

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
{sina.ahmadi, rico.sennrich}@uzh.ch

Abstract

Language identification is a critical preprocessing step for curating multilingual corpora. While current approaches treat language varieties as clearly bounded categories, such discrete labels obscure the continuous nature of linguistic variation in practice. We propose reframing non-standard variety identification as a geographically-grounded task, representing varieties as *geolects*, that is, linguistic data anchored to geographic coordinates. Rather than classifying texts into predefined labels, we place them on a geodesic grid, where predictions are points on the earth’s surface and variety distinctions emerge from spatial structure. We introduce GEOVAR, a fine-grained dataset of 517,985 samples spanning 929 geolects across 20 languages and seven dialect continua in Europe and the Middle East, and develop a grid-based framework with three geolect-to-cell mapping strategies of increasing spatial smoothness: centroid, uniform, and Gaussian. We compare various geolocation methods, with the best model reaching 0.97 accuracy on compact continua such as Basque and Caspian. Accuracy decreases with the geographic dispersion of varieties, with North Germanic being the hardest case. Our model handles non-standard variety text more reliably than current LID tools, while additionally operating at sub-language granularity.



1 Introduction

Language identification (LID), determining the language of a given text, has a long history in natural language processing (NLP). Although interest in standalone LID declined as end-to-end multilingual models learned to handle language implicitly, the task has recently resurfaced as a critical component of data filtering pipelines for training large language models (LLMs) on clean, multilingual corpora (Oepen et al., 2025).

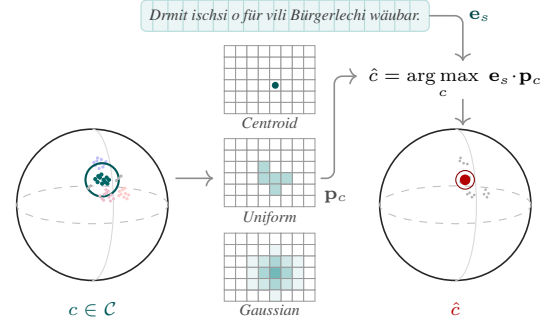


Figure 1: Our methodology with an input example in Swiss German. The encoded text is matched against cell prototypes built from one of our presented strategies. The predicted cell (right) lands back in the West Germanic cluster, highlighted on the left globe.

Conventional LID operates over a fixed inventory of discrete labels, treating “language” as a categorical attribute that can be assigned unambiguously. This framing presupposes that every variety has a standardized name and clear boundaries, assumptions that break down once we move beyond standard languages, for instance the Alemannic varieties of Bern and Basel versus those of Swabia and Upper Bavaria. Dialects, regional vernaculars, and other non-standardized varieties often lack agreed-upon nomenclature entirely, leaving no label to assign when neighboring villages differ linguistically. For such varieties, the label set itself becomes the bottleneck.

We argue that a more natural formulation is geographic rather than categorical. Instead of asking *what language is this?*, we ask *where on earth could this text have been produced?* Under this framing, a prediction is a point (or distribution) on the earth’s surface, and the notion of a “variety” emerges from the spatial structure of the model’s outputs rather than being imposed *a priori*. This perspective aligns with the empirical reality in which linguistic variation is gradient and spatially structured rather than neatly partitioned.

In this paper, we operationalize this idea by casting LID as text geolocation on a geodesic grid. We discretize Earth’s surface into cells and train classifiers to place text samples into cells using multilingual sentence embeddings. Because cells carry coordinates, the same model supports both variety classification and geographic error analysis, revealing trade-offs invisible to purely categorical metrics. Our contributions are as follows:

- GEOVAR, a fine-grained dataset of 517,985 samples spanning 929 geolects across 20 languages and seven dialect continua in Europe and the Middle East
- A geodesic grid framework with three geolect-to-cell mapping strategies: centroid, uniform, and Gaussian
- A systematic comparison of baselines along with logistic regression, prototype networks, and fine-tuned encoders evaluated on both variety accuracy and geographic error

Throughout the paper, we use *variety* as a neutral term for named linguistic systems such as dialect, regional vernacular, or sociolect (Platzgumer et al., 2026). We use *geolect* to denote a variety paired with the geographic location(s) at which it is attested in our data. The location is part of the geolect’s identity, not metadata about it.

2 Related Work

LID for Non-Standard Varieties. LID has been broadly studied with approaches covering a wide range of languages and many considering it a largely solved problem for well-resourced standard languages (Caswell et al., 2020). However, the task remains far from trivial for low-resource and non-standardized varieties, particularly dialects. The problem has recently regained momentum as accurate LID is essential for creating clean, language-specific partitions of large corpora used to train LLMs (Fedorova et al., 2026), prompting continued methodological innovation (Meister et al., 2026; Foroutan et al., 2026).

A growing body of work now targets LID for non-standard varieties (Alam et al., 2024; Chifu et al., 2024), with efforts spanning Swiss German (Plüss et al., 2023), German (Blaschke et al., 2025), South Slavic (Ljubešić et al., 2018), Kurdish (Ahmadi et al., 2024), and Scandinavian (Fedorova et al., 2025), *inter alia*.

Much of this work, however, targets spoken language (O’Shaughnessy, 2025). Closest to our setting, Shim and Plank (2025) evaluate speech-to-text performance across geographically contiguous Italian dialect areas. They show that linguistic similarity correlates with system performance and that providing geographic information can improve interpolation to unseen varieties.

Text-based Geolocation. Text-based geolocation aims to predict the geographic origin of a document, typically from linguistic features but sometimes also from user-profile metadata (Rahimi et al., 2015; Utomo et al., 2018; Rosenfeld and Hinrichs, 2023). The availability of geotagged material from platforms such as Wikipedia, X (Twitter), and Flickr drove substantial research activity in the 2010s, particularly at the intersection of NLP and sociolinguistics (Han et al., 2014; Graham et al., 2014; Wing, 2015). A persistent challenge in the field is the scarcity of geotagged data, motivating work on leveraging non-geotagged sources (Han et al., 2012) and annotating geolocation datasets from existing corpora (DeLozier et al., 2016).

A variety of modeling approaches have been proposed for text geolocation, including bag-of-words classification on geodesic grids (Wing and Baldridge, 2011; Roller et al., 2012), topic models (Eisenstein et al., 2010), logistic regression (Wing and Baldridge, 2014), neural network classifiers (Miura et al., 2017; Lourentzou et al., 2017; Rahimi et al., 2017), knowledge-based methods leveraging entities and relations (Miyazaki et al., 2018), and sequence-to-sequence formulations (Zhang et al., 2023). Kernel density estimation on geodesic grids, as proposed by Hulden et al. (2015), addresses the data sparsity inherent in fine-grained grids by smoothing feature distributions with Gaussian kernels; we build on this framework in our grid representation. More recently, geolocation has been extended to the speech modality; Foley et al. (2024) train models on 4,000 hours of radio data using self-supervised pretrained representations, achieving localization accuracy of approximately 650 km.

Existing Gaps. Our work addresses several gaps in prior research. First, existing geolocation studies typically operate within a single language, whereas we adopt a multilingual perspective. Second, recent efforts to geographically model dialect variation have focused predominantly on

speech (Shim and Plank, 2025; Foley et al., 2024) but not much on the text modality. Third, instead of treating geography as an auxiliary signal or evaluation metric, we place it at the center of the model by representing varieties as geolects.

3 GEOVAR

3.1 Data Sources

One of the principal challenges in working with non-standard language varieties is data scarcity. With the exception of a few language groups for which linguistic variation has been studied extensively, as for Scandinavian languages (Johannessen, 2011), data for varieties is rare and mostly documented within projects focusing on syntax or phonetics rather than as standalone corpora. Recent efforts in NLP have begun to address this gap, with dedicated collection campaigns for Italian (Giordano and Rinaldi, 2025), Basque (Ben-goetxea et al., 2026), and more languages (Sun et al., 2025). To bring these resources together, we compile GEOVAR from a diverse set of existing dialect resources spanning 20 languages across seven dialect continua, shown in Figure 2, containing 517,985 samples of 929 geolects with over 5.3M tokens. Our data selection criteria prioritize varieties that form geographically contiguous dialect areas, allowing spatial structures that capture linguistic gradation.

3.2 Geographical Label Enrichment

Although the collected datasets contain labels and location names, they typically lack the latitude and longitude coordinates required for geographical data representation. Moreover, such labels as recorded in variety corpora are often difficult to resolve unambiguously: they may reflect particular writing (e.g. St_Anna for Sankt Anna in the Östergötland archipelago, Sweden), overly broad descriptions (e.g. *East Icelandic*), or ambiguous references (e.g. *Mellame d’Arsiè* for a hamlet not easily located via standard gazetteers). To handle this systematically, we enrich each location label or variety name with structured geographic data from Wikidata using a semi-automatic approach with manual verification, resolving each to coordinates or a bounding region.

Appendix A presents complete details on the construction of GEOVAR along with the full inventory, grouped by dialect continua.

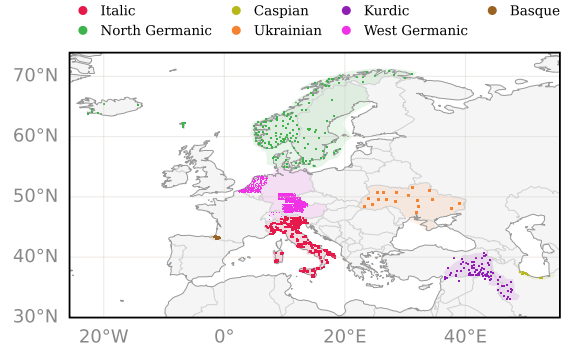


Figure 2: Geographic distribution of variety regions in GEOVAR, colored by continuum. Each point on the map marks a sampled location.

