UNIVERSITY OF BIRMINGHAM

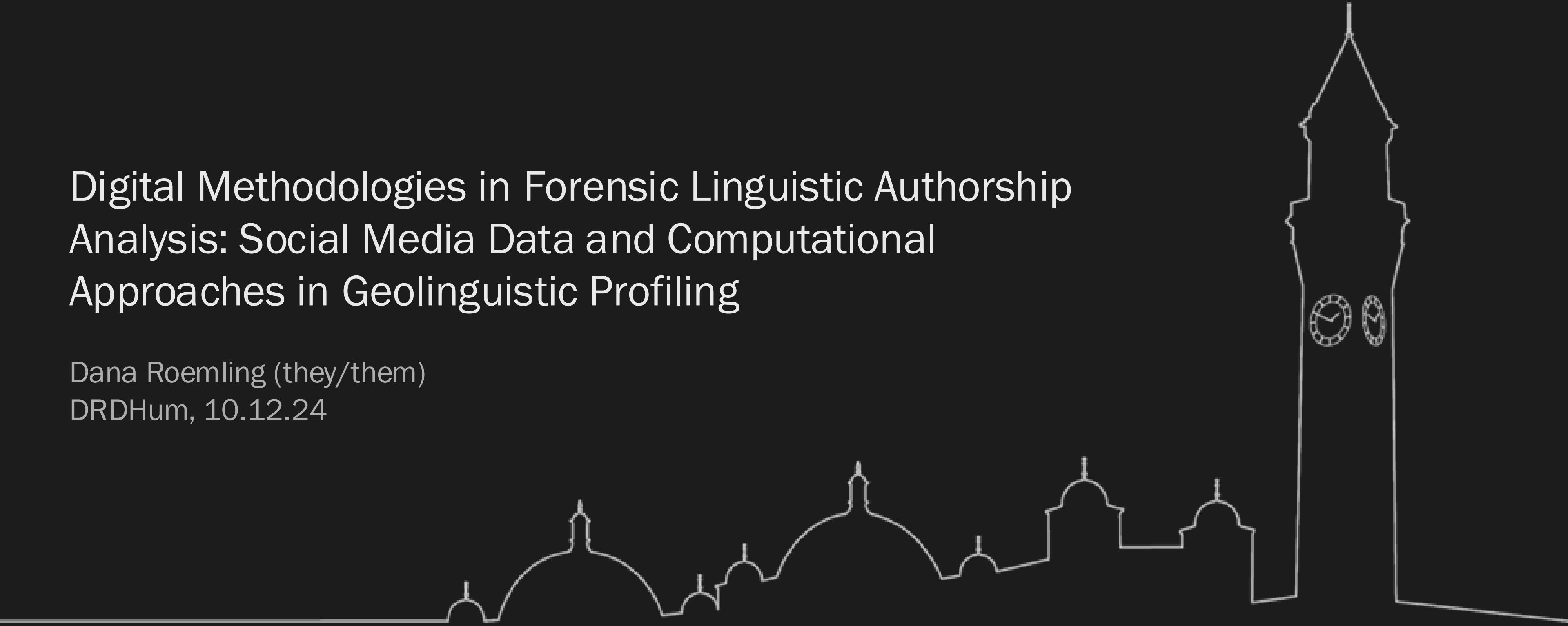
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UKRI Economic and Social Research Council

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

Dana Roemling (they/them)

DRDHum, 10.12.24

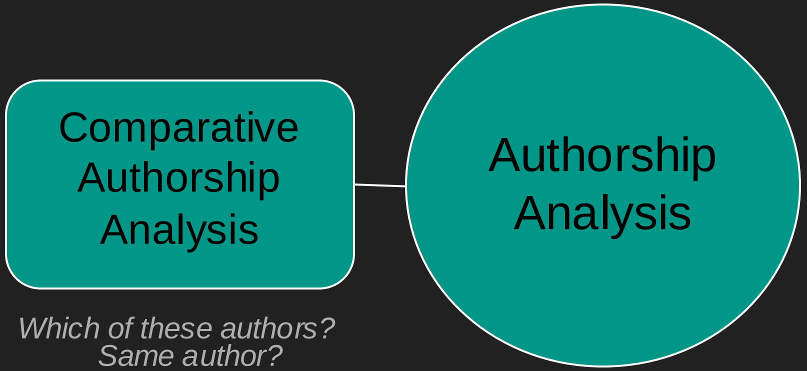


Today

- Brief Intro:
  - A bit of case work
  - Authorship Analysis & Profiling
- Dialect / Geolinguistic Profiling:
  - Corpus & Regional Variation
  - Geostatistics / Map Interpolation
  - Forensic Application

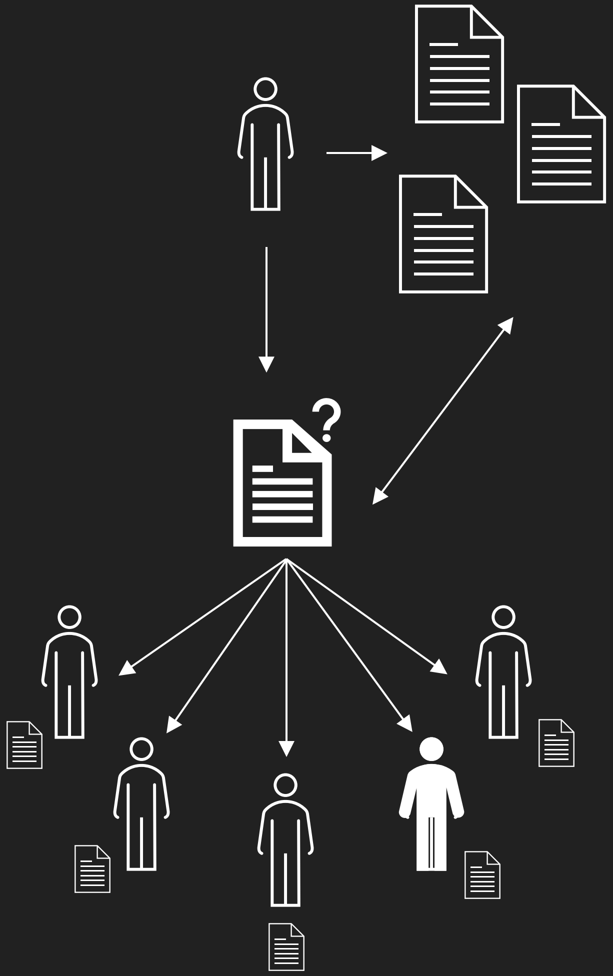
- The application of linguistic methods, approaches and knowledge to investigative, criminal or legal contexts
- Language as evidence or language of the legal process
- Examples: ransom note, text messages in murder investigation, suicide note
- One focus area: authorship analysis

