UNIVERSITY OF BIRMINGHAM

Geolinguistic Profiling in Forensic Linguistic Authorship Analysis

Dana Roemling (they/them)
Language Technology HY, 29.02.24

Today

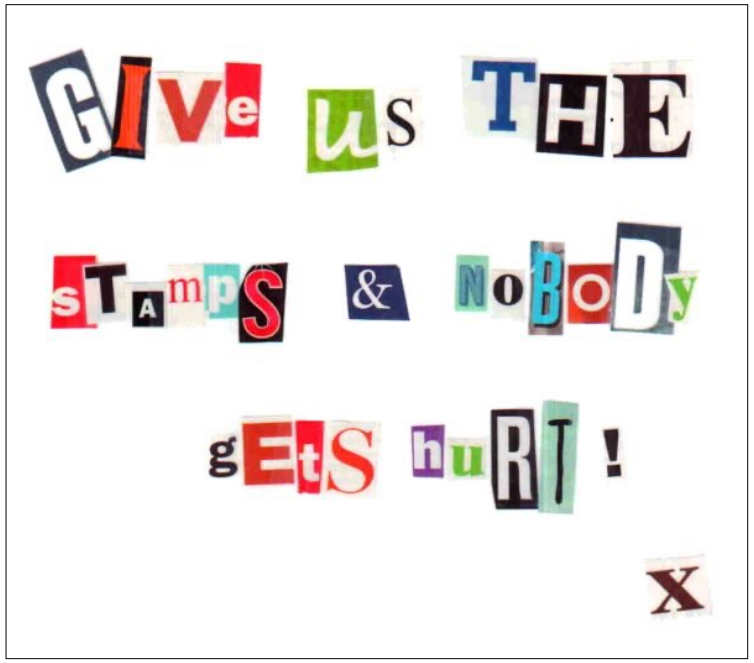
- Intro:
  - Forensic Linguistics &
  - Authorship Analysis
- Dialect / Geolinguistic Profiling:
  - Corpus
  - Analysis & Preliminary Results
  - Forensic Application & Outlook
- Discussion:
  - Project Ideas for HY Research Visit

Content Warning

Forensic Linguistics

Language as evidence & language in the investigative process

for example: ransom note, text messages in a murder investigation



Language of the law & language in the legal process

for example: statutory interpretation, the influence of power & hierarchy in court

2 § Murha.

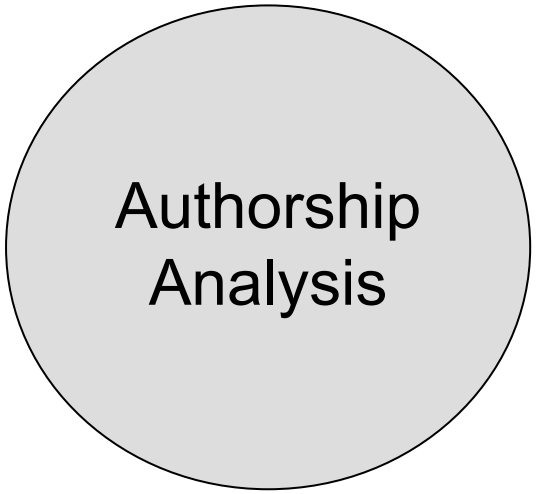
Jos tappo tehdään

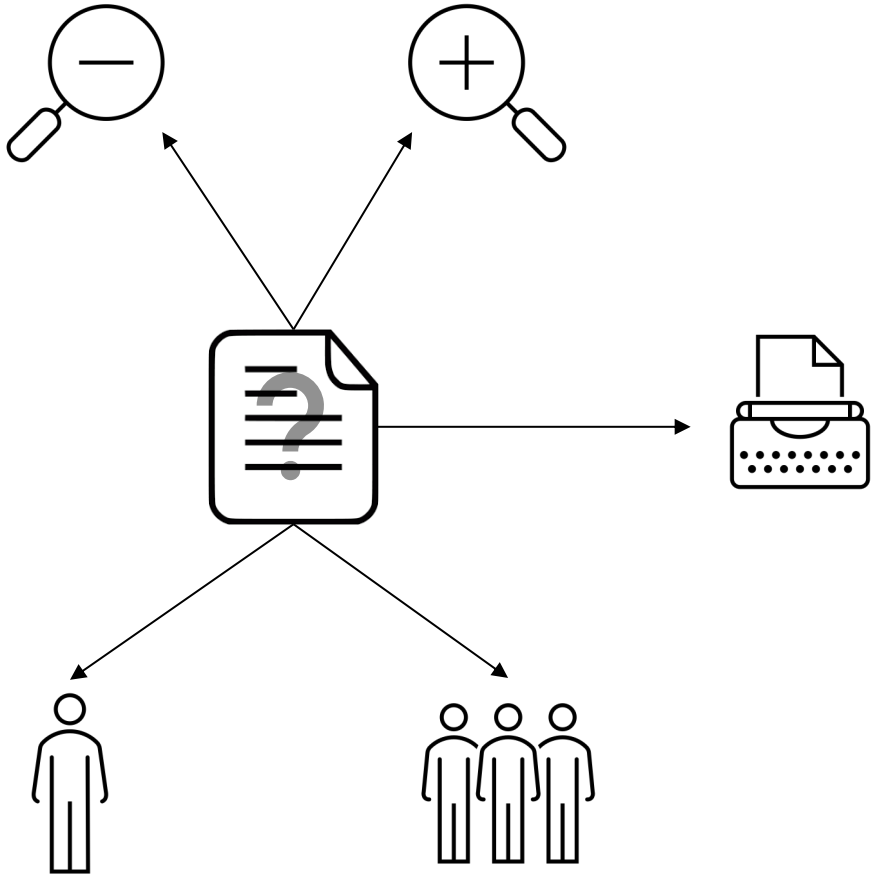
1. vakaasti harkiten,
2. erityisen raa'alla tai julmalla tavalla,
3. vakavaa yleistä vaaraa aiheuttaen tai
4. tappamalla virkamies hänen ollessaan virkansa puolesta ylläpitämässä järjestystä tai turvallisuutta taikka virkatoimen vuoksi

ja rikos on myös kokonaisuutena arvostellen törkeä, rikoksentekijä on tuomittava murhasta vankeuteen elinkaudeksi.

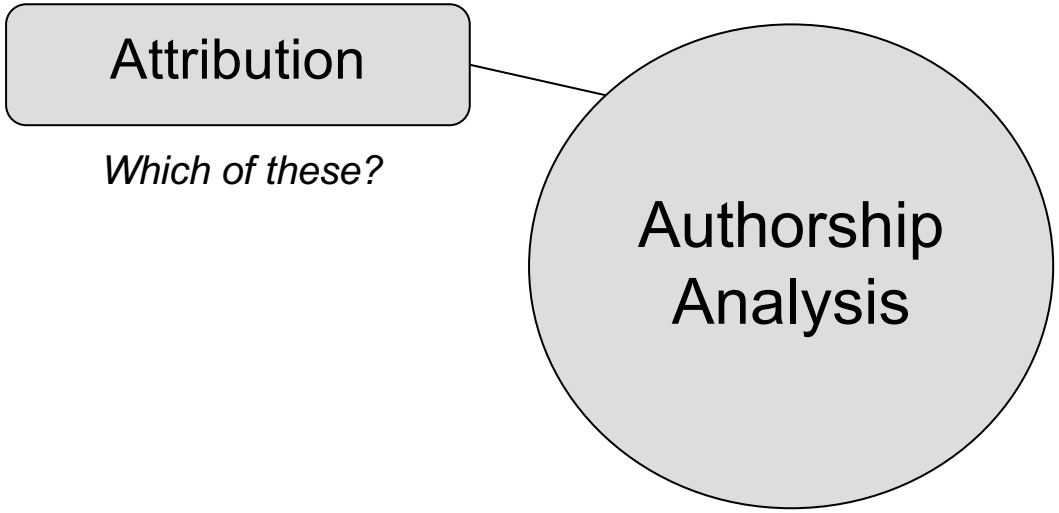
Yritys on rangaistava.

