Quantifying mutual intelligibility gradients in Turkic languages using language models

Moldir Baidildinova, Shiva Upadhye, Austin Wagner, Connor Mayer, Richard Futrell
{mbaidild,upadhyes,wagnera3,cjmayer,rfutrell}@uci.edu
Department of Language Science
University of California, Irvine

1 Introduction

Mutual intelligibility (MI) refers to the ability of speakers of one language to understand another language, often due to genealogical relatedness that results in lexical and grammatical correspondences. MI is a gradient phenomenon because it exists on a continuum from near-total comprehension to no understanding, often showing asymmetry between related language pairs (Chambers and Trudgill, 1998; Tang and van Heuven, 2007; van Bezooijen and Gooskens, 2005). A prime example is the Turkic language family, which is often described as highly similar from a historical-typological perspective, with substantial overlap across multiple linguistic dimensions, which is commonly assumed to support MI within the family (Kornfilt, 2018; Tambovtsev, 2001; Baskakov, 1988; Lindsay, 2010). These languages share substantial lexical similarity, characterized by a high percentage of shared cognates, as well as grammatical structures, word order and phonetic-phonological overlap, as shown in Table 1. Also, languages within the same branch tend to show higher levels of MI, partly because of their close proximity within the genealogical tree, or what has been discussed in terms of cophenetic distance (Gooskens et al., 2018). In other words, the closer two languages are on the tree, the higher their intelligibility rates tend to be. For instance, Turkish and Azerbaijani are generally more mutually intelligible (Salehi and Neysani, 2017) than Turkish and Kazakh or Turkish and Uyghur because they belong to the same Oghuz branch; similar patterns can be observed across other language pairs as well.[1] Taken together, these patterns suggest that MI is best understood as a multidimensional phenomenon, shaped by the combined effects of lexical, grammatical, and phonetic-phonological similarity.

Language	Branch	IPA transcription
Turkish	Oghuz	(ben biɾ) kitap okujoɾum.
Azerbaijani		(mæn) kitɑb oxujuɾɑm.
Kazakh	Kipchak	(men) kɪtɑp oqəp ʒɑtərmən.
Kyrgyz		(men) kitep okup ʒɑtɑm.
Uzbek	Karluk	(men) kitɔb oqijapman.
Uyghur		(mæn) kitɑb oquwɑtimæn.

Table 1: MI among Turkic languages illustrated by the sentence: *I am reading a book.*

However, existing computational approaches typically quantify MI using lexical or phonetic distances computed over hand-curated cognate sets or high-frequency word lists, and most of such work has focused on Indo-European languages (Gooskens and van Heuven, 2021; Nieder and List, 2024). Here, we propose a neural language modeling-based approach to investigating MI within the Turkic language family. Rather than relying on hand-curated features, our approach learns these regularities directly from naturalistic text. We then ask to what extent the resulting similarity estimates track MI among Turkic languages.

2 Current study

We hypothesize that greater overlap across multiple linguistic dimensions facilitates cross-lingual transfer and contributes to higher MI between languages within the same Turkic subgroup. Under this view, a model trained on one Turkic language should generalize more successfully to closely related target languages than to distantly related ones, thereby reflecting the gradient nature of MI. To test this hypothesis, we fine-tune these pre-trained models on target languages that vary in their degree of

[1] To the best of our knowledge, systematic human data on MI patterns within the Turkic family remain limited. Existing observations are often anecdotal, based on reports from Turkologists, native speakers or L2 learners who note, for example, that knowledge of Uyghur may facilitate understanding of Uzbek more than Kazakh.

