Aspects of Selecting the Right ASR Training Languages for Under-Resourced Languages

J. Elizabeth Liebl, Summer Chambers, Matthew C. Kelley, and Géraldine Walther
George Mason University
Fairfax, Virginia, USA
{jliebl, schamb3, mkelle21, gwalthe}@gmu.edu

Abstract

We investigate how training languages should be selected for cross-lingual IPA ASR on unseen languages. Using Common Voice audio and Vox Communis phonetic transcripts, we train multilingual IPA-based ASR models for Upper Sorbian, Luganda, and Tatar under three linguistically motivated selection strategies: genealogical relatedness, geographic proximity, and phonological inventory overlap. We compare these strategies to a random baseline and evaluate performance with phone error rate. Linguistically informed selection generally improves transfer, but no single strategy is consistently optimal. Geographic proximity performs best for Luganda, phonological overlap is slightly best for Tatar, and none of the proposed strategies outperform random selection for Upper Sorbian. The results suggest that linguistic similarity aids low-resource ASR transfer, but that the most useful dimension of similarity varies by target language.

1 Introduction

Language documentation faces a persistent transcription bottleneck: while transcription is essential for the effective use of audio data, it remains slow and costly, often requiring hundreds of hours of expert labor (Anastasopoulos and Chiang, 2018; Bird, 2021; Geng et al., 2025; Liang and Levow, 2025). Automatic speech recognition (ASR) offers a partial solution, and even imperfect output can be useful if correcting it is faster than transcribing from scratch (Fort and Sagot, 2010). This makes ASR particularly attractive for linguist-in-the-loop documentation workflows, where automatic transcripts are iteratively corrected and reused for further training.

Multilingual ASR models that directly predict International Phonetic Alphabet (IPA) transcriptions are especially promising for low-resource and previously unseen languages, because they are not hampered by changes in orthography between languages. MultIPA (Taguchi et al., 2023) is one such model, but its zero-shot phone error rates remain high, limiting its immediate utility for documentation. One reason is data availability: although Common Voice provides broad multilingual speech coverage (Ardila et al., 2020), verified transcripts are generally orthographic rather than phonemic, restricting the set of languages that can be used for IPA-based training. The release of Vox Communis (Ahn and Chodroff, 2022), which provides machine-generated phonetic transcriptions for many Common Voice languages, changes this situation by making larger-scale multilingual phonetic training more feasible.

This raises a practical question for unseen-language ASR: how can training languages be selected in a principled way to reduce error on a target language? In this work, we present an initial comparison of three linguistically motivated training-language selection strategies—genealogical relatedness, phonological inventory overlap, and geographic proximity (as a weak proxy for synchronic language contact)—against a random baseline for unseen-language IPA ASR. Using a pool of 75 candidate training languages, we evaluate transfer to three unseen target languages and model phone error counts over individual audio clips with Poisson mixed-effects regression. We find that across these three target languages, linguistically informed selection often improves over random choice, but no single strategy is uniformly best: the most effective criterion appears to depend on the target language.

2 Related Work

Prior multilingual ASR work suggests that transfer is often stronger when training and target languages are linguistically similar (Zampieri et al., 2020; Kuparinen et al., 2023; Bafna et al., 2024), but similarity can be defined in different ways, in-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 148–156
July 4, 2026 ©2026 Association for Computational Linguistics

cluding genealogical relatedness, phonological inventory overlap, and language contact. The problem of systematically selecting transfer languages has received considerable attention in text-based NLP. Lin et al. (2019) frame it as a ranking problem and evaluate features including genetic distance, geographic distance, and phonological inventory distance—alongside data-driven measures such as word overlap and corpus size—across machine translation, POS tagging, entity linking, and dependency parsing. A central finding is that no single linguistic feature reliably identifies the best transfer language, and that data-driven features are often more predictive than linguistic ones in isolation. Rice et al. (2025) extend this analysis to pretrained multilingual models for POS tagging, finding that combining dataset-dependent and fine-grained typological features yields the strongest rankings, and that genealogical distance remains consistently important across model architectures. These selection strategies have not previously been compared in the ASR domain. Because our setting targets languages for which no or limited labeled data, such as phonetic transcripts, are yet available, we restrict our comparison to data-agnostic linguistic strategies, extending the evaluation of Lin et al. (2019) and Rice et al. (2025) to speech-to-IPA transfer.