Authorship Analysis

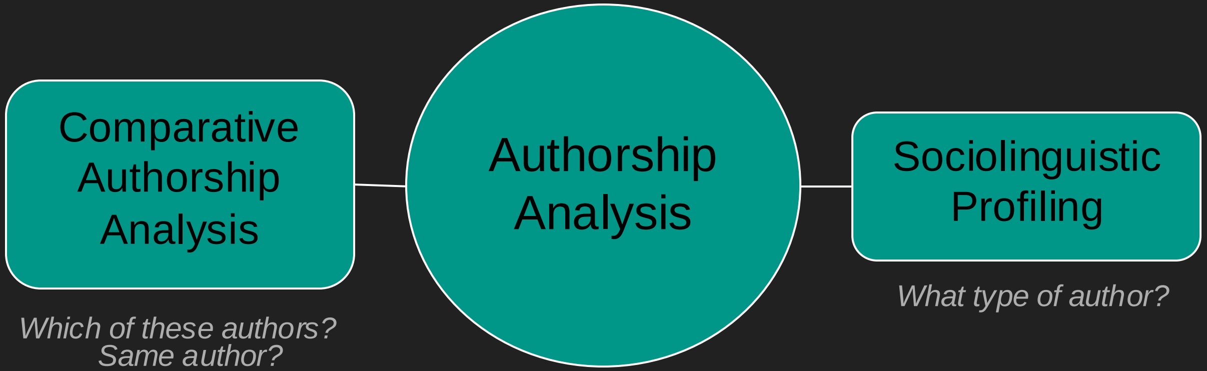


Which of these authors?
Same author?

Wright, 2020; Grant, 2022



Authorship Analysis



Comparative Authorship Analysis

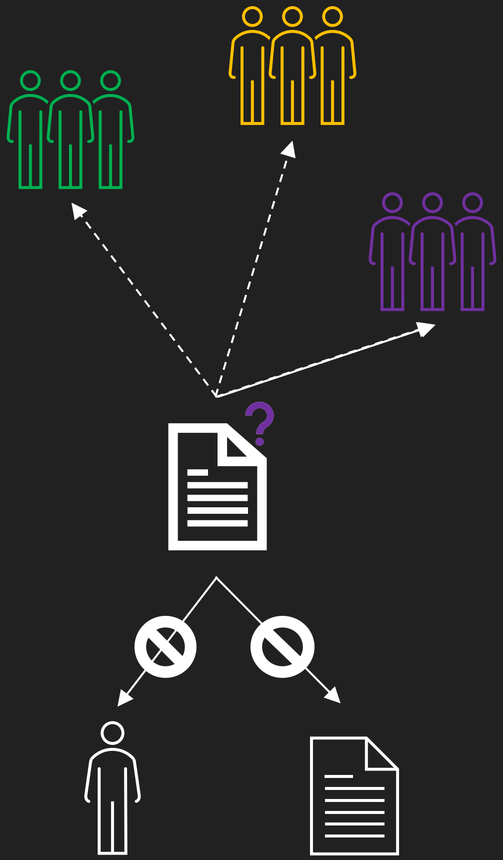
Authorship Analysis

Sociolinguistic Profiling

What type of author?

Which of these authors?
Same author?

Wright, 2020; Grant, 2022



Analysis of texts to infer characteristics about an author

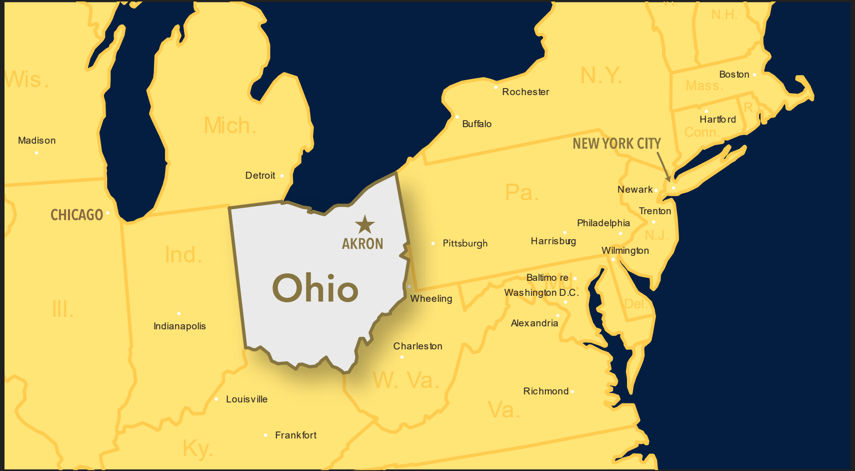
- e.g. age, gender, native language influence, region

“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!!!”

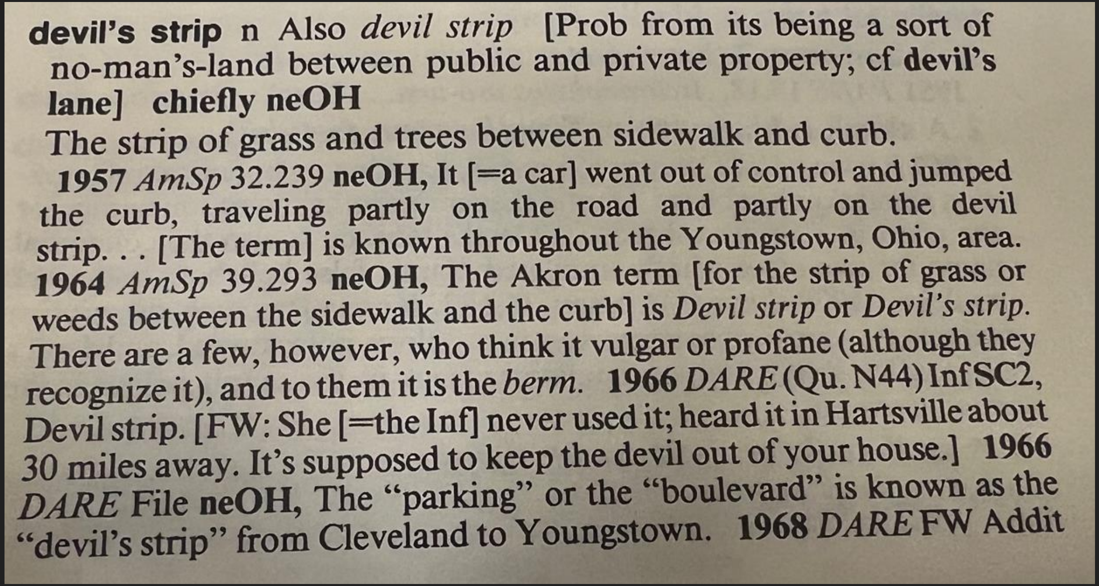
(Shuy, 2001, p.1)