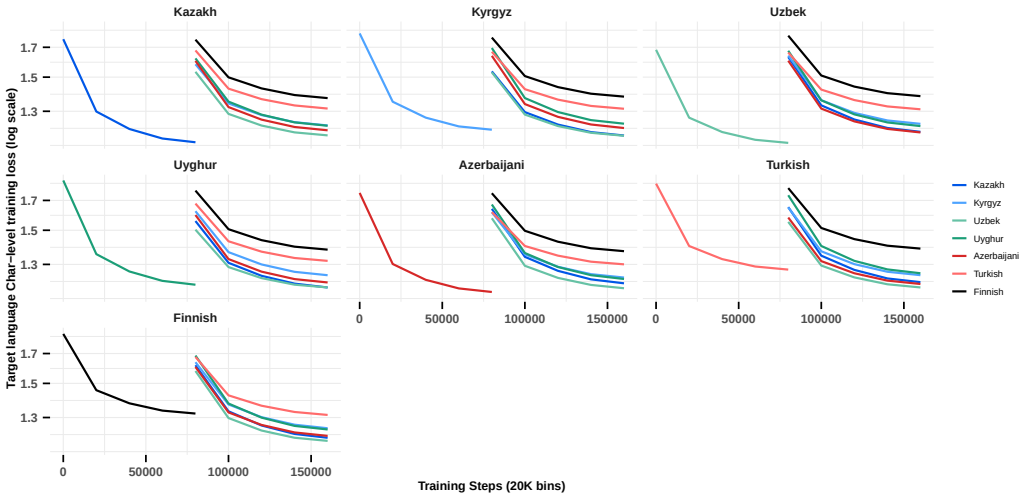

Figure 1: Character-level CE loss across training steps; models trained on source languages, then fine-tuned on target languages.

genealogical relatedness. We operationalize generalizability using character-level *cross-entropy (CE) loss*, which quantifies how surprised the model is, on average, by the phoneme sequences it encounters. Lower training and test CE loss are therefore expected to reflect stronger cross-lingual transfer, which we interpret as evidence of greater linguistic overlap between closely related language pairs. This provides a more fine-grained measure of cross-lingual transfer than curated cognate lists, which are limited to lexical overlap. In addition, we expect closer language pairs to show better overall learning dynamics across training, measured by *the Area Under the Curve (AUC)*, where smaller AUC values indicate better learning. We also expect more distant language pairs to show higher *rates of change (ROC)* during the early stages of training, due to larger gaps in lexicon and sound patterns. Taken together, our findings provide preliminary evidence that CE loss, AUC and ROC can approximate MI patterns across six Turkic languages.

3 Method

3.1 Training Datasets

We extracted raw text for closely related Turkish-Azerbaijani, Kazakh-Kyrgyz, and Uzbek-Uyghur language pairs from the OSCAR[2] corpus, filtered with FastText language identification (≥ 0.95 confidence) to reduce contamination (Abadji et al., 2022; Joulin et al., 2016). To approximate pronunciation across Turkic languages and avoid script inconsistency issues within the language family (Mirzakhalov et al., 2021), we transcribed all texts into broad IPA using established language-specific rule sets[3] (Zimmer and Orgun, 1992; Mokari and Werner, 2017; McCollum and Chen, 2021; McCollum, 2020; Ido, 2025; Mayer, 2021; McCollum, 2021). For each language, the transliteration output was verified to match exactly the IPA rules established in prior phonological analyses.

3.2 LSTM model

To mitigate tokenizer design issues in highly agglutinative, low-resource languages (Toraman et al., 2023), we trained a 2-layer character-level LSTM with 128-dimensional embeddings and 256 hidden units (Hochreiter and Schmidhuber, 1997). The model training followed a sequential bilingual training regimen (Arnett et al., 2025): a model was first trained and validated on a source language, then fine-tuned on a target language (Arnett et al., 2025) for each of the six languages. We also include Finnish as a distractor language, since it is typologically similar (agglutinative and vowel-harmonic) but genealogically unrelated (Suomi et al., 2009) to Turkic languages. In addition, we also manipulated exposure order (e.g., Kazakh-Kyrgyz, Kyrgyz-Kazakh), which yielded a total of 42 models. All models were exposed to 1 million tokens per language, split 75-25 into training and validation sets.

3.3 What is driving model convergence?

To shed light on which factors track the patterns observed in CE loss, AUC, and ROC, we examined four predictors of cross-lingual similarity: *cophenetic distance, lexical similarity, weighted*

[2]Open Super-large Crawled Aggregated coRpus

[3]The IPA rules are available on GitHub. The transliteration suite is available on Hugging Face.