4 Geospatial Framework

We discretize the earth’s surface into a geodesic grid of square cells $\mathcal{C}$. Given a grid resolution parameter g (in degrees), the grid consists of $\frac{360}{g} \times \frac{180}{g}$ cells. Each cell $c \in \mathcal{C}$ is indexed by integer coordinates (x, y) , where a point at longitude λ and latitude ϕ maps to:

$$x = \left\lfloor \frac{\lambda + 180}{g} \right\rfloor, \quad y = \left\lfloor \frac{\phi + 90}{g} \right\rfloor \quad (1)$$

The midpoint of cell (x, y) in world coordinates is recovered as:

$$\lambda_{\text{mid}} = x \cdot g - 180 + \frac{g}{2}, \quad \phi_{\text{mid}} = y \cdot g - 90 + \frac{g}{2} \quad (2)$$

Let $S_d = \{s_1, \dots, s_{n_d}\}$ denote the set of sentences for geolect d in dialect continuum, with corresponding embeddings $\{\mathbf{e}_{s_1}, \dots, \mathbf{e}_{s_{n_d}}\}$. Each geolect d is associated with a geographic region R_d , constructed from Wikidata bounding coordinates clipped to country polygons as described in A.1, and represented here as the set of grid cells covering that region. Each cell c is associated with a *prototype embedding* $\mathbf{p}_c \in \mathbb{R}^D$, computed using three strategies for mapping geolect embeddings onto the grid, forming a spectrum from the most discrete to the smoothest spatial representation.

Centroid Mode. Each geolect maps to a single cell c_d corresponding to the centroid of its region:

$$c_d = \left(\left\lfloor \frac{\bar{\lambda}_d + 180}{g} \right\rfloor, \left\lfloor \frac{\bar{\phi}_d + 90}{g} \right\rfloor \right) \quad (3)$$

where $\bar{\lambda}_d$ and $\bar{\phi}_d$ are the unweighted arithmetic mean of the midpoint longitudes and latitudes of the cells in R_d . That is, the centroid cell is determined by averaging the midpoint coordinates of

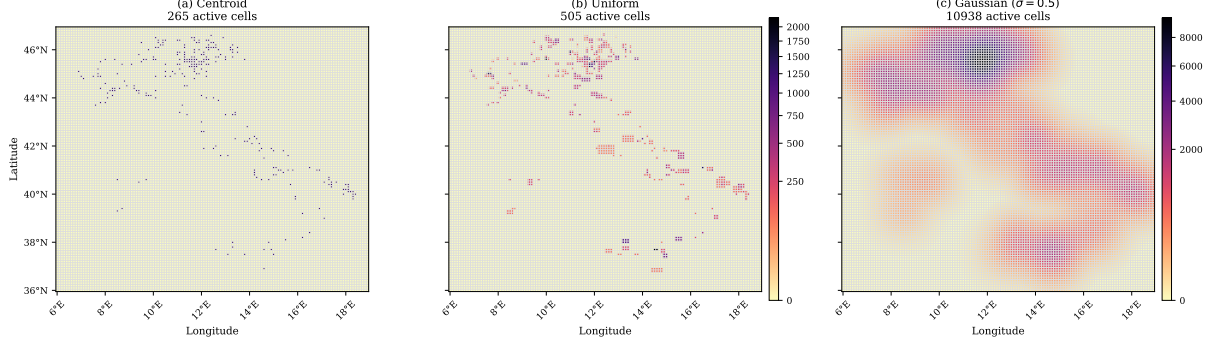


Figure 3: Comparison of the three strategies for mapping the 71,127 samples’ embeddings of 313 Italic geolects to grid cells (grid size 0.1°). Color intensity reflects the total sentence mass assigned to each cell, summed over all contributing geolects. All three panels use a power-law normalization ($\gamma = 0.4$) to make low-mass cells visible; panel (c) additionally uses bilinear interpolation to render the smooth Gaussian falloff at print size.

all cells in the geolect’s region R_d and discretizing back to the grid. Because this average is taken over cells rather than data, geolects with spatially larger regions will have their centroid pulled toward areas of greater cell density rather than greater sentence density. The cell prototype is then the mean of all sentence embeddings from geolects whose centroid falls in c :

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{d \in \mathcal{D}_c} \sum_{i=1}^{n_d} \mathbf{e}_{s_i} \quad (4)$$

where $\mathcal{D}_c = \{d \mid c_d = c\}$ is the set of geolects whose centroid cell is c , and $N_c = \sum_{d \in \mathcal{D}_c} n_d$ is the total number of sentences from those geolects. Note that the average is taken over individual sentences rather than over geolects, so a high-resource geolect sharing a cell with a low-resource one contributes proportionally more to the prototype.

Uniform Mode. Rather than mapping a geolect to a single cell, this mode maps each geolect uniformly to all cells in its region R_d . Every cell $c \in R_d$ receives the full set of sentence embeddings from geolect d with equal weight, following the same prototype in Equation 4. The prototype formula is unchanged, but the cell-to-geolect mapping $\mathcal{D}_c = \{d \mid c \in R_d\}$ is now many-to-many. As such, a single geolect contributes to every cell in R_d , and a cell lying in a region overlap pools sentences from all geolects whose regions cover it. This extends centroid mode by preserving the geographic extent of a geolect region, yielding a uniform distribution over the region with hard boundaries.

Gaussian Mode. Each geolect’s embedding mass is distributed across nearby cells via a two-

dimensional isotropic Gaussian kernel centered on the region’s centroid. The kernel weight assigned to cell c with midpoint (ϕ_c, λ_c) for geolect d is:

$$w_{d,c} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(\lambda_c - \bar{\lambda}_d)^2 + (\phi_c - \bar{\phi}_d)^2}{2\sigma^2}\right) \quad (5)$$

where σ is specified in degrees and controls the spatial spread of the kernel. The prototype for cell c is a weighted mean of the geolect mean-embeddings:

$$\mathbf{p}_c = \frac{\sum_{d \in \mathcal{D}_c} w_{d,c} \cdot \bar{\mathbf{e}}_d}{\sum_{d \in \mathcal{D}_c} w_{d,c}} \quad (6)$$

where $\bar{\mathbf{e}}_d = \frac{1}{n_d} \sum_{i=1}^{n_d} \mathbf{e}_{s_i}$ is the mean embedding of geolect d , and $\mathcal{D}_c = \{d \mid w_{d,c} > \epsilon\}$ for a minimum density threshold ϵ . Unlike the uniform mode, Gaussian mode produces a smooth, graded distribution where cells near the center of the geolect region receive the most mass, while influence gradually decays with distance and extends beyond the region boundary. Note that, in contrast to centroid and uniform modes, the outer average here is taken over geolect mean embeddings rather than over individual sentences.

In all three modes, the final prototype is L_2 -normalized: $\mathbf{p}_c \leftarrow \mathbf{p}_c / \|\mathbf{p}_c\|$. Together these modes, as illustrated in Figure 3, span a spectrum from sharp points (centroid) through bounded regions (uniform) to smooth Gaussian falloff, reflecting the empirical reality that dialect boundaries are transitional rather than sharp.

5 Models and Learning

Given the cell prototypes $\mathbf{p}_c$ constructed in the geospatial framework, we implement several

methods for predicting the grid cell $\hat{c}$ of a given sentence s with embedding $\mathbf{e}_s$. We first establish two baselines that involve no learning on the grid, then present three methods with learnable parameters trained to optimize cell prediction.

5.1 Baselines

B1: Logistic Regression (no grid). As a non-geographic baseline, we train a multinomial logistic regression classifier directly on the sentence embeddings $\mathbf{e}_s$ to predict variety labels. This baseline does not use the geodesic grid and thus has no associated geographic error.

B2: Nearest Centroid. Following the nearest-centroid classifier (also known as Rocchio classification; Schütze et al., 2008, p. 292), we predict the cell whose prototype is most similar under cosine similarity:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} \mathbf{e}_s \cdot \mathbf{p}_c \quad (7)$$

Since both $\mathbf{e}_s$ and $\mathbf{p}_c$ are L_2 -normalized, the dot product equals the cosine similarity.

5.2 Learnable Methods

For all learnable methods, each training sentence is assigned a single target cell c_d corresponding to the centroid of its geolect’s region, regardless of the mapping mode used for building prototypes. This provides a consistent single-label supervision signal across all settings.¹

M1: Logistic Regression on Cells. We train a multinomial logistic regression classifier on the sentence embeddings to predict grid cells:

$$P(\hat{c} = c \mid s) = \frac{\exp(\mathbf{w}_c \cdot \mathbf{e}_s + b_c)}{\sum_{c' \in \mathcal{C}'} \exp(\mathbf{w}_{c'} \cdot \mathbf{e}_s + b_{c'})} \quad (8)$$

where $\mathcal{C}' \subseteq \mathcal{C}$ is the set of centroid cells of the training geolects, and $\mathbf{w}_c \in \mathbb{R}^D$, $b_c \in \mathbb{R}$ are learnable parameters for each cell. At test time, the predicted cell is $\hat{c} = \arg \max_c P(\hat{c} = c \mid s)$. Unlike B1, this method enables geographic error evaluation because classes correspond to spatial locations on the grid.

M2: Prototype Network (ProtoNet). Inspired by Prototypical Networks (Snell et al., 2017), we learn a linear projection that maps sentence embeddings into a space where same-cell sentences

cluster more tightly, rather than using fixed mean embeddings as prototypes. Given a projection matrix $W \in \mathbb{R}^{D \times D}$, the projected embedding is $\mathbf{z}_s = W\mathbf{e}_s$. Classification is performed by cosine similarity to projected prototypes:

$$\hat{c} = \arg \max_{c \in \mathcal{C}'} \frac{\mathbf{z}_s \cdot \mathbf{q}_c}{\|\mathbf{z}_s\| \|\mathbf{q}_c\|} \quad (9)$$

where $\mathbf{q}_c = \frac{1}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} W\mathbf{e}_s$ is the projected prototype for cell c , computed from the training sentences $\mathcal{S}_c$ assigned to that cell. The model is trained by minimizing the cross-entropy loss over the similarity scores:

$$\mathcal{L} = - \sum_s \log \frac{\exp(\tau \cdot \cos(\mathbf{z}_s, \mathbf{q}_{c_s}))}{\sum_{c'} \exp(\tau \cdot \cos(\mathbf{z}_s, \mathbf{q}_{c'}))} \quad (10)$$

where c_s is the target cell for sentence s and τ is a learnable temperature parameter that controls the sharpness of the distribution. The per-cell averages are computed once from the training embeddings and held fixed; only their projections $\mathbf{q}_c$ through W are refreshed each epoch.

M3: Fine-tuned Encoder. In the most expressive setting, we fine-tune the sentence encoder end-to-end. Training uses a standard classification head where a linear layer maps the pooled encoder output to cell logits, and the model is trained with cross-entropy loss over centroid cell labels:

$$\mathcal{L}(\theta, W_{\text{cls}}) = - \sum_s \log \frac{\exp(\mathbf{w}_{c_s} \cdot \mathbf{h}_s + b_{c_s})}{\sum_{c'} \exp(\mathbf{w}_{c'} \cdot \mathbf{h}_s + b_{c'})} \quad (11)$$

where $\mathbf{h}_s$ is the mean-pooled representation of sentence s over non-padding tokens from the encoder with parameters θ , and $W_{\text{cls}} = \{\mathbf{w}_c, b_c\}$ is the classification head. This reshapes the embedding space so that geolect-distinguishing features, such as morphosyntactic markers, are amplified relative to content features shared across geolects.

