Field Matters

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics

March 29, 2026

ISBN 979-8-89176-376-0

Preface

Field Matters is a workshop focused on the various applications of NLP methods to field linguistics and the analysis of field data. The primary pursuit of linguistic fieldwork is to document and describe languages. The former typically involves building a corpus and other resources for the language community, the latter ideally aims to produce a reference grammar. Advances in technology have enabled vast quantities of media to be recorded. These recordings (sound and/or video) require annotation and analysis for further linguistic research or resource development. This is often done manually. This processing bottleneck can be significantly sped up with computational methods. NLP research focuses on developing methodology for different tasks that show significant performance in high-resource languages, allowing the automation of various routine tasks. The processing burdens faced by field linguists present a natural opportunity to marry NLP practices with the workflow of a field linguist. Similarly, the future development of NLP methods could gain from the linguistic diversity and unique tasks encountered during the description/documentation efforts.

With these in mind, *Field Matters* aims to provide a platform to deepen the dialogue between Computational and Field Linguists. Our workshop is hosted by the 19th Conference of the European Chapter of the Association for Computational Linguistics in Rabat, Morocco.

Field Matters 2026 continued to provide field linguists expert reviews, a distinct feature of the review process introduced two years ago. Each paper was assigned a field linguist alongside minimally two computational linguists. Analyzing the difference in reviews of field linguists and NLP researchers, we have seen that reviewers provide different perspectives and give more diverse and fruitful feedback: while field linguists pay attention to how practical this application could be or how well it fits in the idea of the workshop, NLP specialists comment on how relevant and accurate chosen methods are.

After the hard process of reviewing all submissions, the program committee chose seven papers for a poster or oral presentation at the workshop. Accepted papers illustrate the main idea of our workshop: how field linguistics may benefit from using contemporary methods of computational analysis and how the NLP community may evolve its methods with the help of under-resourced languages. More specifically, chosen papers cover the following topics:

- Tools for fieldwork, including a language documentation tool and guidelines for human-computer interaction in the field of sociolinguistics;
- Creation of various corpora (both spoken and written);
- Speech and text processing tools for under-resourced languages and dialect variants;
- Phonology study with machine learning tools.

We are incredibly grateful to the *Field Matters* program committee, who worked on peer review to give meaningful comments for each submission and made this workshop possible. We want to thank the invited speaker, Alexis Palmer, Associate Professor at the University of Colorado Boulder. We would also like to acknowledge all the authors who submitted their papers to our workshop, and we hope to continue to serve as a link between NLP specialists and field linguists.

Organizing Committee

Éric Le Ferrand (Boston College)

Elena Klyachko

Shu Okabe (Technische Universität München)

Ekaterina Voloshina (Chalmers University of Technology, University of Gothenburg)

Oleg Serikov (Palisade Research)

Tatiana Shavrina (Meta)

Ekaterina Vylomova (University of Melbourne)

Table of Contents

Automated Quality Control for Language Documentation: Detecting Phonotactic Inconsistencies in a Kokborok Wordlist
Kellen Parker van Dam and Abishek Stephen 1

What NLP Gets Wrong About Contact: Implications for Field Linguistic Evidence
Manodyna K H 8

Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan
Siyu Liang, Talant Mawkanuli and Gina-Anne Levow 16

Linguistically Informed Tokenization Improves ASR for Underresourced Languages
Massimo Marie Daul, Alessio Tosolini and Claire Bowern 31

Short-form verbal arts as a speech data resource in the field
Matthew Faytak, Tianle Yang, Pius Wuchu Akumbu, Ivo Forghema Njuasi and Éric Le Ferrand 38

Quantitative Lect Description: A Case Study of Lemko from the Field Data of 1920s-1930s
Ilia Afanasev 46

The order of subject, object and verb in Tatyshly Udmurt
Daria Belova and Irina Khomchenkova 60

Conference Program

Automated Quality Control for Language Documentation: Detecting Phonotactic Inconsistencies in a Kokborok Wordlist
Kellen Parker van Dam and Abishek Stephen

What NLP Gets Wrong About Contact: Implications for Field Linguistic Evidence
Manodyna K H

Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan
Siyu Liang, Talant Mawkanuli and Gina-Anne Levow

Linguistically Informed Tokenization Improves ASR for Underresourced Languages
Massimo Marie Daul, Alessio Tosolini and Claire Bowern

Short-form verbal arts as a speech data resource in the field
Matthew Faytak, Tianle Yang, Pius Wuchu Akumbu, Ivo Forghema Njuasi and Éric Le Ferrand

Quantitative Lect Description: A Case Study of Lemko from the Field Data of 1920s-1930s
Ilia Afanasev

The order of subject, object and verb in Tatyshly Udmurt
Daria Belova and Irina Khomchenkova

Automated Quality Control for Language Documentation: Detecting Phonotactic Inconsistencies in a Kokborok Wordlist

Kellen Parker van Dam[1] **and Abishek Stephen**[2]
[1]Chair for Multilingual Computational Linguistics, University of Passau, Germany
[2]Institute of Formal and Applied Linguistics, Charles University, Czech Republic

Abstract

Lexical data collection in language documentation often contains transcription errors and borrowings that can mislead linguistic analysis. We present unsupervised methods to identify phonotactic inconsistencies in wordlists, applying them to a multilingual dataset of Kokborok varieties with Bangla. Using phoneme-level and syllable-level n-gram language models, our approach identifies potential transcription errors and borrowings. We evaluate our methods using hand annotated gold standard and rank the phonotactic outliers using precision and recall at K metric. The ranking approach provides field linguists with a method to flag entries requiring verification, supporting data quality improvement in low-resourced language documentation.

1 Introduction

In linguistic fieldwork, description frequently begins with the collection of lexical data (Chelliah, 2014). This is often done by means of concept lists such as those of Swadesh (Swadesh, 1955). In the early stages of research, initial elicitation sessions commonly produce "messy" data. When lexical items are collected as a preliminary step, the fieldworker may not be fully acquainted with the phonological system of the target language. Which sounds are phonemic as opposed to allophonic variation may be uncertain. As a result, words are often transcribed narrowly which may be inconsistent from entry to entry regarding the underlying phonemes, as well as potentially failing to capture the underlying phonemic contrasts. A lack of systematicity in the transcription can in turn create issues later on in data analysis (Himmelmann, 1998).

For proper documentation it is important to incorporate the speech of multiple participants. However, when data is drawn from multiple speakers in this manner, differences in dialect or accent may be reflected in the forms recorded. Variation between careful and casual speech styles can introduce further irregularities (Chelliah, 2014).

Lexical borrowing constitutes another potential complication. Borrowed terms may have varying degrees of adherence to the underlying phonemic system. Terms may also have been borrowed twice, with an intermediate borrowing of another closely related language having a different set of phonotactic constraints, thus obscuring their borrowed nature. Thus, it is important that borrowings can be readily identified when attempting to understand the phonology of a language. Borrowed forms may enter the dataset without the researcher's awareness, particularly when the donor language is unfamiliar to the fieldworker.

The context of elicitation is also important. Differences in approaches can exert a significant influence on the quality and consistency of the data. The degree of formality in the interaction, the presence of other speakers, and the level of fatigue or attention on the part of the consultant can all affect the data. The fieldworker's own background and expectations also shape the data in subtle but consequential ways (Kelly and Lahaussois, 2021).

For these reasons, having a method for detection of phonological outliers is of great value to the documentary linguist. By identifying potential borrowings or inconsistencies introduced by factors, the end result of any descriptive study is immediately aided in the very first steps of lexical data collection. Having automated flags for "this entry looks phonotactically weird" could save field linguists considerable time, especially when working with under-resourced languages where you can't rely on external data verification. We do this detection using n-gram language modeling based on phoneme and syllable-level analysis.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 1–7
March 26, 2026.

2 Related Work

The current research deals with identifying the phonotactic inconsistencies in a linguistic wordlist which in a different light can be seen to have concordances with spelling checkers or borrowing detection methods. Our work however, is not aimed towards either of them albeit the overt similarities. Worth mentioning are some attempts of borrowing detection using wordlists. Miller et al. (2021) where automatic methods for detecting lexical borrowings from monolingual wordlists, comparing different neural network based architectures. List (2019) presents approaches for detecting language contact and borrowing, focusing on phylogenetic networks, sequence comparison methods for detecting borrowings in multilingual wordlists, and trait-based approaches that distinguish borrowed from inherited features using borrowability arguments.

3 Source Data

We rely on data for Kokborok (Glottocode: tipp1238, Hammarström et al., 2025), an underdescribed Tibeto-Burman language group under the Barish language branch (Delancey, forthcoming).

The consonant inventory is relatively moderate in size, with a notable series of aspirated stops that likely developed through Indo-Aryan influence, as aspiration contrasts are less common in many Tibeto-Burman languages. The language maintains voicing distinctions across bilabial, dental, velar, and palatal places of articulation. Word-finally, however, obstruents typically devoice, a pattern not found in neighboring Bangla. Notably, voiced affricates like /dʒ/ are not native to Kokborok but appear in Bangla loanwords, representing sounds borrowed along with vocabulary. Kokborok strongly prefers open syllables and avoids consonant clusters, reflecting its Tibeto-Burman phonotactic constraints. This contrasts sharply with Bangla, which permits complex consonant clusters both word-initially and word-finally. Where Bangla allows syllables like /bdʒro/ or final clusters like /-sto/, Kokborok maintains simpler CV(C) structures with very limited coda positions.

These phonotactic differences create a clear phonological boundary between Kokborok and Bangla despite their geographic proximity. The result is two typologically distinct systems coexisting in close contact, with Kokborok maintaining its characteristic Tibeto-Burman simplicity in syllable structure even while absorbing lexical material from its Indo-Aryan neighbor.

Our data comes from Kim et al., 2025, a sociolinguistic survey of 306 concepts in 20 Kokborok varieties plus 3 varieties of Garo (garo1247), and standard Bangla as the main contact language. This concept list is a good representation of the language as it covers the majority of basic morphemes which go into lexical construction. As is typical in Tibeto-Burman languages of the region, words are primarily compounds of simpler common morphemes. This set of 306 concepts is considerably larger than the amount that would normally go into computational phylogenetic work, such as the 180 concepts of Sagart et al. (2019) or the 100 of Galucio et al. (2015), and coverts the full range of phonological variation occurring in native forms.

Kokborok data were converted to the Cross-Linguistic Data Format (CLDF; Forkel et al. (2018)) by the authors. Morphological features external to the citation form were removed, as were erroneous repeated diacritics which did not contribute to the transcription. The data were otherwise not modified, leaving ambiguous transcriptions[1] as is. The CLDF dataset is available in a GitHub repository[2]

4 Experiments

The identification of the phonological anomalies or outliers in our dataset proceeds via implementing simple n-gram language models at the phoneme and syllable levels. First, we run the experiments on the phoneme level which aims to capture rare phoneme sequences and then we contrast it with positional phonotactics using the syllable-level analysis that captures more linguistically motivated violations. The source codes are publicly available[3].

The training data has 3055 words after removing duplicates[4]. The gold data contains 555 words marked as borrowings. Our annotations focus exclusively on borrowings, as these were straightforward to identify given the clear phonological and lexical distinctions between Kokborok and Bangla, the primary source of loanwords in the dataset. We treat words in Kokborok that are almost identical to the

[1] For example in the AbengGajni doculect, the verb 'to eat' is erroneously transcribed as tshaʔ with no clear indication if this should be tʃaʔ or tsʰaʔ.

[2] https://github.com/phonemica/kimkokborok. The dataset can also be accessed here-https://doi.org/10.5281/zenodo.17973867.

[3] https://github.com/abishekjs/kokborok-anotect

[4] Since the linguistic varieties or doculects are closely related, the same word forms are used to encode a given semantic concept.

Bangla counterpart phonologically (as explained in § 3) for a given semantic concept as borrowed. For example, the word for 'rainbow' in Bangla and doculect MukchakBarbakpur is ɾɔŋdʰɔnu and hence marked as a borrowing. Transcription errors were not systematically annotated due to the difficulty of distinguishing genuine errors from dialectal variation or unknown phonological processes, making borrowings more reliable for evaluating our methods.

4.1 Phoneme-level N-gram Language Modeling

To identify phonotactic anomalies such as transcription errors and borrowings, we train phoneme bigram and trigram language models on the Kokborok data. Words are padded with boundary markers (e.g., ^noukʰa$), and diacritics are treated as separate characters. We apply Laplace smoothing to handle unseen n-grams.

Our mathematical assumption is that the words with transcription errors or borrowings would have some phoneme sequences which are rare in the language, and using the negative log likelihood (NLL) such words would be flagged when ranking by resulting NLL scores. This can help linguists make quick quality checks, as the stronger outliers would be captured in the top K words. We compute NLL for words using different aggregation methods to capture character-level variations.

Arithmetic Mean captures the expected information content per n-gram. It normalizes for word length enabling fair comparisons between long and short words.

$$\text{Mean NLL} = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p(x_i) \qquad (1)$$

where N is the number of n-grams in the word and x_i represents the i-th n-gram.

Harmonic Mean emphasizes the most typical n-grams within a word, being heavily weighted toward smaller NLL values. This metric is particularly useful for identifying words that contain a core of native phonotactic patterns even when some unusual or rare n-grams are present.