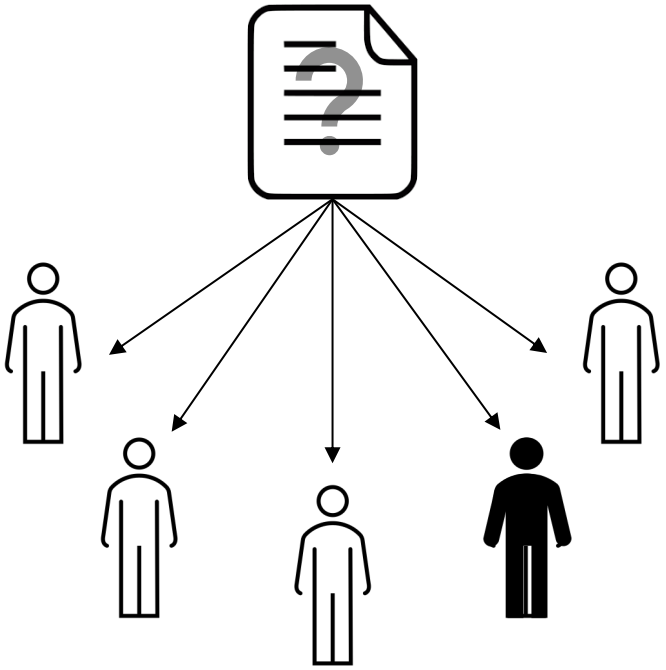
Authorship Analysis – Preliminaries





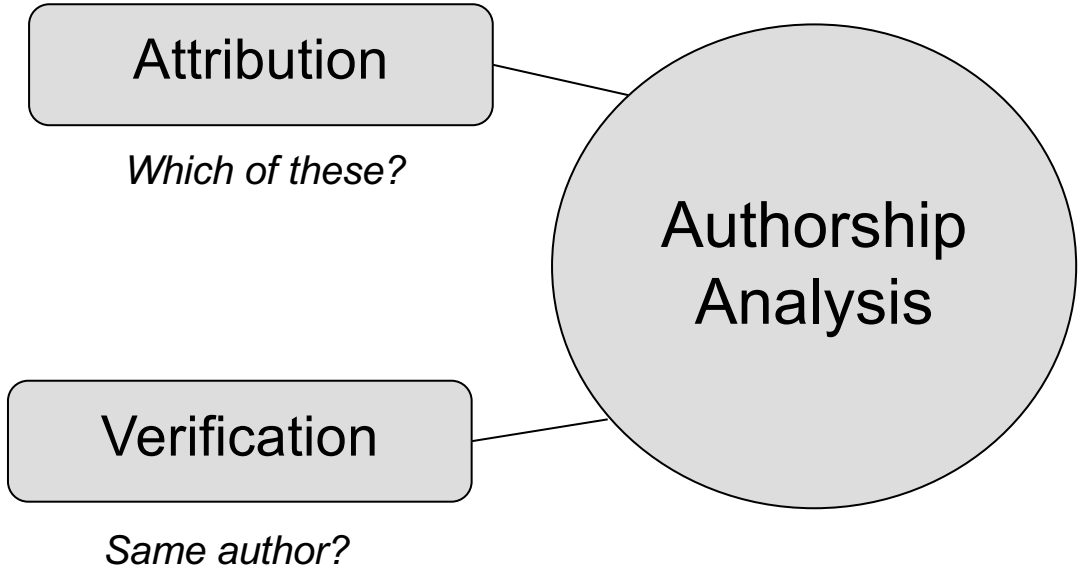
Authorship Analysis



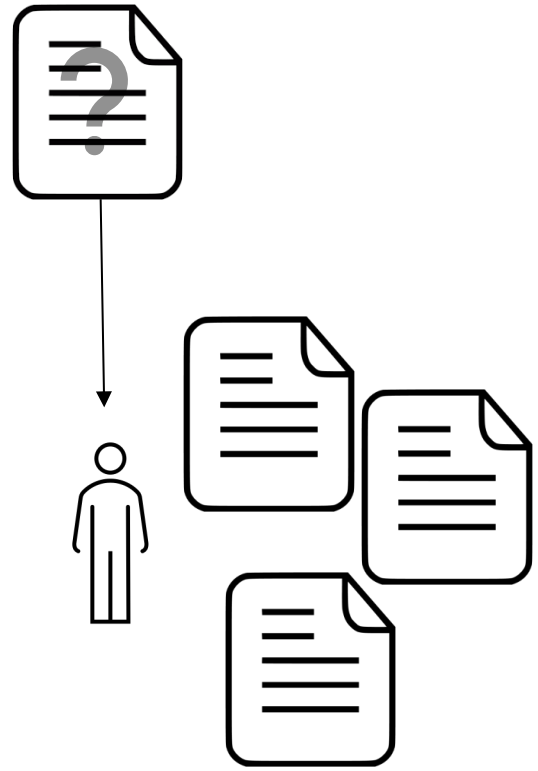


Adapted from Wright, 2020

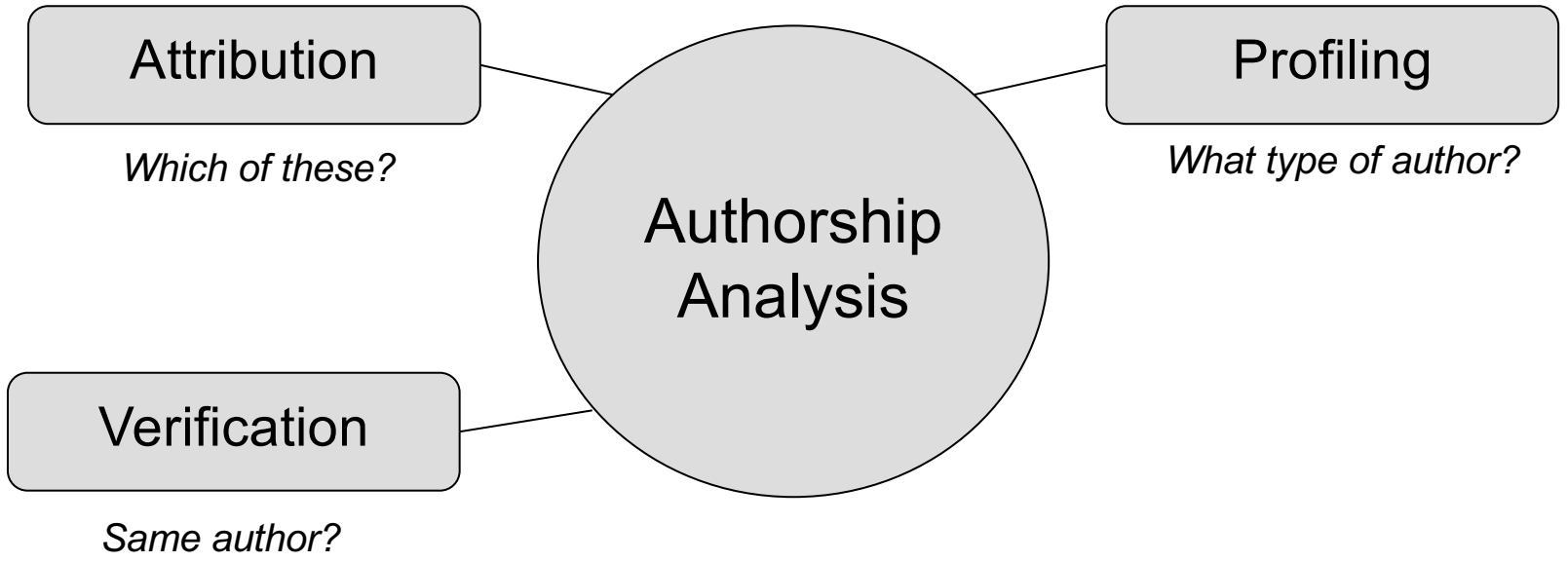
Authorship Analysis



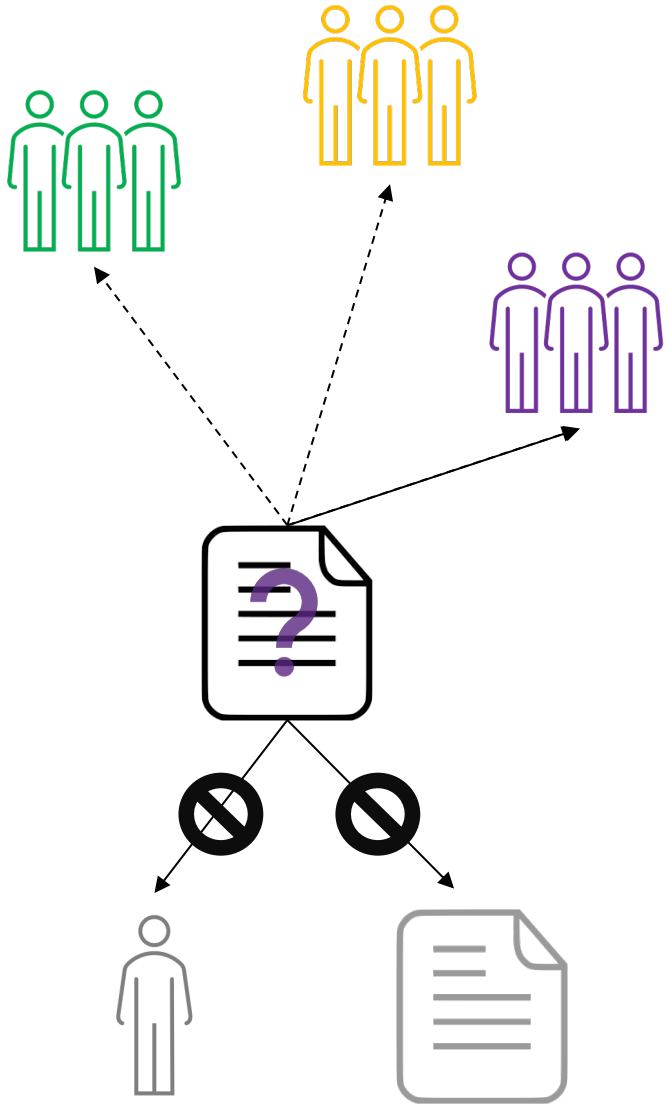
Adapted from Wright, 2020



Authorship Analysis

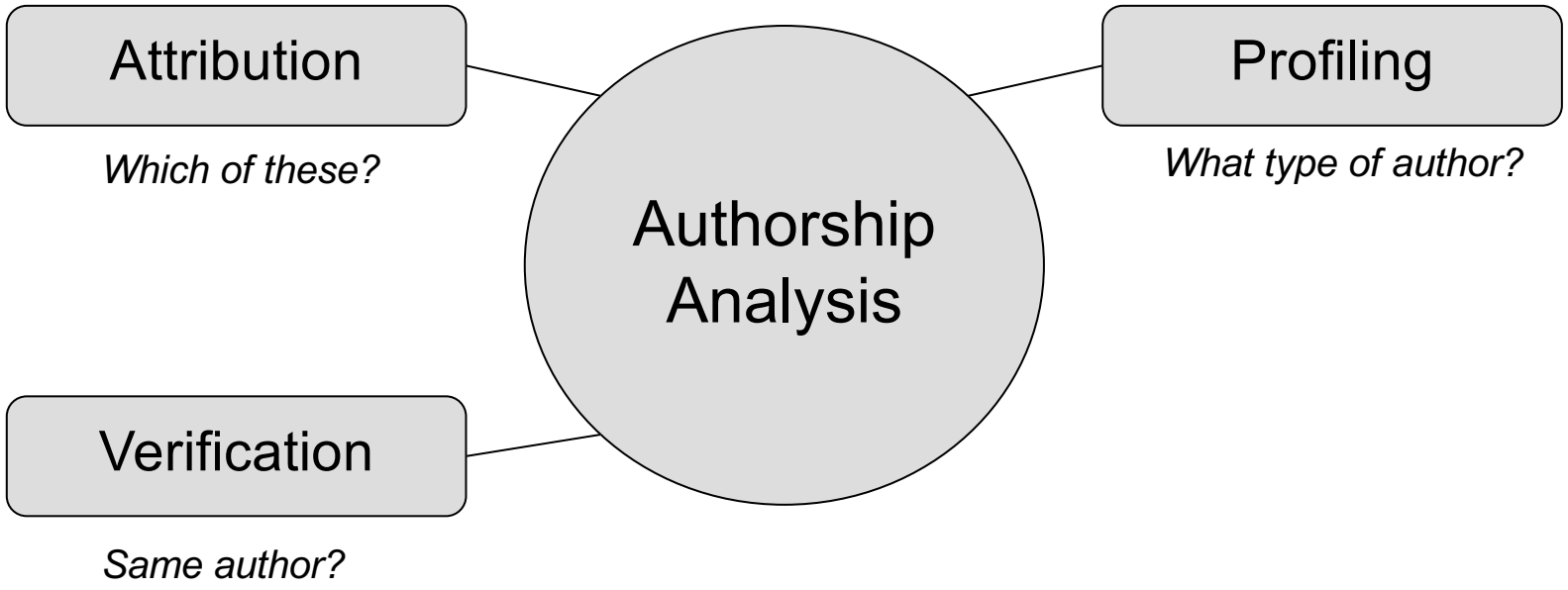


Adapted from Wright, 2020



Authorship Profiling

- Analysis of texts to infer characteristics about an author
  - e.g. age, gender, first language(s) influence



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. […]”

(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 an
- “A style is a vari articular author or social nguistic features that are
- Problem: “Comp necessarily interested in understanding the inner (linguistic) mechanisms of the machine, as long as the accuracy rates are outperforming previous models.”

(Nini, 2018)

Transparency / Replicability / Reliability in the Forensic Context

The Analysis of Writing Style

- Macleod & Grant (2012): Taxonomy for stylistic and statistical authorship analysis
- Ishihara (2014): n-gram modelling for authorship analysis
- Kocher & Savoy (2017): Distance measures for authorship profiling
- HaCohen-Kerner (2022): Survey on age & gender profiling
- PAN workshops: profiling age, gender, native language; authorship analysis
  - Bevendorff et al. (2023): Influence of text type for authorship analysis

The Analysis of Salient Linguistic Features

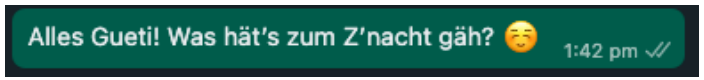
- “[T]he application of sociolinguistic knowledge on a case by case basis to extract *ad hoc* linguistic features that are markers of a certain demographic background”
- “[I]nvolves the linguist’s experience in discovering dialectal or sociolinguistic features that can reveal clues about the background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

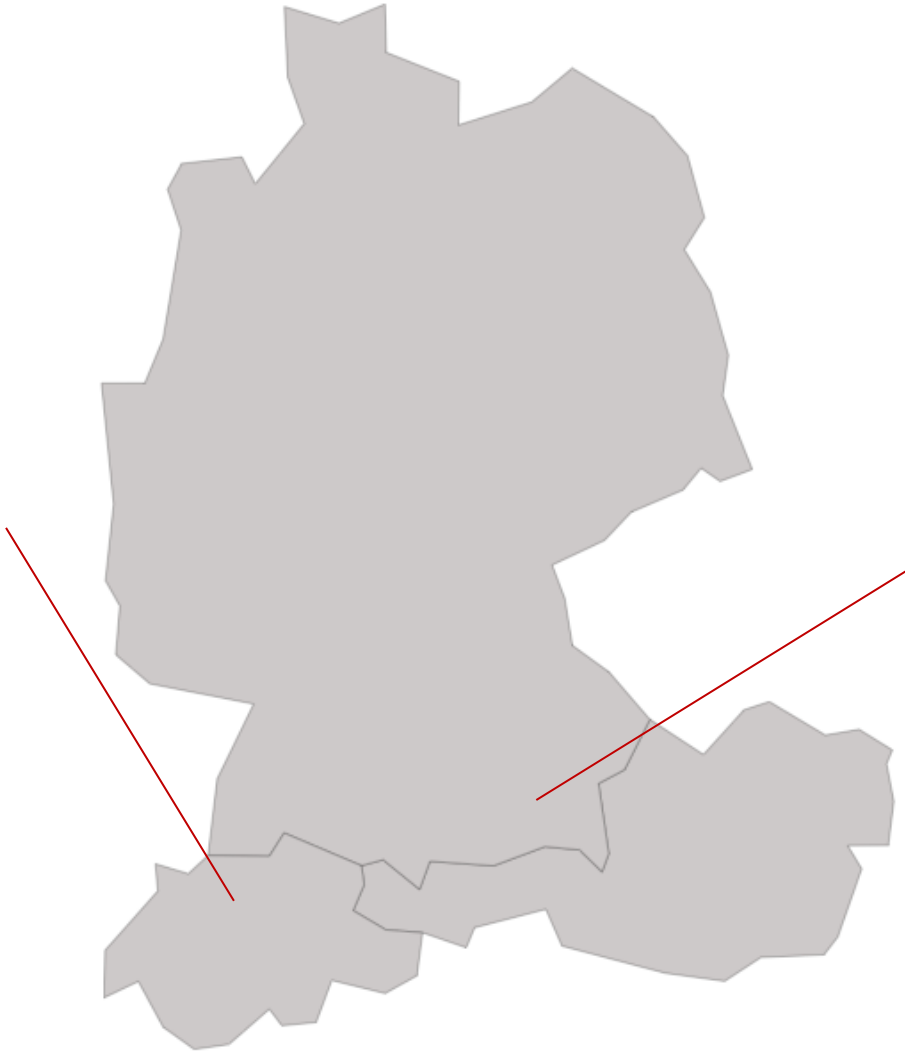
(Nini, 2018)

The Analysis of Salient Linguistic Features

- Chambers (1990): Bear Island Land Claim
- Shuy (2001): Devil strip case
- French et al. (2007): Yorkshire Ripper Hoax
- Leonard et al. (2017): Unabomber

Which regional variety?

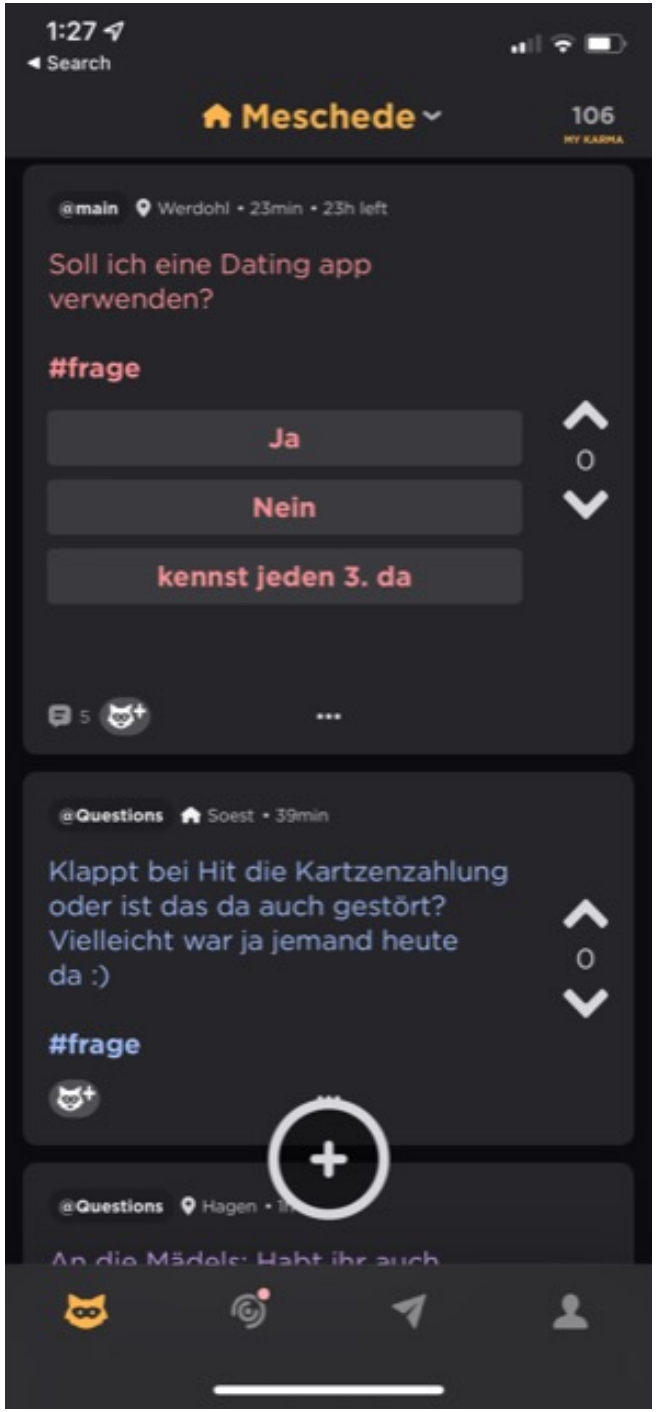




Dialekt, heit oft aa Mundoat, gweanli a 'oatsvaschidnats' Gred oda Gsoad, guit ois a Varietet von ana Sproch, de wo in iran Afkumma, i da raimlichn Asdeanung un im Vaschwinna sozial u historisch o bstimmte Sidlungsraim bundn is. (Wikipedia)

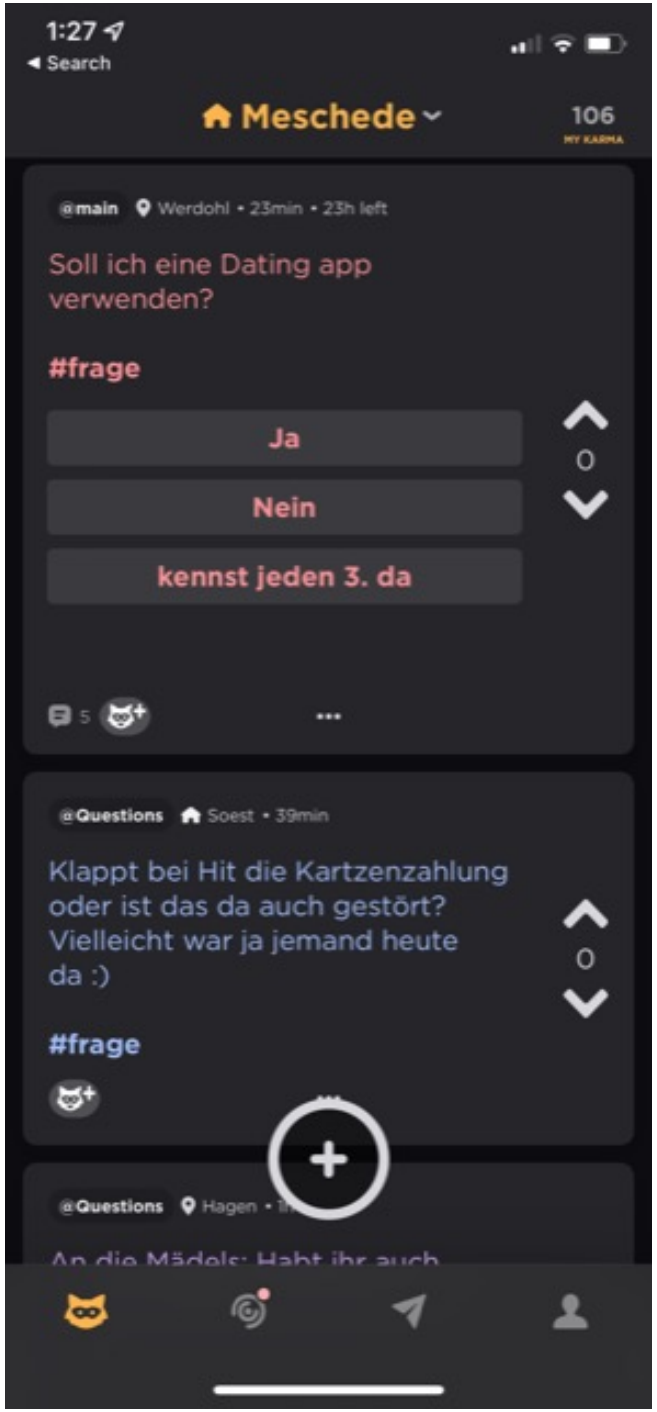
Data

- Jodel: Social Media app
- Interaction in a 10-15km radius around own location
- Collected 2017 by Hovy & Purschke (2018) to represent German-speaking area (= Austria, Germany, Switzerland)