At inference time, however, the classification head is discarded. Instead, we re-embed all training sentences through the fine-tuned encoder to obtain new representations $\mathbf{e}'_s$, and construct cell prototypes from these fine-tuned embeddings using whichever mapping mode (centroid, uniform, or Gaussian) is selected. A test sentence is then classified by nearest-prototype similar to Equation 7:

$$\hat{c} = \arg \max_{c \in \mathcal{C}'} \mathbf{e}'_s \cdot \mathbf{p}'_c \quad (12)$$

¹Our preliminary experiments with mode-consistent targets yielded no gains and complicated training.

Method	Prediction	Recovery
B1: LogReg $\rightarrow$ varieties	—	—
B2: Nearest Centroid	✓	✓
M1: LogReg $\rightarrow$ cells	—	✓
M2: ProtoNet	—	✓
M3: Fine-tuned Encoder	✓	✓

Table 1: Effect of the mapping mode at each stage. B1 is mode-independent as it does not use the grid.

where $\mathbf{p}'_c$ is the L_2 -normalized prototype built from fine-tuned embeddings. This two-stage design, meaning train with a classification head and infer with prototypes, decouples the training objective from the spatial representation used at test time. As a result, the mapping mode affects predictions: centroid, uniform, and Gaussian produce different prototype sets from the same fine-tuned embeddings, yielding different predicted cells and therefore different geographic errors. In contrast, for logistic regression on cells (M1) and the prototype network (M2), the mapping mode does not influence the cell predictions themselves as they are trained and evaluated against the centroid cell label; the mode only affects the cell-to-dialect mapping $\mathcal{D}_{\hat{c}}$ used to recover the variety at evaluation. The fine-tuned encoder is thus the only learnable method for which the mapping mode is an experimental variable.

Table 1 summarizes where each method is sensitive to the mode.

5.3 Evaluation

We evaluate predictions along two complementary axes: *variety identification* and *spatial localization*. Because absolute distances depend on grid resolution and latitude, the grid-relative metrics we define below provide a resolution-independent complement to the kilometer-based ones. Coarser grids inflate variety accuracy by collapsing multiple varieties into shared cells but increase geographic error; finer grids do the opposite. Reporting both axes lets us position each model along this trade-off.

Variety Identification. For each test sentence from geolect d , the model predicts a grid cell $\hat{c}$, which we map back to a geolect via the mode-specific set $\mathcal{D}_{\hat{c}}$ defined in Section 4. To assign a single predicted variety, we use the following rule: if $d \in \mathcal{D}_{\hat{c}}$, the prediction is d ; otherwise, we assign the geolect in $\mathcal{D}_{\hat{c}}$ whose centroid cell is closest in Chebyshev distance (Deza and Deza, 2009, p. 324)

to $\hat{c}$, breaking ties alphabetically. If $\mathcal{D}_{\hat{c}}$ is empty, the prediction is counted as incorrect. This assignment is then scored on:

- **Accuracy:** fraction of sentences whose predicted geolect equals d , i.e. $d \in \mathcal{D}_{\hat{c}}$.
- **Macro-F1:** unweighted mean of per-variety F1 scores, ensuring low-resource varieties count equally.
- **Cell accuracy:** fraction of predictions for which $\hat{c}$ falls within the true geolect’s region R_d , measuring spatial accuracy independently of whether the correct geolect can be recovered from $\hat{c}$.

Spatial Localization. The true location of d is its region centroid $(\bar{\phi}_d, \bar{\lambda}_d)$, the same cell used as the training target c_d .

- **Mean / median error:** haversine distance (km) between $(\bar{\phi}_d, \bar{\lambda}_d)$ and the midpoint of $\hat{c}$, computed as the mean or median over all test sentences.
- **Acc@Nc:** fraction of predictions for which the Chebyshev distance between $\hat{c}$ and the true centroid cell is at most N . Acc@0c is exact-cell accuracy; Acc@1c, Acc@2c, and Acc@3c correspond to within one, two, and three cells.

The B1 baseline predicts varieties directly, without the grid. For spatial evaluation, its predicted variety is mapped to that geolect’s centroid cell, which serves as $\hat{c}$ in all metrics above.

6 Experiments

6.1 Experimental Setup

Rather than training a single model over all varieties or a separate model per language, we train models per dialect continuum. This choice reflects three considerations. First, dialect variation is linguistically continuum-specific: features discriminating Basque varieties differ from those of Kurdish, and a single shared model would need to learn both inventories simultaneously. Second, continua are geographically contiguous, which makes the Acc@Nc metric interpretable; a cell offset N corresponds to a comparable physical distance within a continuum, but to wildly different distances when continua span multiple regions. Third, per-continuum label spaces remain tractable for fine-grained grid resolutions.

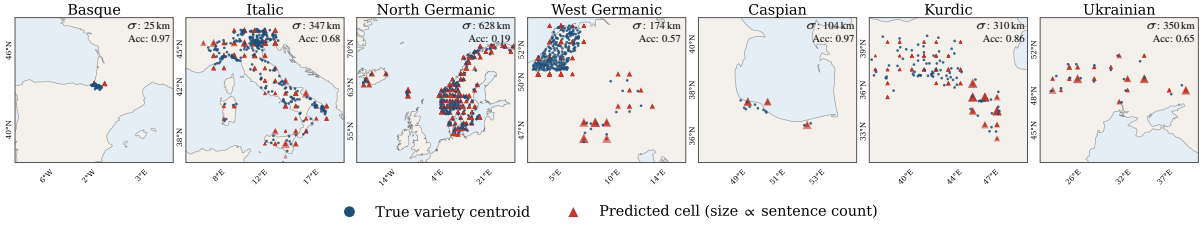


Figure 4: Test set predictions of the best fine-tuned encoder (centroid mode) per continuum. Blue circles mark true variety centroids; red triangles mark predicted cells, sized by sentence count. Each panel is annotated with the standard geographical deviation σ_{geo} and $\text{Acc}@0c$.

		Acc	F1	Median (km)	Acc@1c
Basque	B1	.150	.108	0	.985
	B2	.925	.914	0	.999
	Best	.973	.949	0	1.000
Italic	B1	.242	.110	111	.609
	B2	.206	.197	190	.458
	Best	.683	.679	0	.890
North Germanic	B1	.084	.054	429	.185
	B2	.052	.062	496	.120
	Best	.185	.226	243	.326
West Germanic	B1	.243	.028	111	.675
	B2	.267	.160	130	.600
	Best	.574	.501	0	.912
Caspian	B1	.868	.501	0	.927
	B2	.744	.592	0	.855
	Best	.973	.892	0	.986
Kurdic	B1	.696	.102	0	.922
	B2	.224	.134	111	.695
	Best	.856	.376	0	.962
Ukrainian	B1	.531	.201	0	.548
	B2	.278	.176	246	.334
	Best	.650	.355	0	.672

Table 2: Test set results per continuum for the two baselines (B1 & B2) and the best method (M3), selected by $\text{Acc}@0c$ on the validation split. M3 outperforms both baselines on every metric across all continua.

All methods use CANINE-C (Clark et al., 2022) as the base encoder. We split each variety into 80% train / 10% validation / 10% test with a fixed seed, deduplicating within each variety before splitting. Each learnable method is trained once per continuum with the different hyperparameters; baselines (B1, B2) require no training. Further implementation details, including sentence encoding, training schedules, hyperparameter optimization and compute budgets, are described in Appendix B.

6.2 Results

The fine-tuned encoder (M3) outperforms both baselines on every metric across all seven continua, with accuracy gains ranging from 5 to 50 points (Table 2). Performance varies substantially across continua: Basque and Caspian both exceed 0.97 accuracy, while North Germanic remains the

hardest case at 0.185 accuracy and 243 km median error. The learnable baselines M1 and M2 fall between B2 and M3. Figure 4 illustrates these per-continuum patterns geographically where blue circles mark true variety centroids and red triangles mark predicted locations, exposing both the near-perfect alignment on Basque or Caspian and the large offsets on North Germanic. Full per-mode and per-method numbers are in Appendix C.

Across all seven continua, the centroid mode dominates, suggesting that sharp variety representations outperform smoothed ones for LID-style discrimination. We hypothesize this is because Gaussian smoothing helps for sparse data (cf. Hulden et al. (2015)) but hurts when varieties are densely sampled and need to be distinguished from neighbors.

Rank Correlations We examined Spearman rank correlations between six dataset properties and three model outcomes across the seven continua (full correlation matrix in Appendix C.3). Two dataset properties predict model performance. The spatial standard deviation of variety centroids from the continuum’s geographic center, σ_{geo} , correlates strongly and negatively with macro-F1 ($\rho = -0.89$, $p = .007$). Continua whose varieties span a larger geographic area are harder to classify. The cell collision rate, i.e. the mean number of varieties sharing a centroid cell at 1° resolution, correlates positively with F1. When varieties share cells, the task effectively coarsens, making per-cell predictions easier. The two properties capture different difficulty aspects, namely how much geographic ground the model must cover and how many varieties it must distinguish within each cell.