Min identifies the most typical n-gram in a word, revealing whether the word shares any common phonotactic patterns with the native vocabulary. Even heavily borrowed words may contain some typical n-grams, and this metric helps assess the degree of phonotactic integration of loanwords into the native system. A very low minimum NLL suggests the word has at least partial structural overlap with native phonotactics.

Max identifies the most atypical n-gram in a word, a rare n-gram can be a reminiscent of the source language in case of the word being borrowed. It could also potentially flag off partially integrated loanwords.

4.2 Syllable-level N-gram Language Modeling

We implement automatic syllabification based on the sonority hierarchy and maximum onset principle, where syllable boundaries are determined by identifying sonority peaks and applying language-universal syllable structure constraints. The syllable boundary symbol (.) serves as a structural marker. Here too, we add boundary markers and use the aggregation methods used in the phoneme level analysis.

4.2.1 Analysis Types

We employ three distinct approaches to analyze phonotactic patterns:

- **Within-syllable analysis**: We calculate negative log likelihood of character n-grams that occur strictly within individual syllables, respecting syllable boundaries and focusing on internal syllabic structure.
- **Cross-boundary analysis**: We extract character n-grams that span across syllable boundaries, capturing phonotactic patterns that violate typical syllable constraints and may indicate borrowing or transcription anomalies.
- **Boundary-as-phoneme analysis**: We treat syllable boundaries as legitimate phonemes in the sequence, allowing n-grams to include the syllable boundary symbol and capturing positional sensitivity at syllable edges.

4.3 Results

For the field linguists, it would be highly efficient to discover anomalies based on the ranking of the words following their NLL scores observed for all of the aggregation methods. To facilitate that we use recall and precision at K as our evaluation metric. We use the mean NLL as the baseline for the experiments. The gold data is hand annotated, the words

Table 1: Precision and Recall at K for different NLL aggregation methods and baselines for the phoneme-level n-gram models.

N-gram	Method	P@100	P@500	P@1000	R@100	R@500	R@1000
Bigram	Arithmetic Mean	0.43	0.32	0.26	0.08	0.29	0.47
	Harmonic Mean	0.43	0.31	0.26	0.08	0.28	0.47
	Min NLL	0.36	0.32	0.23	0.06	0.29	0.42
	Max NLL	0.32	0.26	0.23	0.06	0.24	0.42
	Uniform Random	0.15	0.19	0.17	0.03	0.17	0.31
	Stratified Random	0.17	0.20	0.17	0.03	0.18	0.30
Trigram	Arithmetic Mean	0.45	0.33	0.28	0.08	0.30	0.51
	Harmonic Mean	0.46	0.32	0.29	0.08	0.28	0.52
	Min NLL	0.40	0.32	0.27	0.07	0.29	0.49
	Max NLL	0.20	0.29	0.25	0.04	0.26	0.46
	Uniform Random	0.20	0.17	0.19	0.04	0.15	0.34
	Stratified Random	0.32	0.21	0.19	0.06	0.19	0.34

borrowed from Bangla are labeled as borrowings. The current dataset do not have any transcription errors, but the assumption and also the strong caveat of our method also ensures the flagging of such errors.

We establish two random sampling baselines to evaluate whether our n-gram phonotactic models perform better than chance. The uniform random baseline samples K words randomly from the wordlist without any prior assumptions, supporting the hypothesis where all words are equally likely to be anomalies. The stratified random baseline samples words proportionally by length, controlling for the possibility that transcription errors or borrowings may be biased toward longer or shorter words.

4.4 Phoneme-level Results

Table 1 demonstrates that our phoneme-level n-gram phonotactic models substantially outperform random baselines in identifying phonotactic anomalies. Trigram models achieve the strongest performance, with precision at 100 reaching 0.46 and recall at 1000 reaching 0.52 using harmonic mean aggregation. The superiority of trigrams over bigrams suggests that richer phonotactic context is crucial for capturing constraints violations, while the consistent performance of arithmetic and harmonic mean aggregation indicates that anomalies are characterized by sustained phonotactic unusualness across the entire word rather than isolated rare n-grams. At K=500, our best model achieves 33% precision, meaning that approximately one in three flagged items is a genuine anomaly. However, the plateau at 52% recall suggests that roughly half of the gold anomalies are phonotactically well-formed, indicating they may represent semantic borrowings or transcription errors that do not violate native phonological constraints.

The best performing model based on trigram-harmonic mean identifies words like ʤʰaɽu, oʃud, ʈɪkʈɪki, ɔnɛk, mɛgʰgɔɹʤon and so on in the top 100 words being flagged as anomalous. In the top 500 words like poɾibar, murgi, gɔrom get flagged. This ranking pattern reflects the model's sensitivity to different degrees of phonotactic deviations, top 100 of the flagged words typically contain phonemes or phoneme combinations that are extremely rare or absent in native Kokborok vocabulary (such as retroflex consonants and aspirated affricates), while words ranked in the top 500 exhibit more subtle violations involving less frequent but attested phoneme sequences, suggesting partial phonological adaptation common for borrowings.

4.5 Syllable-level Results

Our results (Table 3, see Appendix) reveal distinct performance patterns across three phonotactic modeling approaches. Within-word analysis achieves the strongest overall performance with precision at 100 of 0.47 for bigrams and recall at 1000 of 0.49 for trigrams, effectively capturing internal phonological structure violations. Boundary-as-phoneme analysis shows competitive results, particularly at lower K values where trigram models reach precision at 100 of 0.49, indicating that syllable boundary constraint violations are highly predictive of anomalies. In contrast, cross-boundary analysis substantially underperforms, with precision rarely exceeding 0.30 and recall at 1000 capped at 0.42, suggesting that phonotactic violations are better characterized by position-specific patterns

Table 2: Precision and Recall at K for bigram models with arithmetic mean aggregation across different syllable analysis types.

Analysis	P@100	P@500	P@1000	R@100	R@500	R@1000
Within	0.47	0.32	0.26	0.08	0.29	0.47
Cross	0.21	0.20	0.18	0.04	0.18	0.33
Boundary	0.50	0.29	0.24	0.09	0.26	0.43
Uniform Random	0.15	0.19	0.17	0.03	0.17	0.31
Stratified Random	0.17	0.20	0.17	0.03	0.18	0.30

within syllable constituents rather than by boundary-crossing transitions alone.

Words like mɛgʰ and morɪʧ are caught early on at top 100 using the boundary-as-phoneme bigram arithmetic mean setup (Table 2). These words were flagged off in the top 500 of the phoneme trigram harmonic mean setup. This demonstrates the complementary strengths of syllable-aware modeling. The boundary marker (.) itself being present in some sequence is highly influential on results in that it closely reflects syllable position i.e. the phonemes preceding (.) indicate syllable onset or nucleus, following (.) indicates coda position and so on.

5 Conclusion

Our study addresses transcription errors and unidentified borrowings that can skew typological analysis in wordlist-based language documentation. Designed for the initial stages of data collection, these methods provide field linguists with systematic tools to identify entries requiring closer inspection. Our results reveal that phoneme-level n-gram models capture most anomalies also flagged by syllable-level models, suggesting that while incorporating explicit phonotactic knowledge through syllabification provides some benefit, raw phoneme sequence modeling alone achieves comparable performance. This indicates that computationally simpler approaches may be sufficient for practical anomaly detection in fieldwork settings, though the syllable-level analysis offers additional interpretability by identifying specific constraint violations.

Limitations

Due to the limited size of documentary wordlists, data-intensive approaches such as neural language models cannot be applied, though such methods might yield superior performance in high-resource settings. Borrowing detection is inherently challenging making gold standard creation labor-intensive. However, the statistical nature of n-gram models ensures they capture phonotactic inconsistencies providing field linguists with a tool for identifying entries that deviate from expected patterns and require closer inspection.

In terms of the method's usefulness in borrowing detection, this relies heavily on having a phonotactically more restrictive language borrowing from one with a greater possibility of sounds, or simply sounds which are not found in the borrowing language. Detecting Kra-Dai borrowings into a phonologically similar Tibeto-Burman language would not be possible in this case. However, the method is still useful in detecting transcription errors, novel sound changes, dialectal variation, or other cases which still warrant further investigation by the linguist even if not the result of borrowing.

Finally, as the fieldworker's expectations also shape how they transcribe data (Kelly and Lahaussois, 2021), application of this approach too quickly or without further investigation into the language could result in further over-application of mistakes. One should not blindly assume anomalies are errors. Rather, they are points to be investigated further and confirmed.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was supported by the Charles University student project GA UK No. 101924 and partially supported by SVV project number 260 698.

References

Shobhana Chelliah. 2014. Fieldwork for language description. *Research methods in linguistics*, pages 51–73.

Scott Delancey. forthcoming. The barish languages. In K. Hildebrandt, Y. Modi, D. Peterson, and H. Suzuky, editors, *The Oxford Guide to the Tibeto-Burman Languages*. Oxford University Press, Oxford.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5:180205.

Ana Vilacy Galucio, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2015. Genealogical relations and lexical distances within the tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10:229–274.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. *Glottolog 5.2*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at http://glottolog.org, Accessed on 2025-08-08.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Barbara Kelly and Aimée Lahaussois. 2021. Chains of influence in Himalayan grammars: Models and interrelations shaping descriptions of Tibeto-Burman languages of Nepal. *Linguistics*, 59(1):207–245.

Amy Kim, Palash Roy, Mridul Sangma, and Seung Kim. 2025. Cldf dataset derived from kim et al's "the tripura of bangladesh: A sociolinguistic survey" from 2011. Version v1.0.0, published December 18, 2025.

Johann-Mattis List. 2019. Automated methods for the investigation of language contact, with a focus on lexical borrowing. *Language and Linguistics Compass*, 13(10):e12355.

John Miller, Emanuel Pariasca, and Cesar Beltran Castañon. 2021. Neural borrowing detection with monolingual lexical models. In *Proceedings of the student research workshop associated with RANLP 2021*, pages 109–117.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J Ryder, Valentin Thouzeau, Simon J Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

A Appendix

Table 3: Precision and Recall at K for different analysis types and n-gram sizes for the syllable-level n-gram models.

Analysis	N-gram	Method	P@100	P@500	P@1000	R@100	R@500	R@1000
Within	Bigram	Arithmetic Mean	0.47	0.32	0.26	0.08	0.29	0.47
		Harmonic Mean	0.43	0.32	0.26	0.08	0.29	0.48
		Min NLL	0.35	0.32	0.23	0.06	0.29	0.42
		Max NLL	0.35	0.28	0.23	0.06	0.26	0.41
	Trigram	Arithmetic Mean	0.45	0.31	0.27	0.08	0.28	0.49
		Harmonic Mean	0.45	0.31	0.27	0.08	0.28	0.48
		Min NLL	0.40	0.32	0.27	0.07	0.29	0.49
		Max NLL	0.37	0.29	0.23	0.07	0.26	0.42
Cross	Bigram	Arithmetic Mean	0.21	0.20	0.18	0.04	0.18	0.33
		Harmonic Mean	0.21	0.20	0.19	0.04	0.18	0.33
		Min NLL	0.22	0.19	0.17	0.04	0.17	0.30
		Max NLL	0.34	0.21	0.19	0.06	0.19	0.34
	Trigram	Arithmetic Mean	0.30	0.26	0.21	0.05	0.23	0.37
		Harmonic Mean	0.30	0.26	0.21	0.05	0.23	0.37
		Min NLL	0.30	0.26	0.21	0.05	0.24	0.38
		Max NLL	0.26	0.27	0.23	0.05	0.24	0.42
Boundary	Bigram	Arithmetic Mean	0.50	0.29	0.24	0.09	0.26	0.43
		Harmonic Mean	0.44	0.30	0.25	0.08	0.27	0.45
		Min NLL	0.41	0.25	0.18	0.07	0.22	0.33
		Max NLL	0.36	0.28	0.23	0.06	0.26	0.42
	Trigram	Arithmetic Mean	0.48	0.29	0.26	0.09	0.26	0.47
		Harmonic Mean	0.49	0.29	0.26	0.09	0.26	0.47
		Min NLL	0.44	0.29	0.25	0.08	0.26	0.46
		Max NLL	0.37	0.30	0.24	0.07	0.27	0.44
Baseline	Uniform Random		0.15	0.19	0.17	0.03	0.17	0.31
	Stratified Random		0.17	0.20	0.17	0.03	0.18	0.30

What NLP Gets Wrong About Contact: Implications for Field Linguistic Evidence

Manodnya K H
CLASIC,
University of Colorado Boulder
Boulder, CO, USA
manodynak@gmail.com

Abstract

Field linguistics increasingly relies on computational tools to organize, analyze, and preserve linguistic data, yet the classificatory assumptions embedded in these tools are rarely examined. A pervasive assumption is that languages can be treated as discrete, genealogically defined units, with relatedness modeled as tree-structured descent. We argue that this assumption misrepresents linguistic evidence in contact-heavy regions and risks distorting the computational mediation of field linguistic data. Focusing on South Asia, we show that widely assumed boundaries—such as the Indo-Aryan–Dravidian divide—collapse in long-standing contact zones characterized by convergence, dialect continua, and institutional multilingualism. Through historically grounded case studies including Kannada–Telugu and Tamil–Malayalam, we demonstrate how convergence, script-mediated distance, and post-hoc standardization reshape how field data is segmented, compared, and interpreted when organized through genealogical labels. We argue that contact-aware, relational models of linguistic relatedness are necessary if NLP tools are to support, rather than distort, the documentation and analysis of linguistic diversity.