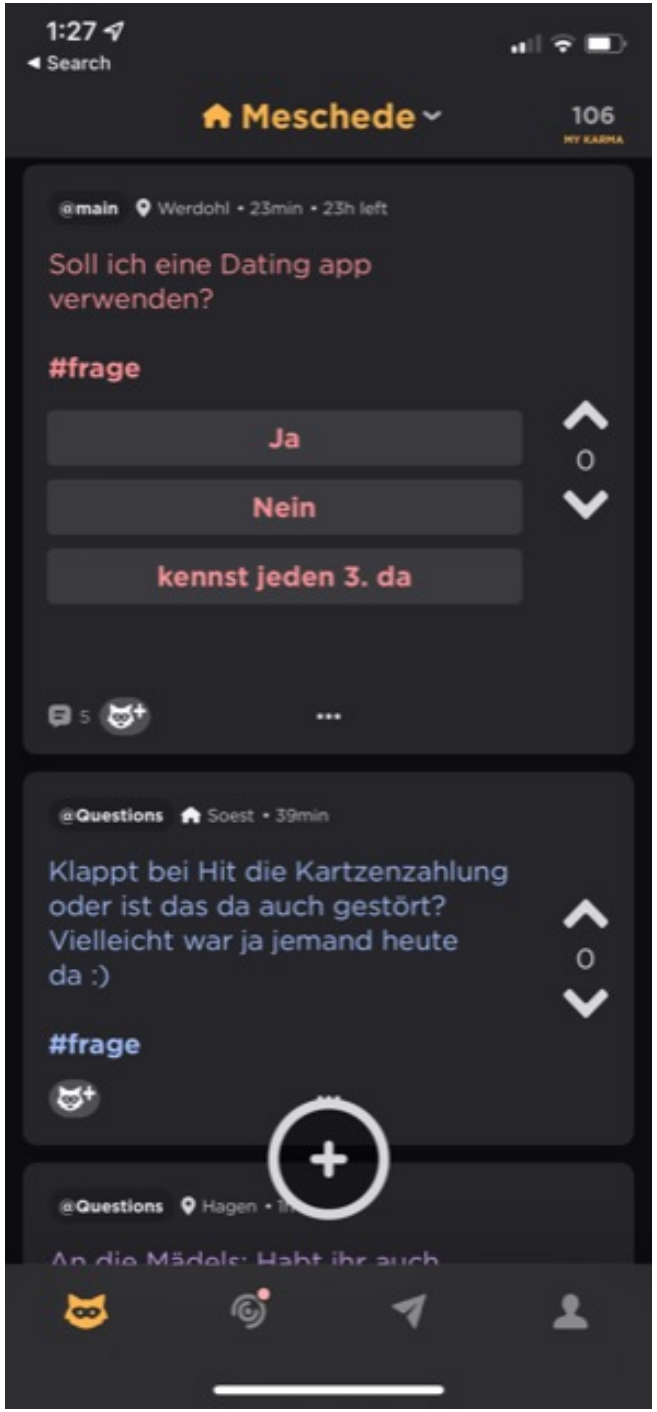
Data

- No accounts / profiles, so “anonymous” interaction
- Demographics:
  - 18-26y → 72%, 27+y → 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

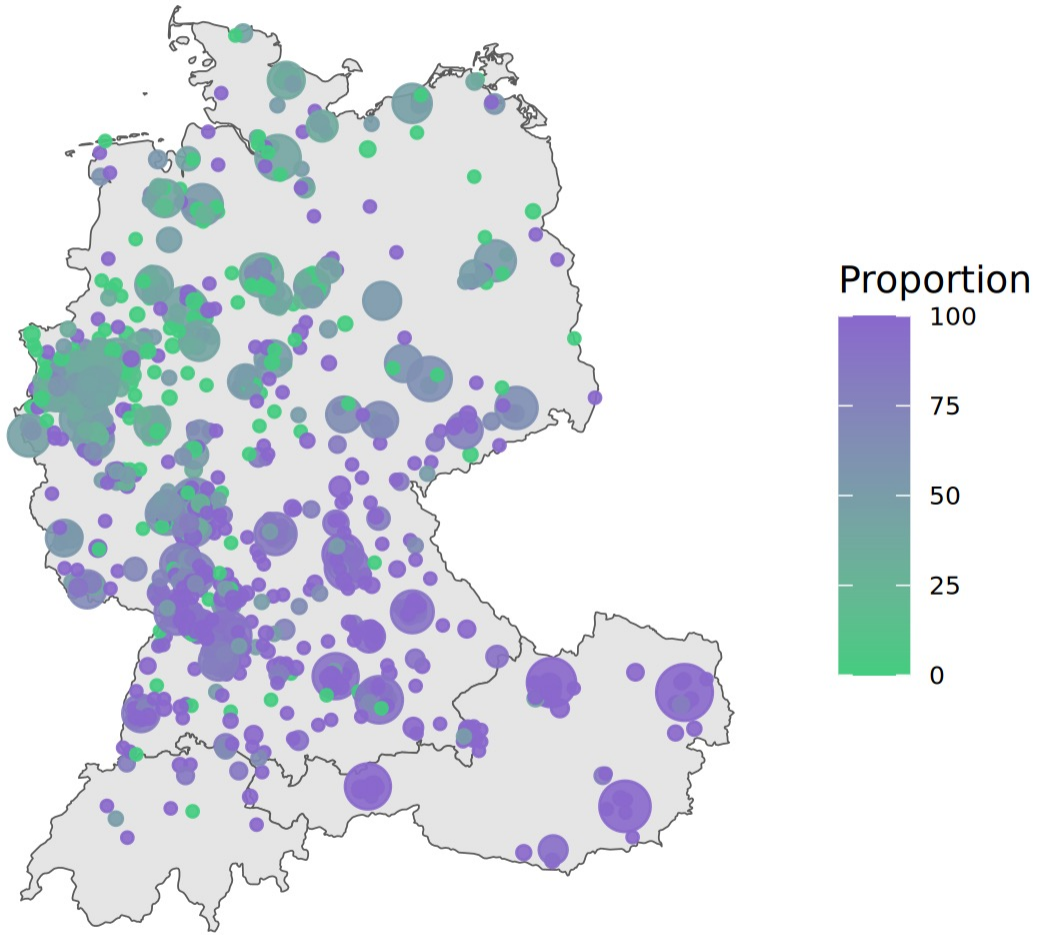


Sample

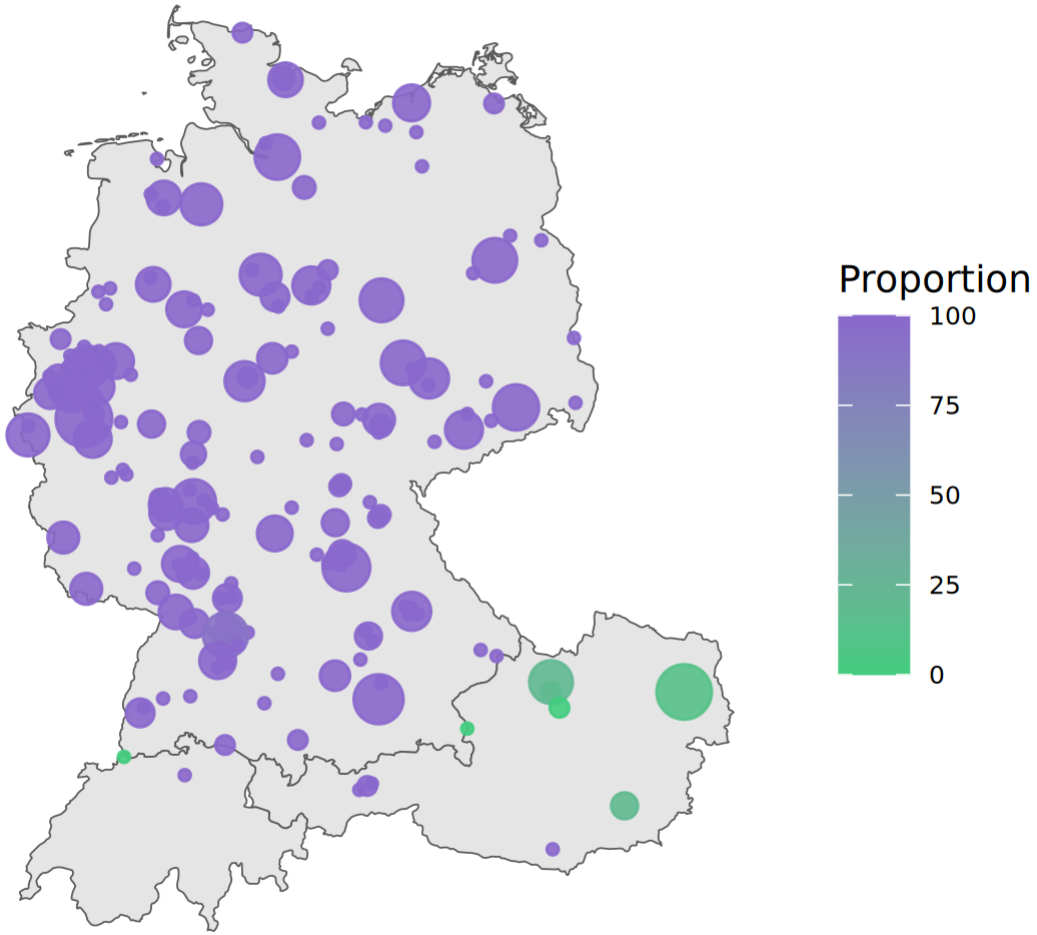
Message	Creation Timestamp	Location	Post ID
Semesterferien: grillen schlafen grillen bar schlafen repeat.. English: *semester break: bbq sleep bbq pub sleep repeat..*	2017-04-04T 22:29:41.814Z	Berlin	58e41e5512e80 a3f0cb6f66b
Ich bin grade leicht verwirrt Werdet ihr Mädels so emotional, kurz bevor ihr 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

Regional Distribution?

schau & *guck* in the GSA

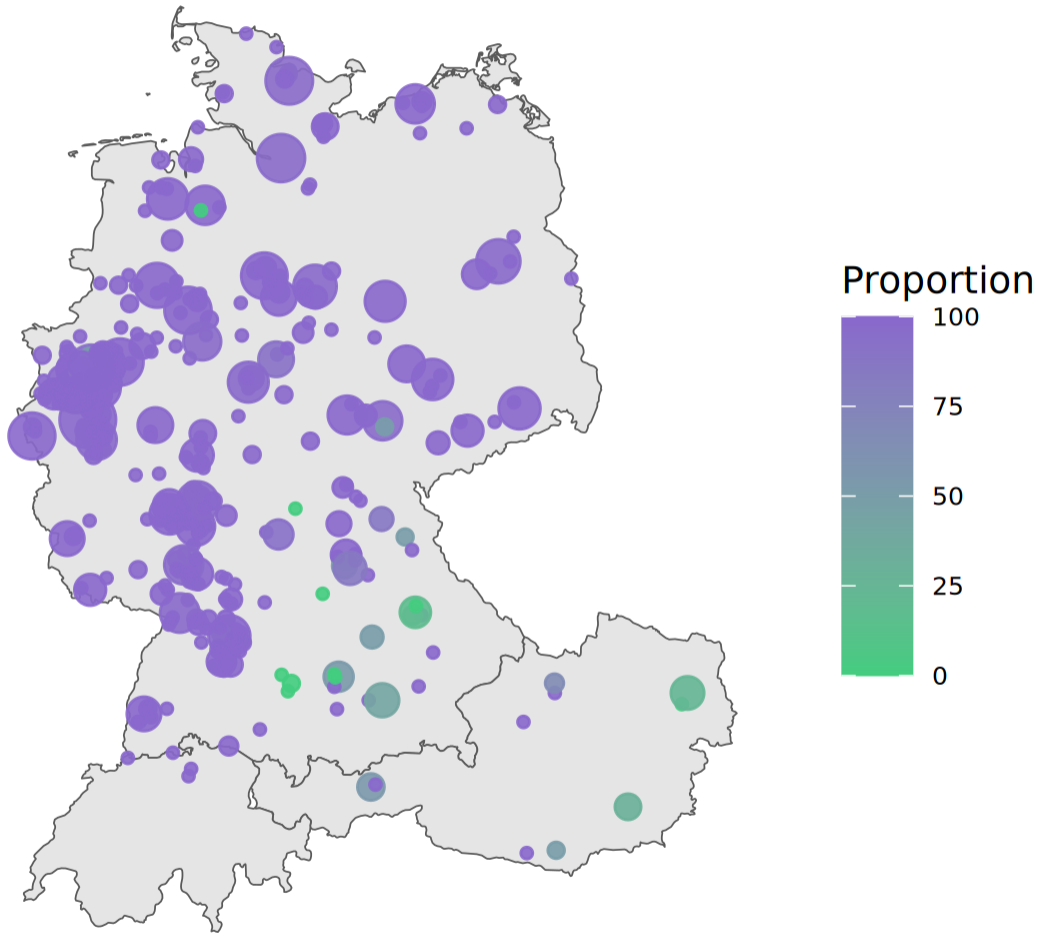


mülleimer & *mistkübel* in the GSA

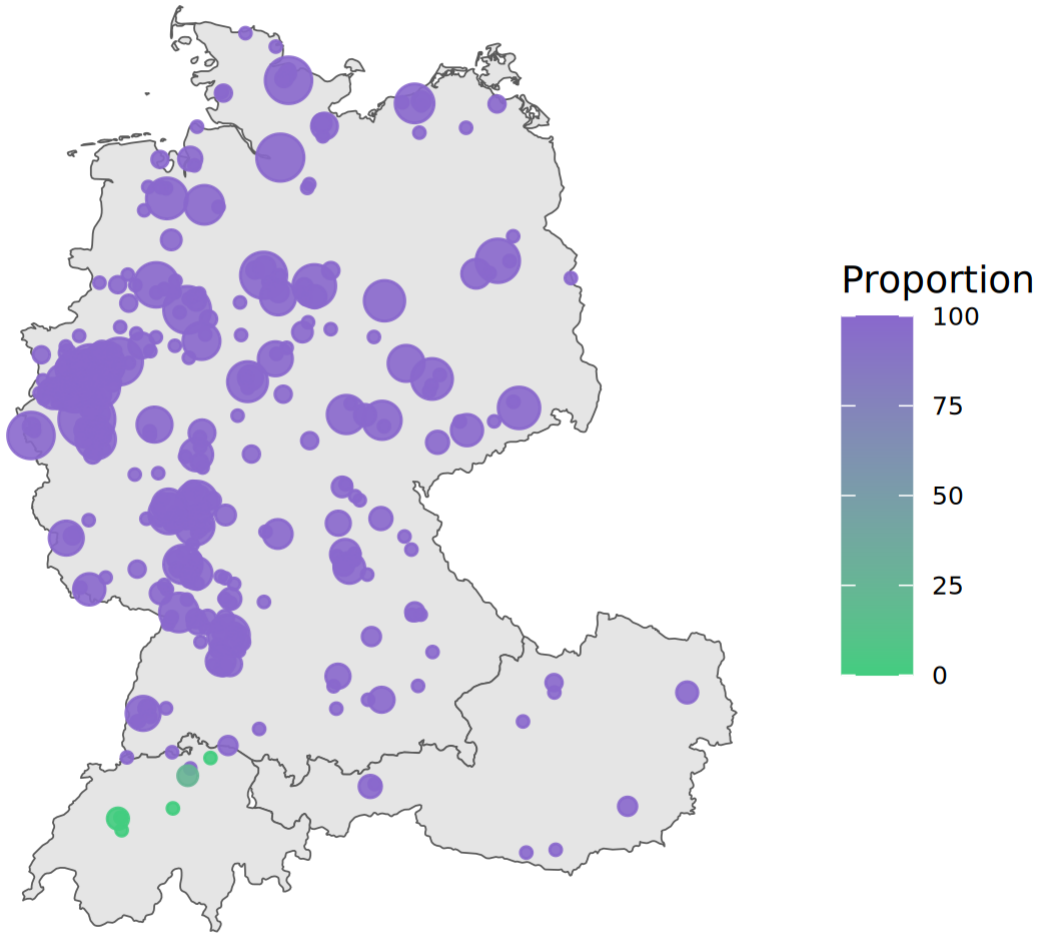


Regional Distribution?