“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

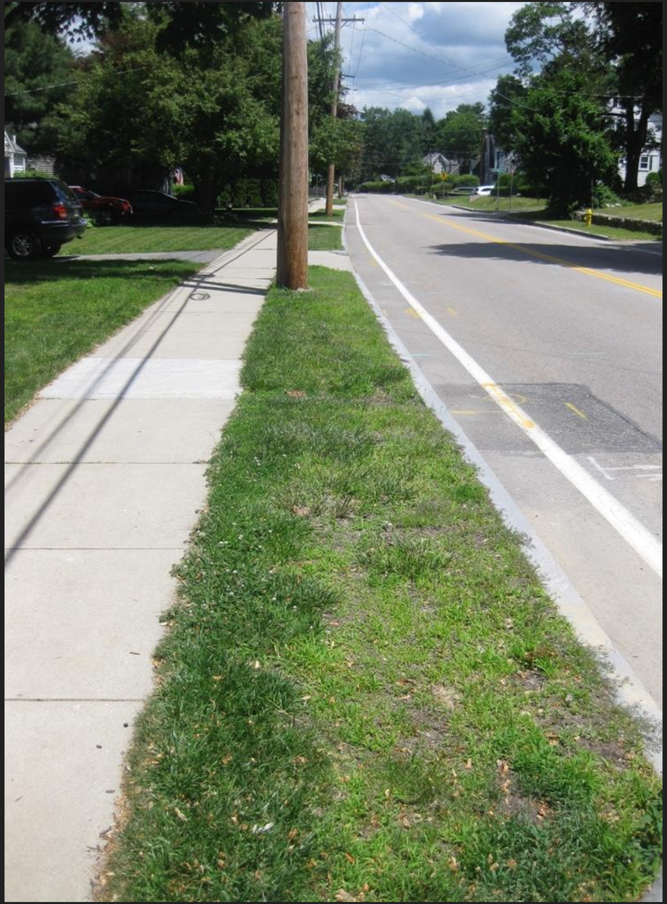


devil's strip n Also *devil strip* [Prob from its being a sort of no-man's-land between public and private property; cf **devil's lane**] **chiefly neOH**

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



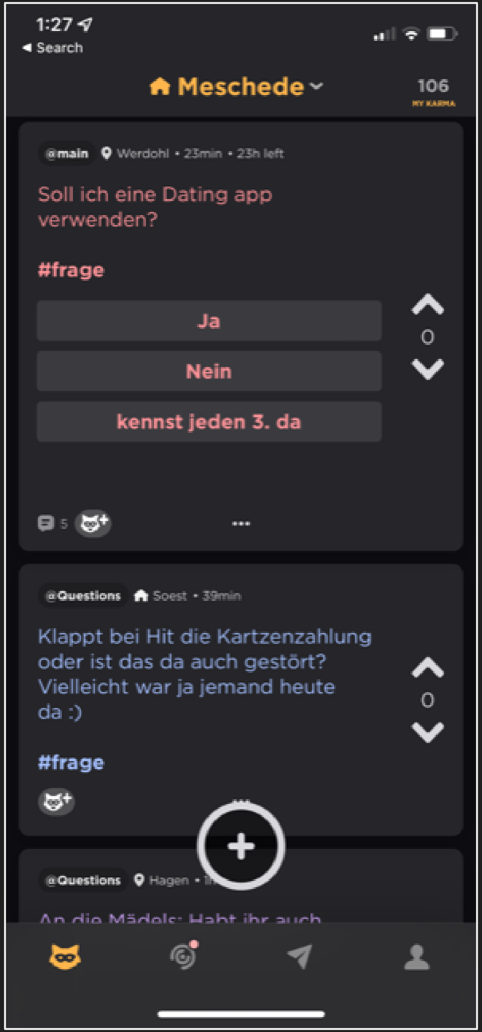
Regional Authorship Profiling

- Based on the expert's knowledge of regional variation
- Based on previous work in dialectology / sociolinguistics
  - Works with elicited, often not digitised data (surveys, interviews)
  - Data takes considerable time to be gathered and analysed and is often outdated given how rapidly language and regional variation changes

- This is why I focus on both social media data and geostatistical methods

The Jodel Corpus

- 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)
- No accounts / profiles, so “anonymous” interaction
- The corpus has 239,151,815 tokens at ~8000 unique locations in the GSA



@main Werdohl • 23min • 23h left

Soll ich eine Dating app verwenden?

#frage

Ja

Nein

kennst jeden 3. da

@Questions Soest • 39min

Klappt bei Hit die Kartzenzahlung oder ist das da auch gestört? Vielleicht war ja jemand heute da :)

