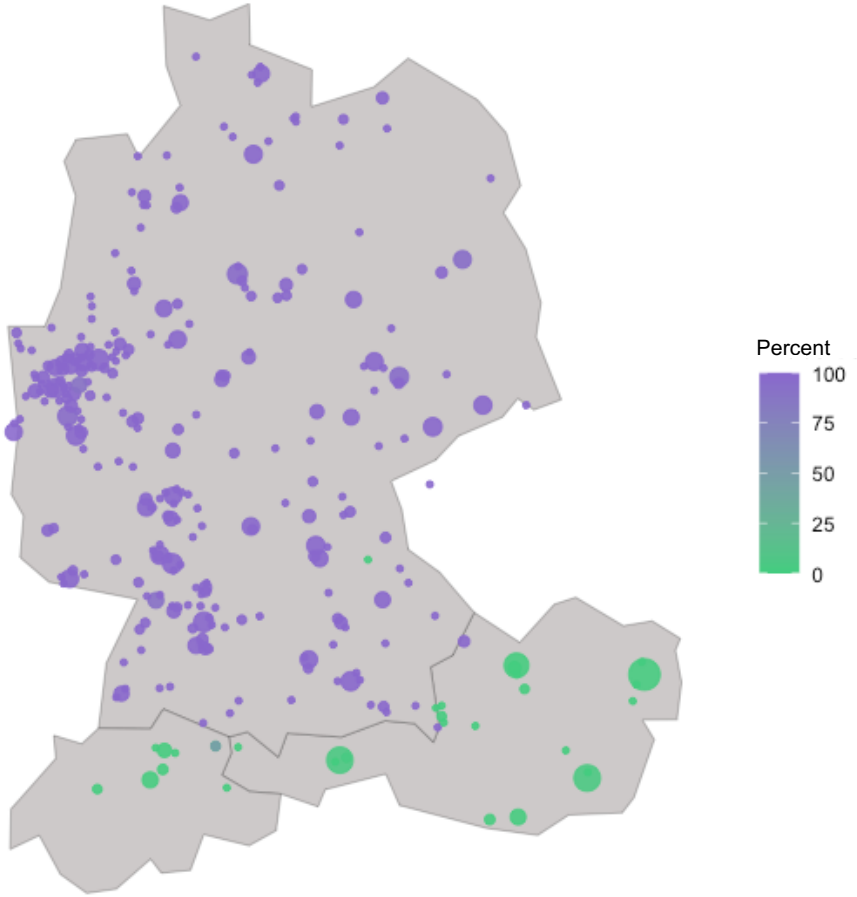
German-speaking area
adapted from
Wiesinger (1983) & König (2019)

Qualitative Forensic Application

- Can I narrow down the region a post might be from given this data set?
  - north/south, east/west, national borders, city level?

