

Language Identification for Non-Standard Varieties

Language Varieties as Geolects on a Geodesic Grid

Sina Ahmadi

Rico Sennrich

Department of Computational Linguistics
University of Zurich

October 26, 2026



Background

Background

- Language identification (LID) → filtering corpora for pretraining LLMs (Oepen et al., 2025)
- Solved for standard varieties, **not for non-standard ones** (Caswell et al., 2020)
- It assumes a **fixed inventory** of discrete labels with names and clear boundaries
- Non-standard varieties have **no agreed name** and often no ISO code!
- The **label set itself** is the bottleneck

Let's Redefine LID!

“Drmit ischsi o für vili Bürgerlechi wäubar.”

(This also makes her electable for many bourgeois voters.)

What language variety is this?



Swiss German (gsw)

reframe



Where on earth could this text have been produced?

⇒ a **geolect**



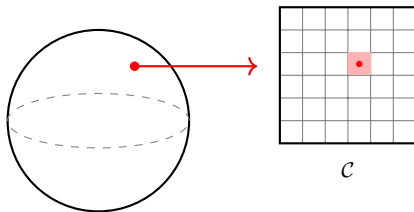
● **Location (*Bern*)**

The Idea: A Geodesic Grid

- Discretize the earth into a grid of **square cells** $\mathcal{C}$
- Given grid resolution g (degrees):

$$x = \left\lfloor \frac{\lambda + 180}{g} \right\rfloor, \quad y = \left\lfloor \frac{\phi + 90}{g} \right\rfloor$$

- A text $\rightarrow$ a **cell**, not a float
- Classification over cells, not labels
- Every prediction is a **location**: cells map directly to coordinates



Related Work in Brief

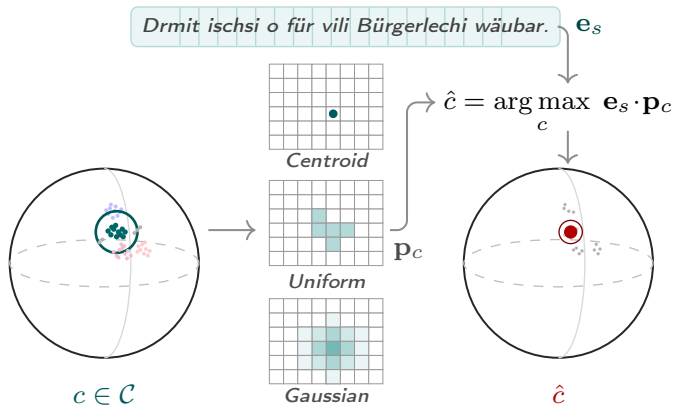
- Document classification into grid cells with bag-of-words features (Wing and Baldrige, 2011; Roller et al., 2012)
- Smoothing sparse grid counts with Gaussian kernels for fine-grained cells (Hulden et al., 2015)
- Neural classifiers for tweet geolocation (Miura et al., 2017)
- Geolocating speech from radio broadcasts via self-supervised representations (Foley et al., 2024)
- Interpolating ASR performance across Italian dialect areas using geographic proximity (Shim and Plank, 2025)

The gaps we address:

Prior work is mostly **single-language** and often **speech**. We bring the continuous, geographic perspective to **text** across **multiple language families**, with geography **at the center** of the model.