More generally, prior work in documentation-oriented ASR has shown that language choice can substantially affect multilingual ASR performance (van der Westhuizen et al., 2021), and Jimerson et al. (2023) argue that low-resource ASR design decisions are often language-dependent. In documentation settings, this makes training-language selection a practical early design decision, since small multilingual seed models may need to be built before enough corrected target-language data exist to support more tailored retraining.

3 Methods

We selected Upper Sorbian, Tatar, and Luganda as target languages because they overlap with the unseen-language evaluation setup in Taguchi et al. (2023), had sufficient Common Voice audio and Vox Communis TextGrids for the present experiments, and provided contrasting relationships to the candidate training pool under the similarity measures used here. Hakha Chin was excluded because corresponding Vox Communis TextGrids were unavailable.

Using the set of Common Voice languages with more than 2800 clips transcribed by Vox Communis, we used genealogical classification data and geographic language-center coordinates from Glottolog (Hammarström et al., 2026) to identify the candidate languages closest to each target in genealogical and geographic terms. In this study, geographic distance was used as a weak proxy for synchronic language contact, although in future studies it may be more prudent to consider language contact in a diachronic sense.

To compare phonological overlap, we used Phoible (Moran and McCloy, 2019) inventories[1] which had been collapsed down to columns reflecting our transcript preprocessing method to assess the amount of phonetic overlap between languages.

For each training language, the audio files were downloaded from Common Voice and the associated TextGrids were downloaded from Vox Communis. Transcripts were generated from the TextGrids and preprocessed to replace multi-character affricates with their associated single-character ligatures. Diacritics were also removed. Training, development, and evaluation splits were randomly selected from the available clips for each language, and selected clips were downsampled to 16 kHz. Notably, for languages which appeared in multiple models the splits were identical for each model.

For each target language, we trained multilingual ASR models under four training-language selection conditions: genealogical relatedness, geographic proximity, phonological inventory overlap, and a single, shared random baseline. The three target languages were first removed from the candidate metadata table so that no target language could be selected as its own training language.

For the geographic condition, we extracted the latitude and longitude of each target language and computed geodesic distance in kilometers from that target language to every remaining candidate language using `GeoPy` (Lopez Gonzalez-Nieto et al., 2020). Candidate languages were then ranked in ascending order of distance, and the nearest languages were selected for model training.

For the genealogical condition, each language's Glottolog classification string was converted into a set of lineage nodes. For each candidate language, we then computed the overlap between the

[1] A full list of language inventories we consulted are available, along with our code, on our GitHub: `https://github.com/ellie-liebl/ASR_ComputEL`.

candidate and training languages using the number of overlapping items between the sets, percentage of the candidate language's set in the overlap, and Jaccard overlap as a percentage, following the workflow in Figure 1. An example of the tie-break process is shown with Tatar in Table 1. For the phonological condition, we used the PHOIBLE inventories, represented as a set of segments for each language. The set-based workflow demonstrated in Figure 1 was then repeated over these sets. [2]

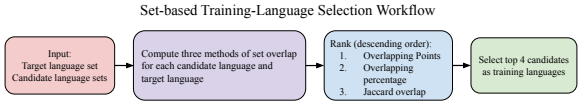

Figure 1: The workflow for determining the 4 closest languages from the candidate languages for the Genealogical and Phonological conditions. The three methods of set overlap are number of points in the set overlap, the percentage of the set that is in the set overlap, and the Jaccard overlap as a percentage.

Candidate	Number of Points	Percentage	Jaccard Overlap
Bashkir	7	100.00	100.00
Kyrgyz	4	57.14	44.44
Uzbek	3	42.86	33.33
Uyghur	3	42.86	30.00

Table 1: Example of genealogical ranking for Tatar. Candidates are ordered first by number of shared lineage nodes, then by percentage of the target classification path covered, and finally by Jaccard overlap as a percentage. Uzbek ranks above Uyghur only at the third step, illustrating the tie-breaking procedure.