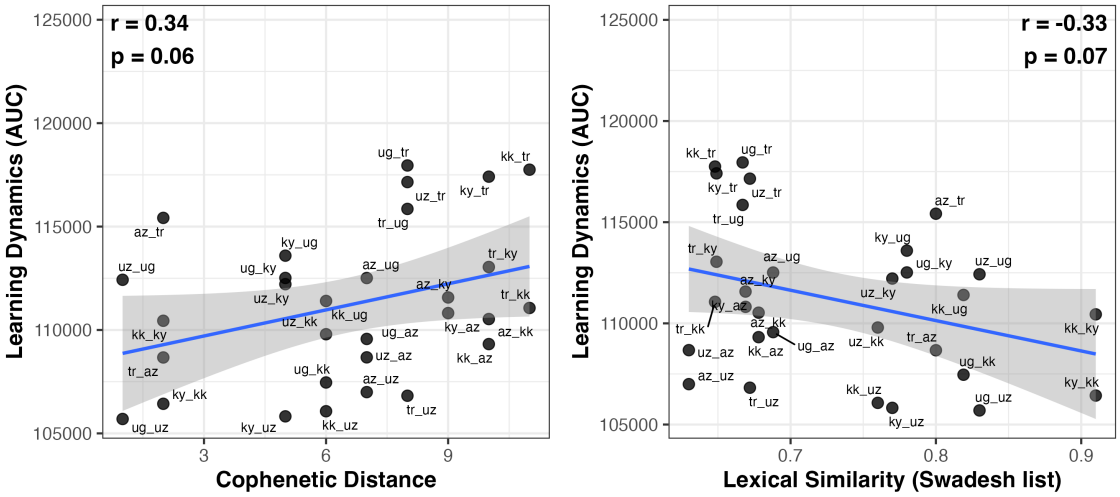

Figure 2: The AUC plotted against the cophenetic distance and lexical similarity.

trigram frequency overlap, and *vowel harmony index*. Cophenetic distance was computed following Gooskens et al. (2018) as the number of nodes separating two languages in the genealogical tree via their lowest shared ancestor node; the Turkic language tree was based on the reconstruction by Savelyev and Robbeets (2020). Lexical similarity was derived from the Swadesh-based estimates in Lindsay (2010). Weighted trigram frequency overlap was computed as a proxy for phonotactic similarity by extracting IPA-based character trigrams from size-normalized corpora, retaining frequent trigrams, and comparing their frequency counts across language pairs using weighted Jaccard similarity (Niwattanakul et al., 2013). We also computed a vowel harmony index for each training dataset following Harrison et al. (2004) and derived difference in the vowel harmony index between source and target languages. Finally, we correlated CE loss, AUC, and ROC with these predictors to evaluate which dimensions of linguistic similarity best account for model convergence across Turkic language pairs.

4 Results

Across experiments, CE loss broadly aligned with MI gradients: closer Turkic pairs showed smaller loss spikes, faster convergence, and lower training loss. Transfer was better from Kyrgyz to Kazakh than to Turkish, from Turkish to Azerbaijani than to Kazakh, and from Uyghur to Uzbek than to Kazakh, as shown in Figure 1. Although these differences were modest, they preserved the expected MI ordering, suggesting that CE loss is sensitive to different degree of overlap across Turkic pairs, with lower CE loss observed for closely related pairs within the same branch, such as Kazakh–Kyrgyz within the Kipchak branch, Turkish–Azerbaijani within the Oghuz branch, and Uzbek–Uyghur within the Karluk branch. Most importantly, the models appear to exhibit sensitivity to directional asymmetries between language pairs. For instance, Kazakh seems to converge to Kyrgyz more slowly than Kyrgyz converges to Kazakh, with similar patterns observed for Azerbaijani $<$ Turkish and Uzbek $<$ Uyghur. In addition, Turkish appears to be the least advantageous target language for most source models, whereas Uzbek shows the opposite pattern, serving as the most advantageous target language across all five source models. Finnish yielded the highest test loss in all transfer directions, consistent with its genealogical distance from the Turkic languages. We also correlated CE loss with cophenetic distance and lexical similarity. The correlations were positive but non-significant, tentatively suggesting that more cophonetically distant pairs may incur higher loss. Differences in the vowel harmony index and trigram frequency overlap did not yield systematic patterns.