6.3 Comparison with Other LID Models

To contextualize our results against existing LID, we evaluate three widely-used tools on the same test sentences used to evaluate M3: fastText

		Strict	Lenient	Variety
Basque	fastText	.702	.702	—
	OpenLID-v3	.908	.908	—
	GlottLID	.940	.940	—
	M3 (ours)	.954	.954	.973
Italic	fastText	.554	.624	—
	OpenLID-v3	.085	.700	—
	GlottLID	.128	.812	—
	M3 (ours)	.982	.982	.683
N. Germ.	fastText	.620	.620	—
	OpenLID-v3	.925	.925	—
	GlottLID	.878	.878	—
	M3 (ours)	.978	.979	.185
W. Germ.	fastText	.565	.567	—
	OpenLID-v3	.468	.612	—
	GlottLID	.716	.915	—
	M3 (ours)	.996	.997	.574
Caspian	fastText	.112	.112	—
	OpenLID-v3	.000	.000	—
	GlottLID	.662	.662	—
	M3 (ours)	.986	.986	.973
Kurdic	fastText	.782	.782	—
	OpenLID-v3	.800	.800	—
	GlottLID	.778	.926	—
	M3 (ours)	.963	.984	.856
Ukrainian	fastText	.810	.810	—
	OpenLID-v3	.901	.901	—
	GlottLID	.826	.826	—
	M3 (ours)	1.000	1.000	.650
All	fastText	.649	.659	—
	OpenLID-v3	.669	.780	—
	GlottLID	.709	.882	—
	M3 (ours)	.978	.985	.572

Table 3: Language-level accuracy of existing LID tools and M3 on the GEOVAR test set. *Strict* requires exact ISO 639 match; *Lenient* additionally credits predictions of recognized sub-varieties. *Variety* is M3’s variety-level accuracy on the same sentences.

lid.176 (Joulin et al., 2017), OpenLID-v3 (Fedorova et al., 2026), and GlottLID (Kargaran et al., 2023). Since these tools predict ISO 639 language codes rather than geolects, we compare at the language level, mapping each variety in GEOVAR to its parent ISO 639 macro-language code and rolling M3’s predicted variety up to the same level. We report two scoring setups: *strict*, requiring the predicted code to match the gold ISO code exactly, and *lenient*, which additionally credits predictions of recognized sub-varieties.

Table 3 shows two dominant failure modes in existing tools. First, of GEOVAR’s 20 languages, Talysh, Ovdalian, and Laki Kurdish are absent from all three tools, and five more (Gilaki, Mazandarani, Hawrami, Southern Kurdish, Swiss German) are recognized only by GlottLID. Second, when strict accuracy drops sharply on Italic and West Germanic, GlottLID is in fact identifying related but distinct varieties rather than the macro-

language, with lenient scoring recovering to 0.812 and 0.915 respectively. OpenLID-v3 yields essentially the same accuracy as OpenLID-v1 (Burchell et al., 2023), confirming that the gap to M3 reflects the categorical LID paradigm itself rather than a specific release. M3 reaches 0.978 strict and 0.985 lenient overall, outperforming existing tools while additionally providing variety-level discrimination at 0.572 accuracy, a granularity at which no existing tool operates.

7 Analysis

7.1 Clusters vs. Varieties

A natural question regarding the geospatial framework is whether the representation that M3 learns reflects a spectrum or rather discrete clusters of varieties. To examine this, we project the per-gelect mean embeddings of the fine-tuned encoder (centroid mode) to two dimensions using UMAP (McInnes et al., 2018) and ask whether geographically adjacent varieties remain adjacent in embedding space. Figure 5 illustrates the case of Italic, where pairwise embedding distances correlate with pairwise great-circle distances at Spearman $\rho = 0.65$ ($p < 10^{-3}$).

First, the color gradient is largely preserved across the projection, with northern, central, and southern varieties occupying contiguous regions of the embedding rather than mixing arbitrarily. This indicates that M3 places geographically adjacent varieties in adjacent regions of embedding space, consistent with the gradient framing that motivates our geolect formulation. Second, varieties associated with geographic discontinuities form visually separated clusters: the Sardinian and southern island varieties (in green and violet) sit apart from the mainland mass despite being projected jointly with it. The embedding geometry is therefore neither a pure continuum nor a discrete partition, but exhibits both regimes within a single continuum.

Geographic Axis via Ridge Regression. The aforementioned UMAP analysis is qualitative and specific to Italic. A complementary analysis using ridge regression to project embeddings onto the dominant geographic axis is reported for all seven continua in Appendix D; it confirms that the embedding-geography correspondence holds broadly for continua with sufficient geographic extent.

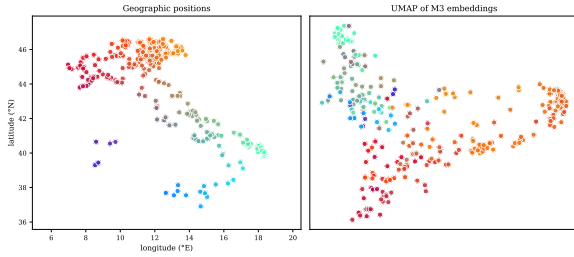


Figure 5: Geographic positions (left) and UMAP projection of M3’s fine-tuned per-geolect mean embeddings (right) for the Italic continuum. Each point is one geolect; color encodes geographic position via a joint latitude-longitude RGB scheme (latitude drives the red-blue axis, longitude drives the green channel). Identical colors in the two panels mark the same geolect.

7.2 Revisiting Continua

To test whether the continuum decomposition we applied during training reflects a modeling necessity, we train one fine-tuned encoder on the union of Italic, West Germanic, and North Germanic. These three form a geographically contiguous block covering 766 varieties, unlike the other four continua which are widely separated. All hyperparameters match the selected per-continuum configuration. Evaluated on the merged test split, the joint model loses 2.2 Acc@0c points on Italic, 1.8 on West Germanic, and 0.6 on North Germanic, with median geographic error unchanged.

Beyond this accuracy drop, which is small despite the joint model facing roughly three times as many varieties, the structure of the errors is interesting. Of the 19,569 misclassified test sentences, only 59 (0.3%) predict a variety from a different continuum than the true one, while the other 99.7% stay within continuum. Even the geographically adjacent Italic-West Germanic pair, where the Alpine boundary places Swiss German and Northern Italian varieties within a few cells of each other, produces only 26 cross-boundary confusions in either direction.

The within-continuum concentration of errors reflects structure the model discovered from the data. That means that the varieties within each continuum form a linguistically coherent region of embedding space distinct from the other two.

8 Conclusion and Discussion

This paper revisits LID for non-standardized varieties and dialect continua by treating them as geolects. Under this framing, the task is no longer tied to a fixed inventory of labels, which

are not systematically available for all varieties, but is instead cast as geolocation. To support this formulation, we construct GEOVAR by gathering samples of 929 geolects in 20 languages across seven continua, and develop a grid-based framework with three mapping strategies (centroid, uniform, Gaussian) combined with classification and prototype-based methods. The fine-tuned encoder outperforms all other methods and surpasses widely-used LID tools at the language level while additionally providing variety-level discrimination. Performance varies systematically across continua, with Basque reaching 0.973 accuracy (0 km median error) while North Germanic does not exceed 0.185 (243 km median error).

Our analyses show that performance is shaped more by how varieties are distributed than by how many there are, with geographic dispersion predicting difficulty. The embedding geometry, in turn, exhibits both continuous and clustered regimes within a single continuum. Existing LID tools, while broadly capable, struggle with non-standard varieties they were not explicitly trained on, and several of GEOVAR’s languages are absent from their label inventories entirely. With finer-grained resources, the geographic framing could compete with these tools while covering varieties that current label inventories do not name.

Future Work. Beyond LID, we see the geospatial representation itself as the most promising direction to build on. Our finding that continuum structure emerges from the embeddings (Sections 7.1 and 7.2) rather than being imposed raises the question of whether geolect representations can serve as a general-purpose signal for representation learning (Hovy and Purschke, 2018). One avenue is to inject geographic grounding into language model pretraining, adaptation, or post-training, and to test whether it improves handling of non-standard varieties downstream. A complementary direction is to replace our fixed geolect-to-cell mapping with a learned location encoder (Wu et al., 2024; Betti et al., 2026), which could yield a smoother and more expressive geographic representation than the centroid, uniform, and Gaussian strategies used here. Another is to move beyond the equirectangular grid to an equal-area geometry such as H3, which would make cells physically comparable across latitudes and could sharpen both the representation and the collision-based analyses.

Limitations

Compiling a comprehensive resource for all non-standard varieties is non-trivial. Although GEOVAR already includes a substantial number of geolects and could be further enriched, achieving broad inclusivity would require a larger coordinated initiative spanning languages not only in Europe but worldwide.

On the methodological side, the kernel density estimation component could be extended with alternative kernels or bandwidth selection strategies beyond the isotropic Gaussian we use here. The current Gaussian formulation also has two notable properties that warrant future refinement. First, geolects with larger regions contribute their sentences to proportionally more cells, which amplifies their aggregate influence across the grid relative to geolects with smaller regions. Second, because the kernel operates directly in degree space rather than on great-circle distances, a given σ corresponds to different physical distances depending on latitude. For instance, at $\sigma = 0.5^\circ$ near the Italian peninsula, the kernel has a standard deviation of roughly 55 km north to south and 42 km east to west. This stems from the fact that kernel weights $w_{d,c}$ are defined at the geolect level, with one Gaussian per region centroid, treating each variety as a single spatially anchored observation. A sentence-level variant would weight every sentence by its geolect’s kernel value, an extension we leave for future work. A further limitation concerns how probability mass is assigned to grid cells during training. In the current formulation, only cells covered by observed varieties receive non-zero probability, which leaves the model unable to place predictions in regions where no training variety is represented.

Ethics Statement

All data sources used in this work are properly cited throughout the paper, and we rely exclusively on existing corpora released by their respective authors for research purposes. We do not foresee any direct harm arising from our work, which aims to support the documentation and computational modeling of linguistic diversity, including non-standardized and under-resourced varieties. We note, however, that a large portion of the underlying material originates from transcribed spoken dialogues, and may therefore contain incidental content that falls outside our control. GEOVAR op-

erates over varieties and regions, not individuals, and our models predict where a text pattern is attested rather than who produced it. We ask that the dataset and models be used for research on linguistic diversity, not for profiling or targeting communities.

Acknowledgements

Sina Ahmadi gratefully thanks the support of the UZH Grant (reference number 269093).

References

- Sina Ahmadi, Daban Q. Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. [Language and speech technology for Central Kurdish varieties](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10034–10045. ELRA and ICCL.
- Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Jalal Hamaamin, Saya Kamal Hama, Ardeshtir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Zamani Roodsari. 2025. [PARME: parallel corpora for low-resourced Middle Eastern languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30032–30053. Association for Computational Linguistics.
- Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. [CODET: A benchmark for contrastive dialectal evaluation of machine translation](#). In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian’s, Malta, March 17-22, 2024*, volume EACL 2024 of *Findings of ACL*, pages 1790–1859. Association for Computational Linguistics.
- Sjef Barbiers, Leonie Cornips, and Jan Pieter Kunst. 2007. The Syntactic Atlas of the Dutch Dialects (SAND): a corpus of elicited speech and text as an online dynamic atlas. In *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 54–90. Springer.

- Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2026. [A catalog of Basque dialectal resources: Online collections and standard-to-dialectal adaptations](#).
- Livia Betti, Sebastian Rieke, Ivica Obadic, Adam J. Stewart, and Esther Rolf. 2026. [What’s in an Earth embedding? an explainability analysis of location encoders](#).
- Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. [A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation](#). In *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Stergios Chatzikyriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis, and Efthymia Sakellariou. 2023. [GRDD: A dataset for Greek dialectal NLP](#). *CoRR*, abs/2308.00802.
- Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. [VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.
- Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.
- Grant DeLozier, Benjamin Wing, Jason Baldrige, and Scott Nesbit. 2016. [Creating a novel geolocation corpus from historical texts](#). In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016, LAW@ACL 2016, August 11, 2016, Berlin, Germany*. The Association for Computer Linguistics.
- Michel Marie Deza and Elena Deza. 2009. Encyclopedia of distances. In *Encyclopedia of distances*, pages 1–583. Springer.
- Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. [SwissDial: Parallel multidialectal corpus of spoken Swiss German](#). *CoRR*, abs/2103.11401.
- Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. 2010. [A latent variable model for geographic lexical variation](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA. Association for Computational Linguistics.
- Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. [OpenLID-v3: Improving the precision of closely related language identification—an experience report](#). *arXiv preprint arXiv:2602.13139*.
- Mariia Fedorova, Jonas Sebulon Frydenberg, Victoria Handford, Victoria Ovedie Chruickshank Langø, Solveig Helene Willoch, Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum, and David Samuel. 2025. [Multi-label Scandinavian language identification \(SLIDE\)](#). In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 179–189, Tallinn, Estonia. University of Tartu Library, Estonia.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Patrick Foley, Matthew Wiesner, Bismarck Odoom, Leibny Paola García-Perera, Kenton Murray, and Philipp Koehn. 2024. [Where are you from? Geolocating speech and applications to language identification](#). In *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.
- Negar Foroutan, Jakhongir Saydaliev, Grace Kim, and Antoine Bosselut. 2026. [ConLID: Supervised contrastive learning for low-resource language identification](#). In *Proceedings of the 19th Conference of*

- the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 6693–6708. Association for Computational Linguistics.
- Yuliia Frund and Sina Ahmadi. 2026. A dialectal corpus for Ukrainian: Collection, classification, and standardization. In *Proceedings of the Workshop on Dialects in NLP—A Resource Perspective*, Palma de Mallorca, Spain. Workshop at LREC 2026.
- Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhainen, Tommi Jauhainen, Krister Lindén, Nikola Ljubesic, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. [A report on the VarDial evaluation campaign 2020](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2020, Barcelona, Spain (Online), December 13, 2020*, pages 1–14. International Committee on Computational Linguistics (ICCL).
- Marco Giordano and Claudia Rinaldi. 2025. [A survey on spoken Italian datasets and corpora](#). *IEEE Access*, 13:29190–29205.
- Mark Graham, Scott A Hale, and Devin Gaffney. 2014. Where in the world are you? geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. [Geolocation prediction in social media data by finding location indicative words](#). In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India. The COLING 2012 Organizing Committee.
- Bo Han, Paul Cook, and Timothy Baldwin. 2014. [Text-based Twitter user geolocation prediction](#). *J. Artif. Intell. Res.*, 49:451–500.
- Dirk Hovy and Christoph Purschke. 2018. [Capturing regional variation with distributed place representations and geographic retrofitting](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.
- Mans Hulden, Miikka Silfverberg, and Jerid Francom. 2015. [Kernel density estimation for text-based geolocation](#). In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 145–150. AAAI Press.
- Janne Bondi Johannessen. 2011. The Nordic Dialect Corpus—a joint research infrastructure. *Oslo Studies in Language*, 3(2).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218.
- Nathaniel Vaughn Kelso and Tom Patterson. 2010. Introducing Natural Earth data—[natureearthdata.com](#). *Geographia Technica*, 5(82-89):25.
- Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, 6(2):100–124.
- Nikola Ljubesic, Tanja Samardzic, and Curdin Derungs. 2016. [TweetGeo - A tool for collecting, processing and analysing geo-encoded linguistic data](#). In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3412–3421. ACL.
- Ismiini Lourentzou, Alex Morales, and ChengXiang Zhai. 2017. Text-based geolocation prediction of social media users with neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 696–705. IEEE.
- Yaron Matras. 2019. Revisiting Kurdish dialect geography: Findings from the manchester database. *Current issues in Kurdish linguistics*, 1:225.
- Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Clara Meister, Ahmetcan Yavuz, Pietro Lesci, and Tiago Pimentel. 2026. [What language is this? ask your tokenizer](#). In *Forty-third International Conference on Machine Learning*.
- Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. [Unifying text, meta-data, and user network representations with a neural network for geolocation prediction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 1260–1272. Association for Computational Linguistics.
- Shogo Miwa and Atsuhiko Kai. 2023. [Dialect speech recognition modeling using corpus of Japanese dialects and self-supervised learning-based model XLSR](#). In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023*, pages 4928–4932. ISCA.
- Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. [Twitter geolocation using knowledge-based methods](#). In *Proceedings of*

- the 4th Workshop on Noisy User-generated Text, NUT@EMNLP 2018, Brussels, Belgium, November 1, 2018, pages 7–16. Association for Computational Linguistics.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. 2025. 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*.
- Douglas O’Shaughnessy. 2025. *Spoken language identification: An overview of past and present research trends*. *Speech Communication*, 167:103167.
- Diego Pescarini and Giorgio Maria Di Nunzio. 2010. Il database dell’Atlante Sintattico d’Italia (ASIt) [the database of the syntactic atlas of Italy (ASIt)]. *Quaderni di lavoro ASIT*, 10:63–81.
- Verena Platzgummer, John McCrae, and Sina Ahmadi. 2026. *Versteasch du mi? computational and socio-linguistic perspectives on GenAI, LLMs, and non-standard language*. *arXiv preprint arXiv:2603.28213*.
- Michel Plüss, Jan Milan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardzic, Manfred Vogel, et al. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772.
- Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017. *A neural model for user geolocation and lexical dialectology*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 209–216. Association for Computational Linguistics.
- Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. 2015. *Exploiting text and network context for geolocation of social media users*. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1362–1367. The Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. *Supervised text-based geolocation using language models on an adaptive grid*. In *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.
- Alex Rosenfeld and Lars Hinrichs. 2023. *Capturing fine-grained regional differences in language use through voting precinct embeddings*. *Computational Linguistics*, 49(4):883–942.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.
- Ryan Soh-Eun Shim and Barbara Plank. 2025. *Dialecto, ma Quanto Dialecto? transcribing and evaluating dialects on a continuum*. In *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.
- Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. *Prototypical networks for few-shot learning*. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.
- Jiugeng Sun, Rita Sevastjanova, Sina Ahmadi, Rico Sennrich, and Mennatallah El-Assady. 2025. *DiaLingle: A gamified interface for dialectal data collection*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 148–158, Vienna, Austria. Association for Computational Linguistics.
- Larraitx Uria and Ricardo Etxepare. 2011. BASYQUE: Aplicación para el estudio de la variación sintáctica. *Linguamática*, 3(1):35–44.
- Muhammad Nur Yasir Utomo, Teguh Bharata Adji, and Igi Ardiyanto. 2018. Geolocation prediction in social media data using text analysis: A review. In *2018 International Conference on Information and Communications Technology (ICOIAC)*, pages 84–89. IEEE.
- Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272.
- Benjamin Wing and Jason Baldridge. 2011. *Simple supervised document geolocation with geodesic grids*. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, Oregon, USA. Association for Computational Linguistics.

- Benjamin Wing and Jason Baldridge. 2014. [Hierarchical discriminative classification for text-based geolocation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 336–348. ACL.
- Benjamin Patai Wing. 2015. [Text-based document geolocation and its application to the digital humanities](#). PhD dissertation, The University of Texas at Austin.
- Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, X. Angela Yao, Hongxu Ma, Lan Mu, Stefano Ermon, Tanuja Ganu, Akshay Nambi, Ni Lao, and Gengchen Mai. 2024. [TorchSpatial: A location encoding framework and benchmark for spatial representation learning](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jingyu Zhang, Alexandra DeLucia, Chenyu Zhang, and Mark Dredze. 2023. [Geo-Seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 4778–4794. Association for Computational Linguistics.

A GeoVar

A.1 Linking to Wikidata

Our selected data sources provide text with meta-data including the name of the varieties and in some cases, the name of the locations. For our geolect modeling, we need to map this information to latitude and longitude. To do so, we first query Wikidata via SPARQL using location labels and country names, then manually verify the results. As a knowledge graph, Wikidata provides not only administrative metadata such as country (P17) and administrative territory (P131), but also geographic coordinates (P625) and, for larger regions, bounding directional coordinates.

For locations on Wikidata, locations spanning a certain geographical region (e.g. cantons or provinces) come with bounding directional coordinates: northernmost (P1332), southernmost (P1333), easternmost (P1334), and westernmost point (P1335). When all four are available, we construct a bounding box and clip it to the corresponding country polygon using Natural Earth data (Kelso and Patterson, 2010), yielding a set of grid cells that approximates the variety’s geographic footprint. When only a point coordinate from P625 is available, the variety maps to a single grid cell. For example, the Bavarian region *Oberpfalz* (Q10555) resolves to a bounding box spanning 49.0°–50.0°N and 11.7°–13.0°E, while the Norwegian municipality *Nes* (Q488026) resolves to a single point at (60.131°N, 11.489°E).

Following this, manual verification proved necessary in several cases, as location names sometimes resolve to non-settlement entities (e.g. a church or a geographic feature) or require disambiguation among multiple candidate entities.

A.2 Text Preprocessing

We apply a uniform set of preprocessing steps across all corpora. We filter out particularly short or uninformative utterances, including samples containing only numerals or single-word responses. Many corpora, particularly those originating from transcribed speech, contain conventional transcription symbols such as # for pauses or brackets for annotator comments; we remove such markers to avoid introducing unintentional biases into identification. For corpora with very long utterances, notably Ukrainian, we split sentences exceeding 100 characters at word boundaries, merging overly short trailing fragments back

into the preceding chunk. Finally, we deduplicate sentences within each variety to prevent repeated formulaic expressions from dominating the representation.

A.3 Data Statistics

Table 4 reports corpus-level statistics per language with the number of geolects, the average number of samples per geolect (Avg/Geo), the type-token ratio (TTR), and the average sentence length in space-delimited words ($|\overline{\text{Word}}|$).