#frage

@Questions Hagen •

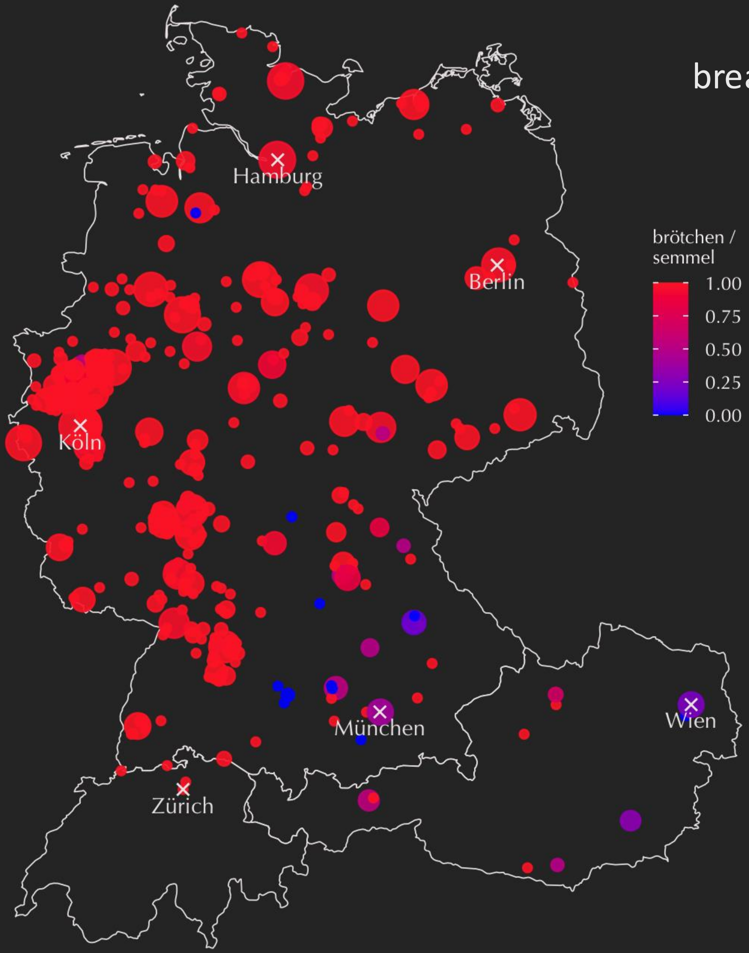
An die Mädels: Habt ihr auch

The Jodel Corpus: Sample

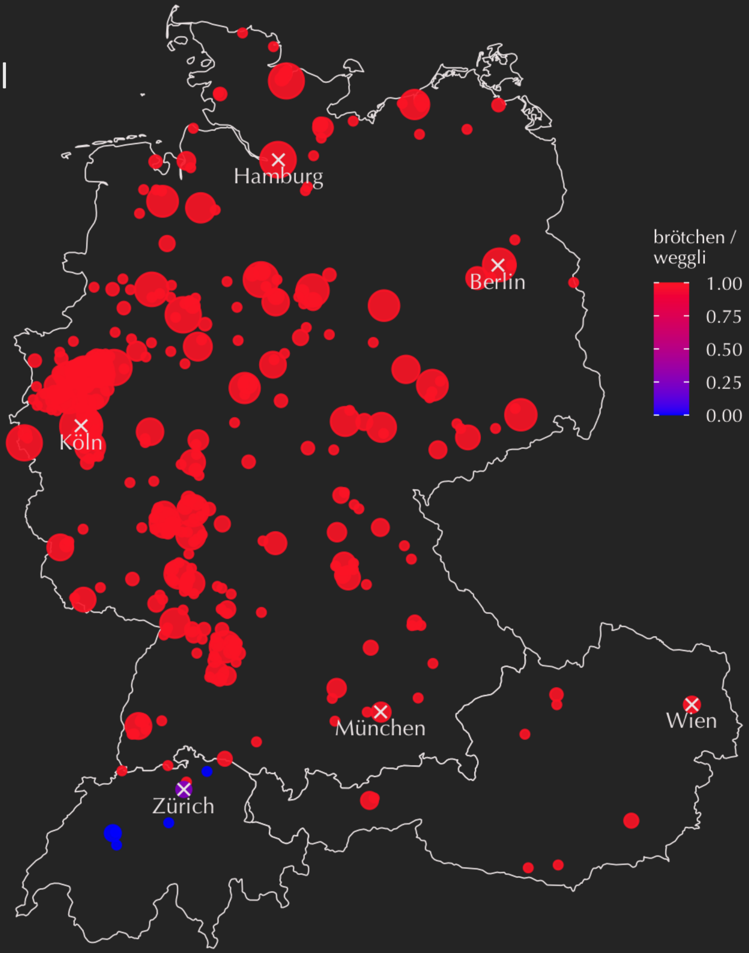
Message	Creation Timestamp	Location	Post ID
Semesterferien: grillen schlafen grillen bar schlafen repeat.. English: *semester break: bbq sleep bbq pub sleep repeat..*	2017-04-04T 22:29:41.814Z	Berlin	58e41e5512e80 a3f0cb6f66b
Ich bin grade leicht verwirrt Werdet ihr Mädels so emotional, kurz bevor ihr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

F

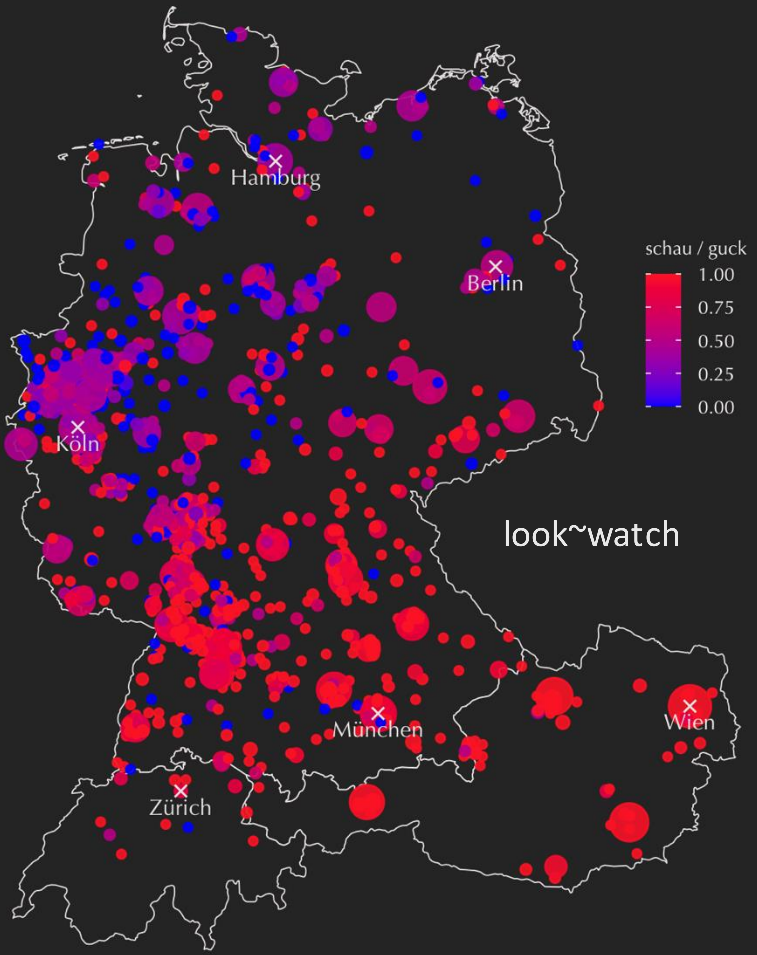
bread roll



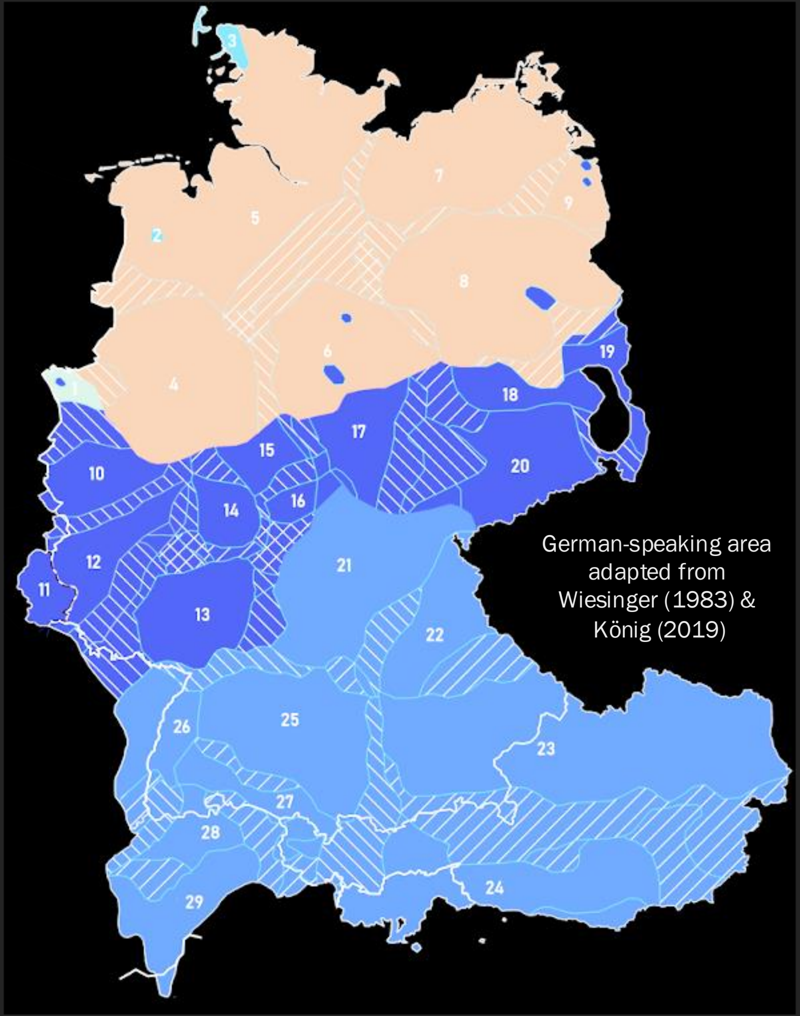
Hamburg
Berlin
Köln
München
Wien
Zürich
brötchen / semmel
1.00
0.75
0.50
0.25
0.00



Hamburg
Berlin
Köln
München
Wien
Zürich
brötchen / weggli
1.00
0.75
0.50
0.25
0.00