brötchen & *semmel* in the GSA

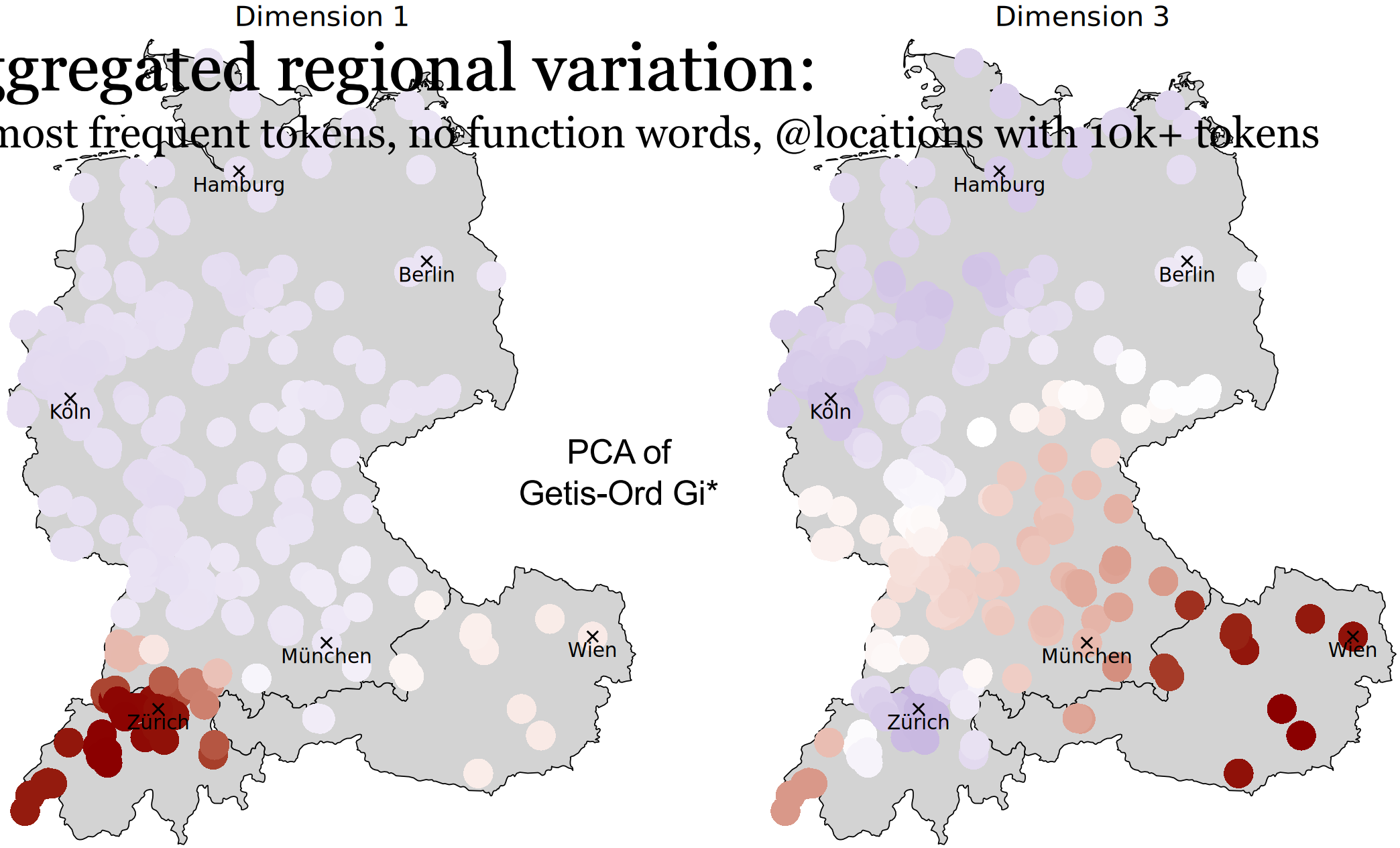


brötchen & *weggli* in the GSA

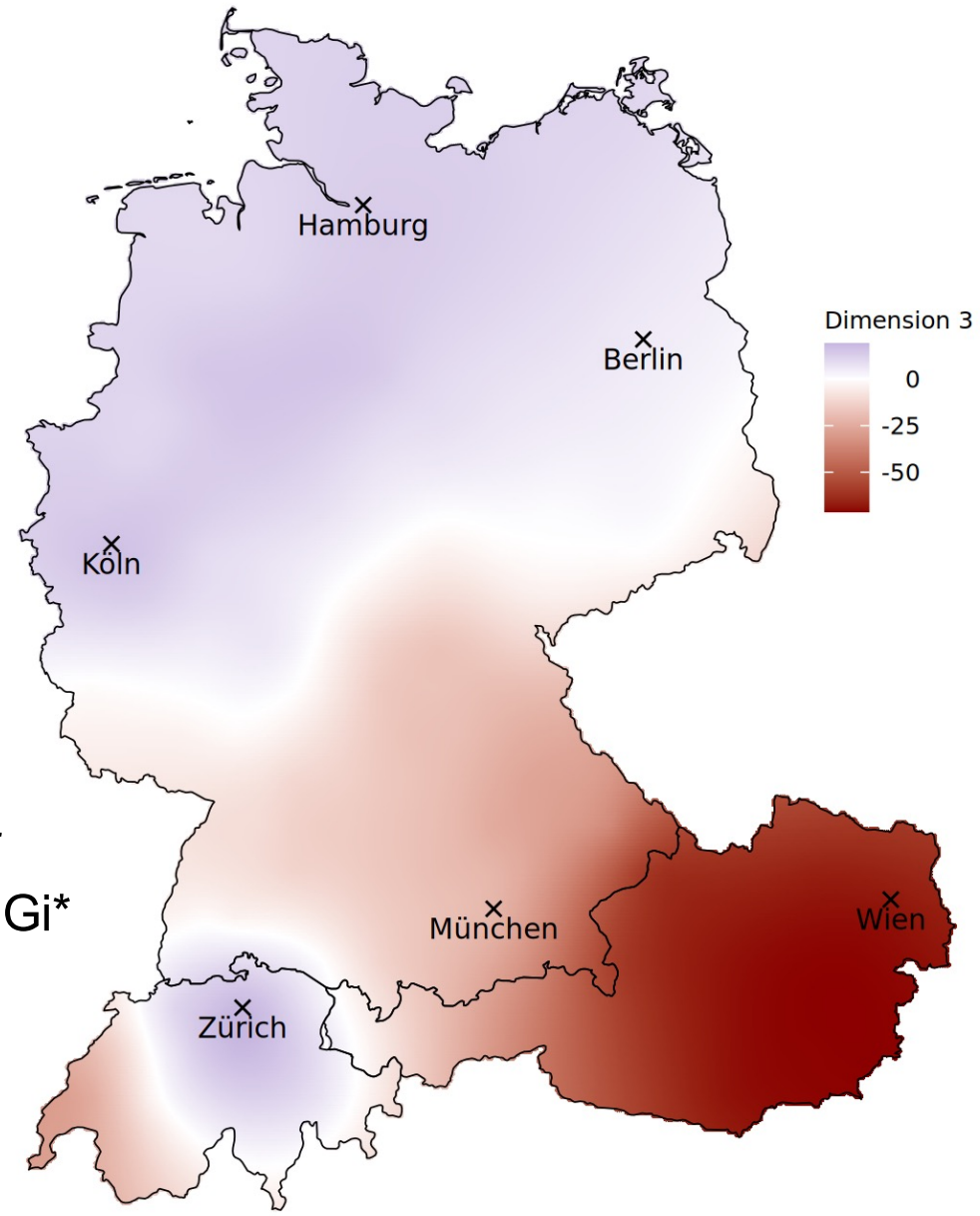


Aggregated regional variation:

2k most frequent tokens, no function words, @locations with 10k+ tokens



kriged
PCA of
Getis-Ord Gi*

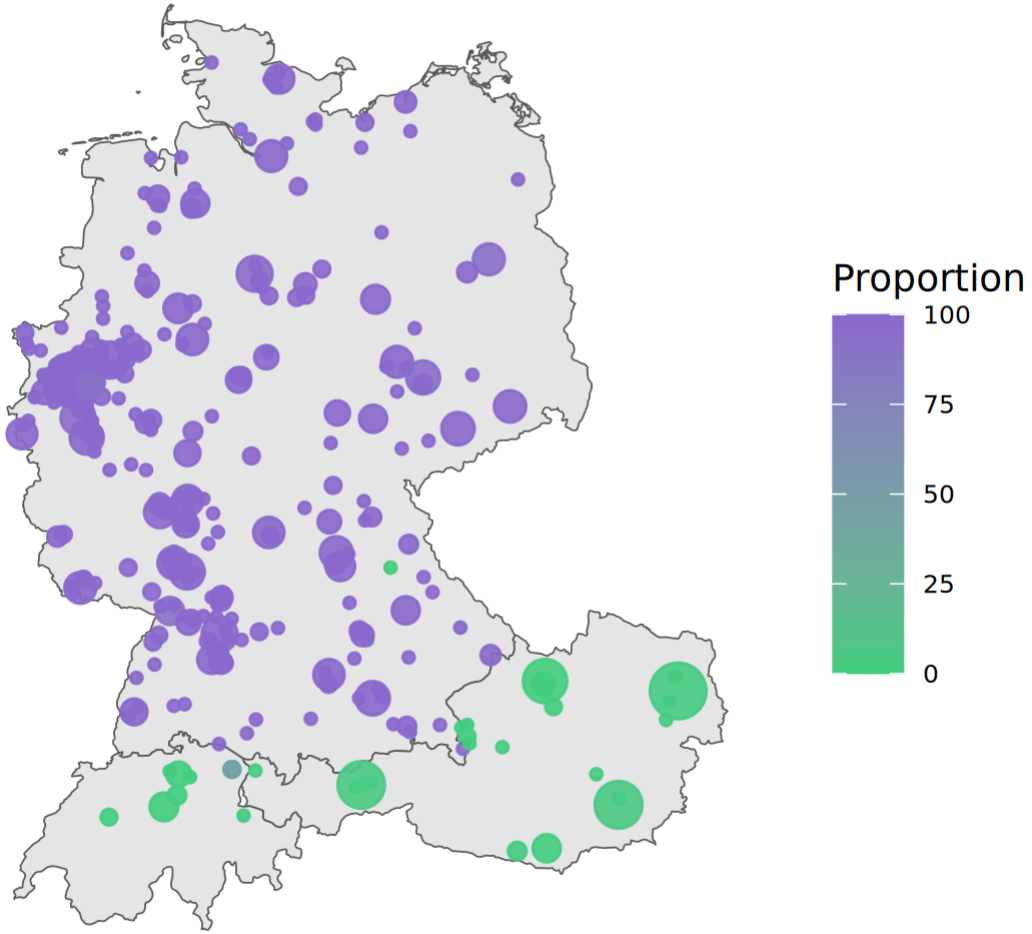


German-speaking area
adapted from
Wiesinger (1983) & König (2019)

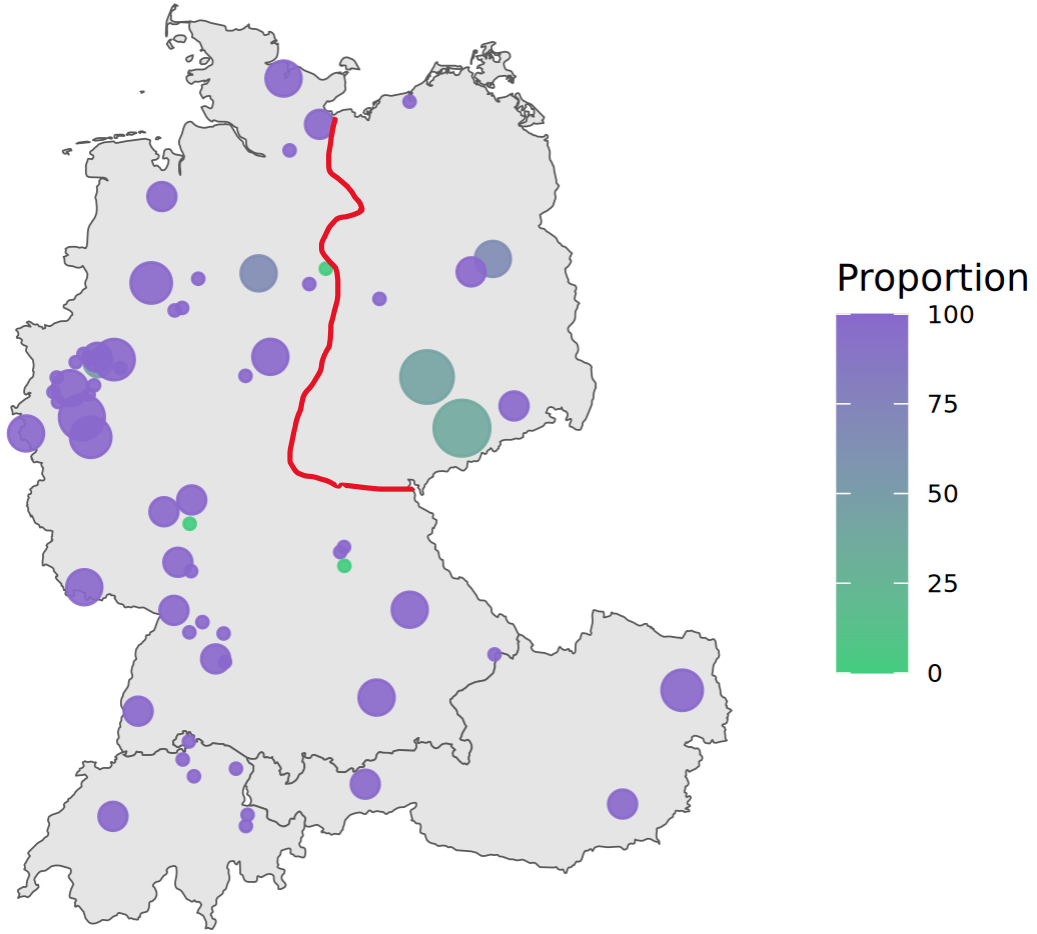
Qualitative Forensic Application

- Is this data able to narrow down the region a post might be from?
- Maybe classic distinctions like north/south, east/west, national borders?