1 Introduction

Genealogical classification has long framed linguistic relatedness in terms of divergence, modeling languages as splitting and branching along tree-structured lineages. This view has been foundational for historical reconstruction and typology, but it is increasingly embedded—often implicitly—into computational tools used to organize, analyze, and preserve linguistic data. In contact-heavy regions such as South Asia, this abstraction captures only part of linguistic reality. Alongside divergence, the region exhibits persistent patterns of convergence across families and branches, produced through centuries of contact, cohabitation, and institutional multilingualism (Emeneau, 1956; Masica, 1976; Southworth, 2005). When computational systems inherit genealogical labels as organizing primitives, these dynamics are systematically obscured.

The Indo-Aryan–Dravidian divide illustrates this problem clearly. Although typological and official classifications present these families as distinct, extensive linguistic work shows that the boundary collapses in transitional regions shaped by sustained contact (Gumperz and Wilson, 1971; Thomason and Kaufman, 1988; Bashir, 2016). In such zones, cross-boundary alignment frequently outweighs internal divergence within standardized languages, giving rise to dialect continua and areal nexuses that resist discrete classification (Masica, 1991; Nichols, 1992). Treating these regions through rigid taxonomic lenses risks misrepresenting the linguistic systems documented through fieldwork.

These issues are not peripheral to field linguistics. Courtly bilingualism, shared literary registers, and long-standing diglossia enabled grammatical constructions, lexemes, and pragmatic conventions to circulate across what are now treated as firm genealogical divides (Ramanujan and Masica, 1969; Krishnamurti, 2003). Scriptal divergence, often taken as evidence of linguistic separation, frequently lags behind structural convergence and instead reflects post-hoc identity formation driven by political and administrative standardization (Southworth, 2005; Trautmann, 2006). When field linguistic data is computationally organized through genealogical categories, these historical processes directly shape how linguistic evidence is segmented, compared, and interpreted.

This paper argues that, in contact-heavy settings, genealogical classification functions poorly as a computational organizing principle for field linguistic data. Using South Asia as a critical case, we show that many distortions introduced by NLP-assisted analysis arise not from data sparsity or

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 8–15

modeling limitations, but from the imposition of tree-based abstractions onto contact-driven linguistic systems. Recognizing this mismatch is essential if computational tools are to support, rather than distort, the documentation and analysis of linguistic diversity.

2 Genealogical Classification as a Computational Prior

In computational workflows applied to linguistic documentation, genealogical classification rarely appears as an explicit analytical choice. Instead, it enters indirectly through the labels, resources, and organizational frameworks inherited from linguistic surveys, census practices, and standard-language corpora. ISO language codes, census categories, script conventions, and standardized orthographies together function as a de facto ontology of linguistic relatedness. When NLP tools are used to organize, compare, or archive linguistic materials, this ontology quietly structures what counts as a language, what is comparable, and what variation is treated as noise. In contact-heavy regions such as South Asia, these inherited categories encode assumptions of discreteness and divergence that are poorly aligned with the linguistic record (Grierson, 1903; Masica, 1991; Southworth, 2005).

Genealogical trees, originally developed as heuristic tools for historical reconstruction, thus come to function as computational ground truth in the processing of field and documentary data. Linguistic materials are partitioned into discrete labels; similarity is inferred across pre-defined language units; and cross-corpus comparison is constrained by presumed family membership (Campbell, 2003; Garrett, 2006). In this way, genealogical classification does not merely describe linguistic structure—it actively shapes how linguistic evidence is grouped, stored, and interpreted by computational systems.

This framing becomes especially problematic in regions characterized by dialect continua and areal convergence. As Emeneau's formulation of a linguistic area made clear, South Asia exhibits widespread structural diffusion across family boundaries that cannot be reduced to inheritance alone (Emeneau, 1956). Subsequent scholarship has documented how features such as retroflexion, case marking, clause-finality, and evidential strategies circulate across Indo-Aryan and Dravidian languages through sustained contact rather than descent (Masica, 1976; Thomason and Kaufman, 1988; Bashir, 2016). When computational tools continue to treat these languages as maximally distinct once genealogical boundaries are crossed, they impose a classificatory logic that obscures precisely the forms of continuity most salient in field linguistic evidence.

Scriptal differentiation further reinforces this abstraction. In South Asia, script is frequently conflated with language identity, despite extensive historical evidence that scriptal divergence often lags behind linguistic convergence and functions as an identity marker rather than a structural delimiter (Southworth, 2005; Trautmann, 2006). When scripts are treated as proxies for language boundaries in data organization and analysis, surface distance is amplified while deeper grammatical and lexical alignment is suppressed. This script-mediated distance is then reified in tokenization practices, representational spaces, and similarity measures, shaping how linguistic relatedness is inferred from field data.

The consequences of these assumptions are visible in computationally mediated linguistic inquiry. Transitional varieties are routinely misattributed to dominant standards or excluded altogether (Nichols, 1992). Measures of similarity overestimate distance across administrative or scriptal boundaries while underestimating divergence within standardized languages that exhibit substantial internal variation (Nichols, 1997). Cross-corpus comparison and reuse of field data consequently privilege the wrong affinities while overlooking structurally aligned contact zones (Kunchukuttan and Bhattacharyya, 2020). In such cases, distortion arises not from the linguistic data itself, but from the classificatory scaffolding through which that data is processed.

Importantly, these effects cannot be attributed solely to limitations of data or model capacity. They reflect a prior commitment to genealogical abstraction as the organizing principle of computational representation. Recent computational studies have begun to expose this mismatch. Dialect-level embeddings and geospatial clustering reveal similarity structures that align more closely with areal proximity and contact history than with genealogical family membership (Arora et al., 2021, 2022, 2023). Large-scale reanalyses of survey data likewise question whether genetic classification adequately captures South Asian language relationships (Borin et al., 2021). These findings do not

challenge historical linguistics; they highlight how selectively genealogical structure has been operationalized in computational treatments of linguistic evidence.

In this paper, we therefore treat genealogical classification not as neutral background context, but as an active computational prior. In contact-heavy settings, this prior systematically flattens networked, historically entangled linguistic ecologies into administratively convenient abstractions, shaping what computational tools can recognize as structure or variation in field linguistic data.

3 Contact Zones and Areal Nexuses in South Asia

South Asian linguistic structure cannot be adequately captured through discrete language units alone. Instead, it is organized around zones of sustained contact—areal nexuses in which linguistic features circulate across genealogical boundaries through prolonged multilingualism, shared institutions, and overlapping communicative domains. The concept of South Asia as a linguistic area, first articulated by Emeneau (Emeneau, 1956), foregrounded convergence as a co-equal force alongside divergence. Subsequent work has confirmed that areality is not a peripheral phenomenon, but a defining characteristic of the region's linguistic ecology (Masica, 1976, 1991; Southworth, 2005; Hook, 1987).

Areal convergence in South Asia operates at multiple linguistic levels. Phonological features such as retroflexion and vowel harmony, morphosyntactic patterns including postpositional case marking and clause-finality, and pragmatic strategies such as honorific alignment and evidential marking diffuse across Indo-Aryan and Dravidian languages through contact rather than inheritance (Ramanujan and Masica, 1969; Hook, 1976; Thomason and Kaufman, 1988). These shared structures do not eliminate genealogical distinctions, but they routinely outweigh them in actual language use, particularly in regions characterized by dense multilingual interaction (Nichols, 1992, 1997).

Dialect continua are most visible in transitional zones that lie between major phylogenetic divisions. Rather than exhibiting sharp boundaries, varieties in these regions form gradients of mutual intelligibility, lexical overlap, and structural alignment. Classical dialectological work has long documented such continua in South Asia, particularly along the Indo-Aryan–Dravidian interface and within Eastern Indo-Aryan (Grierson, 1903; Southworth, 2005). More recent regional studies of Kannada, Marathi, Konkani, Odia, and related varieties reinforce the view that linguistic distance increases gradually rather than categorically across space (Sridhar, 1990; Rane, 2010; Behera, 2006; Patnaik, 2015).

Institutional multilingualism has played a central role in sustaining these contact zones. Precolonial courts, religious institutions, and administrative systems routinely operated across multiple languages and registers, enabling grammatical constructions, lexemes, and stylistic conventions to circulate widely (Talbot, 2001; Zvelebil, 1973). Courtly bilingualism in particular fostered stable patterns of registeral alignment, where literary and bureaucratic norms were shared across languages without being perceived as foreign (Gumperz and Wilson, 1971). These practices produced layered linguistic ecologies in which speakers navigated multiple codes without rigid boundaries.

Scriptal differentiation, often treated as a proxy for linguistic separation in computational pipelines, must be understood within this institutional context. Historical evidence shows that scriptal divergence frequently follows linguistic convergence, crystallizing only when political, religious, or educational regimes seek to formalize identity (Southworth, 2005; Trautmann, 2006). In South Asia, the consolidation of distinct scripts for languages such as Kannada, Telugu, Tamil, and Malayalam reflects processes of standardization rather than deep structural rupture. As King and Pollock have shown in different contexts, scripts often function as symbolic markers of authority and identity rather than transparent reflections of linguistic distance (King, 1994; Pollock, 2006).

Colonial and postcolonial language administration further intensified this process. Large-scale surveys and gazetteers, while invaluable as documentary resources, imposed classificatory grids that privileged discrete languages over continua and standardized forms over local practice (Hunter, 1881, 1885). Educational policy and state formation in the twentieth century hardened these boundaries, aligning language, script, and territory in ways that obscured long-standing zones of overlap (Annamalai, 2001; Mohanty, 2019). Linguistic state reorganization after 1956 represents a particularly consequential moment, transforming fluid contact zones into administratively policed borders

(Trautmann, 2006; Patnaik, 2015).

For field linguistics and computational documentation, these historical processes are not merely background context. They determine how corpora are labeled, how scripts are segmented, and how linguistic relatedness is encoded in machine-readable form. When areal nexuses and dialect continua are flattened into discrete categories, computational representations mischaracterize similarity and erase contact-driven structure. The case studies that follow examine this dynamic in detail, showing how specific South Asian contact zones expose the limits of tree-based assumptions and motivate contact-aware, network-oriented approaches to computationally mediated linguistic analysis.

4 Case Studies: Convergence Against Classification

The following case studies illustrate how genealogical classification collapses in South Asian contact zones. Rather than presenting exhaustive historical surveys, each case foregrounds a specific mode of entanglement—morphological, scriptal, or registeral—that exposes the limits of tree-based models and motivates an areal, network-oriented account of relatedness.

4.1 Kannada–Telugu: Courtly Bilingualism and Morphological Interflow

While modern classifications assign Kannada and Telugu to distinct Dravidian subgroups—South Dravidian and South-Central Dravidian respectively (Krishnamurti, 2003)—their historical record reveals prolonged co-evolution rather than divergence. From the ninth to the fourteenth centuries, both languages functioned as courtly and inscriptional media under the Western Chalukyas, Hoysalas, and the Vijayanagara Empire, producing overlapping literary registers and shared administrative conventions (Talbot, 2001; Gumperz and Wilson, 1971).

Epigraphic evidence from regions such as Hampi, Bagali, Molkalmuru, and Anantapur demonstrates scriptal hybridity prior to the formal divergence of Kannada and Telugu scripts in the thirteenth century (Southworth, 2005). Contemporary spoken varieties along the Molkalmuru–Anantapur belt retain interoperable lemma inventories, case marking, verb-final constructions, and honorific systems, with variation largely restricted to phonological realization rather than grammatical function (Ramanujan and Masica, 1969; Krishnamurti, 2003). These patterns reflect a shared grammatical substrate differentiated only later through scriptal standardization and administrative boundary formation.

4.2 Tamil–Malayalam: Selective Divergence and Script-Mediated Distance

Tamil and Malayalam are often cited as a canonical example of genealogical divergence within Dravidian. Yet this divergence is highly stratified. Early Malayalam inscriptions and literary texts remain closely aligned with contemporaneous Tamil in core morphosyntax, with differentiation concentrated in lexicon, script, and register (Menon, 1933; Krishnamurti, 2003). Border varieties in regions such as Palakkad and Kanyakumari preserve high degrees of mutual intelligibility and shared grammatical structure (Ramanujan and Masica, 1969; Hook, 1976).

Much of the perceived distance between modern standard Tamil and Malayalam reflects post-medieval processes of Sanskritization, orthographic consolidation, and literary standardization. Scriptal divergence in particular amplifies surface distance while masking deeper structural continuity, functioning as an identity marker rather than a reliable indicator of grammatical separation (Southworth, 2005; Trautmann, 2006).

4.3 Marathi–Konkani–Tulu: Littoral Continua and Registeral Layering

Along the western coast, Marathi, Konkani, and Tulu participate in a littoral contact zone characterized by intense multilingualism and registeral layering. Konkani varieties in particular exhibit extensive lexical and morphosyntactic overlap with both Marathi and Kannada/Tulu, reflecting sustained contact rather than clear genealogical alignment (Rane, 2010; Sridhar, 1990). Script choice—Devanagari, Roman, or Kannada—further fragments representation without corresponding structural divergence.