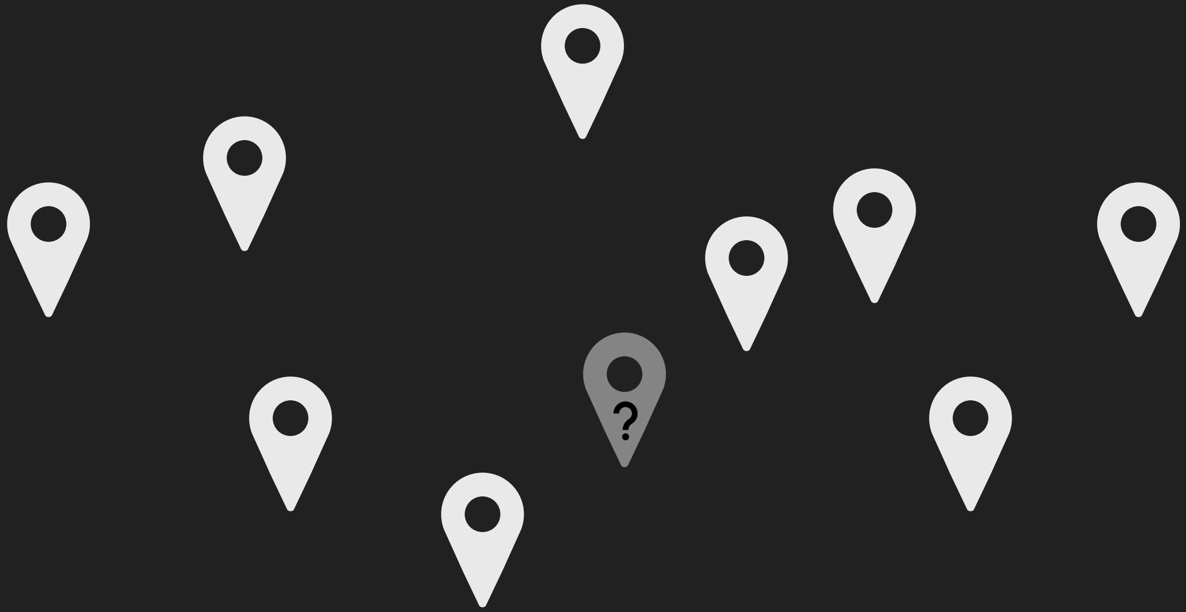
look~watch



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

Accounting for unobserved locations

Ordinary Kriging



Similarity

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