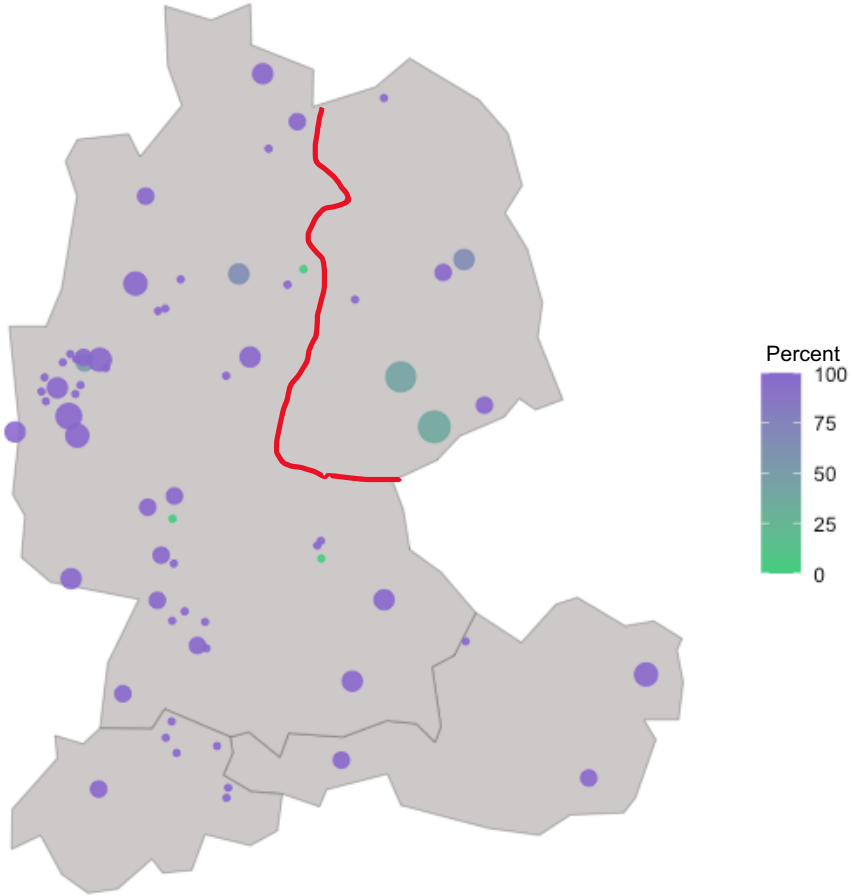
abitur vs. matura in the GSA

= *A-levels*



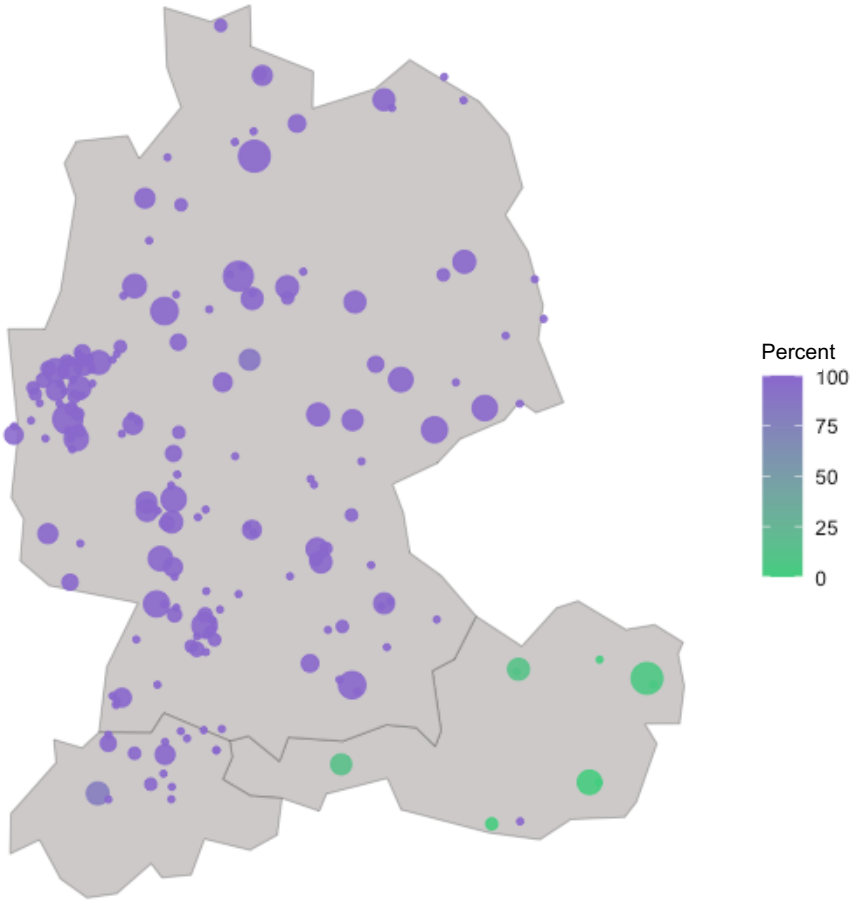
astronaut vs. kosmonaut in the GSA

= *space traveller*



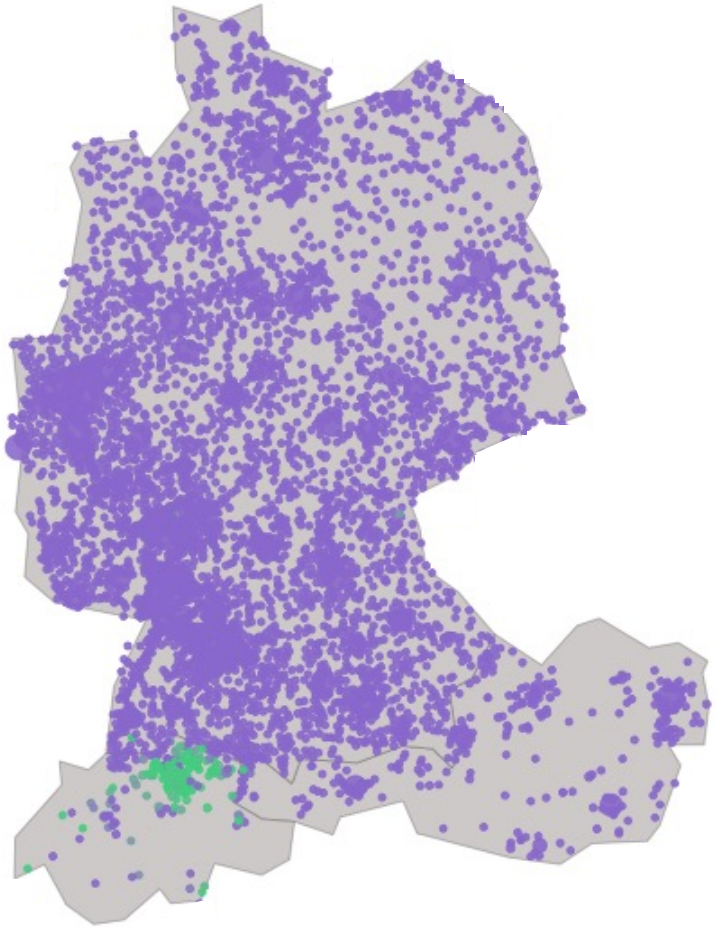
januar vs. jänner in the GSA

= *January*



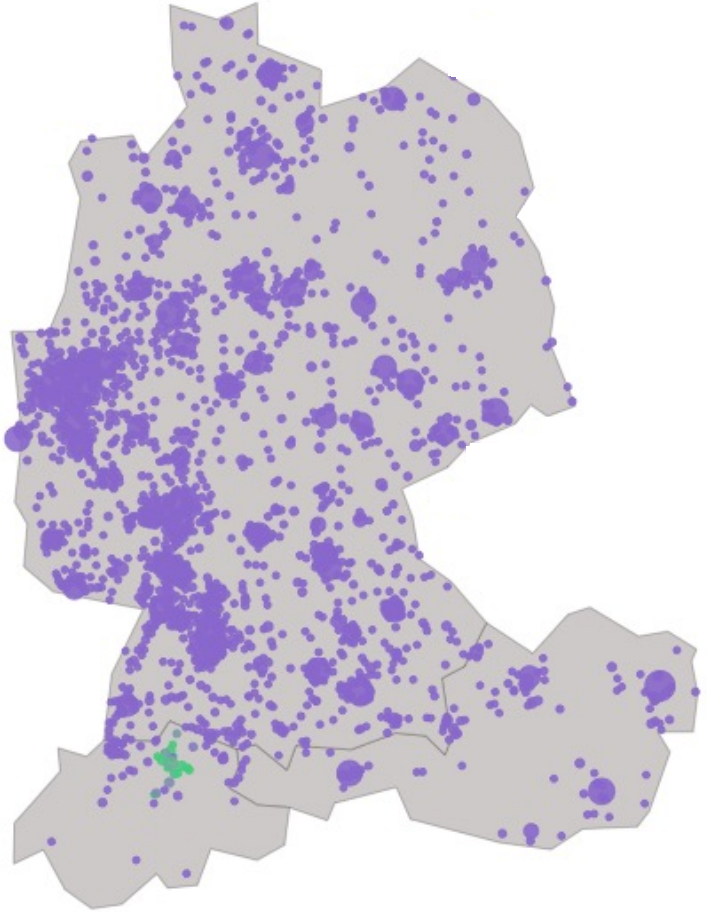