Target Language	Strategy	Training Languages
Luganda	Geographic Proximity	Dholuo, Kinyarwanda, Swahili, Basaa
	Genealogical Relatedness	Kinyarwanda, Swahili, Basaa, Yoruba
	Phonetic Overlap	Dholuo, Dutch, Indonesian, Basaa
Tatar	Geographic Proximity	Chuvash, Russian, Bashkir, Estonian
	Genealogical Relatedness	Bashkir, Kyrgyz, Uighur, Uzbek
	Phonetic Overlap	Bashkir, Hindi, Uighur, Polish
Upper Sorbian	Geographic Proximity	Czech, Polish, Slovak, Slovenian
	Genealogical Relatedness	Czech, Polish, Slovak, Bulgarian
	Phonetic Overlap	Bulgarian, Lithuanian, Romanian, Ukrainian
All	Random Selection	Swedish, Ukrainian, Abkhazian, Romanian

Table 2: Training languages presented by target language and strategy.

All models were fine-tuned from `wav2vec2-large-xlsr-53` (Baevski et al., 2020) to predict IPA transcriptions from audio using `Wav2Vec2ForCTC` (Wolf et al., 2020). Audio was downsampled to 16 kHz before training. For each language, clips were partitioned into training, development, and evaluation sets via random sampling. Each model was trained on 8,000 clips, 2,000 clips from each training language. Four training languages were used per model because Luganda has no fifth genealogical relative among the candidate languages, making four the largest number consistent across all three target languages and all selection strategies.

We used 2,000 clips per training language to keep the comparison across selection strategies computationally controlled, while also following prior evidence that multilingual speech-to-IPA transfer can perform well with relatively small per-language training sets and may not improve monotonically with additional data (Taguchi et al., 2023). All models were trained for 10 epochs with a learning rate of 1×10^{-4} and a batch size of 4, using CTC loss reduction set to `mean` following Taguchi et al. (2023); the learning rate and batch size reflect standard practice for fine-tuning large pre-trained speech encoders (Baevski et al., 2020).

We first evaluated each model on its own training languages as a baseline assessment and then on the corresponding unseen target language. Performance was measured as phone error rate (PER), a variant of character error rate (CER) applied to phonetic transcriptions. Since both training and evaluation transcripts are drawn from the Vox Communis pipeline rather than human-verified annotation, PER in this study measures agreement with that pipeline's phonetic representations rather than accuracy against a human standard. Differences in PER across strategies should therefore be interpreted as reflecting how well each model learns to replicate Vox Communis output, and the practical value of the approach for documentary linguistics depends in part on the quality and consistency of those representations. We do not report word error rate (WER). WER is not appropriate here because word boundaries are not preserved in the pipeline. Further, studies have shown that CER-like measures more closely align with human judgement than WER when evaluating ASR systems specifically (K et al., 2024).

Aggregate PER values summarize overall model performance but do not account for variance across individual clips or allow formal inference about whether strategy differences exceed clip-level noise. To compare strategy effects statistically, we fit a Poisson generalized linear mixed-effects model to predict the PER for each clip. Because PER is a normalized rate, we converted it to an error count by

[2] Appendix A lists the top five candidate training languages identified for each target language.

multiplying each clip's error score by its gold standard transcript length in number of characters. The model included fixed effects for Strategy, Target Language, and their interaction, an offset term for $\log(\text{Clip Length})$, and a random intercept for Clip. The factor variables were treatment coded and their reference levels were Genealogical for Strategy and for Upper Sorbian for Target Language.

4 Results

We first evaluated each model on its own training languages before testing transfer to the unseen targets. Most strategy-based models achieved relatively low seen-language PERs (6.69–11.38), whereas the shared random baseline was notably worse (17.40). The main exception was the Tatar geographic model (17.47), driven largely by high error on Chuvash. These seen-language results indicate that the models generally learned their training distributions, so differences on the target languages are unlikely to reflect outright training failure.