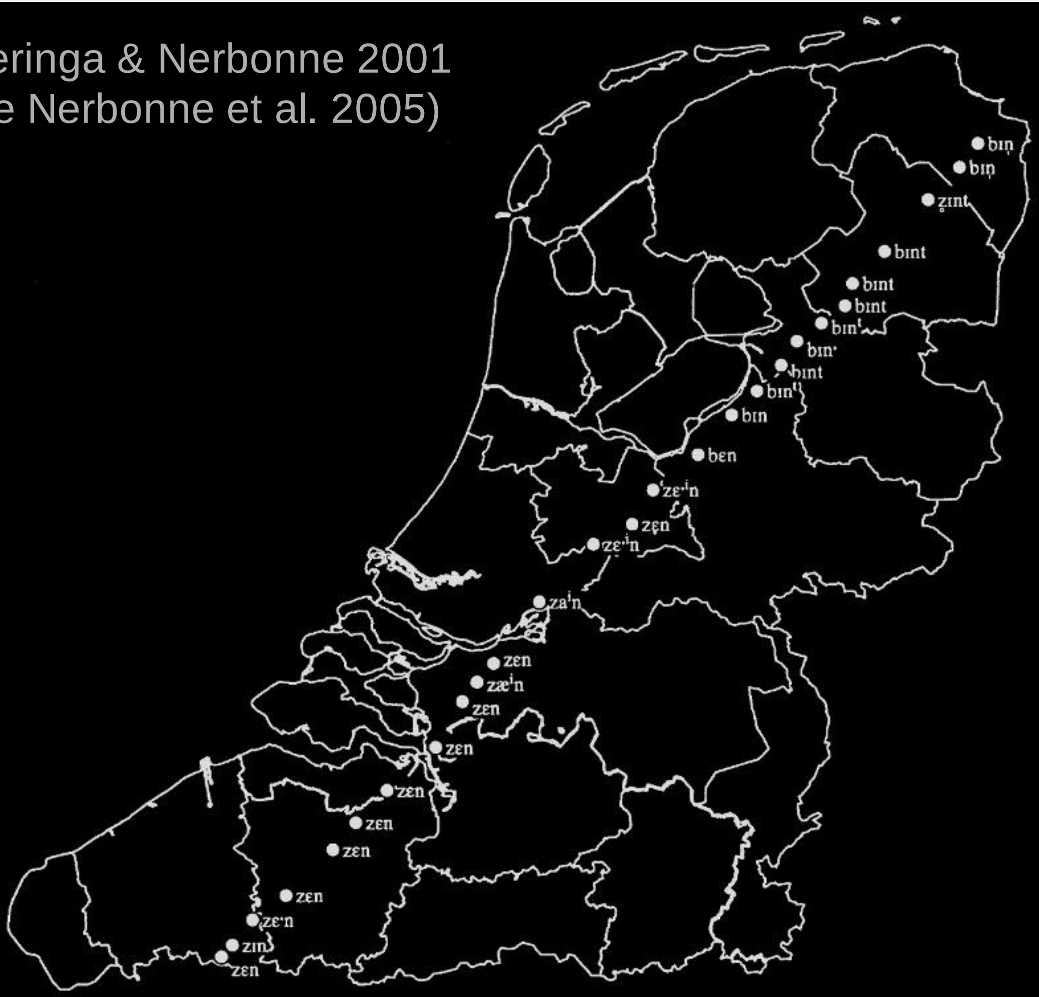
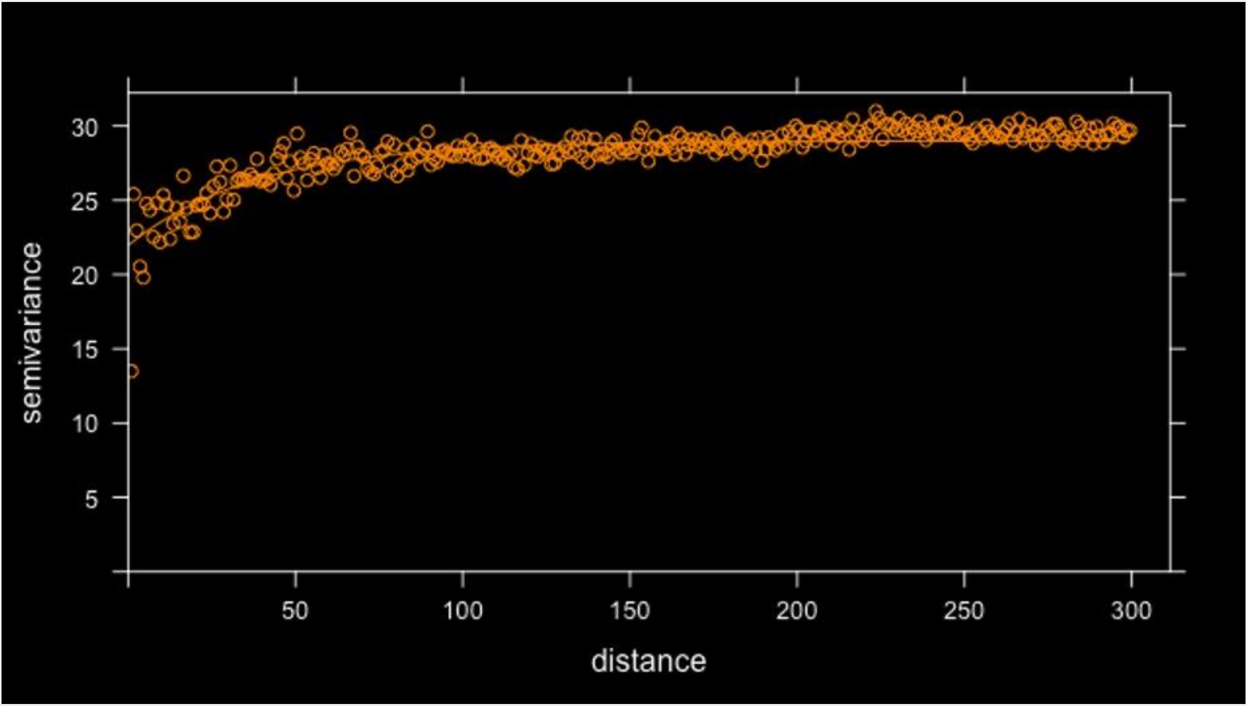
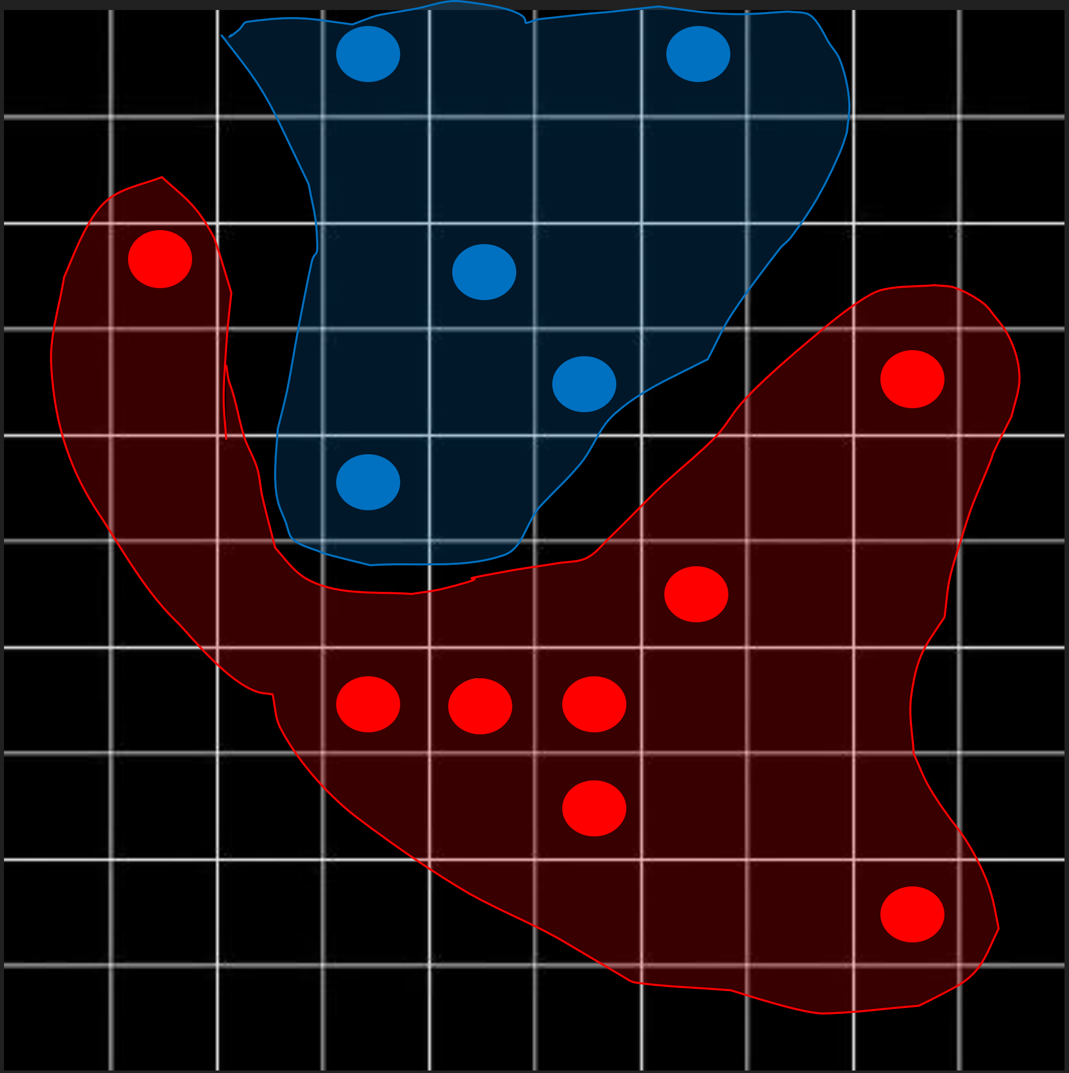
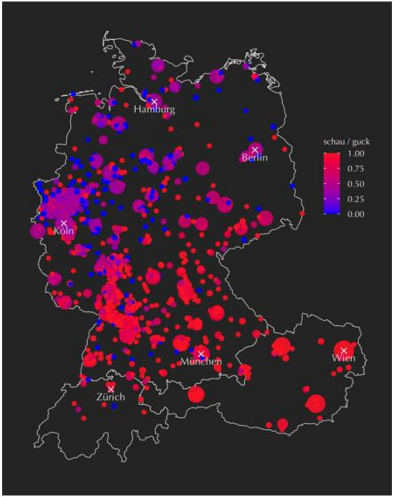


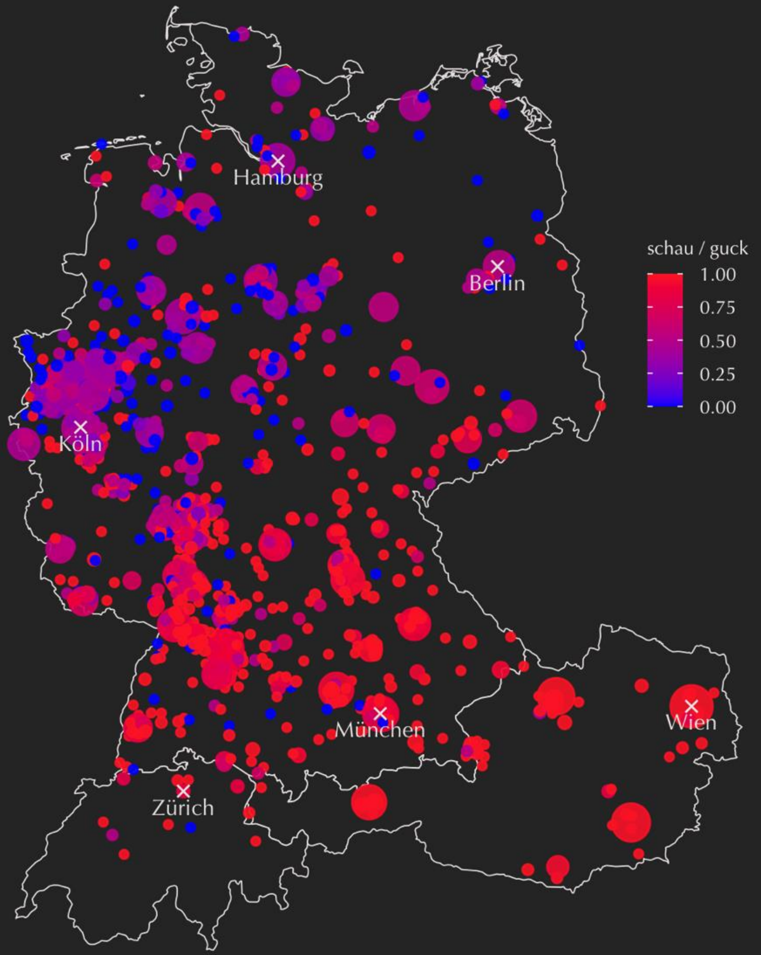
FIGURE 4. Variants of *zijn* ‘to be’ in IPA.