nicht vs. nöd in the GSA

= *not*

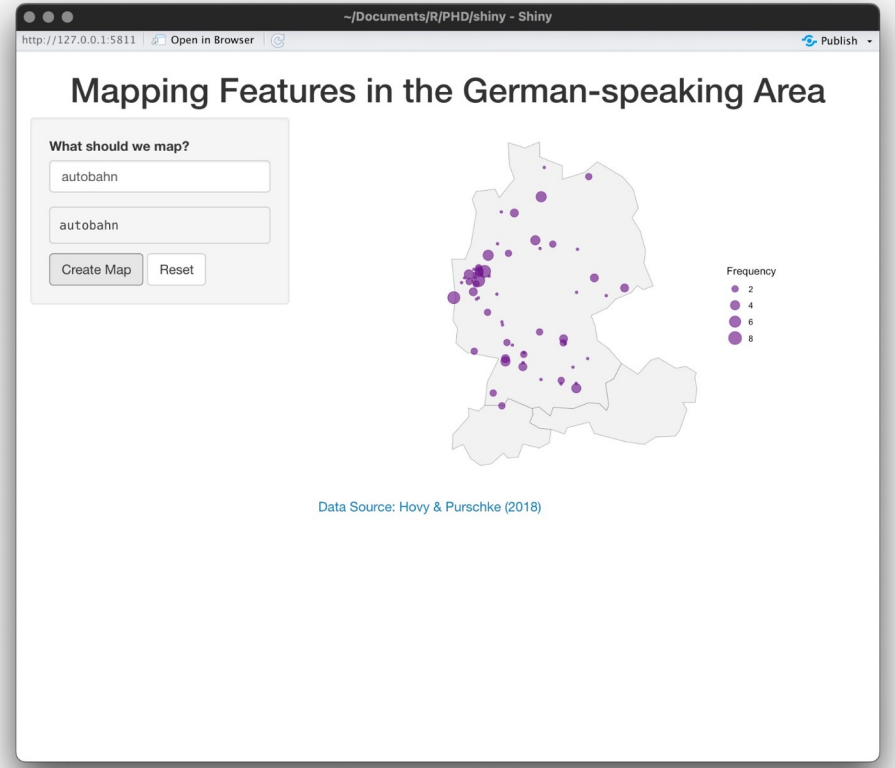


wirklich vs. wüki in the GSA

= *really*



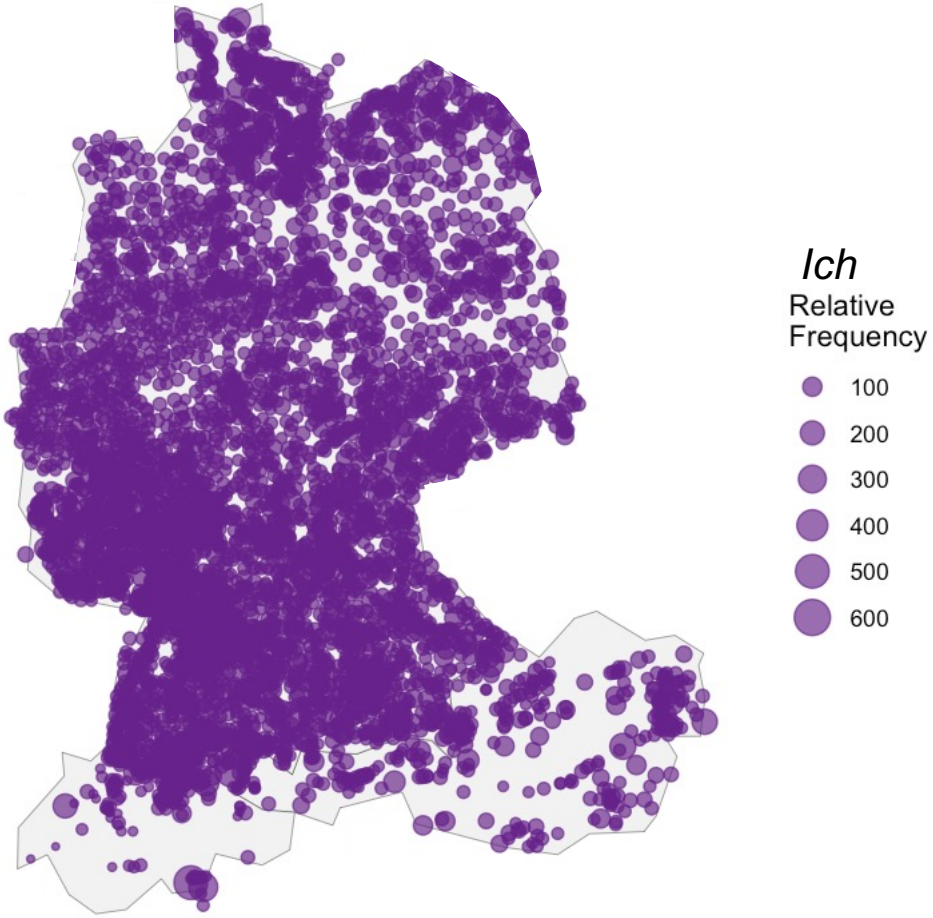
Making this useful in the forensic setting:



Conclusion so far

- The corpus already allows for mapping regional variation in online communication, which means it can be used as a reference tool in forensic investigations to understand geolinguistic distribution
- The corpus follows expected dialectal patterns in the German-speaking area and even in computer-mediated communication there is enough dialectal textualisation to be mapped