Here, standard-language gravity pulls varieties toward administratively dominant centers, while everyday usage preserves hybrid systems that resist categorical classification. Computationally, this produces unstable language identification and distorted similarity judgments when script or standard form is treated as primary signal.

4.4 Bangla–Odia: Eastern Indo-Aryan and Administrative Separation

Bangla and Odia, both Eastern Indo-Aryan languages, share extensive phonological and morphosyntactic structure, particularly in transitional regions such as Medinipur and Ganjam (Masica, 1991; Behera, 2006). Their contemporary separation is reinforced by scriptal differentiation and colonial-era administrative boundaries rather than deep structural rupture.

As with Dravidian contact zones, dialect continua across the Bangla–Odia interface exhibit gradual rather than categorical change. When treated as maximally distinct units, these varieties are artificially distanced in computational representations despite substantial grammatical alignment.

4.5 Hindi/Hindustani Continua: Internal Diversity Under a Single Label

The Hindustani continuum illustrates a complementary failure mode: internal diversity collapsed under a single standardized label. Varieties such as Bhojpuri, Awadhi, Maithili, and Dakhani differ systematically in morphosyntax and lexicon, yet are routinely subsumed under "Hindi" in corpora and NLP pipelines (Grierson, 1903; Southworth, 2005). This flattening erases meaningful internal structure and produces predictable errors in language identification, similarity modeling, and transfer.

5 What This Distorts in Computationally Mediated Field Linguistics

The case studies above show that genealogical classification fails descriptively in South Asian contact zones. When computational tools are used to document, annotate, organize, and compare linguistic data, however, this failure acquires broader consequences. Treating genealogical labels as ground truth does not merely simplify linguistic reality—it reshapes how linguistic evidence is partitioned, aligned, and rendered legible within computational workflows supporting field linguistics. What follows is not a catalog of technical errors, but an account of how a classificatory prior propagates through computational mediation, with direct consequences for how linguistic structure is recorded and interpreted.

5.1 Language Identification as Documentary Misattribution

Computational tools used in field linguistics often presuppose that linguistic material can be unambiguously mapped to discrete language categories. In South Asia, this assumption collapses in contact zones and dialect continua. Varieties spoken along interfaces such as Kannada–Telugu, Bangla–Odia, or within the Hindustani continuum are not marginal or noisy instantiations of a single language, but stable hybrid systems shaped by long-term contact.

When such material is forced into genealogical labels during annotation or corpus organization, linguistic proximity is treated as error and administrative categories as signal. The result is not merely misclassification, but misattribution: texts, utterances, and speakers are reassigned to categories that obscure the linguistic systems they instantiate. Documentation practices that enforce categorical assignment further entrench this distortion, penalizing representations that capture genuine overlap while rewarding conformity to inherited taxonomies. Apparent inconsistency in field data thus reflects classificatory mismatch rather than deficiencies in the data itself.

5.2 Similarity Modeling and the Production of Artificial Distance

Computational analysis of field linguistic data frequently relies on similarity and distance measures to organize corpora, align varieties, or infer relatedness across datasets. When genealogical boundaries are assumed to define similarity space, these measures inherit a distorted geometry. Distance is exaggerated across scriptal or administrative boundaries and suppressed within standardized languages, even where internal variation is substantial.

In South Asia, where script-mediated distance often masks deep grammatical continuity, surface divergence becomes over-weighted. Representational spaces trained on standardized corpora encode this bias, producing similarity structures that mirror census categories more closely than contact history. As a result, varieties that share morphosyntactic and pragmatic structure are rendered artificially distant, while internally heterogeneous standards are treated as coherent units. The distances inferred through computational analysis are thus not discovered in the data, but produced by classificatory assumptions.

5.3 Cross-Corpus Comparison and Misplaced Affinities

Computational workflows increasingly support the reuse, aggregation, and comparison of field linguistic data across projects and corpora. In practice, genealogical classification often serves as a proxy for determining which materials are comparable or transferable. In contact-heavy settings, this proxy misaligns linguistic evidence.

Structurally aligned varieties that cross genealogical boundaries are excluded from comparison, while aggregation within standardized languages suppresses meaningful internal diversity. In South Asian contact zones, convergence-driven affinities are systematically overlooked, while administratively consolidated categories dominate corpus structure. Apparent incompatibilities across datasets therefore reflect misplaced assumptions about where similarity resides, rather than intrinsic differences in linguistic structure.

5.4 Annotation and Organization Against Administrative Abstractions

The most consequential effects of genealogical abstraction emerge at the level of annotation and data organization. When computational tools treat standardized language labels as ground truth, they structure field data according to administrative artifacts rather than linguistic evidence. In such regimes, capturing gradient similarity, overlap, or hybridity becomes difficult or impossible within available annotation schemes.

This has direct implications for field linguistics. Datasets that inherit rigid labels from census categories or ISO standards conflate linguistic adequacy with taxonomic conformity. Over time, tools, annotation practices, and corpora co-evolve around the same abstractions, reinforcing a closed loop in which classificatory schemes are reproduced rather than questioned. Linguistic variation documented in the field is thereby flattened, misaligned, or rendered invisible.

5.5 From Technical Limitation to Epistemic Misalignment

Taken together, these distortions point to a deeper issue. Persistent challenges in the computational handling of field linguistic data are not primarily technical. They reflect an epistemic misalignment between how linguistic relatedness is conceptualized and how linguistic evidence is organized for analysis. In South Asia, where convergence, areality, and institutional multilingualism are foundational, tree-based classification is not a neutral simplification but an interpretive distortion.

Recognizing this misalignment reframes the role of computational tools in field linguistics. The goal is not merely to scale existing annotation and organization practices, but to interrogate the classificatory assumptions that structure them. Without this shift, increasingly sophisticated tools risk producing ever more precise representations of increasingly impoverished abstractions, undermining the very diversity field linguistics seeks to preserve.

6 Discussion

The preceding analysis reframes a problem often treated as technical or task-specific in computational linguistics. Rather than viewing South Asian linguistic diversity as an unusually difficult setting for computational analysis, we argue that many persistent failures arise from a deeper misalignment between linguistic reality and classificatory abstraction. Genealogical labels, inherited from historical linguistics and institutionalized through census practice, script standardization, and corpus design, function as silent priors that shape how linguistic evidence is organized, annotated, and interpreted in computationally mediated fieldwork.

What distinguishes South Asia is not merely the presence of diversity, but the centrality of convergence. Linguistic structure in the region is shaped by sustained multilingualism, registeral layering, and institutional bilingualism, producing contact zones in which relatedness is relational rather than categorical. When computational tools assume divergence as the default and treat convergence as noise, they invert this reality. Hybrid and transitional varieties are mischaracterized precisely because they reflect the linguistic ecologies in which they emerge.

This perspective clarifies why incremental technical improvements often fail to resolve longstanding challenges in computational support for field linguistics. Improved models, better tokenization, or expanded datasets cannot compensate for annotation schemes and data structures that encode administrative abstractions as ground truth. As long as computational workflows privilege rigid labels over contact-driven structure, they will continue to obscure the very forms of variation that field linguistics seeks to document.

More broadly, the South Asian case exposes a general vulnerability in computationally mediated linguistic inquiry. Whenever genealogical classification is treated as exhaustive rather than partial—when trees are mistaken for terrain—computational tools risk organizing classification systems rather than linguistic evidence. For field linguistics, this distinction is not incidental: it determines what kinds of structure become visible, preservable, and interpretable.

7 Conclusion

This paper has argued that genealogical classification, while indispensable for historical reconstruction, becomes a liability when operationalized as computational ground truth in contact-driven linguistic ecologies. In South Asia, where convergence across families and branches is foundational rather than exceptional, tree-based assumptions systematically misrepresent linguistic relatedness, continuity, and variation.

Through historically grounded case studies, we showed how morphological, scriptal, and registeral entanglement undermines the assumptions embedded in computational treatments of language. These failures are not edge cases, nor are they artifacts of insufficient data or modeling capacity. They are the predictable consequences of imposing administrative and phylogenetic abstractions onto linguistic systems shaped by long-term contact.

We conclude that computational approaches supporting field linguistics must move beyond tree-based notions of linguistic organization. In contact-heavy regions, languages and varieties are better understood as nodes in overlapping, historically sedimented networks. Aligning computational tools with this reality is a prerequisite for documenting linguistic diversity without erasing it.

8 Future Work

Future work should translate this diagnostic account into computational practices that are explicitly contact-aware. Promising directions include annotation schemes that permit overlap and gradient membership, similarity measures grounded in areal proximity, and data models that represent continua rather than discrete endpoints.

Field-facing tools should also distinguish misannotation from faithful representation of hybridity, enabling documentation practices that preserve contact-driven structure rather than suppress it. Beyond South Asia, similar analyses should be extended to other regions characterized by long-term contact and dialect continua, including the Balkans, the Arabic-speaking world, and Romance dialect spaces.

Finally, sustained collaboration between NLP practitioners and field linguists is essential. Without close engagement with the epistemic commitments of fieldwork, computational tools risk reproducing inherited abstractions rather than supporting the preservation and analysis of linguistic diversity.

9 Limitations

This study is intentionally diagnostic and theory-driven. While it draws on a broad body of linguistic and computational scholarship, it does not introduce new datasets, tools, or empirical experiments. Its contribution lies in identifying structural failure modes in the computational mediation of linguistic evidence rather than in proposing immediate technical solutions.

Our focus on South Asia reflects the region's extreme linguistic density and long history of contact. Although analogous dynamics are likely present elsewhere, the extent to which these conclusions generalize beyond South Asia remains an empirical question.

Finally, this paper does not reject genealogical classification wholesale. Genealogy remains indispensable for many linguistic purposes. Our claim is narrower: in contact-heavy settings, genealogical labels are insufficient as organizing principles for computational documentation. Determining how genealogical and areal signals should be balanced in field-facing computational tools remains an open problem.

References

E. Annamalai. 2001. *Managing Multilingualism in India: Political and Linguistic Manifestations*. Sage, New Delhi.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2021. Bhasacitra: Visualizing the dialect geography of South Asia. arXiv.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. arXiv.

A. Arora, A. F. Farris, S. Basu, and S. Kolichala. 2023. JAMBU: A historical linguistic database for South Asian languages. arXiv.

Elena Bashir. 2016. Contact and convergence. In Hans Henrich Hock and Elena Bashir, editors, *The Languages and Linguistics of South Asia: A Comprehensive Guide*, pages 123–145. De Gruyter Mouton, Berlin and Boston.

D. Behera. 2006. The Odia language movement: A linguistic assertion. *Orissa Review*, 62(1):18–27.

Lars Borin, Anju Saxena, Bernard Comrie, and Shafqat Mumtaz Virk. 2021. A bird's-eye view on South Asian languages through LSI: Areal or genetic relationships? *Journal of South Asian Languages and Linguistics*, 7(2):203–237.

Lyle Campbell. 2003. How to show languages are related. In Roger D. Woodard, editor, *The Cambridge Encyclopedia of the World's Ancient Languages*, pages 108–120. Cambridge University Press, Cambridge.

Murray B. Emeneau. 1956. India as a linguistic area. *Language*, 32(1):3–16.

Andrew Garrett. 2006. Convergence in the formation of Indo-European subgroups: Phylogeny and the challenge of contact. In A. L. Sims, editor, *Historical Linguistics 2005*, pages 64–75. John Benjamins, Amsterdam and Philadelphia. Verify editor/booktitle details against your copy; chapter metadata varies by edition.

George A. Grierson. 1903. *Linguistic Survey of India*. Government of India Press. Published 1903–1928; Vols. 1–11.

John J. Gumperz and Robert Wilson. 1971. Convergence and creolization: A case from the Indo-Aryan/Dravidian border in India. In Dell H. Hymes, editor, *Pidginization and Creolization of Languages*, pages 151–167. Cambridge University Press, Cambridge.

Peter E. Hook. 1976. Case marking in South Asian languages: A survey. *Indian Linguistics*, 37:45–78.

Peter E. Hook. 1987. Linguistic areas: Getting at the grain of history. In George Cardona and Norman H. Zide, editors, *Festschrift for Henry Hoenigswald*, pages 155–168. Narr, Tübingen.

William Wilson Hunter. 1881. *The Imperial Gazetteer of India*, 1 edition. Trübner and Co., London. 9 vols.

William Wilson Hunter. 1885. *The Imperial Gazetteer of India*, 2 edition. Oxford University Press, Oxford. Published 1885–1887; Vols. 5–24.

Christopher R. King. 1994. *One Language, Two Scripts: The Hindi Movement in Nineteenth Century North India*. Oxford University Press, Oxford.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge.

A. Kunchukuttan and P. Bhattacharyya. 2020. Leveraging language relatedness to improve low-resource machine translation. arXiv.

Colin P. Masica. 1976. *Defining a Linguistic Area: South Asia*. University of Chicago Press, Chicago.

Colin P. Masica. 1991. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge.

T. R. Menon. 1933. *A Primer of Malayalam Literature*. Asian Educational Services, New Delhi.

A. K. Mohanty. 2019. *Multilingualism, Education and Language Policy in India*. Springer, Singapore.

Johanna Nichols. 1992. *Linguistic Diversity in Space and Time*. University of Chicago Press, Chicago.