Table 5 complements these with geographic and distributional statistics aggregated at the continuum level as follows:

- σ_{geo} : standard deviation, i.e. the spatial standard deviation of variety centroids from the continuum’s geographic center, (in km)
- NND: the mean within-continuum nearest-neighbor distance (in km)
- Cell collision rate (Coll.): the mean number of varieties sharing a centroid cell at 1° grid resolution.
- Samples/Geolect: the mean number of test samples per geolect
- Entropy: the normalized entropy of the sentence-count distribution, where 1.0 indicates a perfectly uniform distribution.
- Area: the convex hull area of variety centroids

The seven continua differ substantially along these dimensions. Basque is the most geographically compact ($\sigma_{\text{geo}} = 25$ km) with the highest cell collision rate (25.2 varieties per cell) and near-perfect sentence balance (entropy 0.990), making it a dense but well-resourced classification target. At the other extreme, North Germanic is the most dispersed ($\sigma_{\text{geo}} = 628$ km) with the lowest collision rate (2.3) despite having 171 varieties, reflecting a wide geographic span that includes the North Atlantic islands. West Germanic and Italic occupy a middle ground geographically (σ_{geo} of 174 and 347 km) but have the largest variety inventories (282 and 312), resulting in high collision rates (20.6 and 18.1) comparable to Basque. Kurdish and Ukrainian are spatially spread ($\sigma_{\text{geo}} > 300$ km) with low collision rates, but differ sharply in data balance: Kurdish has the lowest entropy (0.402), dominated by Central Kurdish, while Ukrainian is moderately balanced (0.684). Caspian is small and geographically tight ($\sigma_{\text{geo}} = 104$ km), making it the simplest classification setting in GEOVAR.

Continuum	Language	Source	# Geolects	# Samples	Samples/Geolect	Tokens	TTR	Word
Basque	Basque	BASYQUE (Uria and Etxepare, 2011)	40	7,610	190.2	37,508	0.167	4.9
North Germanic	Danish	DanDiaSyn (Johannessen, 2011)	13	7,851	603.9	173,658	0.055	22.1
	Icelandic	ScanDiaSyn (Johannessen, 2011)	7	6,122	874.6	70,034	0.105	11.4
	Swedish	SweDiaSyn (Johannessen, 2011)	31	20,793	670.7	292,062	0.056	14.0
	Norwegian	NorDiaSyn (Johannessen, 2011)	108	117,256	1,085.7	1,918,713	0.024	16.4
	Ovdalian	ScanDiaSyn (Johannessen, 2011)	7	1,034	147.7	13,591	0.142	13.1
	Faroese	ScanDiaSyn (Johannessen, 2011)	5	3,641	728.2	61,522	0.096	16.9
West Germanic	Dutch	SAND (Barbiers et al., 2007)	267	55,582	208.2	436,259	0.034	7.8
	German	Bethupferl (Blaschke et al., 2025)	7	5,482	783.1	53,746	0.234	9.8
	Swiss German	SwissDial (Dogan-Schönberger et al., 2021)	8	30,682	3,835.2	348,560	0.191	11.4
Italic	Italic	ASIt (Pescarini and Di Nunzio, 2010)	313	71,127	227.2	416,807	0.125	5.9
Caspian	Mazandarani	PARME (Ahmadi et al., 2025)	2	4,106	2,053.0	37,696	0.173	9.2
	Gilaki	PARME (Ahmadi et al., 2025)	8	5,249	656.1	53,957	0.290	10.3
	Talysh	PARME (Ahmadi et al., 2025)	3	2,032	677.3	15,650	0.342	7.7
Kurdic	Hawrami	PARME (Ahmadi et al., 2025)	5	7,669	1,533.8	64,537	0.354	8.4
	N. Kurdish	Matras’s (2019)	72	10,614	147.4	56,775	0.116	5.3
	C. Kurdish	CORDI (Ahmadi et al., 2024)	5	117,221	23,444.2	789,887	0.127	6.7
	S. Kurdish	PARME (Ahmadi et al., 2025)	7	12,053	1,721.9	107,228	0.254	8.9
	Laki Kurdish	PARME (Ahmadi et al., 2025)	3	4,345	1,448.3	46,260	0.309	10.6
Ukrainian	Ukrainian	Frund and Ahmadi’s (2026)	18	29,329	1,629.4	347,277	0.190	11.8
Total			929	517,985	557.6	5,325,618	0.088	10.3

Table 4: Corpus statistics per language, grouped by dialect continuum.

Continuum	Dataset Properties							Performance		
	# Geolects	σ_{geo} (km)	NND (km)	Coll.	Samples/Geolect	Entropy	Area (km ²)	Acc@0c	F1	Med. (km)
Basque	40	25	5	25.2	20	0.990	2,485	.972	.949	0
Italic	313	347	13	18.1	24	0.945	624,290	.683	.679	0
North Germanic	171	628	45	2.3	93	0.980	2,671,519	.185	.226	243
West Germanic	282	174	12	20.6	33	0.878	312,504	.574	.501	0
Caspian	13	104	19	4.7	95	0.773	8,124	.973	.892	0
Kurdic	92	310	37	3.9	160	0.402	497,400	.856	.376	0
Ukrainian	18	350	117	1.0	164	0.684	339,704	.650	.355	0

Table 5: Per-continuum dataset properties and the performance of the best model (M3).

A.4 Additional Data Sources

We are aware of additional variety corpora, including MADAR for Arabic (Bouamor et al., 2018), COJADS for Japanese (Miwa and Kai, 2023), and GRDD for Greek (Chatzikyriakidis et al., 2023), *inter alia*; we chose not to include them in the current version of GEOVAR due to their wide geographical distribution. In addition, VarDial 2020’s shared task (Gaman et al., 2020) focused on Social Media Variety Geolocation (SMG) with datasets covering German, Swiss German and languages in former Yugoslavia (Ljubesic et al., 2016); these could be added to GEOVAR in the future.

B Experimental Setup

B.1 Sentence Encoding

We represent each sentence s as a dense vector $\mathbf{e}_s \in \mathbb{R}^D$ obtained from a pre-trained multilingual sentence encoder. We use CANINE-C (Clark et al., 2022), a character-level transformer that operates directly on Unicode code points without subword tokenization. This choice is motivated

by the nature of linguistic variation: orthographic differences between closely related varieties often manifest at the character level, and subword tokenizers trained on standardized text may collapse these distinctions. In preliminary experiments on the Italian dev set using the B2 nearest-centroid method, CANINE-C substantially outperformed both Paraphrase-MiniLM (Reimers and Gurevych, 2019) (F1: 0.487 vs. 0.317) and LaBSE (Feng et al., 2022) (F1: 0.487 vs. 0.392), confirming the advantage of character-level representations for variety discrimination. As such, all experiments use CANINE-C.

B.2 Data Splits

For each variety, we partition sentences into 80% train, 10% validation, and 10% test, with splits drawn uniformly at random using a fixed seed. We enforce a minimum of one sentence per split so that even the smallest varieties contribute to all three partitions. Sentences are deduplicated within each variety before splitting, which pre-

vents the same formulaic utterance from appearing in both training and evaluation. Because sentence counts per variety vary by several orders of magnitude across GEOVAR, we do not subsample or reweight during training; macro-F1 ensures that low-resource varieties remain visible in the reported metrics.

B.3 Hyperparameter Optimization

We tune the fine-tuned encoder on the validation split via a three-stage sweep:

- Stage 1: grid size (0.1, 0.25, 0.5, 1.0)
- Stage 2: learning rate (10^{-5} , 2×10^{-5} , 5×10^{-5}) and epochs (3, 5, 10)
- Stage 3: prototype mapping (centroid, uniform, Gaussian) and bandwidth σ (0.1, 0.25, 0.5, 1.0, 2.0)

Each stage holds prior winners fixed. We select by Acc@0c rather than macro-F1 because F1 can be inflated by Gaussian smoothing, which expands the geolect set $\mathcal{D}_{\hat{c}}$ without improving geographic precision. Acc@0c rewards exact-cell localization and is immune to this effect. Figure 6 shows the Stage 1 and Stage 3 sweeps per continuum. All continua converge on 1.0° grid size and centroid-mode prototypes. Notably, the best-performing configurations, as presented in Table 6, use centroid-mode prototypes for every continuum. This means that the predicted cell maps to a single geolect, and Acc, CellAcc, and Acc@0c collapse to the same value. As such, a prediction is correct if and only if the model places the sentence in its variety’s centroid cell.

B.4 Training Details

The fine-tuned encoder is trained with AdamW and cross-entropy loss over centroid-cell labels, with hyperparameters (grid size, learning rate, number of epochs, mapping mode, Gaussian bandwidth) selected on the validation set via the sweep described in Section B.3. The non-fine-tuned learnable baselines (M1 and M2) use fixed settings: L-BFGS with $C=1.0$ for logistic regression, and Adam with $\text{lr}=10^{-3}$ for 50 epochs for the prototype network.

All experiments were run on a single machine with a NVIDIA RTX 4090 GPU. The full hyperparameter sweep for the fine-tuned encoder (Stage

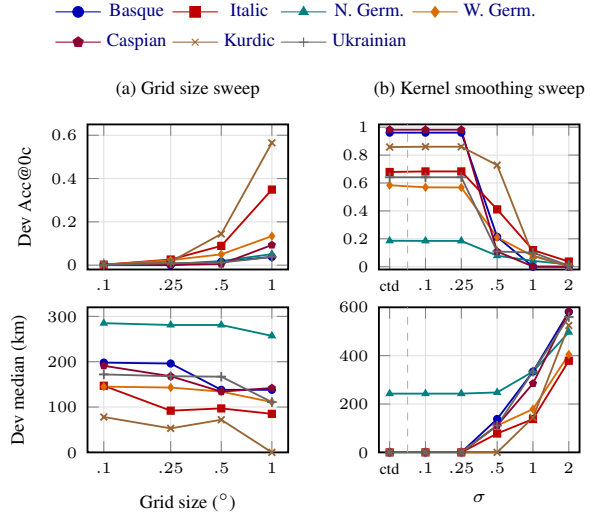


Figure 6: Validation sweeps for the fine-tuned encoder (M3): Acc@0c (top) and median error in km (bottom) across grid resolution (a) and prototype mapping mode (b), where *ctd* is centroid and remaining points vary Gaussian bandwidth σ . Coarser grids and centroid-mode prototypes both increase Acc@0c while keeping median error low.