	Language					
	Luganda		Tatar		Upper Sorbian	
Strategy	S	T	S	T	S	T
Family	8.74	29.12	10.47	34.25	6.77	48.11
Geographic	8.03	**27.81**	17.47	37.88	8.34	51.23
Phonetic	6.69	30.62	9.71	**33.55**	11.38	47.98
Random	17.40	50.35	‘17.40	54.45	17.40	46.47
MultIPA	-	56.69	-	60.00	-	**45.88**

Table 3: Seen-language (S) and target-language (T) PER by target language and training-language selection strategy, including MultIPA (Taguchi et al., 2023) as a comparison. Lower values indicate better performance. Bold marks the best target-language result within each target language.

As shown in Table 3, the effect of training-language selection was target-dependent. For Luganda and Tatar, all linguistically-informed strategies outperformed the random baseline; for Upper Sorbian, none did. This may be because the languages in the random model happened to be more closely related to Upper Sorbian than the other target languages; this is further discussed in Appendix B.

For Upper Sorbian, the phonetic model was best among the strategy-based systems, but all three were slightly worse than the random baseline, and MultIPA performed best overall. For Luganda, the geographic model performed best, with the family and phonetic models close behind. All three outperformed both the random baseline and MultIPA. For Tatar, the three linguistically informed strategies clustered closely together, with a slight advantage for the phonetic model. As with Luganda, all three outperformed both the random baseline and MultIPA.

In the mixed-effects Poisson regression, Luganda and Tatar showed a much larger penalty for the random strategy relative to Upper Sorbian. There was a significant partial effect for the random strategy at the levels of Luganda ($\beta = 0.579$, $SE = 0.022$ $p < .001$) and Tatar ($\beta = 0.536$, $SE = 0.024$ $p < .001$), indicating significantly higher predicted error counts for the random model in these two languages relative to the genealogical strategy used on Upper Sorbian. Among the linguistically informed strategies, differences were generally modest, though not absent. Using emmeans for additional post-hoc comparisons with Tukey p-value adjustments (Lenth and Piaskowski, 2026), in Luganda, the geographic strategy significantly outperformed the phonetic (ratio $= 0.911$, $SE = -0.017$, $p < .001$) and random models (ratio $= 0.559$, $SE = 0.010$, $p < .001$), but did not significantly differ from the family-based model (ratio $= 1.039$, $SE = 0.020$, $p = .189$). For Tatar, the geographic and phonetic strategies remained close to the family-based model. Figure 2 visualizes this interaction between target language and strategy.

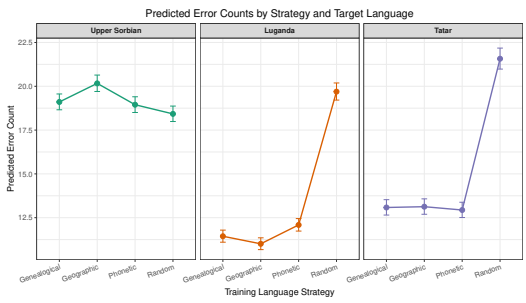

Figure 2: Model-predicted error counts by training-language selection strategy and target language from the Poisson mixed-effects model. Error bars show 95% confidence intervals.

5 Discussion and conclusions

Taken together, the results suggest that linguistically informed training-language selection can improve cross-lingual ASR transfer relative to a random baseline, but the size and form of that benefit are target-dependent. Geographic proximity was most effective for Luganda, phonological overlap was slightly best for Tatar, and none of the pro-

posed strategies improved over random selection for Upper Sorbian. At the same time, geographic and phonological selection avoided the large degradation seen for the random baseline in Luganda and Tatar and remained broadly competitive in Upper Sorbian, making them promising low-cost heuristics for initial training-language selection in documentation-oriented ASR. In these situations, understanding the target language in the context of linguistic similarity may inform the best choice. These findings echo Jimerson et al. (2023)'s observation that design choices in low-resource ASR rarely admit a single universally optimal solution.