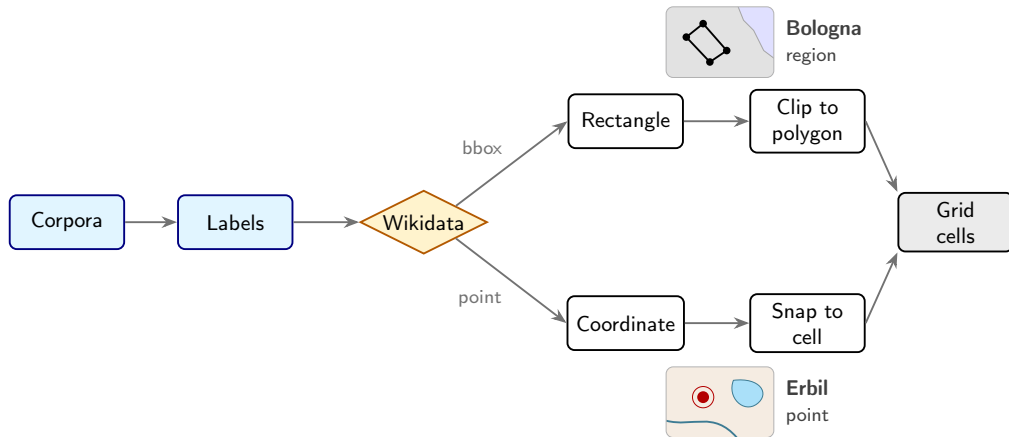
Methodology

Approach

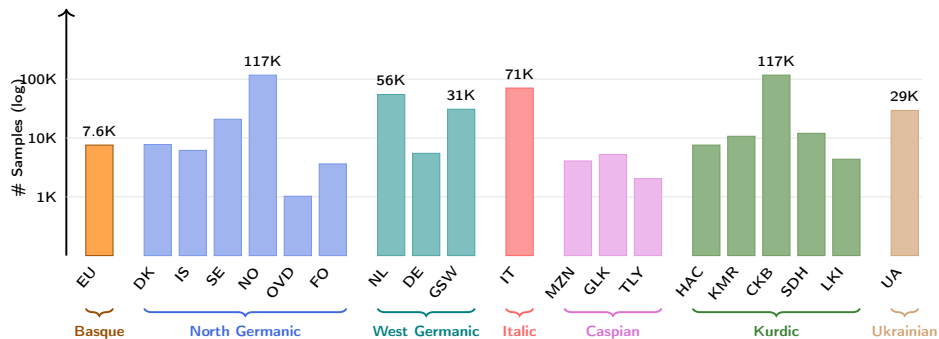


Data: Geographic Anchoring

Location labels are messy, and oftentimes lack geographic coordinates



Data: GeoVar

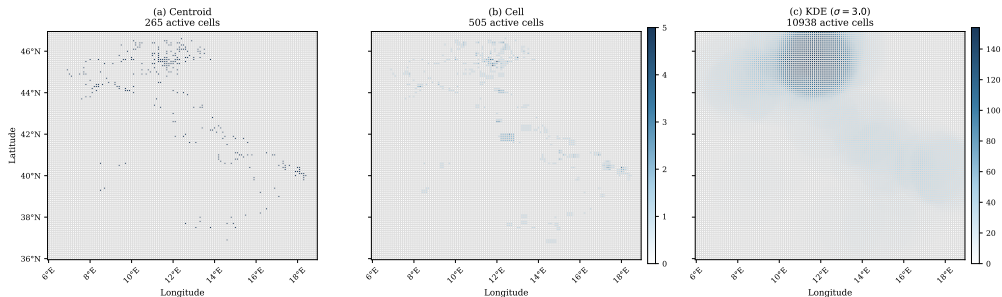


- 20 languages
- 929 geolects
- 517,985 samples
- 7 continua

Geoelect–Cell Mapping

Three strategies forming a spectrum of increasing spatial smoothness:

- **Centroid:** each geoelect maps to a single cell at the center of its region
- **Uniform:** embeddings are distributed uniformly across all cells in the geoelect's region (hard boundaries)
- **Gaussian:** an isotropic kernel spreads mass outward from the centroid with gradual decay beyond the region boundary



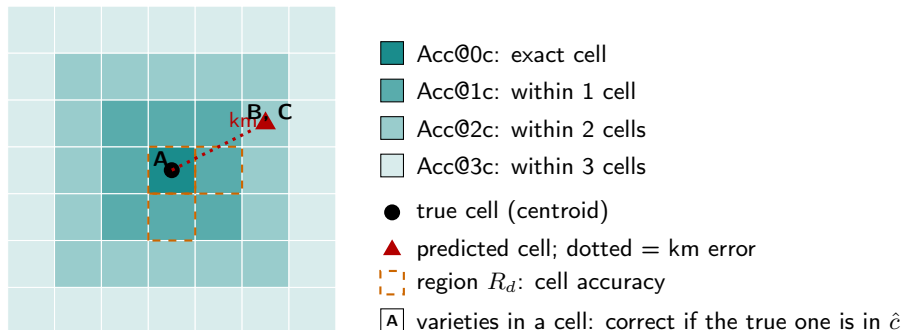
Classification Methods

- All methods use **CANINE-C**, a character-level encoder
- **Baselines**: no learning on the grid
 - **B1**: logistic regression on variety labels, no grid
 - **B2**: nearest centroid, i.e. the closest cell prototype
- **Learnable methods**, from simple to expressive
 - **M1**: logistic regression over cells instead of labels
 - **M2**: ProtoNet, learns a projection that pulls same-cell sentences together
 - **M3**: fine-tunes the encoder, then predicts by nearest prototype
- One model per dialect continuum