As predicted, the AUC pattern was qualitatively lower for closer language pairs and for pairs with greater lexical overlap, as shown in Figure 2. Asymmetric patterns between language pairs are also evident in this plot, and these patterns are the same as those discussed above: Kyrgyz $<$ Kazakh, Azerbaijani $<$ Turkish, and Uzbek $<$ Uyghur. Dif-

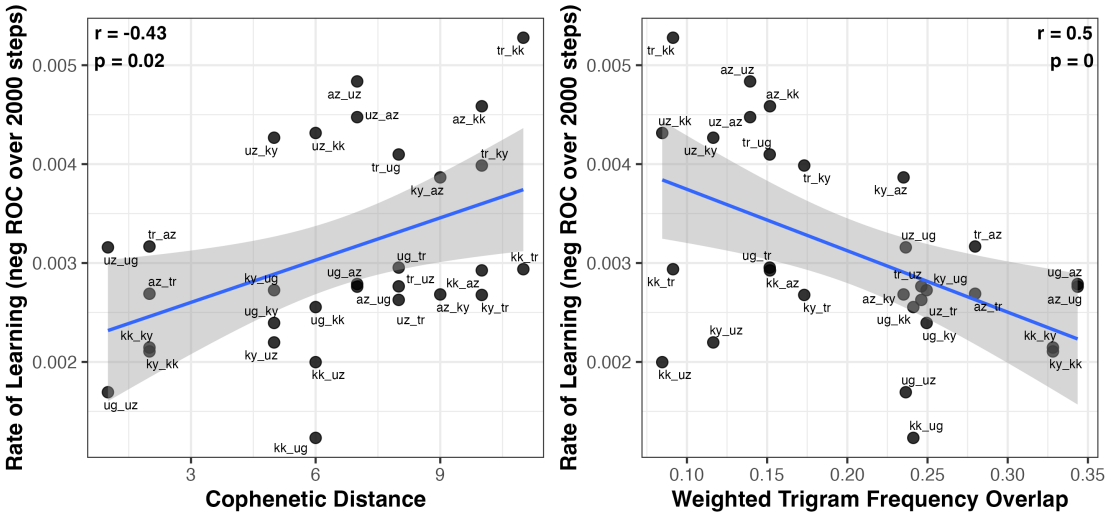

Figure 3: The ROC over 2000 steps of training plotted against cophenetic distance between language pairs and the weighted trigram frequency overlap.

ferences in the vowel harmony index did not show a consistent relationship with AUC across language pairs; a similar pattern was observed for trigram frequency overlap as well. In contrast, greater language distance appears to predict greater learning change: more cophenetically distant pairs show higher ROC, whereas pairs with greater trigram overlap show lower ROC, as shown in Figure 3. Unlike CE loss and AUC, ROC captures early-stage adaptation during training, suggesting that the rate of change may provide an additional indicator of model performance.

5 Discussion

Our results suggest that character-level LSTM models trained on naturalistic text can broadly capture MI patterns within the Turkic family. Across experiments, more closely related pairs generally showed lower CE loss and smaller AUC, consistent with greater cross-lingual generalizability. This suggests that even a relatively simple and scalable modeling approach can track MI gradients from phoneme-level distributional patterns in naturalistic text. The results also suggest that multiple dimensions of linguistic similarity may contribute to model-based MI patterns. We tested several possible predictors of model convergence. Overall, lexical similarity, trigram overlap, and cophenetic distance appeared to be the most informative predictors: models showed better transfer when languages shared more lexical and local phonotactic structure and when they were genealogically closer. In contrast, differences in vowel harmony index did not show a systematic relationship with the model metrics, suggesting that variation in the degree of vowel harmony does not have a strong effect on model performance in this setup.