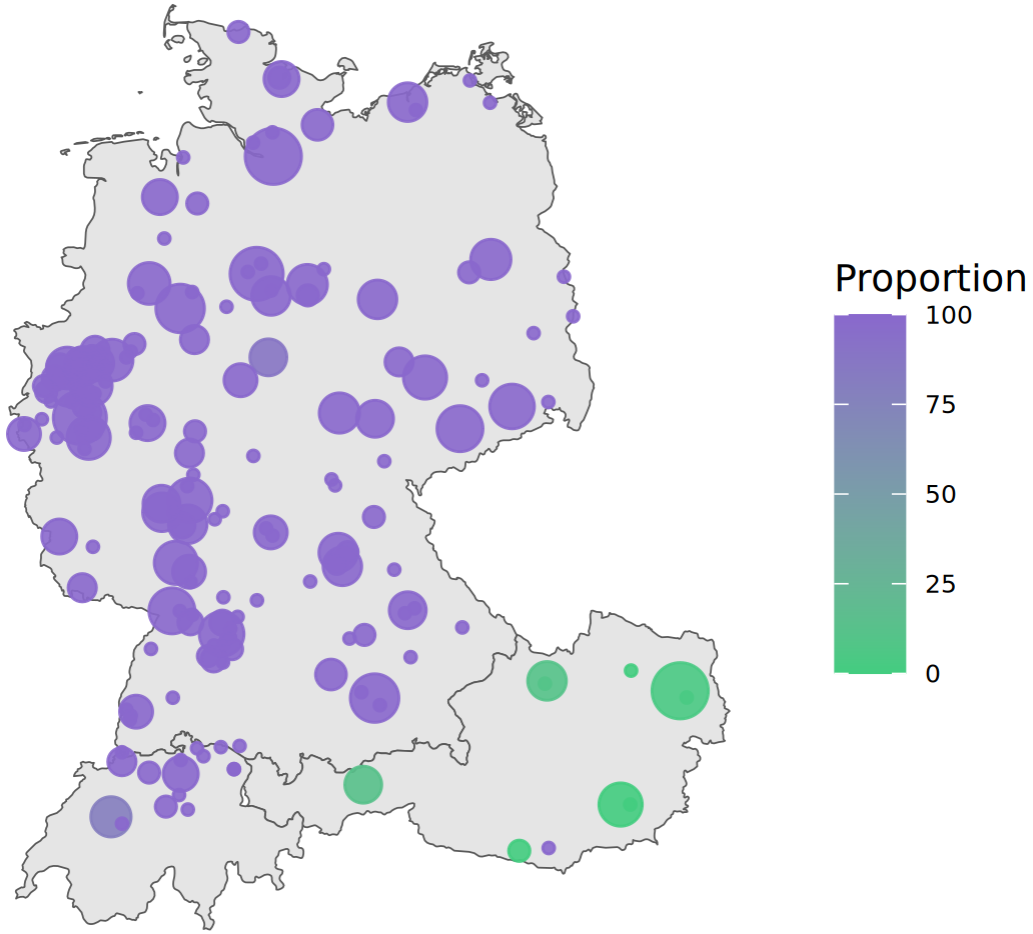
abitur & *matura* in the GSA



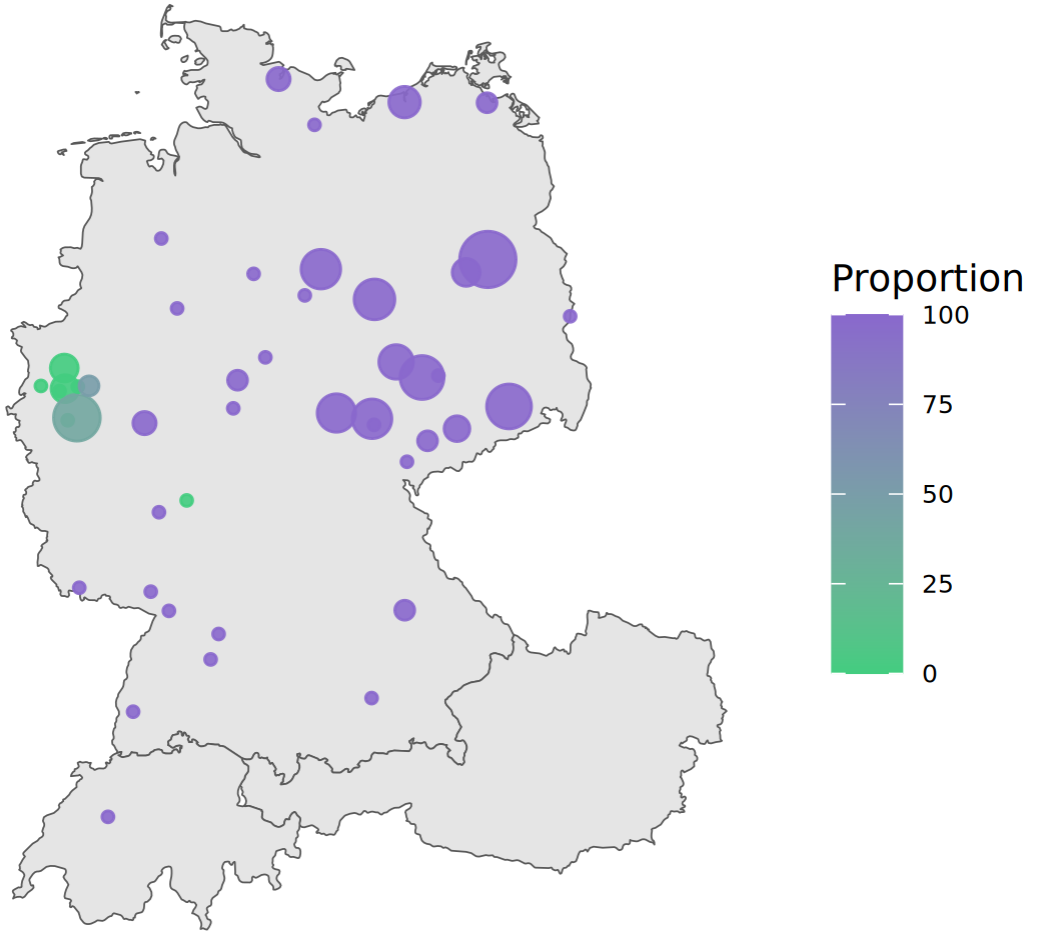
astronaut & *kosmonaut* in the GSA



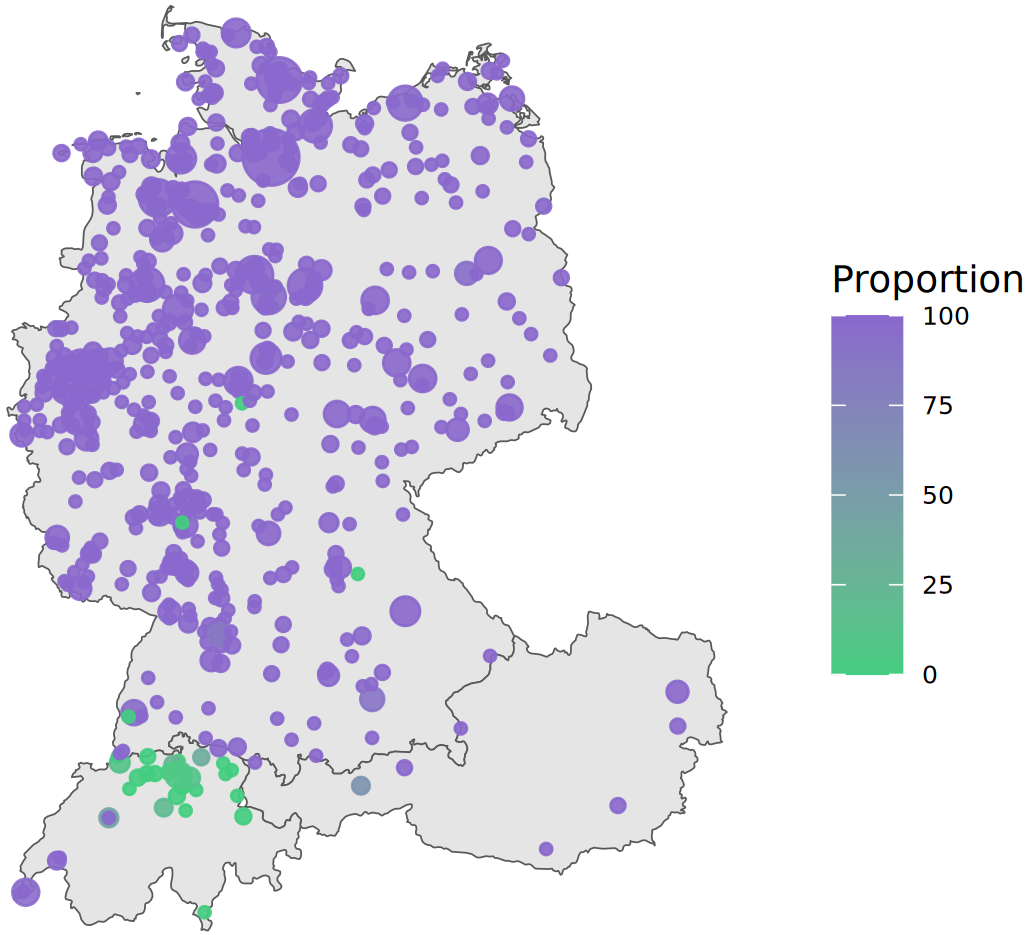
januar & *jänner* in the GSA



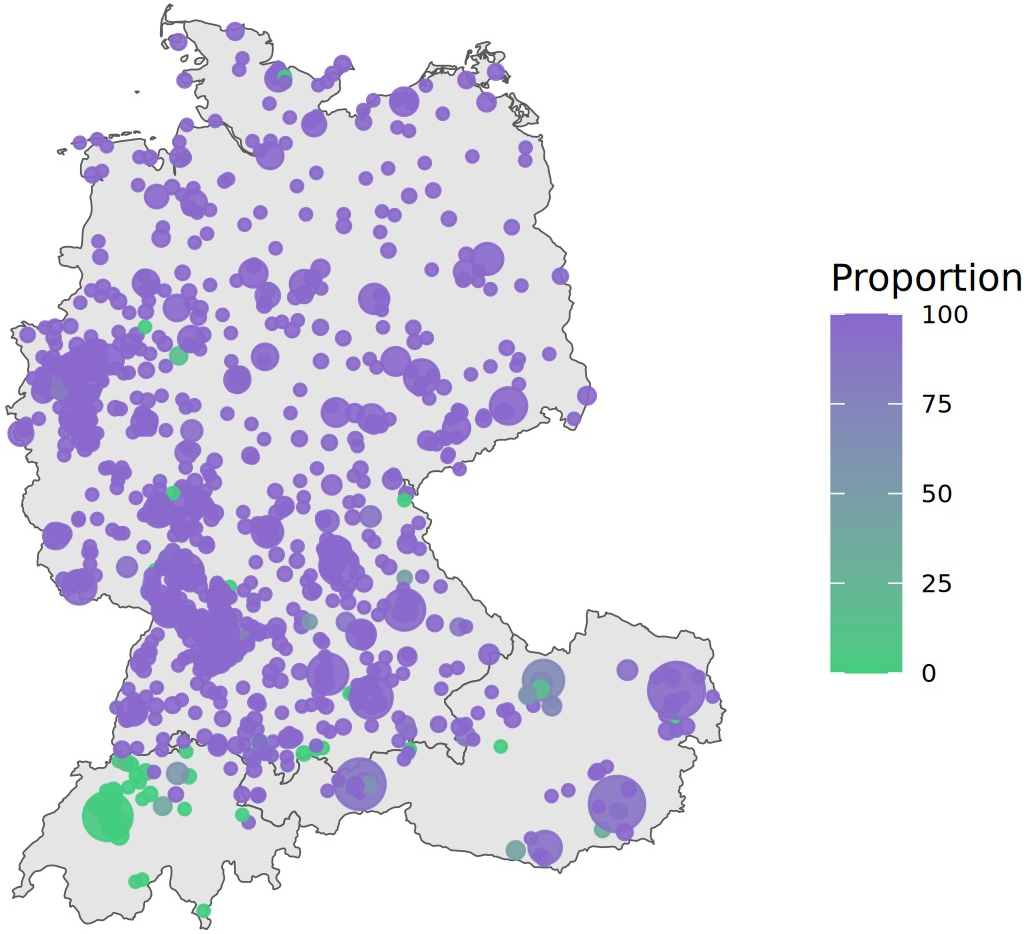
späti & *büdchen* in the GSA



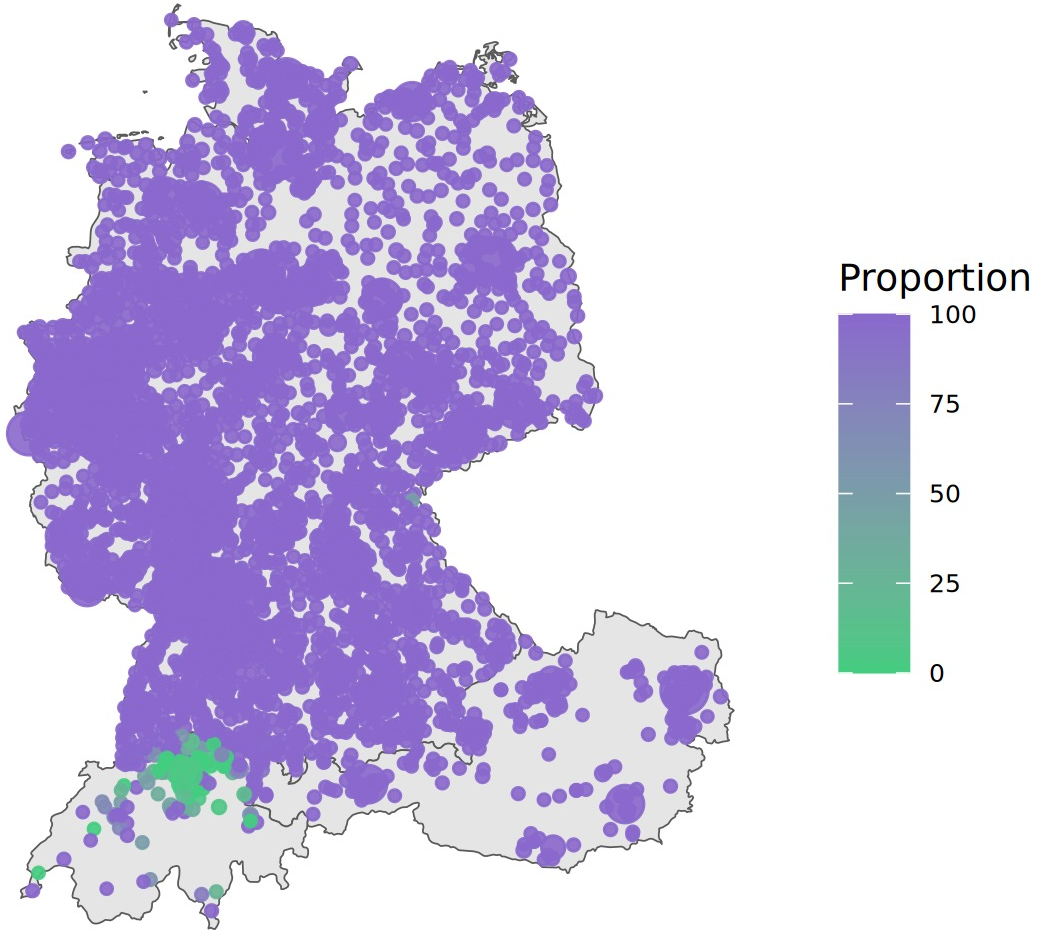
moin & *grüezi* in the GSA



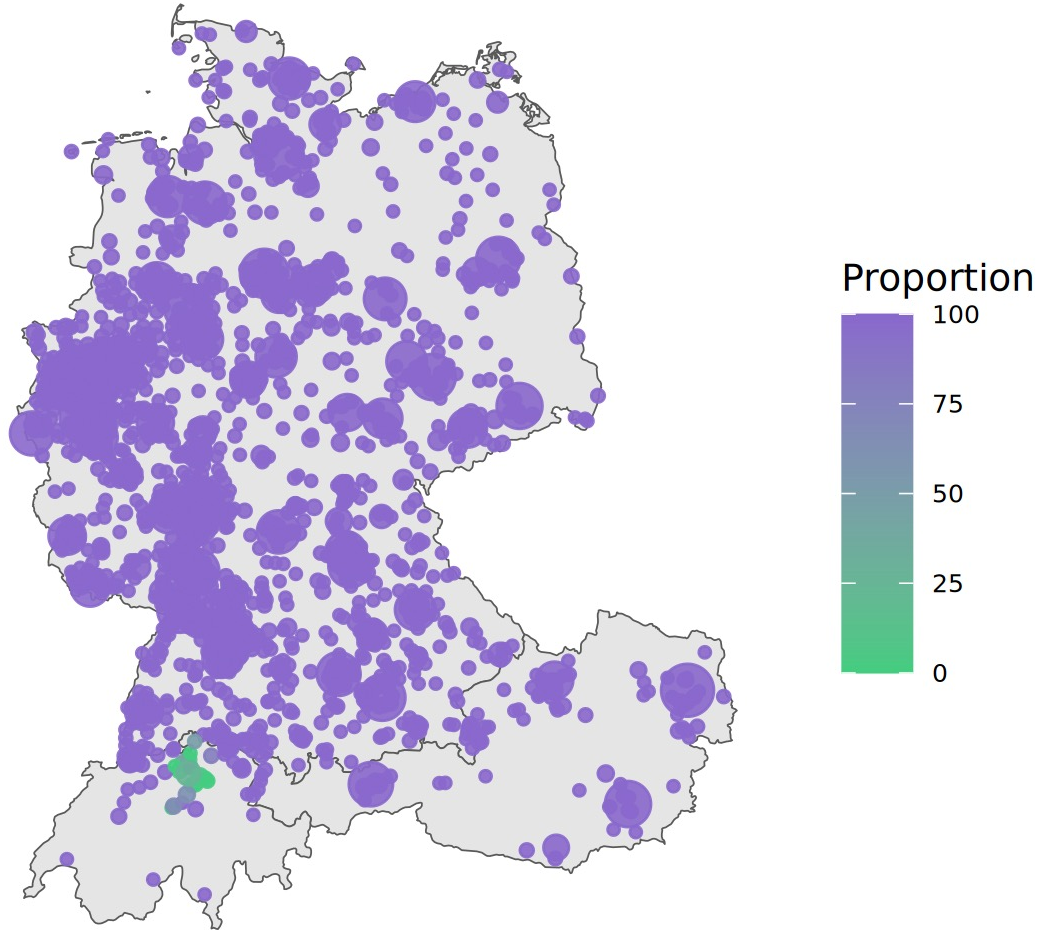
heut & *heit* in the GSA



nicht & *nöd* in the GSA



wirklich & *wüki* in the GSA



Making

~/Documents/R/PHD/shiny_mapping - Shiny

http://127.0.0.1:5550 | Open in Browser | Republish

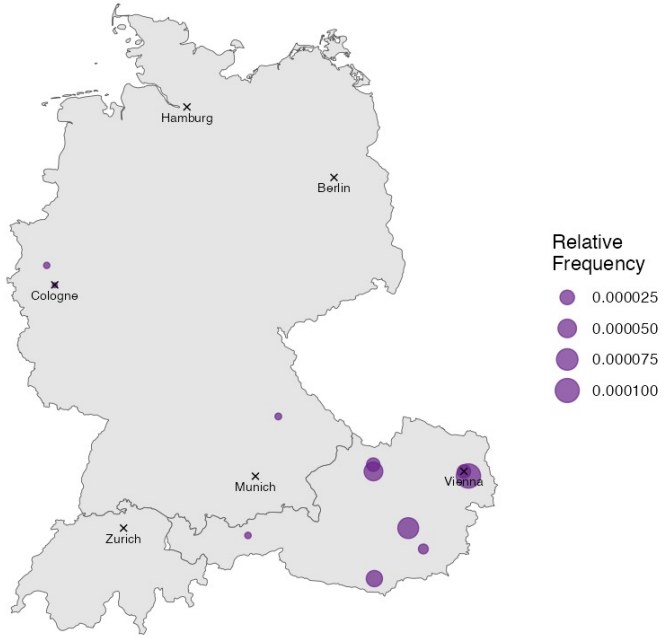
Mapping Features in the German-speaking Area

sackerl

sackerl

Create Map

Data Source
Developer



Conclusion so far

- The corpus follows expected dialectal patterns in the German-speaking area
  - “even though” this is social media data / computer-mediated communication
- This can already serve as a reference tool for forensic investigations to understand geolinguistic distribution