One implication is that different dimensions of cross-linguistic similarity may offer different kinds of practical value. Luganda's results suggest that geographic proximity, used here as a proxy for contact, can identify training languages that transfer well even when genealogical relatedness is not the strongest predictor. In practical terms, this may indicate that nearby or contact-linked languages are a reasonable first place to look when assembling a small seed training set for documentation-oriented ASR. For Tatar, the slight advantage for the phonetic condition is more consistent with the view that inventory-level similarity can facilitate transfer when the downstream task is phonetic transcription. This tentatively suggests that inventory-based selection may be particularly useful when the main objective is to generate an initial phonetic transcript for later correction. Upper Sorbian, however, shows that these heuristics do not guarantee improvement in every case: although the strategy-based models remained broadly competitive with one another, neither geographic nor phonological selection outperformed the broader MultIPA baseline. This in turn suggests that for some targets, heuristic selection should be treated as a starting assumption to test rather than as a reliable predictor of the best training configuration. These results therefore motivate broader evaluation of training-language selection strategies in documentation-oriented ASR.

Limitations

Several limitations follow from the design of this study. First, the analysis evaluates only three unseen target languages, which constrains the generalizability of the observed strategy effects. More specifically, the present results are not sufficient to establish a broader typology of low-resource ASR transfer scenarios, since the target set is too small to support strong claims about when particular training-language selection strategies should be expected to work.

The training-language pool is additionally limited to languages with sufficient Common Voice and Vox Communis coverage, excluding many lower-resource languages and making the candidate set partly dependent on existing dataset availability rather than purely linguistic considerations. This means that the space of possible training languages is shaped not only by linguistic relevance, but also by current corpus coverage, which may under-represent languages and language types most relevant to documentation practice.

Further, the study does not attempt a fuller linguistic or sociolinguistic characterization of the target languages beyond the proxy measures used for selection. Practical multilingual model design may depend on additional information, including known contact relationships, community multilingualism, regional sociolinguistic dynamics, and other descriptive knowledge about the language, so the present comparison isolates a small set of transparent heuristics rather than the full range of factors that may shape transfer in documentation-oriented ASR.

More broadly, the study operationalizes linguistic similarity using only three proxies—genealogical relatedness, geographic proximity, and phonological inventory overlap—which capture important but incomplete aspects of cross-linguistic transfer. Each of these measures offers only a relatively shallow view of relatedness in its domain. Geographic distance is only an imperfect proxy for synchronous contact, since spatial proximity does not necessarily imply ongoing interaction, bilingualism, or borrowing, while substantial contact can persist across larger distances through migration, trade, media, or political institutions.

Likewise, phonological inventory overlap does not capture many potentially important dimensions of speech structure, including suprasegmental properties such as tone, stress, or phonation contrasts, nor does it reflect sequential or distributional properties of segments. In future studies, it may prove useful to look at transition probabilities, which express additional information about the phonological system of a language. In addition, none of the three proxies directly captures typological similarities in areas such as morphology or syllable structure, and languages with relatively rare structural properties

may therefore be especially poorly represented by the present approach.

Finally, the phonetic transcripts used for training are machine generated, and model performance is evaluated on a specific IPA-based ASR setup; the results therefore speak most directly to this training and evaluation pipeline rather than to low-resource ASR more broadly.

These limitations suggest that future work should test a wider range of target languages, incorporate richer linguistic and sociolinguistic characterizations of both targets and candidate training languages, and explore more fine-grained similarity measures that better reflect the complex factors shaping transfer in documentation-oriented ASR.

Ethical Considerations

This study uses existing public speech resources from Mozilla Common Voice and Vox Communis and does not involve new data collection or live ASR deployment. The Common Voice datasets used here are released under CC0-1.0 with additional terms prohibiting attempts to identify speakers and prohibiting re-hosting or re-sharing the data; Vox Communis is distributed under the Mozilla Public License 2.0. Accordingly, we treat these materials as licensed research resources and do not attempt to infer speaker identities.

Because this work targets low-resource languages, an additional ethical concern is uneven model performance across languages. Our results show that no single training-language selection strategy is uniformly effective, so we do not treat the proposed heuristics as universally applicable. In documentation settings, automatically generated phonetic transcripts may reduce transcription effort, but they may also bias later annotation if treated as authoritative. We therefore view these systems as assistive tools for linguist-in-the-loop workflows, not replacements for expert or community transcription. More broadly, because our candidate pool is limited to languages with sufficient Common Voice and Vox Communis coverage, this study may reproduce existing resource imbalances; future work should therefore remain attentive to community priorities and responsible use in low-resource settings.