Johanna Nichols. 1997. Modeling ancient population structures in linguistics. *Annual Review of Anthropology*, 26:427–450.

D. Patnaik. 2015. Language identity and politics in eastern India. *Indian Journal of Linguistics*, 75(3):223–244.

Sheldon Pollock. 2006. *The Language of the Gods in the World of Men: Sanskrit, Culture, and Power in Premodern India*. University of California Press, Berkeley.

A. K. Ramanujan and Colin P. Masica. 1969. Toward a phonological typology of the Indian linguistic area. In Thomas A. Sebeok, editor, *Current Trends in Linguistics, Vol. 5: Linguistics in South Asia*, pages 543–577. Mouton, The Hague and Paris.

J. Rane. 2010. Konkani dialectology: Intracontinental variation. *Journal of Indo-Aryan Studies*, 24:57–74.

Franklin C. Southworth. 2005. *Linguistic Archaeology of South Asia*. Routledge, London and New York.

K. K. Sridhar. 1990. Kannada dialects and multilingualism in India. *International Journal of the Sociology of Language*, 86:99–119.

Cynthia Talbot. 2001. *Precolonial India in Practice: Society, Region, and Identity in Medieval Andhra*. Oxford University Press, Oxford.

Sarah G. Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley and Los Angeles.

Thomas R. Trautmann. 2006. *Languages and Nations: The Dravidian Proof in Colonial Madras*. University of California Press, Berkeley.

Kamil V. Zvelebil. 1973. *The Smile of Murugan: On Tamil Literature of South India*. Brill, Leiden.

Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan

Siyu Liang[1], **Talant Mawkanuli**[2], **Gina-Anne Levow**[1]
[1] Department of Linguistics, University of Washington
[2] Department of Middle Eastern Languages and Cultures, University of Washington
{liangsy, tmawkan, levow}@uw.edu

Abstract

Interlinear glossed text (IGT) creation remains a major bottleneck in linguistic documentation and fieldwork, particularly for low-resource morphologically rich languages. We present a hybrid automatic glossing pipeline that combines neural sequence labeling with large language model (LLM) post-correction, evaluated on Jungar Tuvan, a low-resource Turkic language. Through systematic ablation studies, we show that retrieval-augmented prompting provides substantial gains over random example selection. We further find that morpheme dictionaries paradoxically hurt performance compared to providing no dictionary at all in most cases, and that performance scales approximately logarithmically with the number of few-shot examples. Most significantly, our two-stage pipeline combining a BiLSTM-CRF model with LLM post-correction yields substantial gains for most models, achieving meaningful reductions in annotation workload. Drawing on these findings, we establish concrete design principles for integrating structured prediction models with LLM reasoning in morphologically complex fieldwork contexts. These principles demonstrate that hybrid architectures offer a promising direction for computationally light solutions to automatic linguistic annotation in endangered language documentation.

1 Introduction

Interlinear glossed text (IGT) is essential for linguistic documentation and preservation, aligning language transcriptions with morpheme segmentation, glosses, and translations (Lehmann, 2004a). Despite its importance, IGT creation remains highly labor-intensive, creating bottlenecks in language documentation projects (Chelliah and Reuse, 2010). Recent advances in neural sequence models and large language models offer new possibilities for automated IGT generation, yet each approach has limitations: structured models lack flexibility and world knowledge, while LLMs struggle with consistency and require extensive in-context examples or expensive fine-tuning, all inaccessible in most real-life use cases.

We present a hybrid pipeline that combines a BiLSTM-CRF model for initial gloss prediction with LLM-based post-correction, evaluated on Jungar Tuvan (henceforth Tuvan), a morphologically complex Turkic language. Through systematic experiments across four LLMs and multiple design choices, we demonstrate that this two-stage approach substantially improves over the BiLSTM baseline for most models. Our ablation studies reveal key design principles: retrieval-augmented prompting significantly outperforms random example selection; morpheme dictionaries generally hurt performance for most models; and optimal few-shot parameters range from five to fifteen examples.

Our contributions are the following: (1) we present a hybrid architecture combining structured prediction with LLM reasoning for automatic glossing; (2) we carry out comprehensive ablation studies establishing design principles for retrieval strategies, glossary configurations, and few-shot scaling; (3) we provide comparative evaluation of model performance across generation versus correction tasks, providing evidence-based guidance in fieldwork contexts.

2 Related Work

2.1 IGT and Language Documentation

IGT serves as a standard format in field linguistics, encoding source language transcriptions, morphological segmentation, gloss labels, and free translations (Lehmann, 2004b; Comrie et al., 2008). For many endangered and low-resource languages, IGT represents the primary digitized documentation (Chelliah and Reuse, 2010; Hargus et al., 2020).

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 16–30
March 26, 2026.

Recent work has developed tools for IGT extraction from grammatical descriptions (Schenner and Nordhoff, 2016; Round et al., 2020; Nordhoff and Krämer, 2022) and multi-modal IGT generation from speech (He et al., 2024). The SIGMORPHON 2023 shared task on automatic IGT generation (Ginn et al., 2023) further stimulated interest in computational approaches to glossing, leading to subsequent work on large-scale IGT modeling and evaluation, including pretrained language models for glossing (Ginn et al., 2024b) and LLM-based prompting approaches for low-resource IGT generation (Elsner and Liu, 2025). While these efforts demonstrate steady progress on benchmark datasets, they have not yet displaced existing fieldwork practices; in most documentation projects, IGT creation remains largely manual, relying on tools such as ELAN and FLEx (Wittenburg et al., 2006; International, 2025). Traditional workflows also include rule-based morphological parsers and deterministic dictionary lookup within tools like FLEx, which supports semi-automatic glossing and lexicon building from texts; our approach is intended to complement rather than replace such methods.

2.2 Automatic Morphological Analysis and Glossing

Traditional approaches to automatic glossing employ structured prediction models. Sequence-to-sequence architectures have been applied to morphological segmentation (Ruzsics and Samardžić, 2017; Liu et al., 2021; Rice et al., 2024), while CRF-based models have proven effective for morphological tagging (Buys and Botha, 2016; Malaviya et al., 2018). BiLSTM-CRF architectures in particular balance local pattern recognition with global constraints (Ma and Hovy, 2016; Cotterell and Heigold, 2017), achieving strong performance on sequence labeling tasks in morphologically rich languages.

Recent work specifically targeting IGT generation has explored neural encoder-decoder models with translation data (Zhao et al., 2020), CRF-based approaches for low-resource scenarios (Barriga Martínez et al., 2021; Okabe and Yvon, 2023), and lightweight models using structured linguistic representations (Shandilya and Palmer, 2023). Moeller et al. (2020) demonstrate how IGT can support downstream morphological analysis tasks. However, these models require substantial annotated corpora and struggle with rare morphemes and novel combinations. The recent work by Rice et al. (2025) also identifies significant gaps between computational morphology research outputs and real-world language documentation needs, highlighting the importance of user-centered design.

2.3 LLMs for Linguistic Annotation

Large language models have shown promise for linguistic annotation tasks in low-resource settings. Recent work (Ginn et al., 2024a; Elsner and Liu, 2025) explore LLM-based gloss prediction and prompting strategies for IGT, demonstrating that prompt design and example selection substantially affect performance. Zhang et al. (2024) show that providing dictionaries and grammar sketches enables translation for unseen languages. Yang et al. (2025b) evaluate models on metalinguistic reasoning using reference grammars and IGT, though their benchmark relies on curated reference grammar data without the full fieldwork contexts. LLM post-correction and refinement steps are also widely explored across NLP tasks as cascaded or post-editing stages (Zouhar et al., 2021; Izacard et al., 2023).

Among recent studies, few-shot prompting has emerged as a key technique for adapting LLMs to specialized tasks. Studies demonstrate that careful selection and presentation of in-context examples significantly impacts performance (Logan IV et al., 2022; Winata et al., 2021), with retrieval-based example selection often outperforming random selection (Stahl et al., 2024). However, LLMs face challenges in low-resource settings: they require extensive in-context examples (increasing inference cost), struggle with paradigmatic consistency, and lack the inductive biases of structured sequence models.

2.4 Hybrid and Multi-Stage Architectures

Hybrid approaches combining multiple model types have proven effective across NLP tasks. Retrieval-augmented generation (RAG) enhances LLMs by dynamically incorporating relevant examples (Lewis et al., 2021; Jiang et al., 2023), with recent work exploring specialized RAG architectures for domain adaptation (Siriwardhana et al., 2023; Yu, 2022). Cascaded architectures leverage specialized models for different subtasks (Izacard et al., 2023), while post-processing steps that refine outputs using external knowledge have shown consistent gains (Zouhar et al., 2021). Our work extends these ideas to morphological annotation, proposing a two-stage pipeline where a BiLSTM-

CRF provides initial structure and an LLM refines predictions through contextual inference and consistency checks.

3 Data

3.1 Language and Corpus

Tuvan is a Turkic language spoken in the Republic of Tuva of the Russian Federation, Mongolia, and the Xinjiang Uyghur Autonomous Region of China, with approximately 280,000 speakers across these regions (Harrison and Anderson, 2002). The present study focuses on the variety of Jungar Tuvan, spoken in the Altay region of Xinjiang, China. We treat Jungar Tuvan as a low-resource variety used in documentation contexts; we do not adjudicate its formal endangerment status in this paper. Jungar Tuvan shares the core typological properties of Tuvan—canonical agglutinative morphology, extensive case marking (nominative, accusative, genitive, dative, locative, ablative, comitative), complex aspectual systems, and productive derivational morphology—while also exhibiting vowel harmony and consonant alternations that create allomorphic variation (Mawkanuli, 1999, 2005).

Our corpus comprises 895 IGT-annotated sentences drawn from data collected during fieldwork in Xinjiang, China from the 1987 to 1995. The data span 40 recording sessions across conversational registers and narratives. All IGT annotations were produced manually following a consistent project-specific schema informed by typological conventions (Lehmann, 2004b; Comrie et al., 2008). Glosses distinguish lexical items (e.g., *money*, *give*) from grammatical morphemes (e.g., DAT, 1SG, PRS).

Table 1 summarizes corpus statistics. The tagset comprises 240 unique gramamtical morpheme labels (e.g., 1SG, PST), and 1258 unique content word glosses.

Metric	#
Total sentences	895
Narratives	40
Unique grammatical morphemes	240
Unique content morphemes	1258
Average words per sentence	8.38 (± 5.71)
Average morphemes per word	1.69 (± 0.30)

Table 1: Corpus statistics for the Tuvan fieldwork dataset.

Example 1 illustrates a representative IGT instance from the corpus, demonstrating the alignment structure our models must learn.

(1) jïlgï-nan iyi joon bar
horse-ABL two big EXIST
"(We) have two big horses."

3.2 Data Split

We perform a train-test split at the document level, allocating approximately 85% of sentences (760) to training and 15% (135) to testing. To avoid information leakage, no segments of the same narrative appear in both splits, ensuring that models cannot exploit discourse-level or speaker-specific patterns from related utterances. We further verify the absence of near-duplicate sentences using character-level TF-IDF (term frequency–inverse document frequency) cosine similarity with a threshold of 0.95.

4 Methodology

4.1 Task Formalization

We frame glossing as a structured prediction problem: given a hyphen-segmented Tuvan utterance $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where each x_i represents a morpheme, produce a parallel sequence of gloss labels $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ where each y_i is drawn from a tagset. We assume gold morpheme boundaries and do not use translations in the glossing model; segmentation and glossing are treated as separate steps. We evaluate using token-level accuracy, defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i] \quad (1)$$

where N is the total number of morphemes in the test set, $\hat{y}_i$ is the predicted gloss, and y_i is the reference gloss. This metric directly reflects annotation workload reduction: higher accuracy means fewer manual corrections required.

4.2 BiLSTM-CRF Model

Our baseline employs a two-layer bidirectional LSTM with CRF decoding, widely used for sequence labeling (Lample et al., 2016; Ma and Hovy, 2016; Huang et al., 2015). The model uses 100-dimensional character-level embeddings (randomly initialized and trained from scratch), 128-dimensional hidden layers, and learns to predict gloss labels for segmented morphemes. We

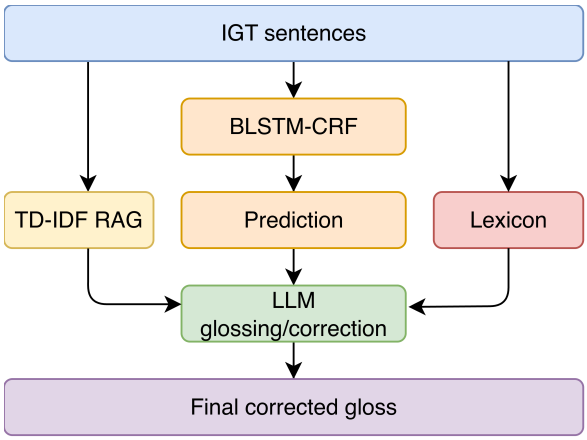


Figure 1: Hybrid pipeline combining BiLSTM-CRF structured prediction with LLM post-correction using retrieval-augmented prompting.

train for up to 100 epochs with early stopping (patience=10) on validation loss. This architecture captures local morphological patterns but cannot leverage broader linguistic knowledge or rare paradigms not well-represented in training data.