Accounting for locations without data

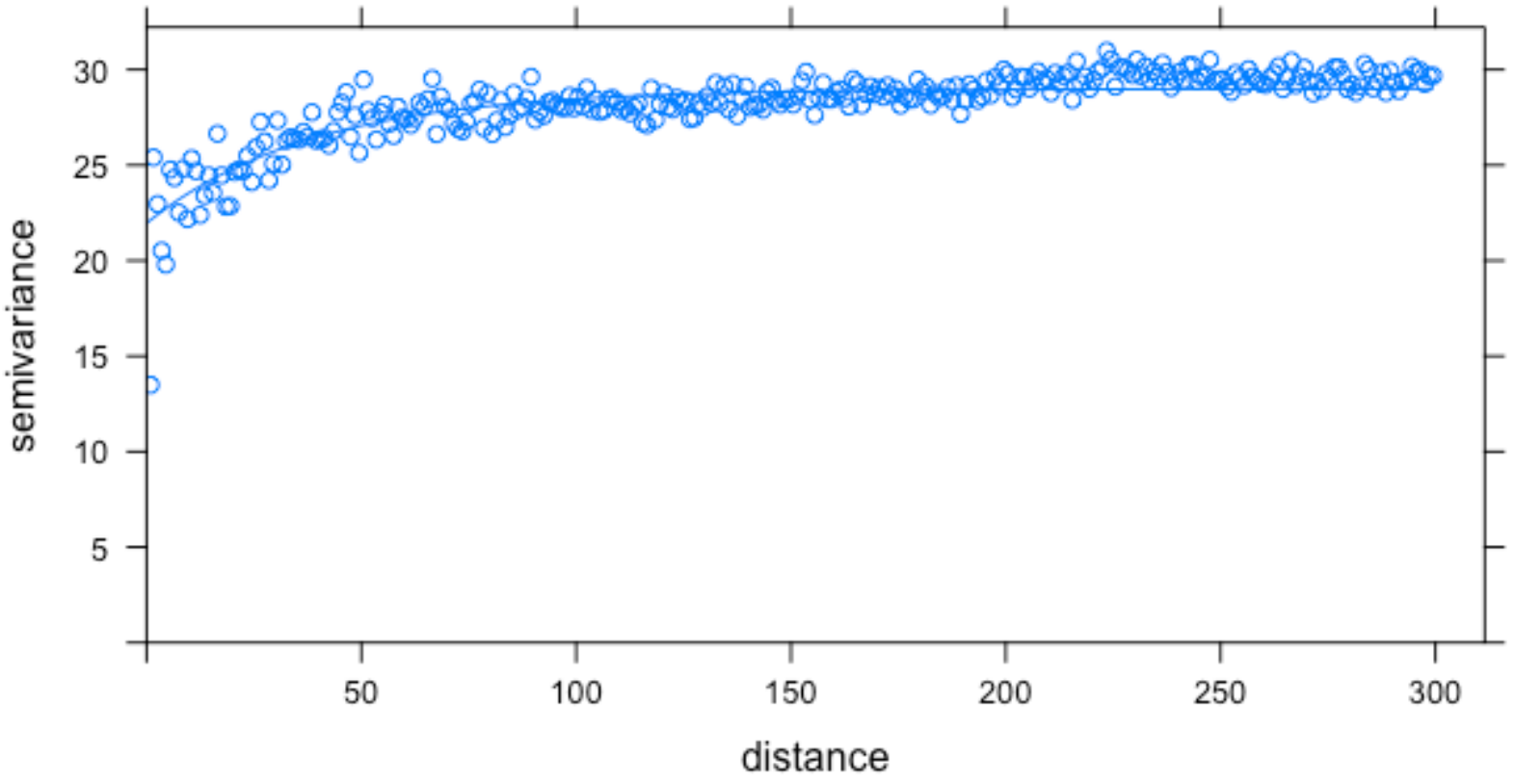
Locations with token “ich”

Similari

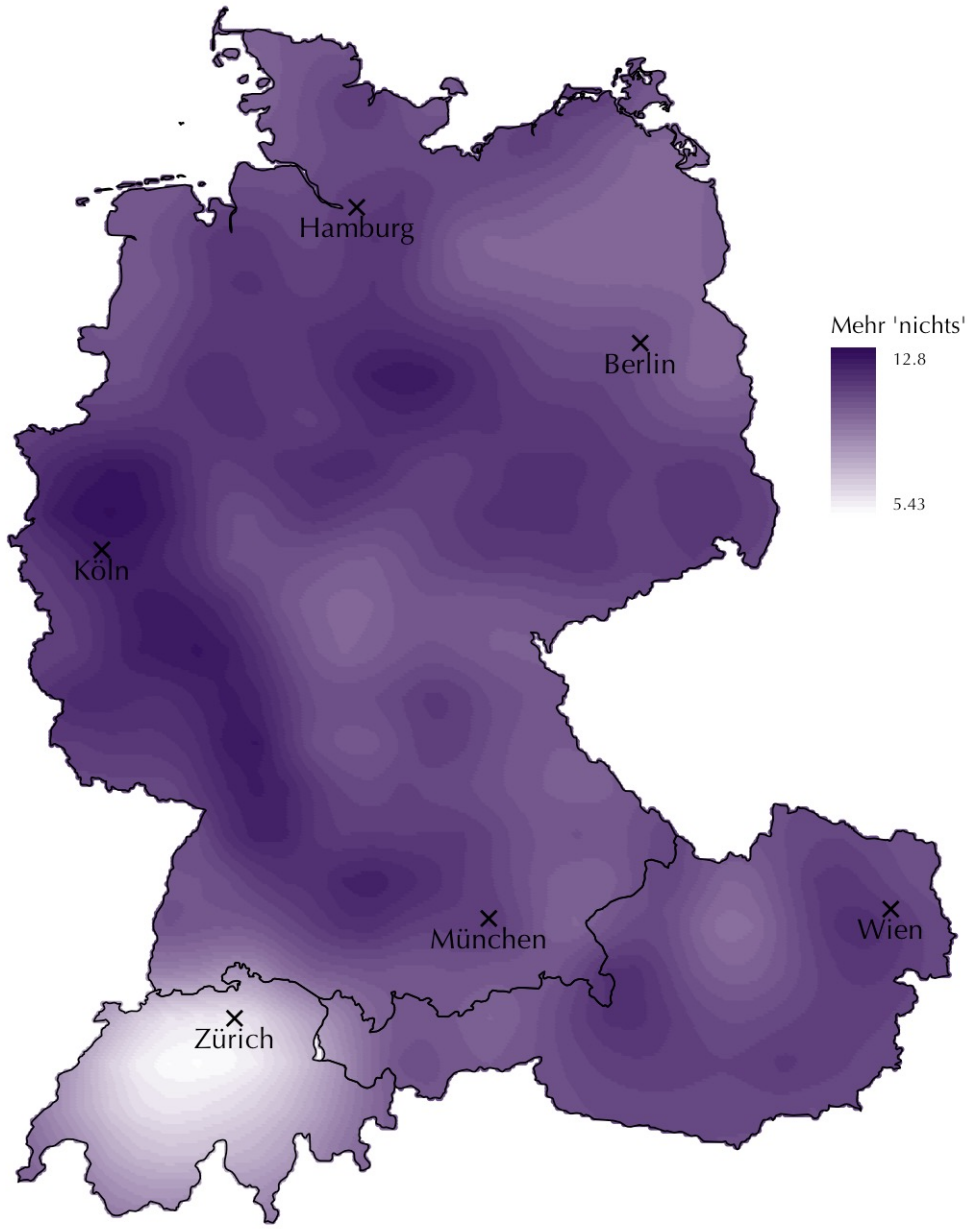
Heeringa & Nerbonne 2001
(cf. Nerbonne et al. 2005)

FIGURE 4. Variants of *zijn* 'to be' in IPA.

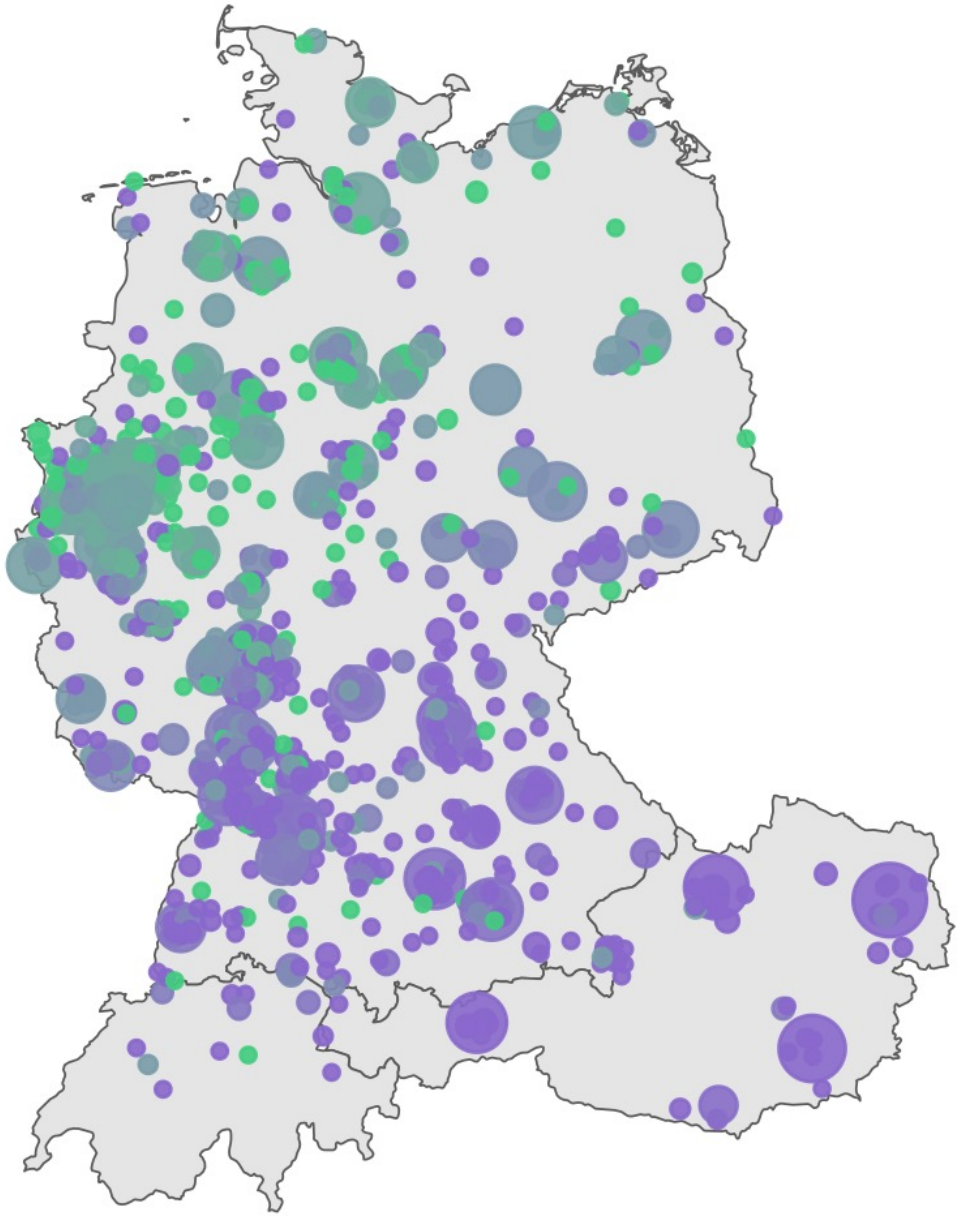
Inferring unobserved values: Variogram models