During the preparation of this work, the authors used ChatGPT (OpenAI, 2026) to reformat Taguchi et al. (2023)'s existing code for use without the original automatic transliteration modules and to ensure compatibility with the high performance cluster the code was run on. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Code edited by ChatGPT is clearly marked in the file name and file description.

Acknowledgments

We would like to thank Chihiro Taguchi for coding assistance and helpful advice throughout this project.

References

Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294, Marseille, France. European Language Resources Association.

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *Preprint*, arXiv:1803.08991.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, and Rachel Bawden. 2024. When your cousin has the right connections: Unsupervised bilingual lexicon induction for related data-imbalanced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17544–17556, Torino, Italia. ELRA and ICCL.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden. Association for Computational Linguistics.

Mengzhe Geng, Patrick Littell, Aidan Pine, Penáć, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOŦEN language documentation efforts with

automatic speech recognition. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. glottolog/glottolog: Glottolog database 5.3.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. Advocating character error rate for multilingual asr evaluation. *Preprint*, arXiv:2410.07400.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Russell V. Lenth and Julia Piaskowski. 2026. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 2.0.2.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning asr models for extremely low-resource fieldwork languages. *Preprint*, arXiv:2506.17459.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

P. Lopez Gonzalez-Nieto, M. Gomez Flechoso, M.A. Arribas Mocoroa, A. Muñoz Martin, M.L. Garcia Lorenzo, G. Cabrera Gomez, J.A. Alvarez Gomez, A. Caso Fraile, J.M. Orosco Dagan, R. Merinero Palomares, and R. Lahoz-Beltra. 2020. Design and development of a virtual laboratory in python for the teaching of data analysis and mathemaics in geology: Geopy. In *INTED2020 Proceedings*, 14th International Technology, Education and Development Conference, pages 2236–2242. IATED.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

OpenAI. 2026. ChatGPT (Mar 14 version). Large language model.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal automatic phonetic transcription into the international phonetic alphabet. In *Interspeech 2023*, pages 2548–2552.

Ewald van der Westhuizen, Trideba Padhi, and Thomas Niesler. 2021. Multilingual training set selection for asr in under-resourced malian languages. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings*, page 749–760, Berlin, Heidelberg. Springer-Verlag.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

A Similarity Measure Scores for Training Languages

These tables report the top five candidate training languages identified for each target language under the three selection criteria used in the study: geographic proximity, genealogical relatedness, and phonological inventory overlap. Ultimately, only the top four from each strategy were selected, due tot he lack of a fifth family member for Luganda, which defaulted to the first option alphabetically. These tables are intended to make the training-language selection procedure more transparent by showing the nearest-ranked candidates under each heuristic. As described in the Methods section, geographic candidates were ranked by geodesic distance, while genealogical and phonological candidates were ranked using set-based overlap measures, with ordering determined first by the number of shared items and then by additional overlap-based tie-breakers.

These appendix tables document the candidate space available to each target language and illustrate the different kinds of similarity captured by the three heuristics. This is useful because the paper's main results show that no single notion of relatedness was uniformly best across Upper Sorbian, Luganda, and Tatar. Presenting the ranked candidate lists in full therefore helps clarify both how the final training sets were derived and why different strategies could plausibly lead to different transfer outcomes across targets.

Language	Geodesic Distance (km)
Upper Sorbian	
Czech	159.8
Polish	300.0
Slovak	434.2
Slovene	555.1
Hungarian	615.4
Luganda	
Dholuo	314.8
Kinyarwanda	372.9
Swahili	1158.8
Basaa	2441.6
Hausa	2832.1
Tatar	
Chuvash	127.7
Russian	360.8
Bashkir	538.4
Estonian	1434.3
Ukrainian	1448.7

Table 4: Top five candidate training languages for each target language under the geographic proximity criterion. Distances are geodesic distances in kilometers.