The Finnish results further support the importance of genealogical relatedness. Although Finnish shares some typological properties with Turkic languages, such as agglutinative morphology and vowel harmony, it consistently produced the highest test loss across transfer directions. The models also appear to be sensitive to directional asymmetries between language pairs, suggesting that transfer is not simply determined by overall pairwise similarity. Instead, some source languages provide more advantageous transfer than others, even within related pairs. Of the measures we used to assess learning, ROC showed the clearest systematic pattern, with more distant pairs showing higher rates of change during the early stages of training, while pairs with greater trigram overlap showed lower ROC. This suggests that ROC may capture early-stage adaptation when the model encounters less familiar phoneme-sequence patterns.

Taken together, these findings suggest that character-level CE loss, AUC, and ROC can approximate MI patterns across Turkic languages, especially when similarity is grounded in lexical overlap, local phonotactic overlap, and historical relatedness. Ongoing collection of native-speaker MI judgments will allow us to test whether human speakers perceive these similarity and asymmetry patterns in the same way as the models.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France. European Language Resources Association.

Catherine Arnett, Tyler A Chang, James A Michaelov, and Benjamin K Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. *arXiv preprint arXiv:2503.03962*.

N. A. Baskakov. 1988. *Istoriko-tipologicheskaya fonologiya tyurkskikh yazykov*. Nauka, Moscow.

J.K. Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Charlotte Gooskens and Vincent J. van Heuven. 2021. Mutual intelligibility. In Marcos Zampieri and Preslav Nakov, editors, *Similar Languages, Varieties, and Dialects: A Computational Perspective*, Studies in Natural Language Processing, pages 51–95. Cambridge University Press, Cambridge.

Charlotte Gooskens, Vincent J Van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in Europe. *International Journal of Multilingualism*, 15(2):169–193.

David Harrison, Emily Thomforde, and Michael O'Keefe. 2004. The vowel harmony calculator. *Online: http://www. swarthmore. edu/SocSci/harmony/public_html*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Shinji Ido. 2025. Uzbek. *Journal of the International Phonetic Association*, 55(1-2):152–168.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jaklin Kornfilt. 2018. Turkish and the Turkic languages. In *The world's major languages*, pages 536–561. Routledge.

Robert Lindsay. 2010. Mutual intelligibility among the Turkic languages. *Beyond Highbrow*.

Connor Mayer. 2021. *Issues in Uyghur backness harmony: Corpus, experimental, and computational studies*. University of California, Los Angeles.

Adam G McCollum. 2020. Vowel harmony and positional variation in Kyrgyz. *Laboratory Phonology*, 11(1).

Adam G McCollum. 2021. Transparency, locality, and contrast in Uyghur backness harmony. *Laboratory Phonology*, 12(1).

Adam G McCollum and Si Chen. 2021. Kazakh. *Journal of the International Phonetic Association*, 51(2):276–298.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, and 1 others. 2021. A large-scale study of machine translation in the Turkic languages. *arXiv preprint arXiv:2109.04593*.

Payam Ghaffarvand Mokari and Stefan Werner. 2017. Azerbaijani. *Journal of the International Phonetic Association*, 47(2):207–212.

Jessica Nieder and Johann-Mattis List. 2024. A computational model for the assessment of mutual intelligibility among closely related languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 37–43.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Mohammad Salehi and Aydin Neysani. 2017. Receptive intelligibility of Turkish to Iranian-Azerbaijani speakers. *Cogent Education*, 4(1):1326653.

Alexander Savelyev and Martine Robbeets. 2020. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53.

Kari Suomi, Juhani Toivanen, and Riikka Ylitalo. 2009. *Finnish sound structure: Phonetics, phonology, phonotactics and prosody*. University of Oulu.

Yuri Tambovtsev. 2001. The phonological distances between Turkic languages based on some phonological features of consonants. *Asian and African Studies*, 10(1):11–43.

Chia-Jung Tang and Vincent J. van Heuven. 2007. The role of typological and lexical similarity in the perception of Chinese dialects. *Journal of Chinese Linguistics*, 35(2):309–326.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Renée van Bezooijen and Charlotte Gooskens. 2005. How easy is it for speakers of Dutch to understand Frisian and Afrikaans, and why? *Lingua*, 115(10):1479–1500.

Karl Zimmer and Orhan Orgun. 1992. Turkish. *Journal of the International Phonetic Association*, 22(1-2):43–45.