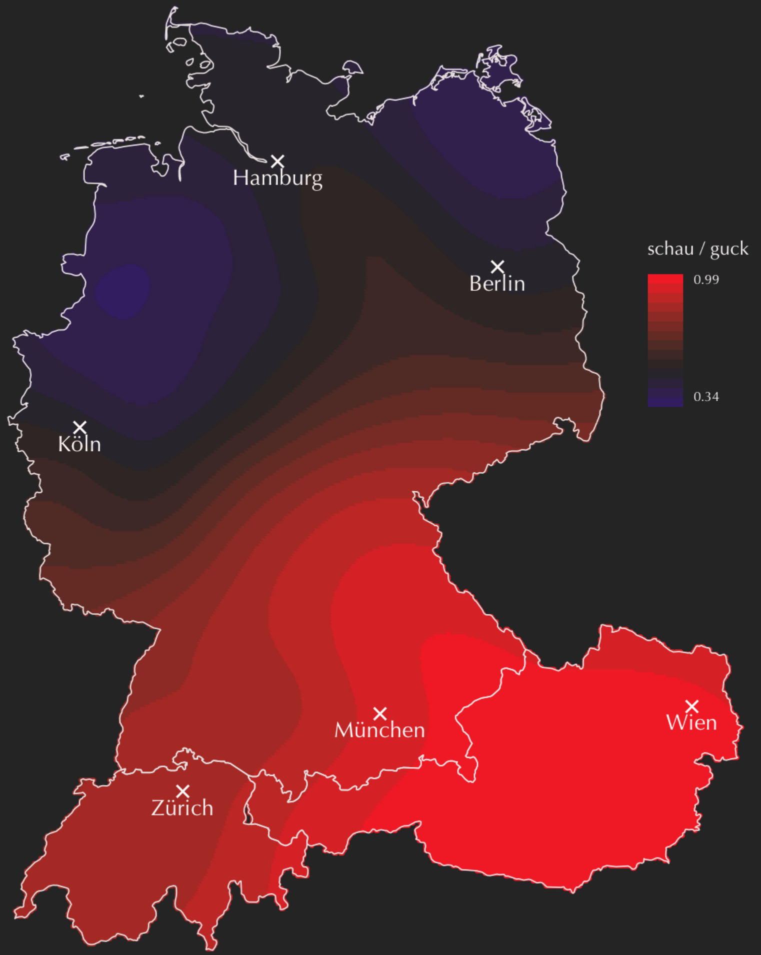
Similarity & Distance






Hamburg
Berlin
Köln
München
Wien
Zürich
schau / guck
1.00
0.75
0.50
0.25
0.00


Hamburg
Berlin
Köln
München
Wien
Zürich
schau / guck
1.00
0.75
0.50
0.25
0.00

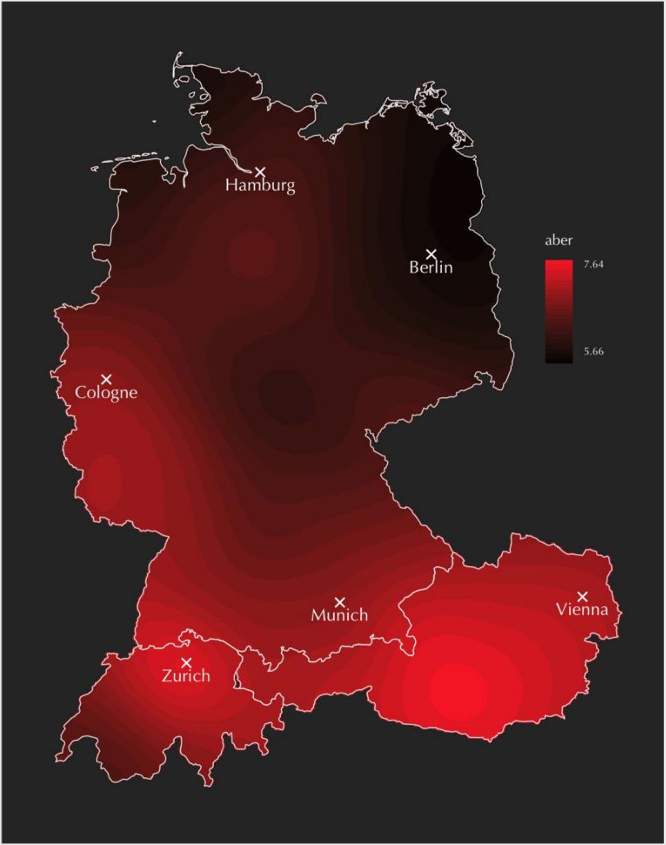

Hamburg
Berlin
Köln
München
Wien
Zürich
schau / guck
0.99
0.34

Using kriged maps for location prediction

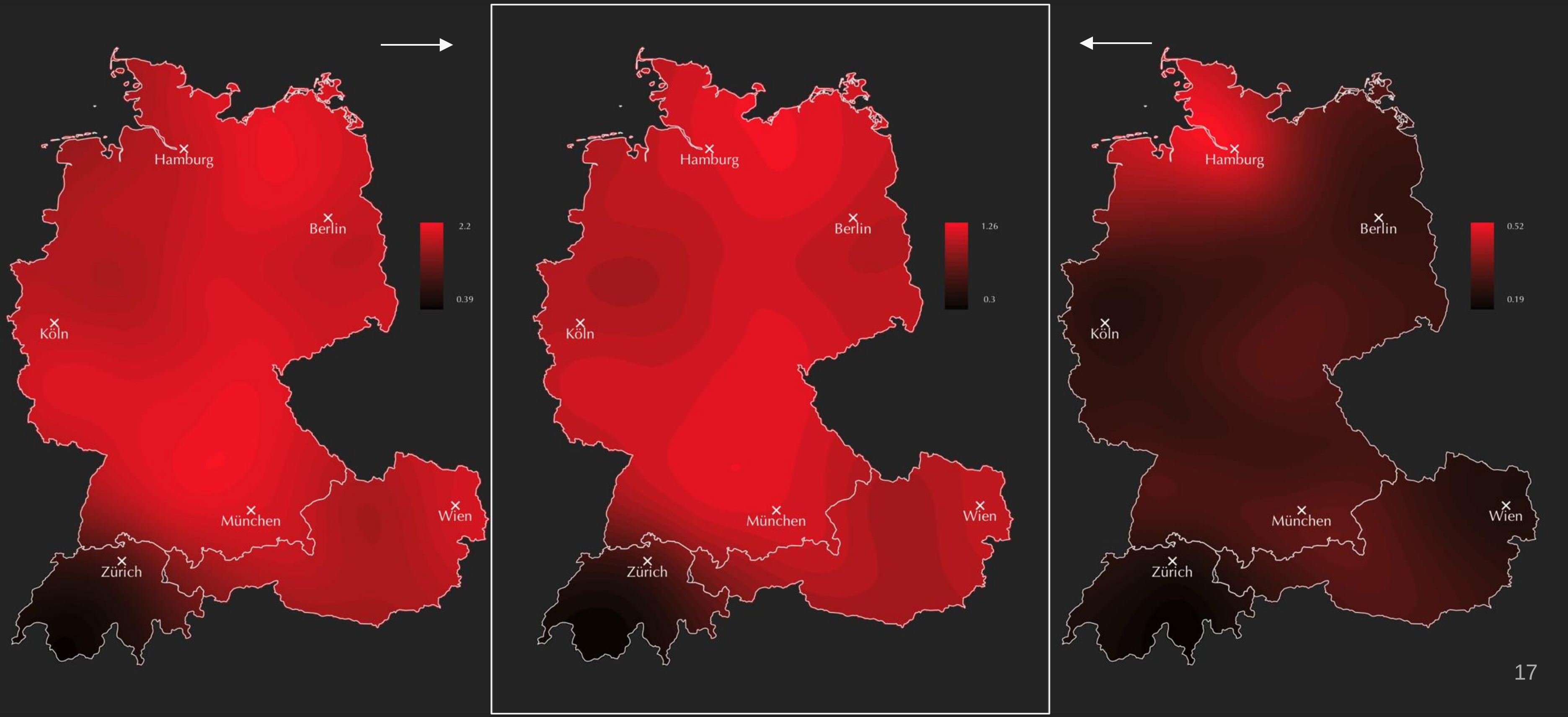
- The maps are a visualisation of areas of high / low feature use (in this corpus)

→ Prediction of country

→ Prediction of location (more narrowly)



“keine polizei“ (*no police*)



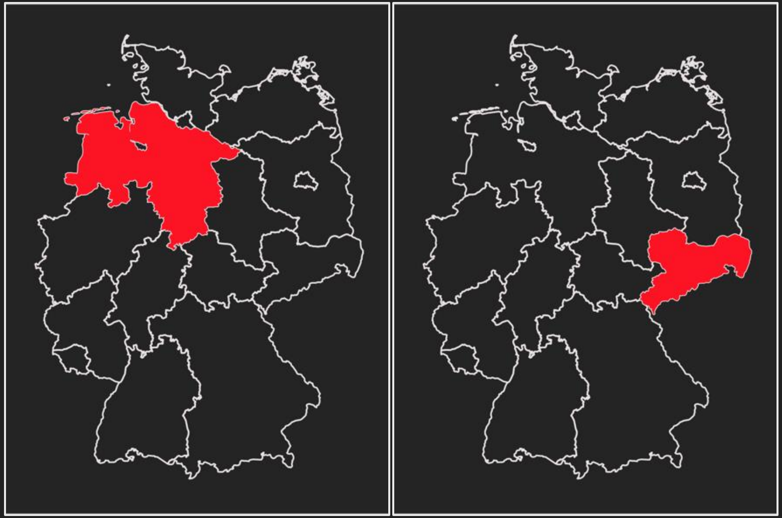
Jodel

kennt ihr diesen Hans Entertainment Moment, wenn du Niveau in einem anonymen messageboard erwartest? ich nicht.

Do you know that Hans Entertainment moment when you expect class on an anonymous message board? I don't.

Country: DE

Within country:

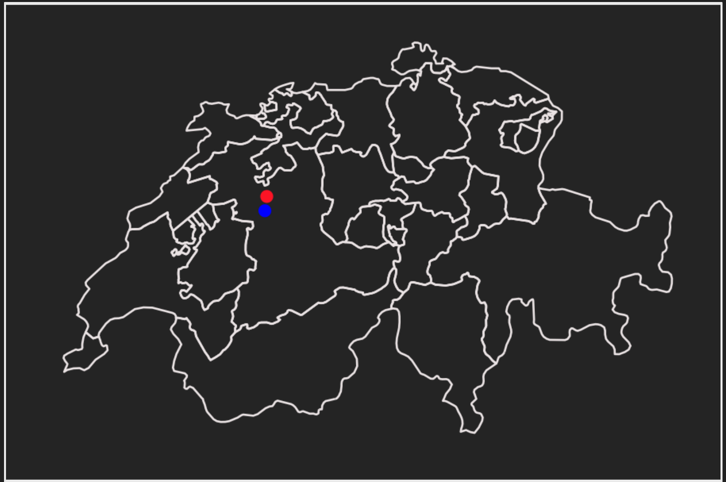


iviu Fieber hesch? Hesch gmässe? Wed Fieber hesch würdi itz auso no chly warte. Sobauds Fieber wäg isch chasch ja ga

Do you have a fever? Did you take [your temperature]? If you have a fever I would wait a bit. As soon as the fever is gone you can go.

Country: CH

Within country:



Next steps

- For some areas regional profiling works better than for others
- Improving the prediction by:
  - Weighting words based on:
    - Lower average distance (between their hotspot and predicted locations)
    - Frequency (higher or lower)
    - Moran’s I values
    - A combination of these

Concluding

- The Jodel corpus and its regional variation can serve as a reference tool for qualitative (forensic) analysis
- Features with high regional variance can be extracted