Language	# Nodes	% Overlap	Jaccard %
Upper Sorbian			
Czech	5	83.3	71.4
Slovak	5	83.3	71.4
Polish	5	83.3	62.5
Russian	4	66.7	57.1
Slovenian	4	66.7	50.0
Luganda			
Kinyarwanda	9	81.8	64.3
Swahili	8	72.7	50.0
Basaa	6	54.5	37.5
Yoruba	3	27.3	15.8
Abkhaz	0	0.0	0.0
Tatar			
Bashkir	7	100.0	100.0
Kyrgyz	4	57.1	44.4
Uzbek	3	42.9	33.3
Uyghur	3	42.9	30.0
Sakha	2	28.6	25.0

Table 5: Top five candidate training languages for each target language under the genealogical relatedness criterion. Rankings are based primarily on the number of shared classification nodes, with additional tie-breaking metrics described in the main text.

Language	# Segments	% Overlap	Jaccard %
Upper Sorbian			
Lithuanian	35	85.4	44.3
Romanian	33	80.5	48.5
Bulgarian	33	80.5	41.8
Ukrainian	31	75.6	55.4
Russian	28	68.3	41.2
Luganda			
Dholuo	25	89.3	42.4
Dutch	25	89.3	28.7
Indonesian	24	85.7	61.5
Basaa	24	85.7	53.3
Catalan	24	85.7	39.3
Tatar			
Bashkir	29	69.0	43.9
Hindi	27	64.3	26.0
Uyghur	26	61.9	47.3
Polish	26	61.9	42.6
Northern Kurdish	26	61.9	33.3

Table 6: Top five candidate training languages for each target language under the phonological inventory overlap criterion. Overlap is computed over segment inventories.

B Relationships to Languages in Random Model

The random baseline appears to have been less mismatched to Upper Sorbian than to Luganda or Tatar because, despite being selected without reference to the target languages, its training set still includes languages that are not especially distant from Upper Sorbian under the similarity measures used here. In particular, Ukrainian shows substantial genealogical overlap with Upper Sorbian and relatively high phonological overlap, while Romanian also shows fairly strong phonological similarity. By contrast, the same random set has no genealogical overlap at all with either Luganda or Tatar, and its geographic distances to Luganda are especially large. In other words, the "random" baseline was not equally unrelated across targets: for Upper Sorbian, it accidentally included languages with moderate structural similarity, whereas for Luganda and Tatar it was a much poorer match overall. This interpretation is consistent with the main results, where the random model remained competitive for Upper Sorbian but was much worse for Luganda and Tatar.

Language	Upper Sorbian Distance (km)	Luganda Distance (km)	Tatar Distance (km)
Abkhazian	2207.3	4780.2	1527.5
Romanian	899.0	5124.4	2016.9
Swedish	971.2	6681.2	1918.5
Ukrainian	1111.3	5448.3	1448.7

Table 7: Geographic relationships between each target language and the four languages used in the random baseline model. Values are geodesic distances in kilometers.

	Upper Sorbian			Luganda			Tatar		
Language	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**
Abkhazian	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Romanian	2	33.3	14.3	0	0.0	0.0	0	0.0	0.0
Swedish	2	33.3	16.7	0	0.0	0.0	0	0.0	0.0
Ukrainian	4	66.7	50.0	0	0.0	0.0	0	0.0	0.0

Table 8: Genealogical relationships between each target language and the four languages used in the random baseline model. Columns report the number of shared lineage nodes, the percentage of the target classification path covered, and Jaccard overlap.

	Upper Sorbian			Luganda			Tatar		
Language	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**
Abkhazian	23	56.1	26.4	17	60.7	21.3	23	54.8	26.1
Romanian	33	80.5	48.5	22	78.6	33.3	25	59.5	32.5
Swedish	17	41.5	27.9	18	64.3	38.3	19	45.2	31.7
Ukrainian	31	75.6	55.4	20	71.4	37.0	23	54.8	35.4

Table 9: Phonological relationships between each target language and the four languages used in the random baseline model. Columns report shared segments, percentage of the target inventory covered, and Jaccard overlap.