Methodology: Evaluation

- **Variety identification:** accuracy $\uparrow$, macro-F1 $\uparrow$, cell accuracy $\uparrow$
- **Spatial localization:** median error in km $\downarrow$, $\text{Acc}@N_c \uparrow$

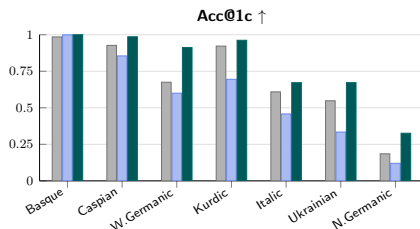
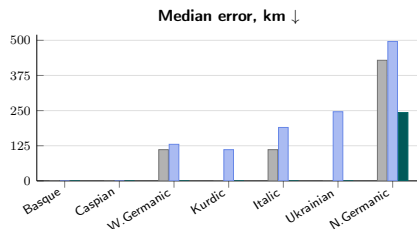
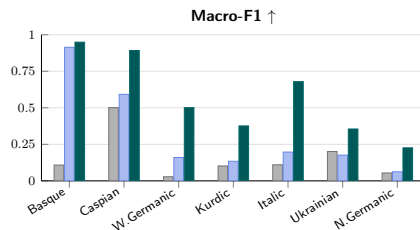
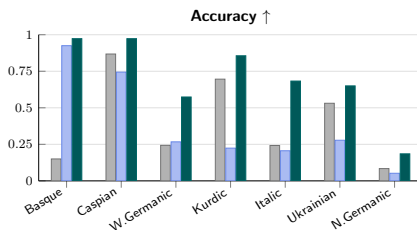
Coarser grids help the first and hurt the second $\Rightarrow$ we report both



Results

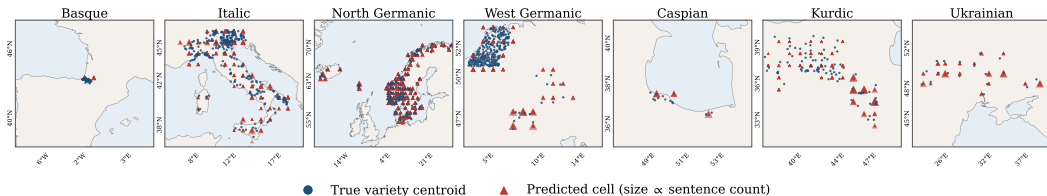
Can a Single Sentence Tell Us Where It Comes From?

■ B1: LogReg, no grid ■ B2: nearest centroid ■ M3: best



Ordered by geographic dispersion σ_{geo}

Why is North Germanic so much harder than Basque?



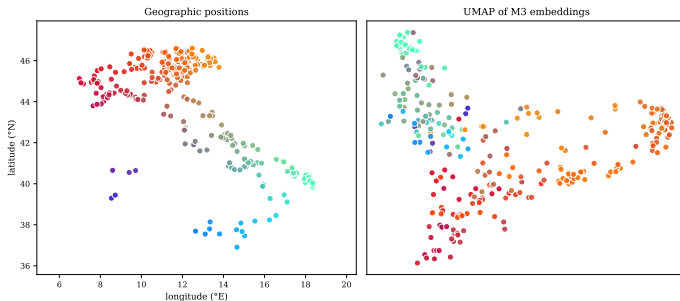
- Difficulty tracks **how far a continuum spreads**, not how many varieties it holds
- More varieties sharing a **cell** go with higher F1, as there are fewer cells to tell apart
- Per continuum:
 - **Basque, Caspian**: compact, above .97 accuracy
 - **North Germanic**: most dispersed, 243 km off
 - **Kurdic**: high accuracy, low F1, as Central Kurdish dominates the data

Which Modelling Choices Actually Matter?