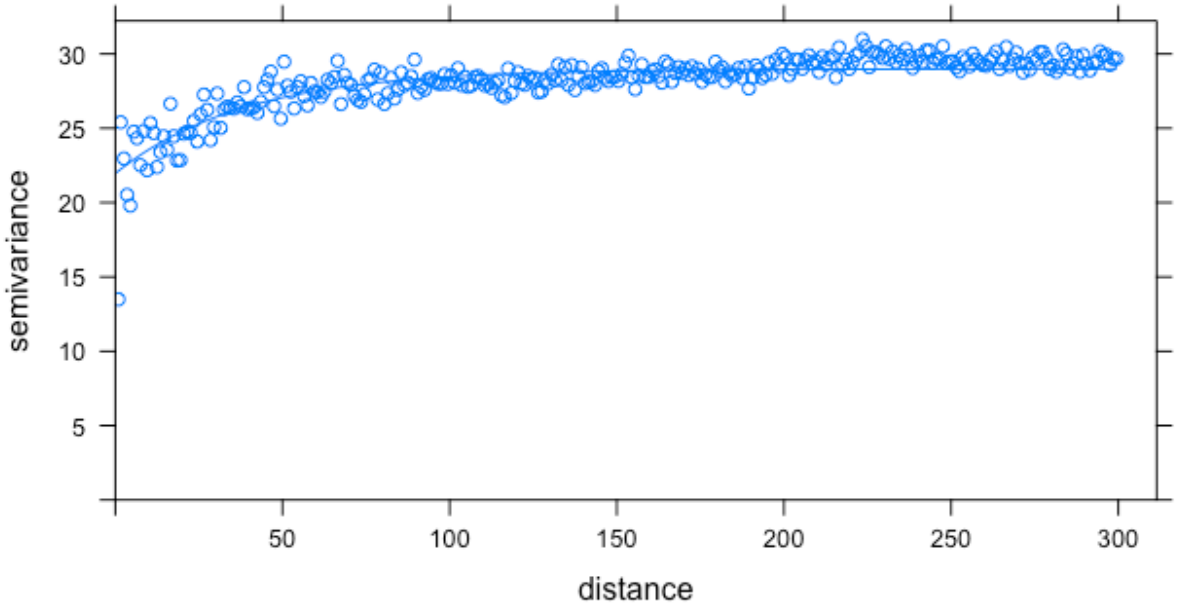
Account for locations without data



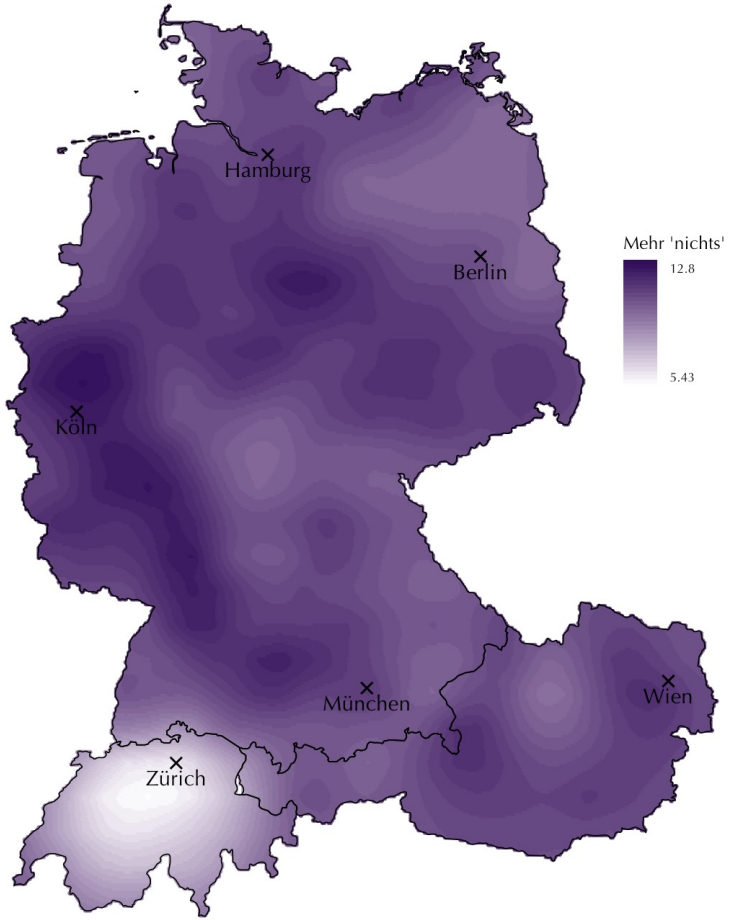
Heeringa & Nerbonne 2001

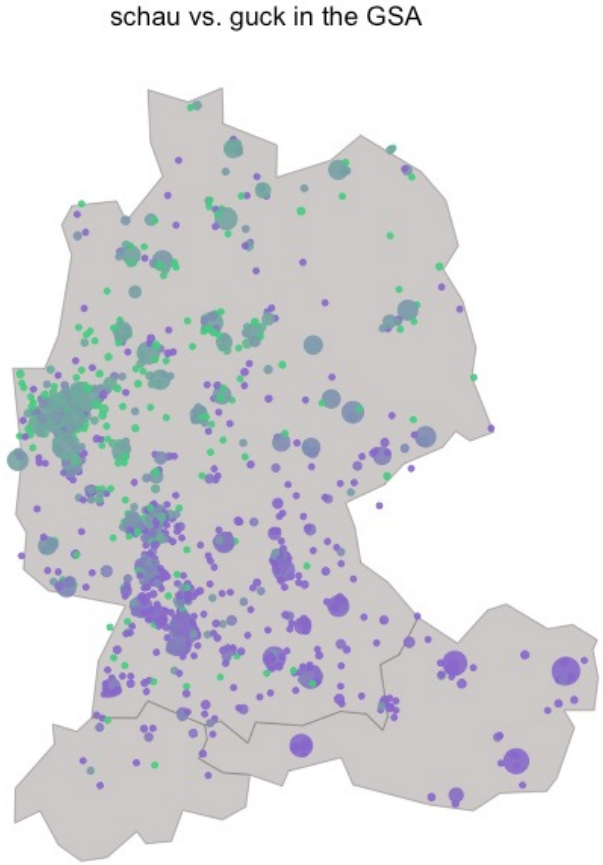
FIGURE 4. Variants of *zijn* ‘to be’ in IPA.

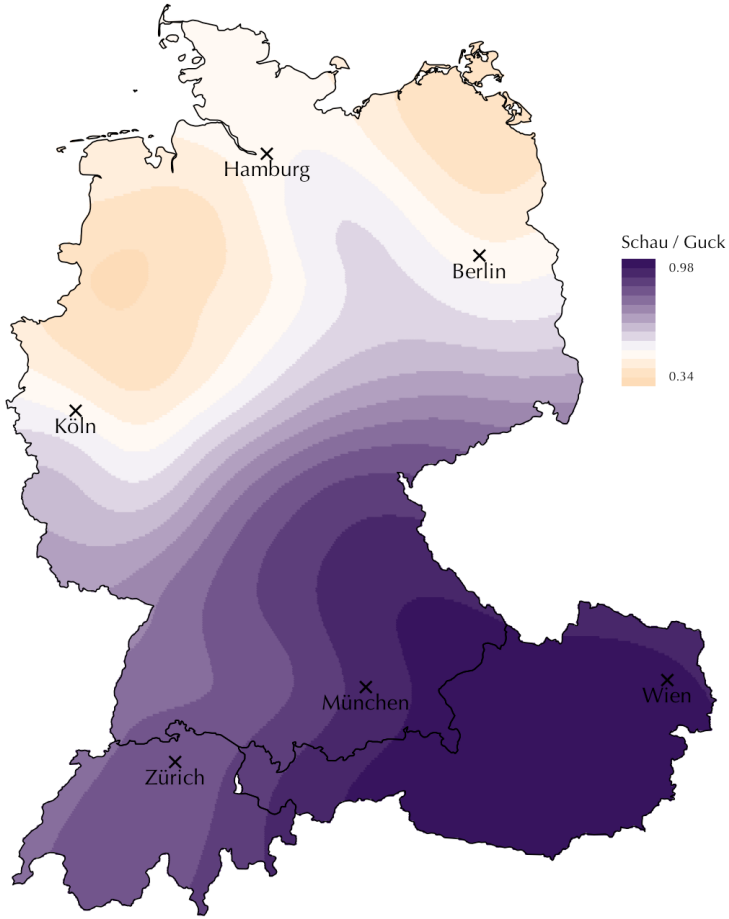
Inferring unobserved values: Variogram models