4.3 LLM Configuration and Prompting

We evaluate four LLMs: `deepseek-v3.2-exp` (DeepSeek-AI et al., 2025), `qwen3-max` (Yang et al., 2025a), `gpt-4o-mini` (OpenAI et al., 2024), and `gemma-3-27b-it` (Kamath et al., 2025). All models use greedy decoding with temperature zero for deterministic outputs. Prompts follow a consistent template: natural language instruction, k retrieved examples showing morpheme segmentation and gloss pairs, optional morpheme dictionary, and the test input. For the hybrid pipeline, prompts additionally include the BiLSTM prediction as a hypothesis to correct, framed as a "rough initial attempt" requiring verification. Complete prompt templates for all experiments are provided in Appendix A. Figure 1 illustrates our hybrid pipeline architecture.

4.4 Experimental Design

We conduct several experiments testing both LLM generation and hybrid correction. In Experiment 1 (Retrieval vs. Random Selection), we compare character-level TF-IDF (term frequency–inverse document frequency) cosine similarity-based retrieval against uniform random sampling for selecting three in-context examples to explore the effect of similarity-based retrieval. Retrieval operates at the sentence level: for each test sentence, we retrieve the most similar training sentences based on their Tuvan source text representations. Experiment 2 (N-Shot Scaling) varies example count from 1 to 20 with RAG and no glossary, mapping the accuracy-cost tradeoff for RAG LLM generation.

Experiment 3 (Glossary Ablation) tests four glossary configurations—none, top-100 most frequent morphemes, all grammatical morphemes, and the entire 1,498-pair dictionary—all using three-shot RAG to reveal whether partial dictionaries help or hinder performance. The glossary is provided to the LLM within the prompt as a plain-text key:value list (Appendix A), rather than as a deterministic lookup table. Finally, Experiment 4 (Hybrid Pipeline) evaluates BiLSTM plus LLM correction with varying n-shot counts including a zero-shot condition that tests whether LLMs can correct predictions without in-context examples. These ablations correspond to fieldwork-relevant choices about retrieval quality, example budget, and availability of lexical resources. Detailed prompt templates for RAG LLM generation (Experiments 1–3) and hybrid correction (Experiment 4) are provided in Appendix A.

5 Results

5.1 Baseline: BiLSTM-CRF Performance

Our BiLSTM-CRF baseline achieves 0.474 token accuracy on the test set with training data of 760 sentences. The model learns frequent morphological patterns (case markers, possessives, tense/aspect) but struggles with infrequent lexical morphemes and combinations of grammatical morphemes not attested in the training data. Error analysis reveals that 0.38 of errors involve lexical items appearing fewer than 5 times in training, and 0.24 involve grammatical morphemes in novel combinations.

5.2 Experiment 1: Retrieval vs. Random Selection

Table 2 shows the effect of retrieval enhancement across all four LLMs using 3-shot prompting without glossary.

Retrieval-augmented generation provides meaningful improvements across all models: +0.388 for `deepseek-v3.2-exp`, +0.319 for `qwen3-max`, +0.293 for `gpt-4o-mini`, and +0.276 for `gemma-3-27b-it`. `deepseek-v3.2-exp` achieves the highest absolute accuracy with RAG (0.506), while all models show substantial gains from retrieval. The consistent gains across

Model	Random	RAG
deepseek-v3.2-exp	0.118	**0.506**
qwen3-max	0.062	0.381
gpt-4o-mini	0.103	0.396
gemma-3-27b-it	0.068	0.344

Table 2: Retrieval-augmented prompting (RAG) vs. random example selection across four LLMs (3-shot, no glossary). All models show meaningful improvement with RAG.

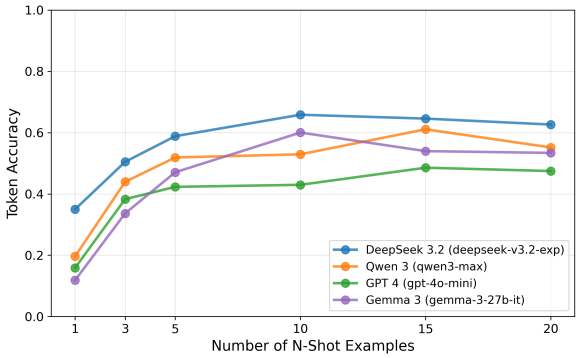


Figure 2: Experiment 2: n-shot scaling curves for RAG LLM generation. Performance scales approximately logarithmically with example count, plateauing around n=10–15 for most models. The BiLSTM baseline (0.474) is provided in the text for reference.

architectures demonstrate that similarity-based example selection is beneficial for morphological glossing tasks.

5.3 Experiment 2: N-Shot Scaling

Using the same TF-IDF-based retrieval from Experiment 1, we vary the number of retrieved examples from 1 to 20 without providing any glossary. Figure 2 shows performance scaling with example count (see Table 5 in Appendix B for detailed values).

Performance scales approximately logarithmically with example count. `deepseek-v3.2-exp` peaks at n=10 (0.658), then slightly declines at n=15 (0.646) and n=20 (0.626), while `qwen3-max` shows continued gains up to n=15 (0.611). `gpt-4o-mini` peaks at n=15 (0.486), and `gemma-3-27b-it` achieves 0.600 at n=10 before declining to 0.534 at n=20. The consistent pattern across models indicates diminishing marginal returns beyond 10 to 15 examples, with some models showing degradation at higher values. This decline may reflect either model saturation (distraction from excessive context) or retrieval quality degradation (less similar examples as the pool expands). We use character-level TF-IDF cosine similarity for retrieval; alternative similarity measures such as edit distance or contextualized embeddings might exhibit different scaling behavior, though we leave this investigation to future work. These results suggest practical operating points around n=5 to 10 for cost-sensitive applications and n=10 to 15 for maximum accuracy.

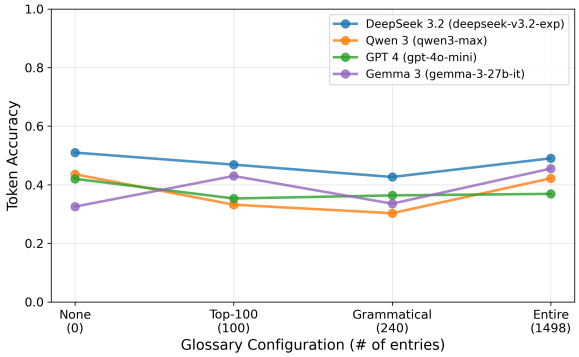


Figure 3: Experiment 3: glossary ablation results. Partial glossaries (Top-100, Grammatical) hurt performance compared to no glossary, while complete glossaries show modest gains. The negative effect suggests models are usually distracted by morphological information. The BiLSTM baseline (0.474) is provided in the text for reference.

5.4 Experiment 3: Glossary Ablation

To assess the impact of morpheme dictionaries on performance, we test four glossary configurations using 3-shot RAG: None (no dictionary provided), Top-100 (the 100 most frequent morpheme-gloss pairs), Grammatical (all 240 grammatical morpheme-gloss pairs), and Entire (the complete 1,498-pair dictionary including both grammatical and lexical morphemes). Figure 3 shows the results (see Table 4 in Appendix B for detailed values).

Counter-intuitively, providing morpheme dictionaries generally hurts performance. Partial glossaries consistently degrade accuracy across all models: Top-100 causes drops ranging from 0.041 to 0.104, while Grammatical shows similar or worse declines. Even the complete 1,498-pair dictionary fails to help for most models, with `deepseek-v3.2-exp`, `qwen3-max`, and `gpt-4o-mini` all performing worse with the entire glossary than with none (losses of 0.019, 0.014, and 0.051 respectively). Only `gemma-3-27b-it` benefits from dictionary information, achieving 0.455 with the entire glossary (+0.130 over None).

We lack direct evidence about whether models

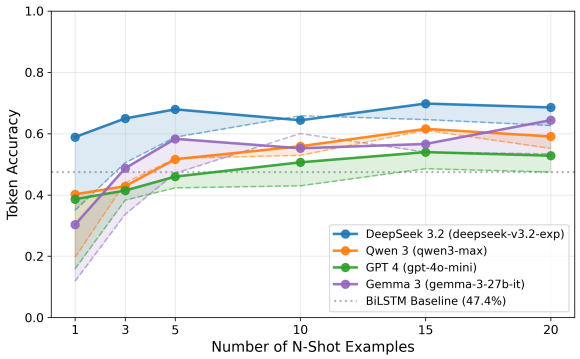


Figure 4: Experiment 4: hybrid pipeline improvement over RAG LLM generation. Solid lines show hybrid accuracy (BiLSTM + LLM correction), dashed lines show pure N-Shot baseline from Experiment 2, and shaded areas indicate improvement. The hybrid approach consistently improves performance across all four models, particularly in low-shot scenarios (n=1–5).

actually consult dictionary entries versus simply becoming distracted by additional prompt material. The degradation could reflect either inappropriate reliance on dictionary lookups for context-dependent glossing decisions, or models' difficulty balancing multiple information sources (retrieved examples, dictionaries, and morphological patterns). The no-glossary condition forces models to extract morphological structure from aligned examples, which appears more effective than dictionary consultation for most architectures, though we cannot definitively isolate the causal mechanism without fine-grained analysis of model predictions.

5.5 Experiment 4: Hybrid Pipeline Performance

We then evaluate the hybrid pipeline using the same TF-IDF retrieval as in Experiments 1 and 2, but providing the BiLSTM-CRF predictions as initial hypotheses for LLM correction. We test with n=1, 3, 5, 10, 15, 20 retrieved examples, all without glossaries. Figure 4 shows improvement from the hybrid pipeline over RAG LLM generation across varying n-shot counts for all four models.

The hybrid pipeline substantially outperforms pure BiLSTM predictions across all models. `gemma-3-27b-it` achieves 0.644 at n=20, a +0.170 gain over the BiLSTM baseline (0.474) and +0.110 improvement over its pure generation performance (0.534 at n=20). `deepseek-v3.2-exp` reaches 0.698 at n=15 (+0.224 over BiLSTM, +0.052 over pure generation), `qwen3-max` achieves 0.616 (+0.142 over BiLSTM, +0.005 over pure generation), and `gpt-4o-mini` reaches 0.540 (+0.066 over BiLSTM, +0.054 over pure generation), demonstrating that the hybrid approach benefits all models across performance tiers.

The improvements are particularly substantial in low-shot scenarios: at n=1, `gemma-3-27b-it` shows +0.185 improvement over pure generation (0.118 $\rightarrow$ 0.303), `deepseek-v3.2-exp` improves by +0.058, and `gpt-4o-mini` by +0.063. This demonstrates that the BiLSTM provides valuable structural guidance when few examples are available. As n increases, the gains diminish but remain consistent across all models, with `gemma-3-27b-it` showing +0.110 improvement even at n=20.

5.6 Error Analysis

To understand where improvements originate, we analyze errors by morpheme type. Comparing BiLSTM baseline against the best-performing hybrid configuration (`deepseek-v3.2-exp` with $n = 10$ RAG examples), we find asymmetric improvements.

The BiLSTM baseline achieves 0.923 accuracy on 168 grammatical morphemes but only 0.479 on 213 lexical morphemes. This confirms that structured models learn morphological paradigms effectively but struggle with lexical gaps. The hybrid pipeline improves lexical accuracy to 0.682 (+0.204 absolute gain) while grammatical accuracy drops slightly to 0.866 (-0.057). The substantial lexical gains outweigh minor grammatical losses, explaining the overall hybrid improvement.

When we stratify by training frequency, the pattern becomes clearer. We categorize each test morpheme by its frequency in the training data: infrequent (1 to 5 occurrences), common (6 to 20), and frequent (over 20). Table 3 shows accuracy by frequency bin.

Frequency	Count	BiLSTM	Hybrid
Infrequent (1 to 5)	69	0.029	0.426
Common (6 to 20)	58	0.448	0.724
Frequent (over 20)	86	0.860	0.859

Table 3: Accuracy by training frequency for lexical morphemes. Hybrid improvements concentrate in infrequent morphemes (+0.397), with no test morphemes completely unseen in training.

Infrequent morphemes show dramatic improvement: BiLSTM achieves only 0.029 accuracy while the hybrid reaches 0.426 (+0.397). Common mor-

phemes improve from 0.448 to 0.724 (+0.276). Frequent morphemes remain stable around 0.86 for both approaches. Notably, no test morphemes are completely unseen in this split. We did not enforce this property; with a document-level split and a small evaluation set, all test morphemes appear at least once in training. This demonstrates that hybrid gains stem from contextual inference on low-frequency items rather than handling zero-shot vocabulary, and that the baseline already captures frequent grammatical markers effectively when sufficient training examples exist.

5.7 Key Findings Summary

Our experiments reveal several important patterns for LLM-assisted morphological glossing. Retrieval-augmented prompting proves essential across all models, with similarity-based example selection dramatically outperforming random selection. Perhaps most surprisingly, providing morpheme dictionaries generally hurts performance: partial dictionaries universally degrade accuracy, while even complete dictionaries fail to help most models (only one of four shows gains). Performance scales approximately logarithmically with the number of in-context examples, typically peaking around ten to fifteen examples before plateauing or declining.