- Interpolation reduces noise, aids visualisation, makes data comparable, and helps explain the methodology for legal use
- Regional profiling is effective, depending on author/area

Thank you!

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

@danaroemling

https://github.com/danaroemling

danaroemling.com

Selected References

- Carver, C. M. (1987). American Regional Dialects. University of Michigan Press.
- Chambers, J. K., & Trudgill, P. (1998). Dialectology (2nd ed). Cambridge University Press.
- Kretzschmar, W. A. (2006). Art and Science in Computational Dialectology. Literary and Linguistic Computing, 21(4), 399–410. https://doi.org/10.1093/llc/fql033
- Grieve, J. (2015). Dialect variation. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 362–380). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.021
- 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
- Kurath, H. (1949). A word geography of the Eastern United States. University of Michigan Press.
- Labov, W., Ash, S., & Boberg, C. (2006). The construction of isoglosses. In The atlas of North American English: Phonetics, phonology and sound change (pp. 42–44). Mouton de Gruyter.
- Pi, T. (2006). Beyond the Isogloss: Isographs in Dialect Topography. The Canadian Journal of Linguistics / La Revue Canadienne de Linguistique, 51(2), 177–184. https://doi.org/10.1353/cjl.2008.0011
- Séguy, J. (1973). La dialectometrie dans l'Atlas linguistique de la Gascogne. Revue de Linguistique Romane, 37, 1–24.
- Shuy, R. W. (1962). The North-Midland Dialect Boundary in Illinois. University of Alabama Press.
- Wackernagel, H. (2003). Ordinary Kriging. In H. Wackernagel, Multivariate Geostatistics (pp. 79–88). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-05294-5_11