Kriging


Hamburg
Berlin
Köln
München
Wien
Zürich
Mehr 'nichts'
12.8
5.43


schau vs. guck in the GSA

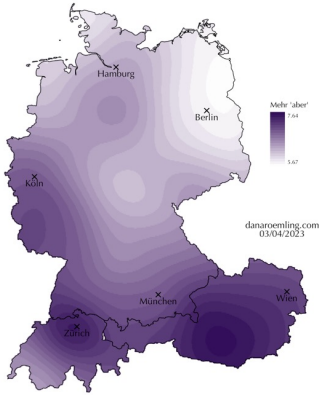

Hamburg
Berlin
Köln
München
Wien
Zürich
Schau / Guck
0.98
0.34

Currently…

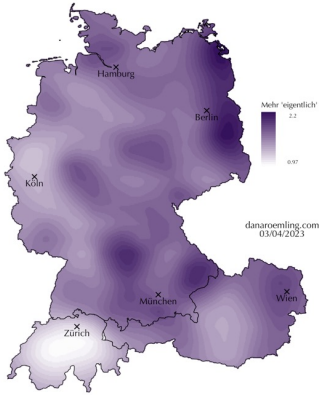
aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😃

1:00 pm

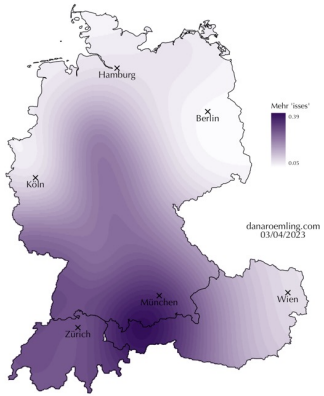
aber



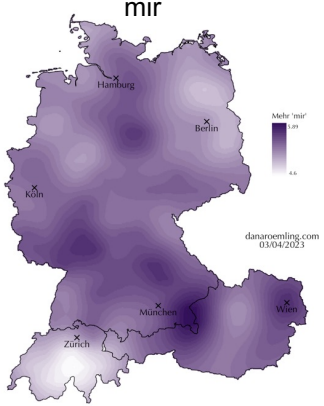
eigentlich



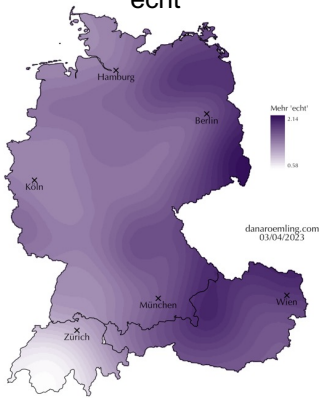
isses



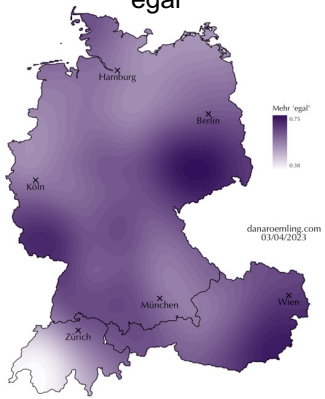
mir



echt



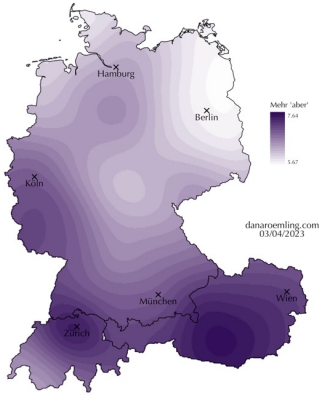
egal

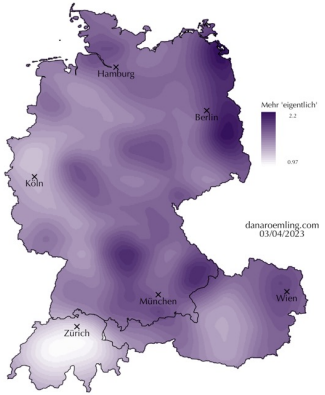


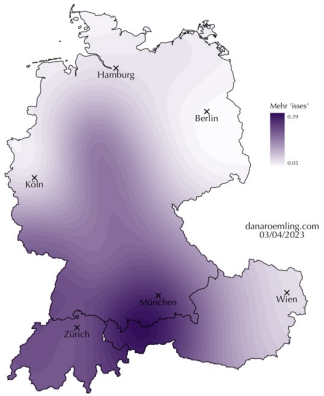
Currently…

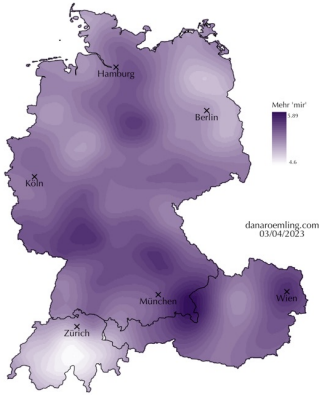
aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😃