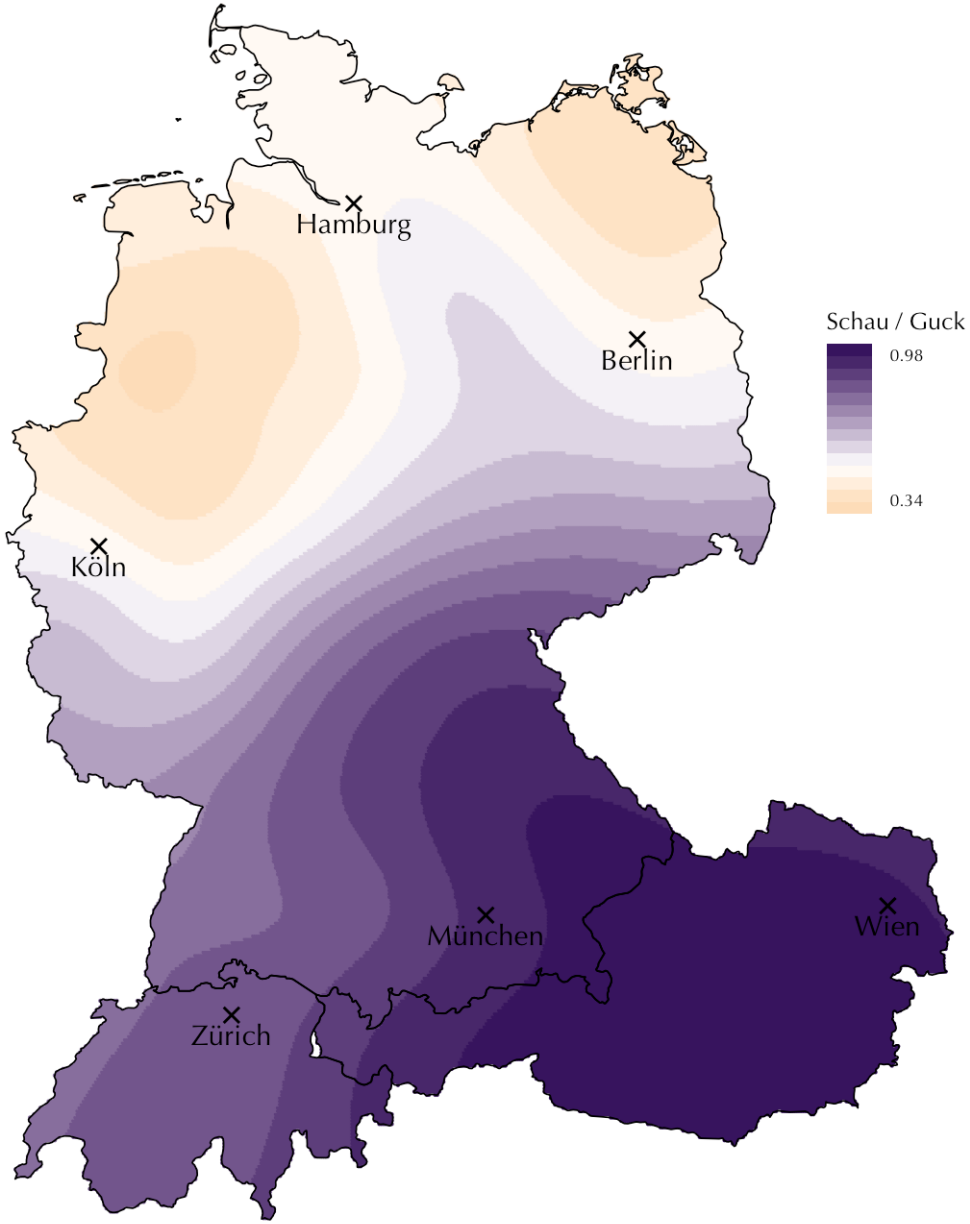


Kriging

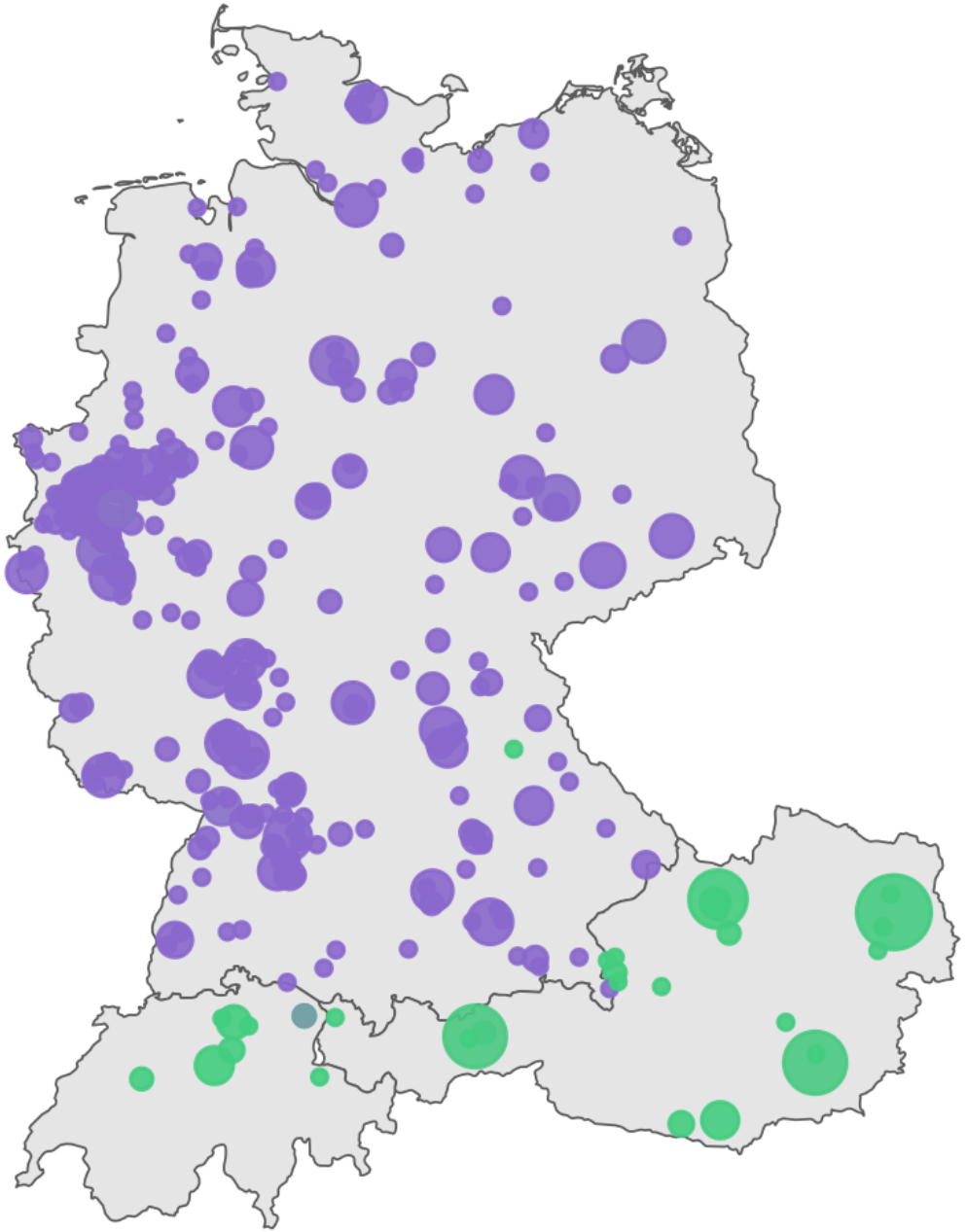

Hamburg
Berlin
Köln
München
Wien
Zürich
Mehr 'nichts'
12.8
5.43

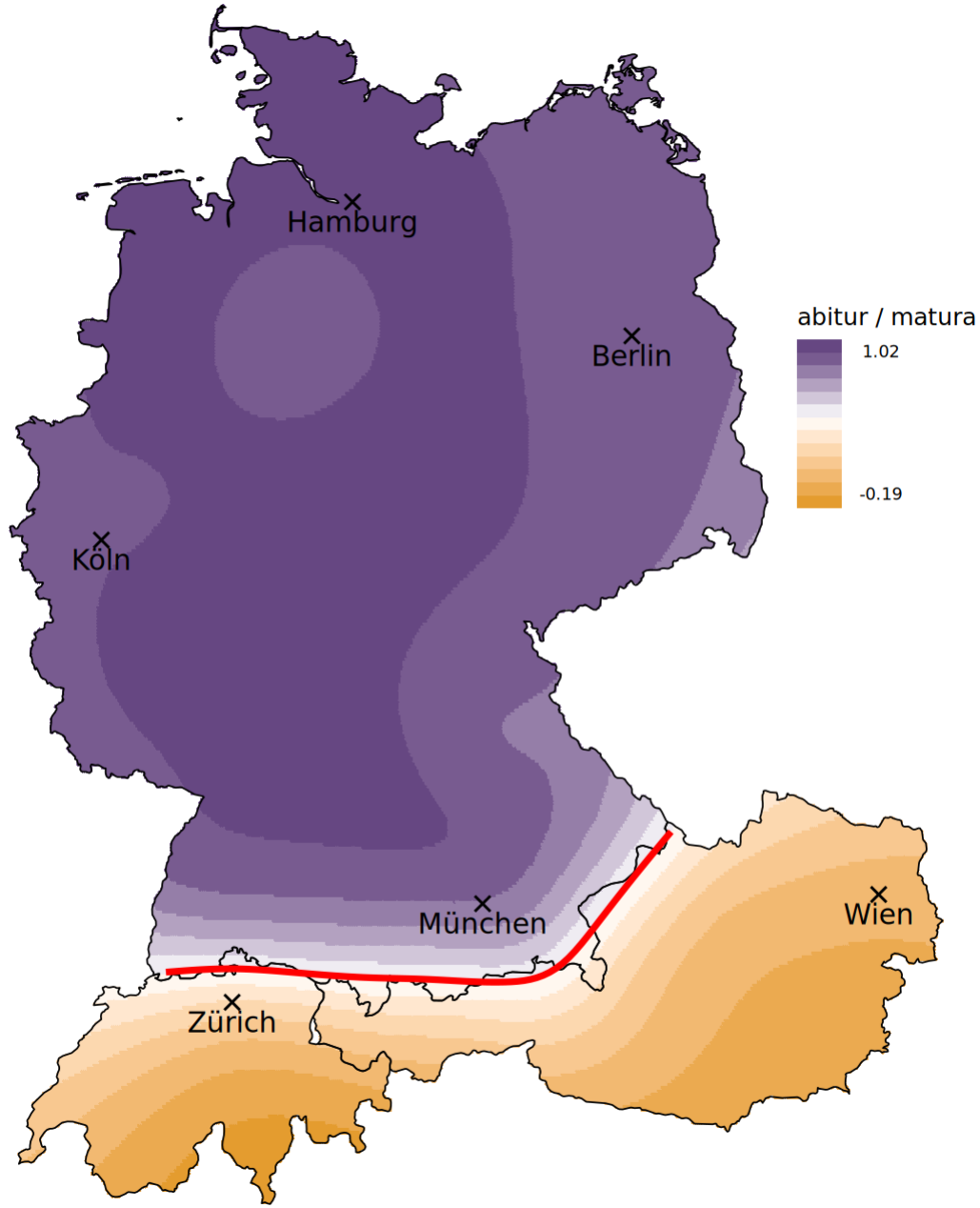
schau & *guck* in the GSA





abitur & *matura* in the GSA



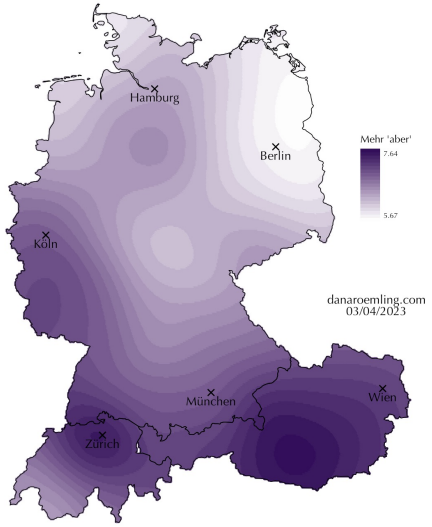


Currently

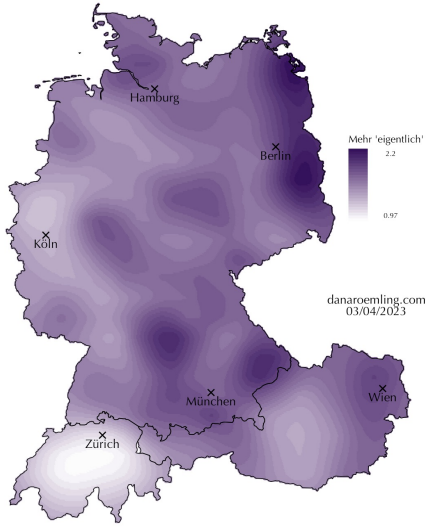
aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😀

1:00 pm

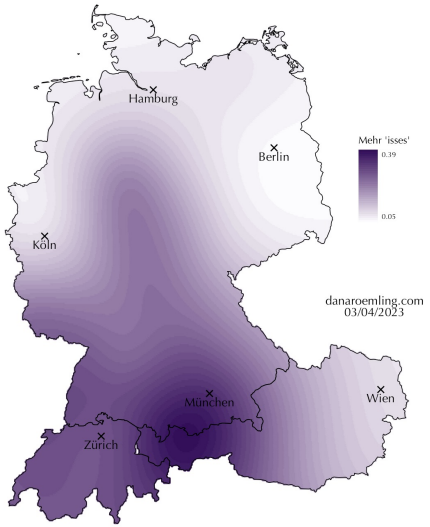
aber



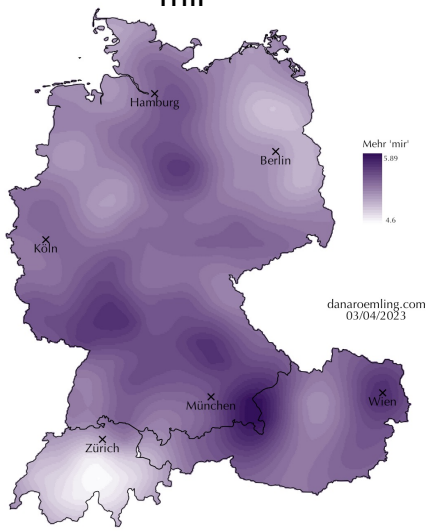
eigentlich



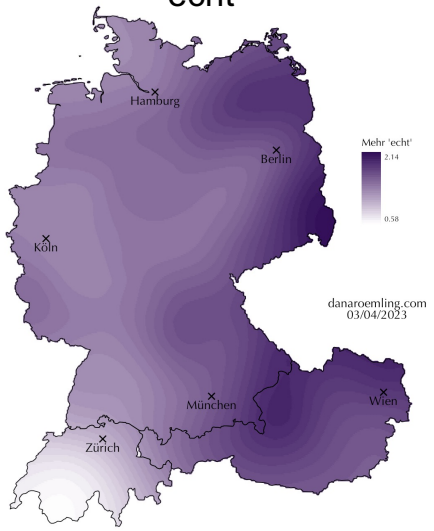
isses



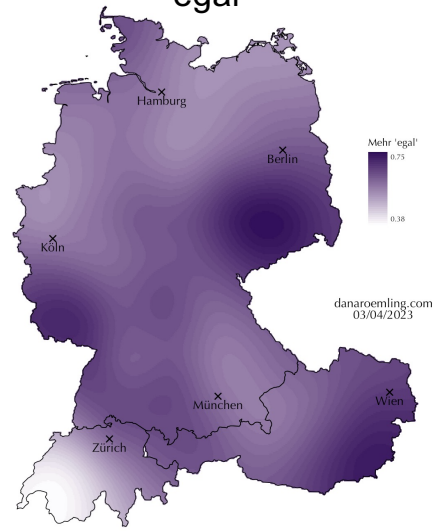
mir



echt



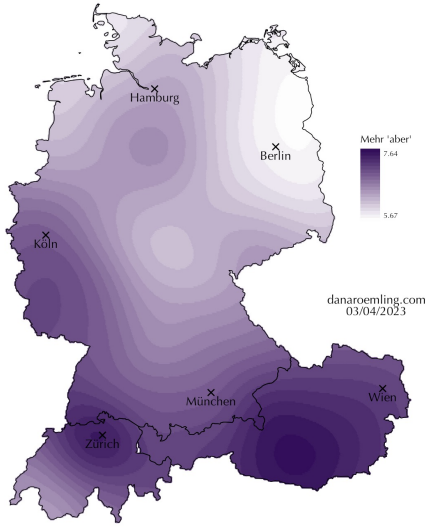
egal

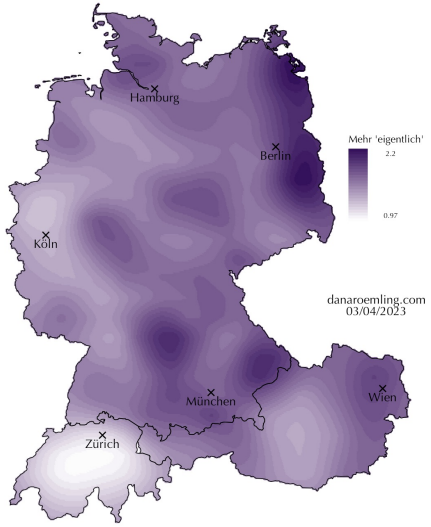


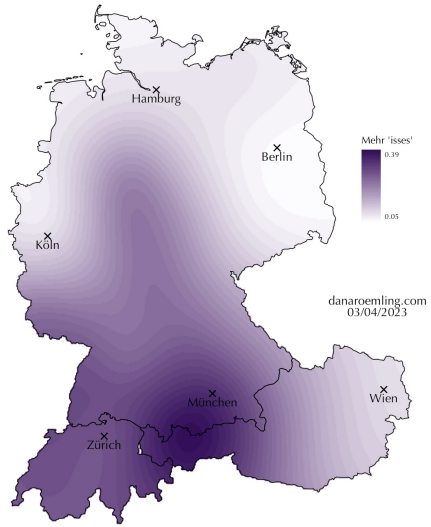
Currently

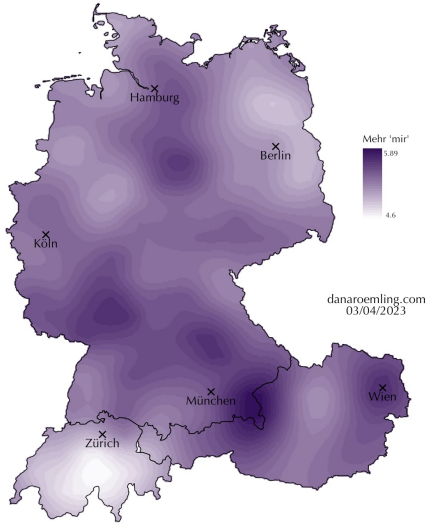
aber eigentlich isses mir echt egal. bin auch offen für kreative Sachen. 😀