Continuum	Grid	lr	Ep.	Mode	Dev Acc@0c
Basque	1.0°	2×10^{-5}	3	ctd	0.961
Italic	1.0°	5×10^{-5}	10	ctd	0.679
N. Germanic	1.0°	5×10^{-5}	10	ctd	0.186
W. Germanic	1.0°	5×10^{-5}	10	ctd	0.584
Caspian	1.0°	5×10^{-5}	3	ctd	0.982
Kurdic	1.0°	1×10^{-5}	10	ctd	0.858
Ukrainian	1.0°	5×10^{-5}	3	ctd	0.641

Table 6: Best fine-tuned encoder configuration per continuum, selected on the validation split by maximizing Acc@0c (exact-cell accuracy). Grid is the geodesic grid resolution, lr and Ep. are the learning rate and epochs, and Mode is the prototype mapping.

13) completed in under 36 GPU-hours per continuum, for a total of approximately 250 GPU-hours across the seven continua.

C Results

C.1 Test Evaluation

Tables 8-14 report per-continuum test set results for all five methods across the three mapping modes. For M1 and M2, the centroid, uniform, and Gaussian rows report identical cell predictions and therefore identical spatial metrics (CellAcc, Median km, Acc@Nc); only the variety-dependent metrics (Acc, F1) vary across modes. This can produce striking jumps. For instance, M1’s F1 rises from .919 (centroid) to .998 (Gaus-

Continuum	# Pairs	# Sentences	Median (km)	Mean (km)	% NN	% Periph. True	% Periph. Pred
Basque	3	18	80.7	80.7	0.0	100.0	100.0
Italic	382	1704	111.2	330.5	1.3	25.8	29.4
North Germanic	2912	10603	337.9	438.1	2.5	33.3	32.0
West Germanic	720	3200	111.2	109.9	2.3	44.1	49.5
Caspian	6	26	288.7	203.9	0.0	84.6	73.1
Kurdic	232	1847	143.7	188.3	6.9	31.4	34.4
Ukrainian	103	961	381.2	449.4	12.2	45.1	24.9

Table 7: Confusion characteristics per continuum (M3, centroid mode). # denotes number.

sian) on Basque (Table 8), not because the model localizes sentences better, but because Gaussian mode’s $\mathcal{D}_c$ pools more geolects per cell, raising the chance that the true geolect is in the pool. For M3 and B2, both the predictions and the variety recovery depend on the mode, so all metrics can vary across rows. It is also worth noting that centroid and uniform coincide for most continua because, at the selected 1° grid, the majority of geolect regions occupy a single cell, so the two mappings induce the same cell-to-geolect sets; they diverge only where regions span multiple cells (e.g. West Germanic).

In addition, we report the performance of off-the-shelf LID tools in comparison to our model in Table 3.

C.2 Confusion Analysis

Table 7 summarizes confusion characteristics per continuum for the best fine-tuned encoder in centroid mode. *Pairs* counts distinct (true, predicted) variety pairs with at least two confused sentences, and *Sentences* gives the total number of misclassified sentences across those pairs. *Median* and *Mean* (both in km) report the haversine distance between the centroids of confused varieties. *%NN* is the fraction of confused sentences predicted as the true variety’s nearest neighbor, while *%Periph. True* and *%Periph. Pred* give the fractions of confusions involving a peripheral true or predicted variety.

C.3 Rank Correlations

We compute Spearman rank correlations between six dataset properties (number of geolects, σ_{geo} , samples per geolect, cell collision rate, entropy, and area as defined in Section A.3) and three model outcomes (Acc@0c, macro-F1, median geographic error) across the seven continua as shown in Figure 7. For each property-outcome pair, we rank the seven continua on both variables and compute ρ using SciPy (Virtanen et al., 2020), which

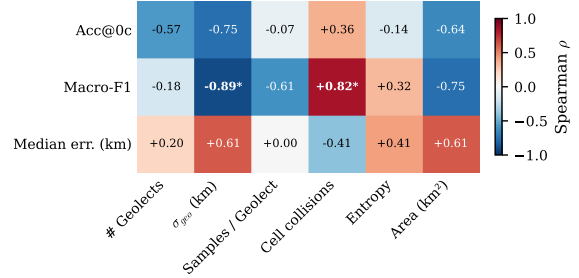


Figure 7: Spearman rank correlations ($n = 7$ continua) between dataset properties (rows) and model outcomes (columns). Cell values are ρ ; * denotes $p < 0.05$.

yields a two-sided p -value under the null hypothesis of no monotonic relationship. With only seven data points, statistical power is limited and only strong monotonic effects reach $p < 0.05$. The significant correlations we observe therefore indicate robust, large-magnitude relationships rather than marginal trends, and the non-significant cells should be read as inconclusive rather than as evidence of no effect.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.150	.889	.108	0	.889	.985	.985	.985
B2: Nearest Centroid	centroid	.925	.925	.914	0	.925	.999	.999	.999
	uniform	.925	.925	.914	0	.925	.999	.999	.999
	Gaussian	.979	.001	.956	696	.001	.010	.024	.048
M1: LogReg $\rightarrow$ cells	centroid	.940	.940	.919	0	.940	.998	.998	.998
	uniform	.940	.940	.919	0	.940	.998	.998	.998
	Gaussian	.998	.940	.998	0	.940	.998	.998	.998
M2: ProtoNet	centroid	.955	.955	.937	0	.955	1.000	1.000	1.000
	uniform	.951	.951	.932	0	.951	1.000	1.000	1.000
	Gaussian	.999	.945	.999	0	.945	.999	.999	.999
M3: Fine-tuned Encoder	centroid	.973	.973	.949	0	.973	1.000	1.000	1.000
	uniform	.973	.973	.949	0	.973	1.000	1.000	1.000
	Gaussian	.998	.243	.997	138	.243	.998	1.000	1.000

Table 8: Test set results for the **Basque** continuum. The selected configuration uses centroid mode; Gaussian rows are shown for comparison. Note that Gaussian inflates F1 via pooled geolect sets $\mathcal{D}_c$ but collapses CellAcc and Acc@0c because its argmax cells sit outside the true region polygon.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.242	.376	.110	111	.376	.609	.728	.787
B2: Nearest Centroid	centroid	.206	.206	.197	190	.206	.458	.606	.700
	uniform	.206	.206	.197	190	.206	.458	.606	.700
	Gaussian	.192	.192	.214	193	.192	.439	.592	.694
M1: LogReg $\rightarrow$ cells	centroid	.427	.427	.408	78	.427	.689	.806	.859
	uniform	.427	.427	.408	78	.427	.689	.806	.859
	Gaussian	.427	.427	.408	78	.427	.689	.806	.859
M2: ProtoNet	centroid	.353	.353	.368	111	.353	.621	.748	.829
	uniform	.365	.365	.378	111	.365	.630	.756	.834
	Gaussian	.354	.354	.369	111	.354	.625	.755	.834
M3: Fine-tuned Encoder	centroid	.683	.683	.679	0	.683	.890	.940	.960
	uniform	.683	.683	.679	0	.683	.890	.940	.960
	Gaussian	.686	.686	.686	0	.683	.890	.939	.960

Table 9: Test set results for the **Italic** continuum (Italian, 313 varieties). Under Acc@0c selection all modes converge to similar performance because centroid, cell, and Gaussian differ negligibly at the chosen grid size for varieties occupying a single cell.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.084	.105	.054	429	.105	.185	.265	.376
B2: Nearest Centroid	centroid	.052	.052	.062	496	.052	.120	.208	.308
	uniform	.052	.052	.062	496	.052	.120	.208	.308
	Gaussian	.099	.000	.149	1046	.000	.001	.005	.010
M1: LogReg $\rightarrow$ cells	centroid	.121	.121	.139	334	.121	.226	.346	.498
	uniform	.121	.121	.139	334	.121	.226	.346	.498
	Gaussian	.723	.121	.786	334	.121	.226	.346	.498
M2: ProtoNet	centroid	.085	.085	.111	351	.085	.195	.316	.433
	uniform	.083	.083	.105	352	.083	.191	.313	.431
	Gaussian	.662	.085	.729	352	.085	.191	.309	.429
M3: Fine-tuned Encoder	centroid	.185	.185	.226	243	.185	.326	.458	.569
	uniform	.185	.185	.226	243	.185	.326	.458	.569
	Gaussian	.318	.079	.379	247	.079	.318	.449	.560

Table 10: Test set results for the **North Germanic** continuum (Danish, Swedish, Norwegian, Icelandic, Faroese, Ovdalian; 171 varieties). This is the hardest continuum in GEOVAR: the large number of varieties, uneven data distribution, and wide geographic span including the North Atlantic islands make fine-grained variety discrimination substantially more difficult. Even the best configuration yields a median error of 243 km.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.243	.332	.028	111	.332	.675	.839	.902
B2: Nearest Centroid	centroid	.267	.267	.160	130	.267	.600	.827	.938
	uniform	.269	.269	.162	130	.263	.603	.828	.939
	Gaussian	.484	.000	.440	776	.000	.005	.028	.058
M1: LogReg $\rightarrow$ cells	centroid	.397	.398	.320	71	.397	.800	.938	.984
	uniform	.398	.398	.321	71	.397	.800	.938	.984
	Gaussian	.999	.398	.999	71	.397	.800	.938	.984
M2: ProtoNet	centroid	.350	.351	.242	111	.350	.738	.908	.975
	uniform	.353	.353	.248	93	.353	.740	.908	.974
	Gaussian	.998	.354	.999	75	.353	.745	.910	.975
M3: Fine-tuned Encoder	centroid	.574	.574	.501	0	.574	.912	.977	.994
	uniform	.576	.576	.502	0	.562	.912	.977	.995
	Gaussian	.847	.220	.829	111	.212	.845	.965	.988

Table 11: Test set results for the **West Germanic** continuum (Dutch, German, Swiss German; 282 varieties). The fine-tuned encoder achieves 57.4% exact-cell accuracy and a median error of 0 km at the selected 1 grid; Gaussian mode still appears higher on F1 because of $\mathcal{D}_c$ pooling but collapses CellAcc.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.868	.901	.501	0	.901	.927	.928	1.000
B2: Nearest Centroid	centroid	.744	.744	.592	0	.744	.855	.862	1.000
	uniform	.744	.744	.592	0	.744	.855	.862	1.000
	Gaussian	.696	.000	.598	783	.000	.000	.000	.001
M1: LogReg $\rightarrow$ cells	centroid	.879	.879	.662	0	.879	.923	.925	1.000
	uniform	.879	.879	.662	0	.879	.923	.925	1.000
	Gaussian	1.000	.879	1.000	0	.879	.923	.925	1.000
M2: ProtoNet	centroid	.912	.912	.707	0	.912	.941	.944	1.000
	uniform	.910	.910	.704	0	.910	.942	.944	1.000
	Gaussian	1.000	.907	1.000	0	.907	.941	.945	1.000
M3: Fine-tuned Encoder	centroid	.973	.973	.892	0	.973	.986	.986	1.000
	uniform	.973	.973	.892	0	.973	.986	.986	1.000
	Gaussian	.984	.131	.918	111	.131	.984	.986	.992