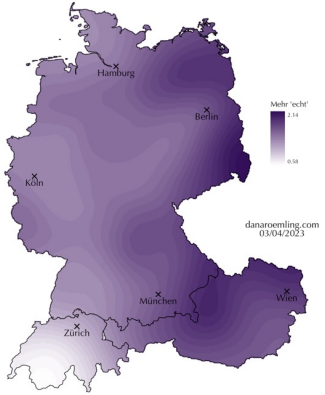
1:00 pm

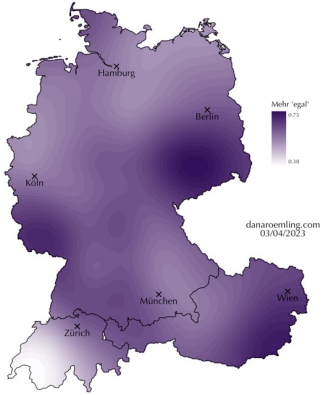








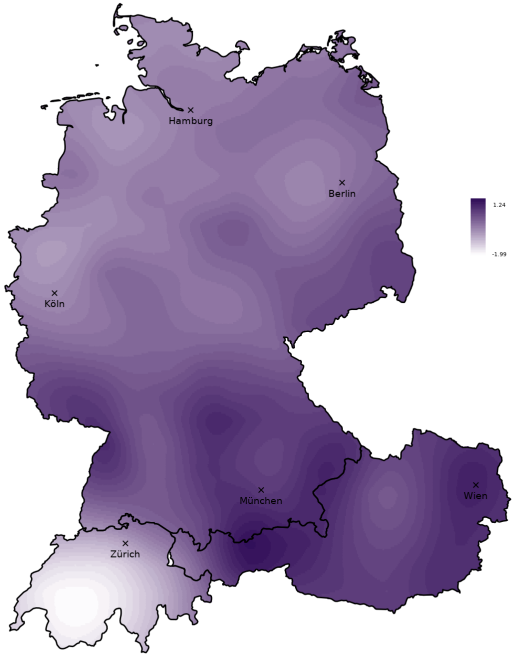




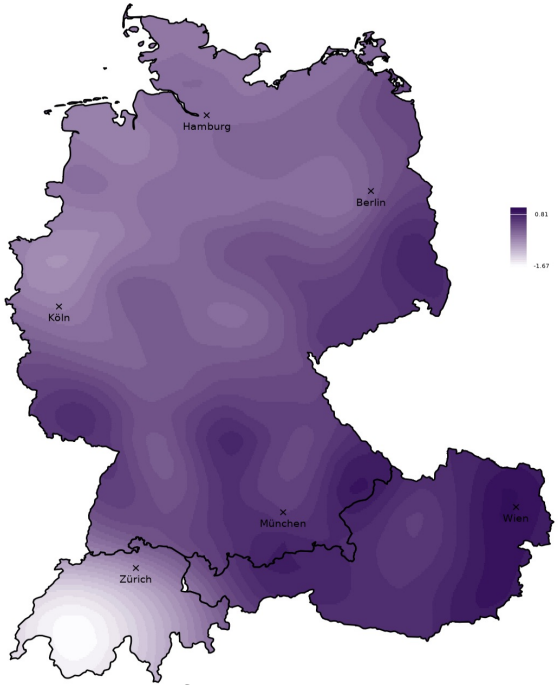
Currently…

aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😃

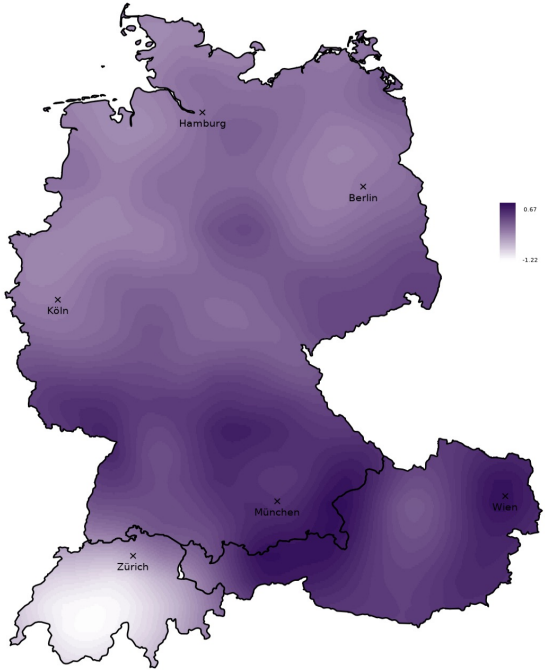
1:00 pm



unweighted



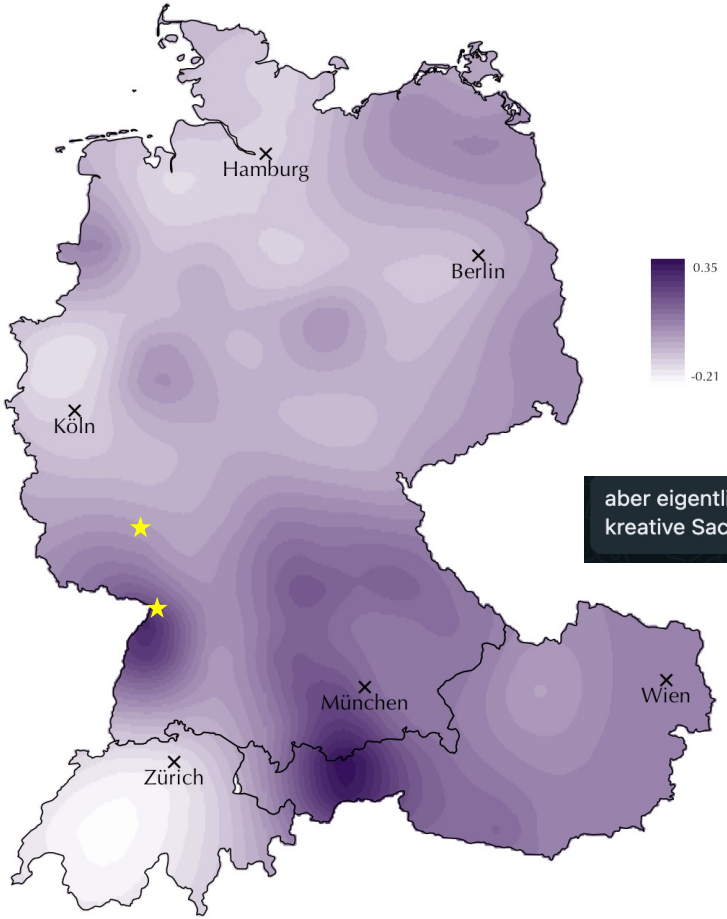
function words



noun(s)

aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😃

1:00 pm



Weighting based on Moran's *I*

aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😃

1:00 pm

Next Steps

- Weighting of features
- … and soon machine learning approaches
- Both for prediction and for feature extraction
- Bearing in mind transparency of methods given the forensic context

Forensic Comparison

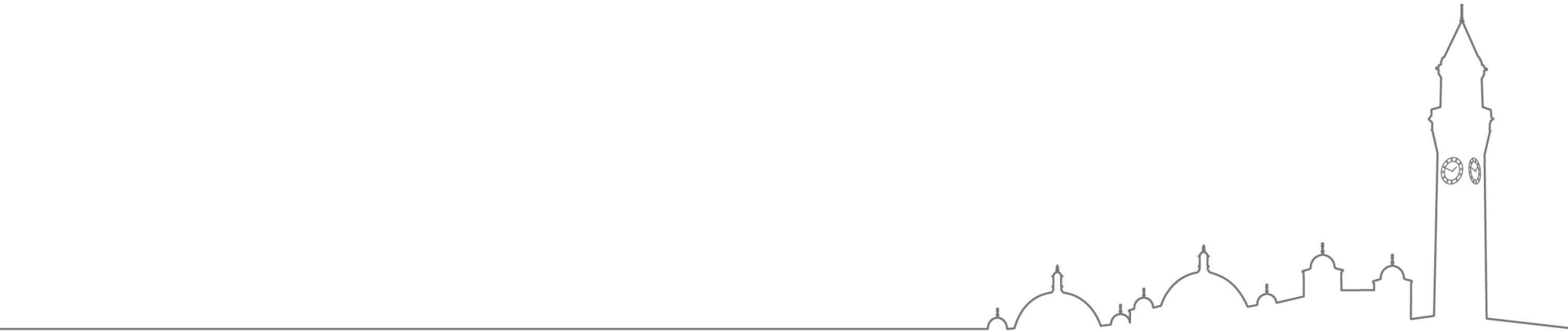
- The last step will be applying what I have found to a corpus of forensic texts from the German federal criminal police office (BKA) to evaluate the validity of the method

Thank you!

danaroemling@gmail.com

@danaroemling

https://github.com/danaroemling



Selected References

- Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119–123. https://doi.org/10.1145/1461928.1461959
- Cassidy, F. G. (1985). *Dictionary of American regional English*. the Belknap press of Harvard university press.
- Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.
- “Dialect" OED Online. Oxford University Press, March 2023. Web.
- Heeringa, W., & Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, *13*(3), 375–400. https://doi.org/10.1017/S0954394501133041
- Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4383–4394. https://doi.org/10.18653/v1/D18-1469
- Hudson, R. A. (1996). *Sociolinguistics* (2nd ed). Cambridge University Press.
- Leonard, R., Ford, J. E. R., & Christensen, T. K. (2017). Forensic Linguistics: Applying the Science of Linguistics to Issues of the Law. *Hofstra Law Review*, *45*(3), 881–898. https://doi.org/10.1017/S0047404508080421
- Mattheier, K. J. (1990). Dialekt und Standardsprache. Über das Varietätensystem des Deutschen in der Bundesrepublik. *International Journal of the Sociology of Language*, *1990*(83). https://doi.org/10.1515/ijsl.1990.83.59
- Nguyen, D., Smith, N. A., & Rose, C. P. (2011). Author Age Prediction from Text using Linear Regression. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 115–123.
- 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. *Dictionary of American Regional English Newsletter*, *4*(3), 1–5.
- Wright, D. (2020). Identifying authors and idiolects using forensic linguistics. Presentation at the University of Mosul, Iraq.

Images

- Akron, Ohio: https://www.uakron.edu/international/images/where-is-akron-01-01.svg
- Devil Strip: https://commons.wikimedia.org/wiki/File:Massachusetts-devils_strip.JPG

- All other images are either cited from their academic publication or have been taken / created by the author

Software

- R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Including (but not limited to) the packages:

- Massicotte, P., & South, A. (2023). *rnaturalearth: World Map Data from Natural Earth* (R package version 0.3.2) [Computer software]. https://CRAN.R-project.org/package=rnaturalearth
- Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009
- South, A., Schramm, M., & Massicotte, P. (2023). *rnaturalearthhires: High Resolution World Vector Map Data from Natural Earth used in rnaturalearth* [Computer software]. https://github.com/ropensci/rnaturalearthhires
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Wickham, H., François, R., & Henry, L. (2020). *dplyr: A Grammar of Data Manipulation* (R package version 1.0.2) [Computer software]. https://CRAN.R-project.org/package=dplyr