The hybrid architecture combining BiLSTM predictions with LLM correction improves performance across all tested models. These gains are particularly pronounced in low-shot scenarios where few examples are available, suggesting that the structured model provides valuable guidance when in-context learning is limited. Even without any examples, LLMs can successfully identify and correct many errors in structured predictions, indicating inherent morphological reasoning capabilities.

6 Discussion

6.1 The Case for Hybrid Architectures

Our results demonstrate that combining structured prediction with LLM reasoning yields consistent improvements across all tested models. The BiLSTM-CRF captures frequent morphological patterns from limited training data but struggles with rare morphemes and novel combinations. RAG LLM generation leverages broader linguistic knowledge and few-shot generalization but varies widely in accuracy depending on model choice.

The hybrid pipeline combines these complementary strengths through a division of labor revealed by error analysis. Structured models are more accurate on grammatical morphemes, which likely reflects their high frequency and regularity in the training data, while LLMs excel at contextual inference for infrequent lexical items, leveraging retrieved examples to handle vocabulary gaps. Hybrid improvements concentrate in low-frequency morphemes where BiLSTM lacks sufficient training signal, while both approaches perform similarly on frequent items. This asymmetry explains why the two-stage architecture succeeds: each component addresses the other's primary weakness. BiLSTM predictions provide structural guidance (particularly valuable in low-shot scenarios) while LLMs refine these predictions using patterns from retrieved examples. Our prompt design frames the BiLSTM output as a fallible hypothesis rather than an authoritative baseline, encouraging critical evaluation while providing useful structural constraints.

While we do not claim that BiLSTM-CRF represents the optimal base model for this task, our results suggest a broader principle: combining any trainable structured predictor with retrieval-augmented LLM post-correction can yield gains over either approach alone. This synergy proves universally beneficial across all performance tiers, with particularly strong gains when in-context learning is limited by few examples or budget constraints.

6.2 The Role of Morpheme Dictionaries

Perhaps our most surprising finding is that providing morpheme dictionaries generally hurts performance compared to providing no dictionary at all. Partial glossaries universally degrade accuracy, while even complete dictionaries fail to help most models. Only one model (`gemma-3-27b-it`) shows substantial gains from the complete dictionary, while three others perform worse with it than without.

These patterns suggest issues with how models integrate dictionary information, though we lack direct analysis of whether models actually consult dictionary entries. The degradation could stem from information overload, inappropriate dictionary consultation, or suboptimal prompt structure. Our approach provides dictionaries as simple unstructured key-value lists; alternative strategies merit investigation, such as organizing entries by morphological class or presenting dictionaries in structured formats.

Importantly, prompt structure can dramatically affect model behavior, and we have not systematically explored alternative formulations. Our findings should therefore be interpreted as demonstrating the ineffectiveness of our specific prompt design for dictionary integration, rather than fundamental limitations of dictionary use in morphological glossing. Future work should conduct systematic prompt engineering studies to identify more effective strategies for incorporating lexical resources.

6.3 Practical Implications for Fieldwork

Our findings suggest several considerations for practitioners working with LLM-assisted IGT annotation, though these patterns may vary across different languages, models, and documentation contexts. In our experiments, similarity-based retrieval consistently outperformed random example selection across all tested models, suggesting that investment in retrieval infrastructure may be worthwhile. The hybrid approach combining structured models with LLM post-correction showed gains across all configurations we tested, indicating potential value in this two-stage strategy.

For morpheme dictionaries, our results suggest caution: simple key-value presentations tended to hurt performance for most models in this setting, though alternative presentation strategies remain unexplored. Regarding few-shot example count, we observed approximately logarithmic scaling with diminishing returns beyond 10–15 examples, though optimal operating points likely depend on task complexity, model capabilities, and cost constraints. These patterns emerged from our specific experimental setup with Tuvan and four general-purpose LLMs; practitioners should validate these findings against their own languages and workflows, as model capabilities and architectural designs continue to evolve rapidly.

6.4 Ethical Considerations

Our experiments use data collected with informed consent under agreements restricting raw material sharing. We report only aggregate statistics and anonymized examples to protect speaker privacy. Speaker communities are not monolithic, and Tuvan speakers are distributed across multiple countries with different sociopolitical contexts. We do not claim community-wide consent; we follow the data agreements and consult the relevant fieldwork partners. Given the cross-border context, we avoid releasing raw data and refrain from identifying individuals or locations. Commercial APIs raise concerns about training data provenance (Bender et al., 2021; Sainz et al., 2023), and even at achieved accuracy levels, uncritical adoption risks introducing errors into the linguistic record. Decisions about automation should be made collaboratively with stakeholders, balancing efficiency gains against concerns about data control and quality.

6.5 Future Directions

Key directions include cross-linguistic evaluation on diverse morphological systems (polysynthetic, templatic, tonal) to test whether our design principles generalize beyond Turkic agglutination. Joint segmentation and glossing would address the limitation that our pipeline assumes gold boundaries. Parameter-efficient fine-tuning could improve performance with minimal language-specific data. Interactive interfaces with confidence estimation would help annotators prioritize review of uncertain predictions, and paradigm-level evaluation would better assess morphological generalization beyond token-level accuracy.

7 Conclusion

We present a hybrid automatic glossing pipeline combining BiLSTM-CRF structured prediction with LLM post-correction, evaluated on Tuvan fieldwork data across four LLMs. The hybrid approach consistently improves performance across all tested models, with particularly strong gains in low-shot scenarios where structural guidance from the BiLSTM proves most valuable. Error analysis reveals that improvements concentrate in lexical morphemes, especially rare vocabulary items, while BiLSTM already captures grammatical paradigms effectively. This demonstrates complementary strengths rather than simple performance stacking.

Our ablation studies reveal key design principles: retrieval-augmented prompting provides substantial gains; morpheme dictionaries generally hurt performance for most models; performance scales logarithmically with examples, plateauing around ten to fifteen; and hybrid correction benefits all models universally. These findings challenge conventional assumptions about prompt engineering, showing that more information is not always better.

While the utility of these accuracy levels for practical annotation workflows remains an open

question dependent on community priorities and annotation contexts, the substantial error rates necessitate careful human oversight. Model outputs should be treated as hypotheses requiring expert validation rather than authoritative annotations.

8 Limitations

Our evaluation focuses on a single language (Tuvan) from one language family (Turkic), leaving generalization to other morphological systems untested. The patterns we observe may not hold for polysynthetic, templatic, or non-concatenative morphology. Additionally, our test set is small, reflecting realistic annotation costs but introducing sampling variance that limits statistical precision.

Our evaluation metrics capture only token-level accuracy rather than higher-order properties like paradigm consistency, morphophonological regularity, or alignment with community language priorities. The pipeline also assumes gold morpheme boundaries and does not address the segmentation problem, which itself requires substantial linguistic expertise in fieldwork contexts.

The prompt engineering component of our study lacks systematic exploration. We tested only one prompt structure for each experiment, leaving unexplored how alternative formulations might affect dictionary effectiveness, example integration, or instruction following. Prompt design choices (information ordering, instruction phrasing, formatting conventions, and the balance of different knowledge sources) can dramatically impact model behavior, yet we lack the controlled comparisons needed to isolate their effects. Our glossary ablation findings in particular should be interpreted as showing that our specific prompt design failed to effectively leverage dictionary information, rather than demonstrating that dictionaries cannot help morphological glossing. We also lack fine-grained analysis of whether models actually consult dictionary entries or become distracted by additional prompt content.

Commercial API access limits transparency about training data and potential contamination from existing Tuvan linguistic resources. We use general-purpose LLMs rather than specialized models like GlossLM (Ginn et al., 2024b), which are explicitly trained on IGT data and may achieve higher absolute accuracy. However, our focus on establishing design principles (retrieval strategies, glossary configurations, hybrid architectures) likely generalizes across model types. Moreover, general-purpose LLMs offer practical advantages for fieldwork: no local infrastructure requirements, operation in few-shot regimes with minimal language-specific data, and accessibility to linguists without machine learning expertise. The tradeoff between specialized performance and practical accessibility merits further investigation.

References

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic Interlinear Glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Jan Buys and Jan A. Botha. 2016. Cross-Lingual Morphological Tagging for Low-Resource Languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany. Association for Computational Linguistics.

Shobhana L. Chelliah and Willem J. de Reuse. 2010. *Handbook of Descriptive Linguistic Fieldwork*. Springer Science & Business Media. Google-Books-ID: d1FfE30hZ7EC.

Bernard Comrie, Martin Haspelmath, and Bickel Balthasar. 2008. The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses. Google-Books-ID: e8B7AQAACAAJ.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual Character-Level Neural Morphological Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui

Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Micha Elsner and David Liu. 2025. Prompt and circumstance":" A word-by-word LLM prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*, pages 1–14, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024b. GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Sharon Hargus, Olga M. Semenova, and Siri G. Tuttle. 2020. Glossing Dene Languages. Publisher: Alaska Native Language Center.

K. David Harrison and Gregory D. S. Anderson. 2002. *A Grammar of Tuvan*. Scientific Consulting Services International. Google-Books-ID: RXnhAQAACAAJ.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating Interlinear Glossed Text from Speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint*. ArXiv:1508.01991 [cs].

SIL International. 2025. FieldWorks Language Explorer.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24(251):1–43.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry,

Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Christian Lehmann. 2004a. Data in linguistics. 21(3-4):175–210. Publisher: De Gruyter Mouton Section: The Linguistic Review.

Christian Lehmann. 2004b. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas, and Wolfgang Kesselheim, editors, *Morphologie*, pages 1834–1857. De Gruyter.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological Segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural Factor Graph Models for Cross-lingual Morphological Tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Talant Mawkanuli. 1999. *The phonology and morphology of Jungar Tuva*. Ph.D., Indiana University, United States – Indiana. ISBN: 9780599667938.

Talant Mawkanuli. 2005. *Jungar Tuvan Texts*. Uralic and Altaic Series. Indiana University, Bloomington.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From Interlinear Glossed Texts to Paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Sebastian Nordhoff and Thomas Krämer. 2022. IMT-Vault: Extracting and Enriching Low-resource Language Interlinear Glossed Text from Grammatical Descriptions and Typological Survey Articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

Shu Okabe and François Yvon. 2023. Towards Multilingual Interlinear Morphological Glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Enora Rice, Ali Marashian, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2024. TAMS: Translation-Assisted Morphological Segmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6752–6765, Bangkok, Thailand. Association for Computational Linguistics.

Enora Rice, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11284–11296, Suzhou, China. Association for Computational Linguistics.

Erich Round, Mark Ellison, Jayden Macklin-Cordes, and Sacha Beniamine. 2020. Automated Parsing of Interlinear Glossed Text from Page Images of Grammatical Descriptions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2878–2883, Marseille, France. European Language Resources Association.

Tatyana Ruzsics and Tanja Samardžić. 2017. Neu-

ral Sequence-to-sequence Learning of Internal Word Structure. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 184–194, Vancouver, Canada. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Mathias Schenner and Sebastian Nordhoff. 2016. Extracting Interlinear Glossed Text from LaTeX Documents. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4044–4048, Portorož, Slovenia. European Language Resources Association (ELRA).

Bhargav Shandilya and Alexis Palmer. 2023. Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–92, Toronto, Canada. Association for Computational Linguistics.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11:1–17. Place: Cambridge, MA Publisher: MIT Press.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN : a professional framework for multimodality research. pages 1556–1559.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025b. LingGym: How Far Are LLMs from Thinking Like Field Linguists? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1314–1340, Suzhou, China. Association for Computational Linguistics.

Wenhao Yu. 2022. Retrieval-augmented Generation across Heterogeneous Knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 52–58, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural Machine Translation Quality and Post-Editing Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experimental Prompt Templates

This appendix provides the complete prompt templates used across all five experiments. All prompts follow a consistent structure with a system message defining the linguistic expert role, followed by task-specific instructions and formatting requirements.

A.1 System Message (All Experiments)

All experiments use the following system message:

> *You are a linguistic expert specializing in morpheme-by-morpheme glossing for an unknown language.*

A.2 Experiments 1–3: RAG LLM generation

Experiments 1 (Retrieval vs. Random), 2 (Glossary Ablation), and 3 (N-Shot Scaling) use RAG LLM generation with the following template:

> **User message:**
>
> Here are some examples of sentences with morpheme boundaries (marked by hyphens) and their glosses:
>
> *[For each example i in 1..k:]*
>
> Example *i*:
>
> Segmented: *[morpheme-segmented source text]*
>
> Gloss: *[corresponding gloss sequence]*
>
> *[If glossary provided:]*
>
> You are also given a morpheme dictionary mapping morphemes to their English glosses. For morphemes in the dictionary, use the provided gloss. For others, infer from context. Some morphemes may have multiple translations; choose the most appropriate for this context.
>
> Morpheme dictionary: *[morpheme1: gloss1, morpheme2: gloss2, ...]*
>
> *[End glossary section]*
>
> Please gloss this sentence:
>
> Segmented: *[test sentence with morpheme boundaries]*
>
> Output the gloss with the same structure (spaces between words, hyphens between morphemes). Enclose your gloss in ###. Example: ###word1-MORPH1 word2-MORPH2###

Experiment-specific variations:

- **Experiment 1:** Uses 3 examples, no glossary. Compares TF-IDF retrieval (RAG) against random sampling.
- **Experiment 2:** Uses RAG, no glossary. Varies $k \in \{1, 3, 5, 10, 15, 20\}$ examples.
- **Experiment 3:** Uses 3 RAG examples. Tests four glossary sizes: none (0 pairs), top-100 (100 pairs), grammatical (240 pairs), entire (1,498 pairs).