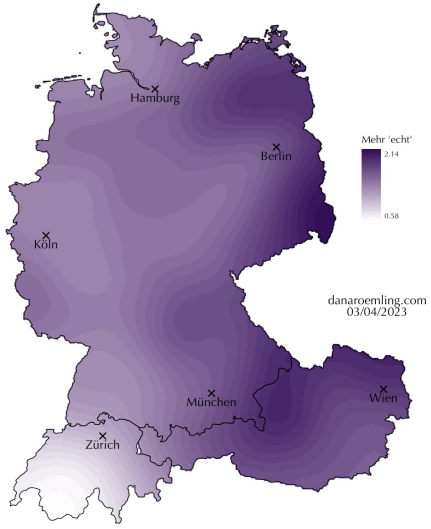
1:00 pm

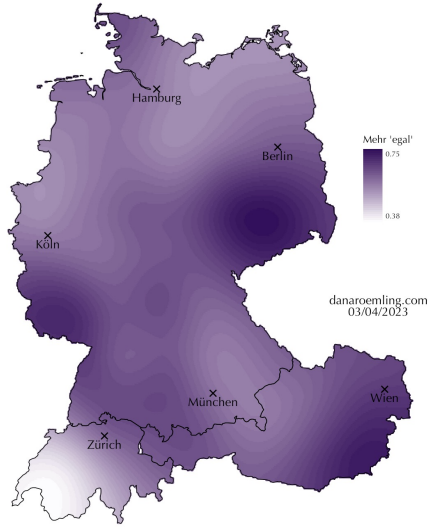


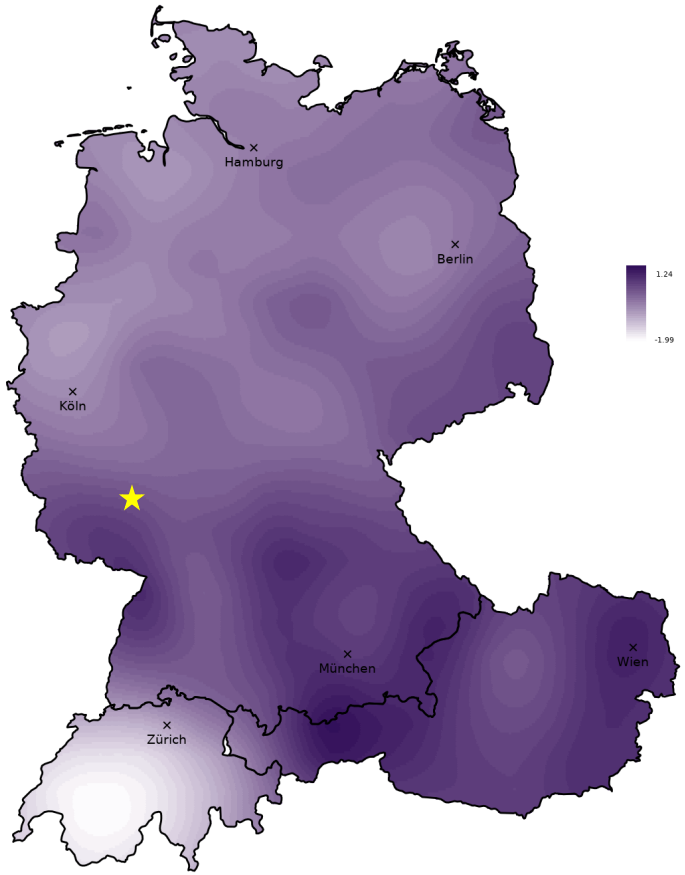




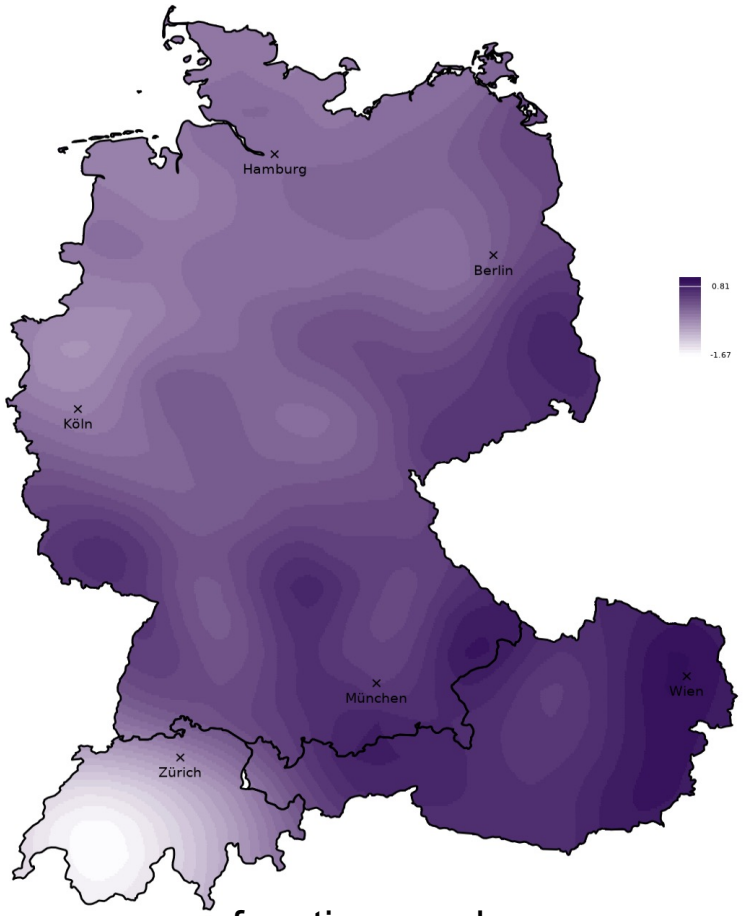




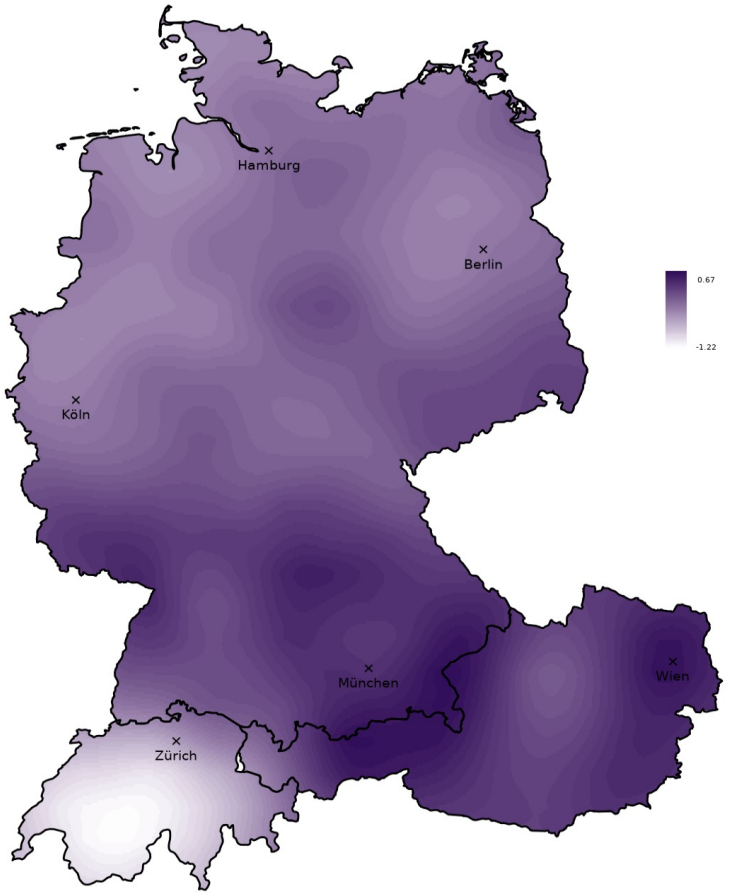




unweighted



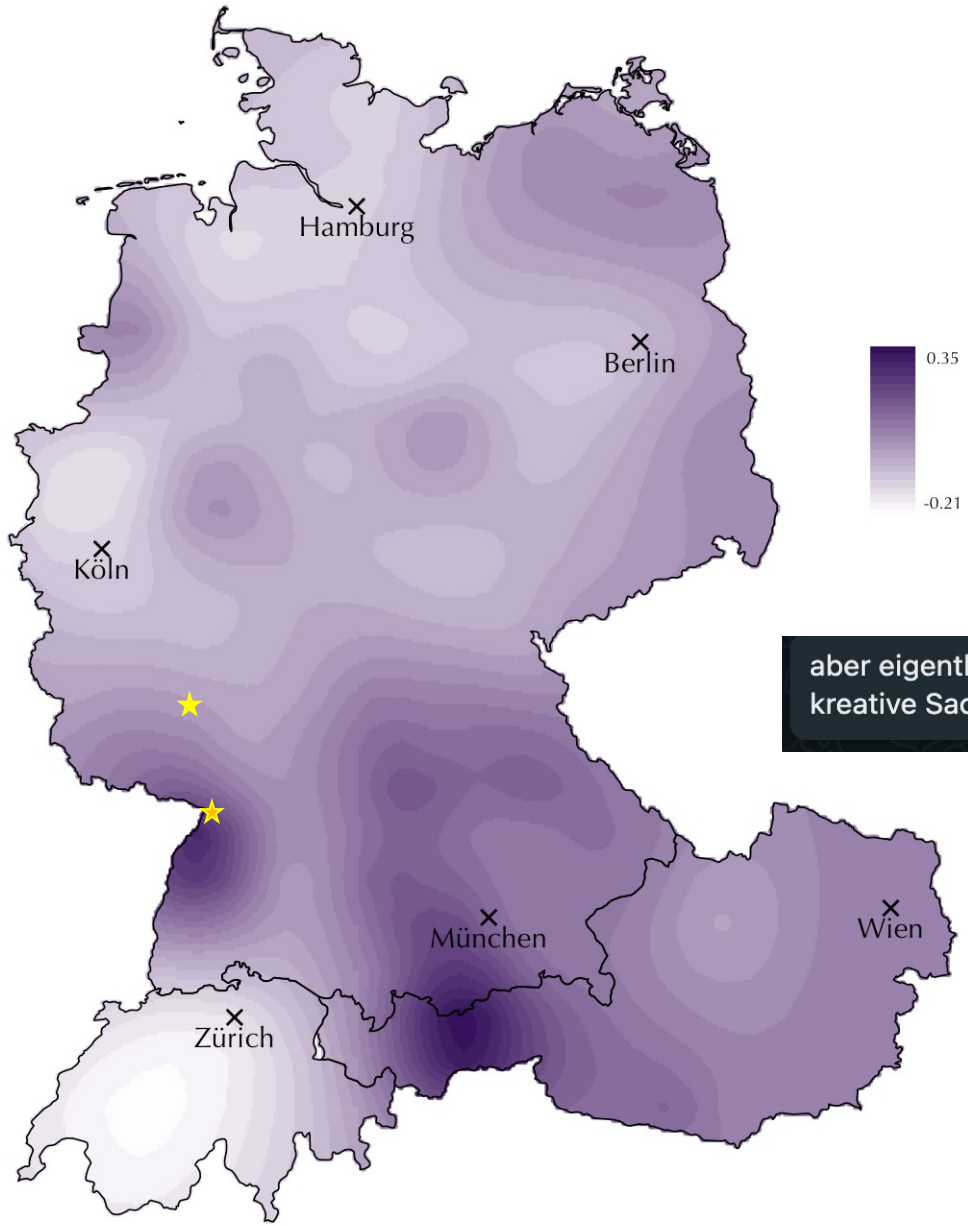
function words



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

The road so far...

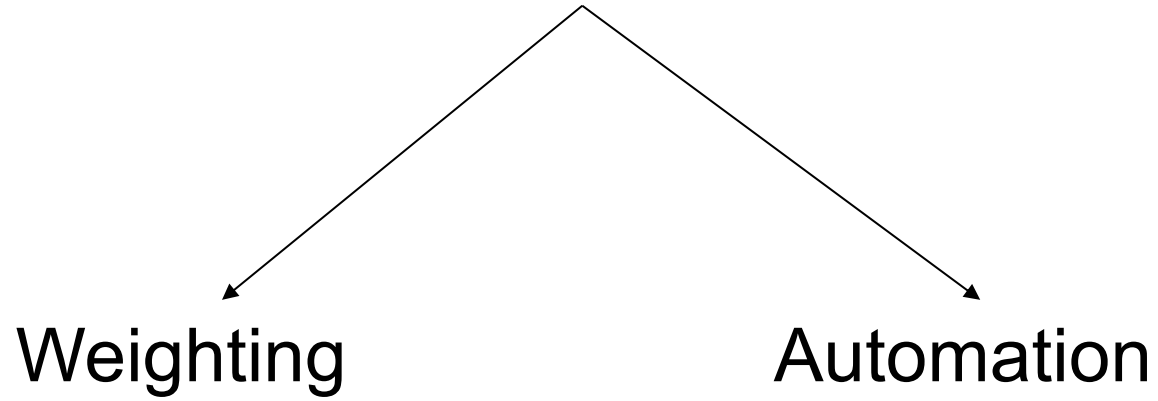
- Jodel corpus & proof of concept
- Cultural / dialect regions in the Jodel corpus
- Handling unobserved locations / kriged maps

→ Corpus as reference tool for qualitative forensic analysis

- Aggregated / weighted maps and (semi-)automated profiling
- Validation: Corpus of forensic texts from the German federal criminal police office (BKA)

What comes next?

- The goal is to produce a profile / inference(s) about an author which help narrow down the suspect list in a transparent / explainable / replicable / reliable way



Discussion

- What comes next, especially during my research visit

More data → a PostDoc plan

- Different German dialect corpora, especially given the text type constraints we know of in authorship analysis

  - Mojedano Batel et al. (2023), Bevendorff et al. (2023):
  Authors are not consistent across text types and authors differ as to which text types are consistent across an author

- Finnish dialect corpora

Weighting

- The corpus is not POS-tagged – unclear how helpful this would be given the dialect data
  - POS instead of lexical items
- Filter by, instead of weight by, Moran's I / Getis-Ord Gi*
- Distance to standard (spelling?) as weight

Automation

- SVM?

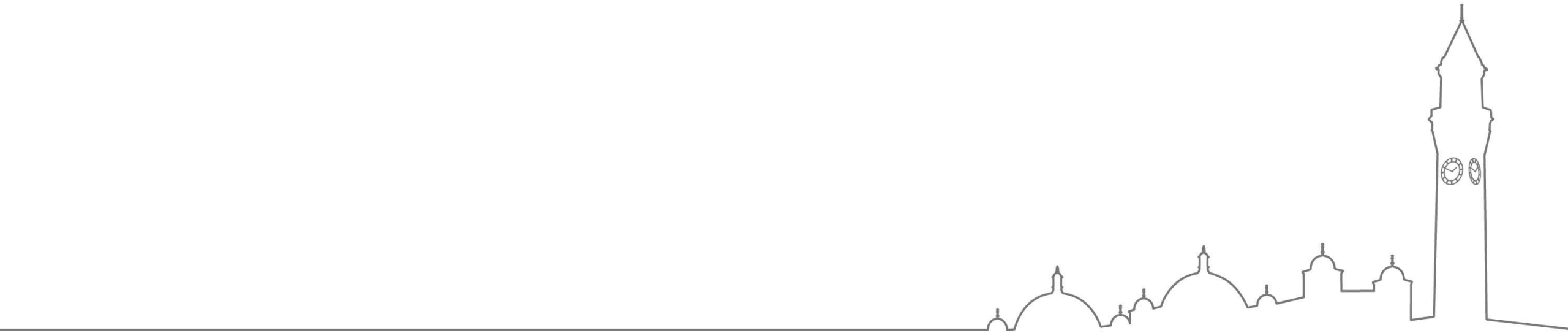
  - Looking at VarDial: SVM perform well and even outperform ML, while being (somewhat) explainable

Thank you!

danaroemling@gmail.com | dana.romling@helsinki.fi

@danaroemling

https://github.com/danaroemling



Selected References

- Bevendorff, J., Chinea-Ríos, M., Franco-Salvador, M., Heini, A., Körner, E., Kredens, K., Mayerl, M., Pęzik, P., Potthast, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., Wiegmann, M., Wolska, M., & Zangerle, E. (2023). Overview of PAN 2023: Authorship Verification, Multi-author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection: Extended Abstract. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, & A. Caputo (Eds.), *Advances in Information Retrieval* (Vol. 13982, pp. 518–526). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28241-6_60
- Cassidy, F. G. (1985). *Dictionary of American regional English*. the Belknap press of Harvard university press.
- Chambers, J. K. (1990). Forensic Dialectology and the Bear Island Land Claim. *Annals of the New York Academy of Sciences*, *606*(1 The Language), 19–31. https://doi.org/10.1111/j.1749-6632.1990.tb37732.x
- Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.
- French, P., Harrison, P., & Lewis, J. W. (2007). R v John Samuel Humble: The Yorkshire Ripper Hoaxer Trial. *International Journal of Speech Language and the Law*, *13*(2), 967. https://doi.org/10.1558/ijsll.2006.13.2.255
- HaCohen-Kerner, Y. (2022). Survey on profiling age and gender of text authors. *Expert Systems with Applications*, *199*, 117140. https://doi.org/10.1016/j.eswa.2022.117140
- 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
- Alberich Buera, N., & Kredens, K. (2023). Estabilidad idiolectal del español a través de cuatro géneros de comunicación: Aportaciones al análisis de autoría forense. *Revista de Llengua i Dret*, *79*, 285–304. https://doi.org/10.58992/rld.i79.2023.3951
- Nerbonne, J., van Gemert, I., & Heeringa, W. (2005). A Dialectometric View of Linguistic “Gravity”. *University of Groningen*, 1–46.
- 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

- Text Icon: https://www.iconsdb.com/black-icons/text-file-5-icon.html
- Ransom Note: https://16sparrows.typepad.com/.a/6a00d834515a1f69e2017d424ea9a3970c-600wi
- 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
- Dialect Regions Wiesinger/König: https://de.wikipedia.org/wiki/Deutsche_Dialekte#:~:text=Die%20deutschen%20Dialekte%20oder%20deutschen,geprägten%20Formen%20der%20deutschen%20Sprache.

- 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