- **Grid size:** coarser cells raise accuracy; every continuum lands on 1°
- **Mapping:**
 - Centroid: best on all seven continua
 - Uniform: same as centroid for most continua, since at 1° most regions fit in a single cell
 - Gaussian: inflates accuracy and F1 by pooling more varieties per cell, but exact-cell accuracy collapses
- **Merging continua?:** barely needed!
 - One model on Italic, West Germanic and North Germanic (766 varieties): accuracy barely drops
 - Only 59 of 19,569 errors cross a continuum, just 26 across the Alps
- **Encoder:** character-level CANINE-C beats subword encoders (LaBSE, MiniLM)

Does the Model Rediscover the Varieties Map?

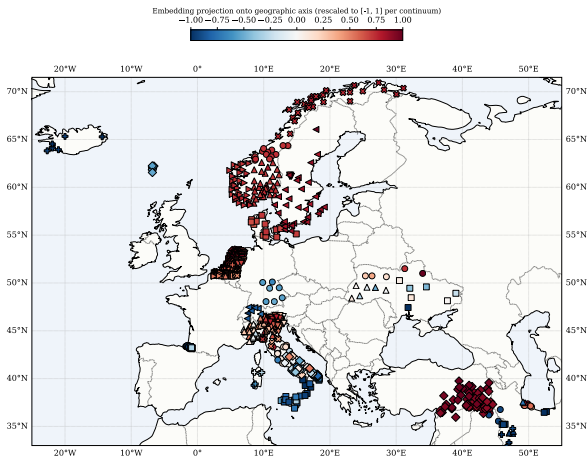
- If variation is gradual, neighbouring varieties should also be neighbours in the embedding
- They are: the geographic gradient is preserved
- Sardinia and the southern islands sit apart as separate clusters
- Both regimes coexist $\Rightarrow$ a continuum with clusters



Italic: geographic position (left) vs. embedding (right); same colour = same geoelect

Does Every Continuum Leave a Geographic Trace?

- If geography is encoded, a variety's location should be readable from its embedding
- It is the case in Italic, both Germanic groups and Kurdish
- Less so in Basque (too compact), Caspian and Ukrainian (too few varieties)
- Geography is encoded wherever the continuum spans real distance



Conclusion

Conclusion

Identify varieties by **where they are attested**, not by a fixed label list

- Circumvents the **labelling bottleneck** we started with: a variety needs a location, not a name
- GEOVAR: **929 geolects** in 20 languages and 7 continua, anchored to coordinates

Openly available at github.com/sinaahmadi/GeoVar



- Difficulty depends on **how far varieties spread**, not how many there are
- The model learns the map: geography and continuum boundaries emerge from the data
- **What's next:**
 - Geography as a training signal
 - Learned location encoders and equal-area grids (e.g. H3)
 - Coverage beyond Europe and the Middle East

Thank you! Questions?

References

- I. Caswell, T. Breiner, D. van Esch, and A. Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.579. URL <https://aclanthology.org/2020.coling-main.579/>.
- P. Foley, M. Wiesner, B. Odoom, L. P. García-Perera, K. Murray, and P. Koehn. Where are you from? Geolocating speech and applications to language identification. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16–21, 2024, pages 5114–5126. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.286. URL <https://doi.org/10.18653/v1/2024.naacl-long.286>.
- M. Hulden, M. Silfverberg, and J. Francom. Kernel density estimation for text-based geolocation. In B. Bonet and S. Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA*, pages 145–150. AAAI Press, 2015. doi: 10.1609/AAAI.V29I1.9149. URL <https://doi.org/10.1609/aaai.v29i1.9149>.
- Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In R. Barzilay and M. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1260–1272. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1116. URL <https://doi.org/10.18653/v1/P17-1116>.
- S. Oepen, N. Arefev, M. Aulamo, M. Bañón, M. Buljan, L. Burchell, L. Charpentier, P. Chen, M. Fedorova, O. de Gibert, et al. HPLT 3.0: Very large-scale multilingual resources for LLM and MT. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*, 2025.
- S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In J. Tsujii, J. Henderson, and M. Pasca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12–14, 2012, Jeju Island, Korea*, pages 1500–1510. ACL, 2012. URL <https://aclanthology.org/D12-1137/>.
- R. S. Shim and B. Plank. *Dialetto, ma Quanto Dialetto?* transcribing and evaluating dialects on a continuum. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 838–849. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.48. URL <https://doi.org/10.18653/v1/2025.findings-naacl.48>.
- B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL <https://aclanthology.org/P11-1096/>.