Table 12: Test set results for the **Caspian** continuum (Gilaki, Talysh, Mazandarani; 13 varieties). Because these three languages occupy a narrow strip along the southern Caspian coast, Acc@3c saturates at 1.0 for most methods: within three cells of any true centroid, the correct variety is almost guaranteed to be among the candidates.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.696	.713	.102	0	.713	.922	.970	.983
B2: Nearest Centroid	centroid	.224	.224	.134	111	.224	.695	.926	.968
	uniform	.224	.224	.134	111	.224	.695	.926	.968
	Gaussian	.299	.299	.137	111	.299	.702	.925	.968
M1: LogReg $\rightarrow$ cells	centroid	.719	.719	.225	0	.719	.935	.977	.990
	uniform	.719	.719	.225	0	.719	.935	.977	.990
	Gaussian	.719	.719	.225	0	.719	.935	.977	.990
M2: ProtoNet	centroid	.456	.456	.168	91	.456	.857	.956	.973
	uniform	.460	.460	.169	91	.460	.859	.958	.975
	Gaussian	.467	.467	.169	91	.467	.852	.955	.974
M3: Fine-tuned Encoder	centroid	.856	.856	.376	0	.856	.962	.982	.989
	uniform	.856	.856	.376	0	.856	.962	.982	.989
	Gaussian	.859	.859	.377	0	.859	.961	.982	.990

Table 13: Test set results for the **Kurdic** continuum (Central, Northern, Southern, Laki Kurdish, Hawrami; 92 varieties). The large gap between accuracy (.856) and macro-F1 (.376) for the fine-tuned encoder reflects class imbalance: Central Kurdish accounts for over 117k of the 147k Kurdish sentences, and high accuracy on it masks weaker performance on the four smaller varieties.

Method	Mode	Acc	CellAcc	F1	Median (km)	Acc@0c	Acc@1c	Acc@2c	Acc@3c
B1: LogReg $\rightarrow$ varieties	—	.531	.531	.201	0	.531	.548	.621	.645
B2: Nearest Centroid	centroid	.278	.278	.176	246	.278	.334	.494	.532
	uniform	.278	.278	.176	246	.278	.334	.494	.532
	Gaussian	.412	.000	.321	818	.000	.009	.025	.043
M1: LogReg $\rightarrow$ cells	centroid	.531	.531	.201	0	.531	.548	.621	.645
	uniform	.531	.531	.201	0	.531	.548	.621	.645
	Gaussian	.757	.531	.764	0	.531	.548	.621	.645
M2: ProtoNet	centroid	.446	.446	.264	145	.446	.479	.611	.644
	uniform	.442	.442	.267	145	.442	.480	.607	.638
	Gaussian	.798	.456	.735	145	.456	.488	.616	.649
M3: Fine-tuned Encoder	centroid	.650	.650	.355	0	.650	.672	.784	.812
	uniform	.650	.650	.355	0	.650	.672	.784	.812
	Gaussian	.693	.113	.435	111	.113	.693	.765	.808

Table 14: Test set results for the **Ukrainian** continuum (Ukrainian; 18 varieties). Acc@ N c increases modestly between $N=0$ and $N=3$ because Ukrainian varieties in GEOVAR are separated by distances larger than three cells at the selected grid resolution, leaving little near-miss regime to recover.

D Embedding Geometry and Geographic Axis

Complementing the UMAP analysis of Italic in Section 7.1, we provide a quantitative analysis across all seven continua of whether the per-geolect mean embeddings of CANINE-C encode geographic position beyond the discrete cell-classification signal. We project each geolect’s embedding onto the direction in embedding space that best predicts its geographic coordinate. For each continuum, we fit a ridge regression ($\alpha = 1$) from the 768-dimensional embeddings to either latitude or longitude, take the L_2 -normalized weight vector, and project each embedding onto it. The reported Pearson r in Figure 8 is computed on out-of-fold predictions via 5-fold cross-validation, so the correlation reflects how well the embedding generalizes geographic position to unseen geolects rather than a training-set fit. Because the dominant geographic axis varies across continua where some are stretched primarily north-south (e.g. Italic) and others east-west (e.g. North Germanic, spanning Iceland to Finnmark), we fit the regression against both latitude and longitude and report whichever yields the stronger absolute cross-validated correlation per continuum.

The projection reveals substantial variation in how strongly the embedding encodes geography across continua. Italic ($r = 0.88$, northward), West Germanic ($r = 0.87$, northward), North Germanic ($r = 0.83$, eastward), and Kurdish ($r = 0.78$, northward) exhibit clear geographic signal, with the colour gradient on the map closely tracking the chosen axis. The dominant axis itself is informative: continua with greater latitudinal extent (Italic, West Germanic, Kurdish) project most strongly along latitude, whereas North Germanic, stretched across more than fifty degrees of longitude from Iceland to Finnmark, projects along longitude. The remaining three continua show weaker correlations. Basque ($r = 0.36$) covers only ~ 25 km of geographic spread, leaving little coordinate variation for the embedding to encode. Caspian ($r = -0.46$) and Ukrainian ($r = -0.53$) have small geolect counts (13 and 18 respectively), and the cross-validation estimate becomes high-variance and can flip sign under small perturbations of the fold assignment.

A second observation concerns the alignment between the embedding axis (colour) and dialectological grouping (shape). Within strong-signal

continua, geolects belonging to the same dialectological group tend to cluster along similar bands of the colour gradient rather than being scattered across it, consistent with the within-continuum substructure visible in the UMAP analysis (Section 7.1).

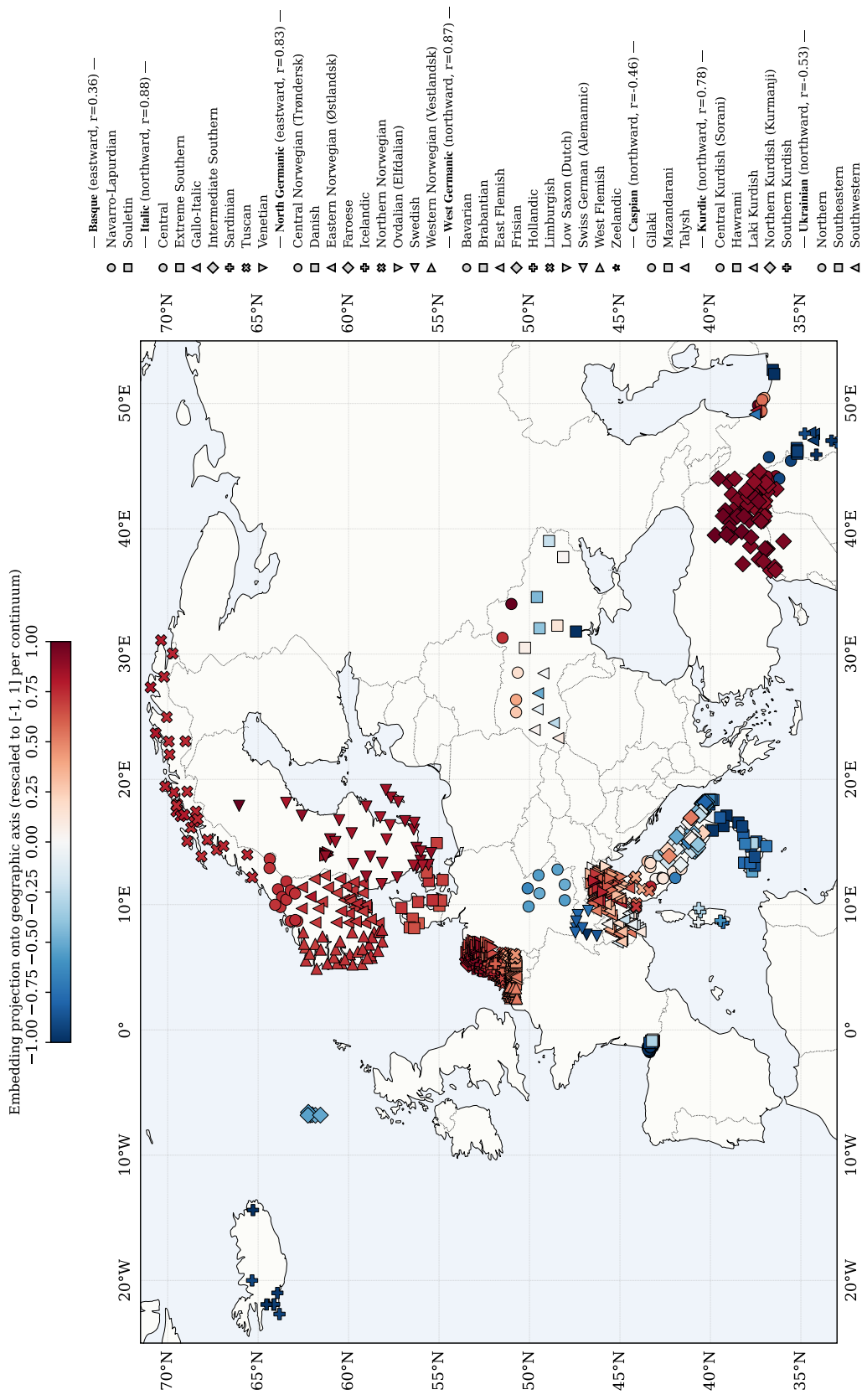


Figure 8: Per-gelect map combining geographic position, embedding's geographic projection (colour), and dialectological group (shape). Each point is one gelect. Position shows geographic coordinates; colour shows the embedding's projection onto the axis (latitude or longitude) with the stronger cross-validated Pearson correlation, rescaled to [-1, 1] per continuum; shape shows the traditional dialectological group. The legend reports, for each continuum, the chosen axis and its cross-validated r . Strong-signal continua (Italic, West Germanic, North Germanic, Kurdic) show clear colour gradients along the chosen axis; weaker continua (Basque, Caspian, Ukrainian) reflect small geographic extent or small gelect counts.