A.3 Experiment 4: Hybrid Pipeline (BiLSTM + LLM)

Experiment 4 uses BiLSTM-CRF predictions as initial hypotheses for LLM correction, with the following template:

> **User message:**
>
> You will be given:
>
> 1. A rough initial glossing attempt from a statistical model (may contain errors)
> 2. Some example sentences with their correct glosses
> 3. A morpheme dictionary
>
> Your task is to produce the correct gloss, using all available information.
>
> Here are some example sentences with correct glosses:
>
> *[For each example i in 1..k (or 0 for zero-shot):]*
>
> Example *i*:
>
> Segmented: *[morpheme-segmented source text]*
>
> Gloss: *[corresponding gloss sequence]*
>
> *[If glossary provided:]*
>
> You also have access to a morpheme dictionary: *[morpheme1: gloss1, morpheme2: gloss2, ...]*
>
> *[End glossary section]*
>
> Now, please gloss this sentence:
>
> Segmented: *[test sentence with morpheme boundaries]*
>
> Initial attempt (from statistical model): *[BiLSTM-CRF prediction]*
>
> This initial attempt may contain errors. Use the examples, dictionary, and linguistic patterns to produce the correct gloss. Maintain the same structure (spaces between words, hyphens between morphemes).
>
> IMPORTANT: Output ONLY the gloss wrapped in ###. Do not explain your

reasoning. Example format: ###word1-MORPH1 word2-MORPH2###

Design rationale: The hybrid prompt presents the BiLSTM prediction as a "rough initial attempt" rather than an authoritative baseline, encouraging the LLM to critically evaluate and correct errors. Clean gold examples (not error-correction pairs) provide paradigmatic context, while the statistical prediction narrows the search space by proposing plausible morpheme boundaries and candidate glosses.

A.4 Output Extraction

All experiments extract model outputs by locating text between ### delimiters. If delimiters are absent (indicating non-compliance), the entire output string is used as the predicted gloss. This extraction method proved robust across all four LLM providers.

B Detailed Results Tables

This section provides detailed numerical results for experiments presented as figures in the main text.

Glossary	deepseek	qwen3	gpt-4o	gemma-3
None	**0.510**	**0.436**	**0.420**	0.325
Top-100	0.469	0.332	0.353	0.430
Grammatical	0.426	0.303	0.364	0.335
Entire	0.490	0.422	0.369	**0.455**

Table 4: Glossary ablation study across four LLMs (3-shot RAG). Partial glossaries consistently degrade performance, while complete dictionaries fail to help most models. Model names abbreviated: deepseek = `deepseek-v3.2-exp`, qwen3 = `qwen3-max`, gpt-4o = `gpt-4o-mini`, gemma-3 = `gemma-3-27b-it`.

N	deepseek	qwen3	gpt-4o	gemma-3
1	0.350	0.196	0.159	0.118
3	0.505	0.439	0.383	0.336
5	0.588	0.519	0.423	0.471
10	**0.658**	0.529	0.430	**0.600**
15	0.646	**0.611**	**0.486**	0.540
20	0.626	0.552	0.475	0.534

Table 5: N-shot scaling for RAG LLM generation (RAG, no glossary). Performance improves logarithmically, with diminishing returns beyond n=10–15. Model names abbreviated as in Table 4.

Linguistically Informed Tokenization Improves ASR for Underresourced Languages

Massimo Daul
New York University
mmd9604@nyu.edu

Alessio Tosolini
McGill University
alessio.tosolini@mail.mcgill.ca

Claire Bowern
Yale University
claire.bowern@yale.edu

Abstract

Automatic speech recognition (ASR) is a useful tool for linguists aiming to perform a variety of language documentation tasks. However, modern ASR systems use data-hungry transformer architectures, rendering them generally unusable for underresourced languages. We fine-tune a wav2vec2 ASR model on Yanhangu, an Indigenous Australian language, comparing the effects of phonemic and orthographic tokenization strategies on performance. In parallel, we explore ASR's viability as a tool in a language documentation pipeline. We find that a linguistically informed phonemic tokenization system substantially improves WER and CER compared to a baseline orthographic tokenization scheme. Finally, we show that hand-correcting the output of an ASR model is much faster than hand-transcribing audio from scratch, demonstrating that ASR can provide significant assistance for underresourced language documentation.

1 Introduction

Automatic Speech Recognition (ASR, also known as speech-to-text) is a natural language processing technology that converts spoken words to text. Most modern ASR systems, such as wav2vec2 (Baevski et al., 2020) and Whisper (Radford et al., 2022) are neural and transformer-based, working well for languages with large amounts of training data but falling short for those without. ASR is used for a broad range of human-computer interaction tasks, but there exists a big resource gap, where only a small number of languages have robust and freely available ASR models. Common Voice (Ardila et al., 2020), for example, covers only 2% of the world's languages. The geographic distribution of these languages is also unequal, with no Indigenous Australian languages represented in this corpus.

One reason for this asymmetry is that for low-resource languages, ASR training data comes from linguistic fieldwork where manual transcription and annotation are both time-consuming and require specialist knowledge. Few documentation projects have the resources to train and pay annotators (Chelliah, 2001), with transcription taking anywhere from 5 minutes to over an hour per minute of speech (Dwyer, 2006). That means that there is both much less training data for underresourced languages, and that the means to acquire such training data is extremely labor-intensive. Accurate ASR would vastly assist in this respect.

ASR models are trained to predict sequences of tokens: discrete textual units. In the context of ASR, tokenization refers to the process of segmenting transcribed speech into meaningful units, such as words, subwords, or characters. Different tokenization strategies may impact vocabulary size, error rates, and adaptability to various languages and domains, with the best tokenization strategies for high-resource languages not necessarily being the same as those for low-resource languages (Adlaon and Marcos, 2024; Bañeras-Roux et al., 2024). Additionally, linguistically informed tokenization strategies – i.e., ones where the tokenization occurs across phonological (Atuhurra et al., 2024; Liao and Shi, 2026) or morphological (Bayram et al., 2025; Hofmann et al., 2021) units – have been shown to improve model performance for some low-resource tasks. This paper investigates whether linguistically informed phonemic tokenization improves ASR accuracy for Yanhangu, and evaluates its practical impact within a fieldwork documentation pipeline. We show that linguistically informed ASR improves error rates and substantially speeds up transcription. Furthermore, linguistically transparent tokenization aligns more closely with fieldwork pipelines by enabling local, interpretable computation, thereby supporting data sovereignty.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 31–37

2 Methodology

2.1 Data origin and preprocessing

The corpus for this test includes selected recordings of the Yan-nhangu language (Glottolog code YANN1237; ISO-639 JAY; Pama-Nyungan). Recordings were made from 5 fluent speakers of Yan-nhangu between 2004 and 2007. Since the Yan-nhangu data originated from different elicitation sessions, the recording quality is variable, though the recordings being made in the field allows the results from this experiment to be generalized to other field environments. The data was preprocessed to ensure consistency. For information regarding speaker demographics, recording and archiving information, and pre-processing see Appendix A.

2.2 Tokenization and Acoustic Models

Yan-nhangu is customarily written with a combination of Latin letters (e.g. *y*, *l*), accented letters (e.g. *ä* and *ṯ*), digraphs (e.g. *nh*, *th*), and characters not otherwise used in English standard segmental orthography (*ŋ*, *'*). There are 25 consonants and 6 vowels. For a complete consonant and vowel chart in IPA and orthography, see Appendix B. Note that for all figures in Section 3.2, capitalized tokens in the phonemic models represent apical consonants (e.g. "*N*" stands for n̪) or long vowels (e.g. "*A*" stands for aː) to emphasize the one-to-one mapping between Yan-nhangu phone and phonemic token. Additionally, whitespaces are represented by underscores.

Two identical models were trained on the same data with two contrasting tokenization methods. In the first class of models, the token was defined as the grapheme, splitting digraphs like *ny* /ɲ/ into *n* and *y*. The orthography is the same as the Yolŋu Matha orthography defined in (Zorc, 1996). In the second class of models, the token was defined as the phoneme, such that orthographic digraphs representing one phone like *ny* /ɲ/ are one token: *ɲ*. The analysis of Yan-nhangu phonology is based on the most detailed existing documentation of Yan-nhangu phonology (Baymarrwaŋa et al., 2006).

We use Wav2Vec2-BERT 2.0 model (Chung et al., 2021) accessed from HuggingFace, which has been pre-trained in a self-supervised manner on Facebook's multilingual corpus. Although this pretraining corpus spans over 140 languages, it does not include Indigenous Australian languages or typologically similar phonological systems, which suggests limited zero-shot transfer and motivates targeted fine-tuning in this setting. Our training dataset consists of up to 156 minutes of speech, and we optimize the model using Connectionist Temporal Classification (CTC) loss with an 80/20 train-validation split. Training is performed in Jupyter notebooks on one RTX-8000 GPU. [1] Each model is trained for 16 epochs using a linear learning rate scheduler, an initial learning rate of 1e-5, and early stopping to prevent overfitting. Training takes less than two hours per model.

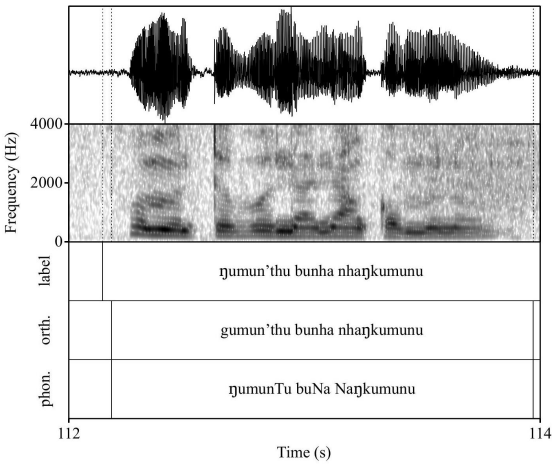


Figure 1: Waveform, spectrograph, and annotations for a sample testpoint.

2.3 Evaluation

Word Error Rate (WER) and Character Error Rate (CER) were used as evaluation metrics. We use CER when selecting top-performing models since post-alignment editing is common in language documentation workflows, meaning a lower CER reflects a greater speedup during manual correction compared to a lower WER. Models with orthographic and phonemic transcription schemes were trained and evaluated on 10, 30, 60, 90, 120, and 156 minutes of data. A qualitative analysis of the automatic transcription errors most common across models was also performed and reported on in Section 3.3. Following previous ASR error analysis research (Errattahi et al., 2018), we look at the Levenshtein distance between a held-out testing set and the best orthographic and phonemic ASR models' transcriptions. Finally, the last author manually corrected four minutes of automatically transcribed Yan-nhangu speech, such as

[1] The codebase is currently under active development as part of a broader open-source tool and will be released publicly following the associated corpus release.

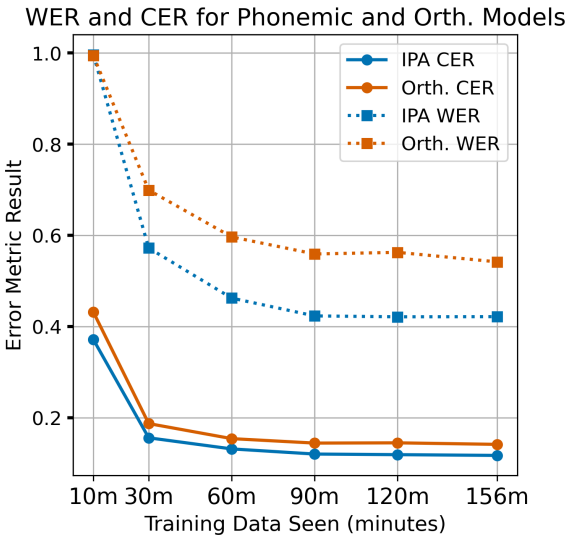


Figure 2: CER and WER metrics for phonemic and orthographic models across training set sizes.

Figure 1. Qualitative interpretations about the errors encountered were documented and the time saved between correcting automatic transcription and manually transcribing from scratch are discussed.

3 Results

3.1 WER and CER Across Models

Our results show that phonemic tokenization improves low-resource ASR performance. Figure 2 shows a consistent gap in WER and CER on held-out validation data, with the phonemic model outperforming the orthographic model across all training set sizes.

3.2 Levenshtein Distance Analysis of Errors

Analyzing the Levenshtein distance between manual annotations and the best orthographic and phonemic ASR outputs allows a more detailed investigation into model differences. The total number of errors made on the validation set are described in Table 1. Although the phonemic model shows lower overall Levenshtein distance from the manually transcribed label than the orthographic model, substitutions seem to be more frequent for the phonemic model.

The bar charts in Figure 3a and Figure 3d show the frequency of each deleted character. For both orthographic and phonemic models, spaces and short vowels are among the most commonly deleted characters. This is consistent with word boundaries being unclear during rapid speech, and

Model Type	Dels	Interts	Subs	Total
Phonemic	433	454	547	1434
Orthographic	624	592	438	1654

Table 1: Summary of Levenshtein distance between human annotated and ASR transcription by error type: deletions, intertions, and substitutions

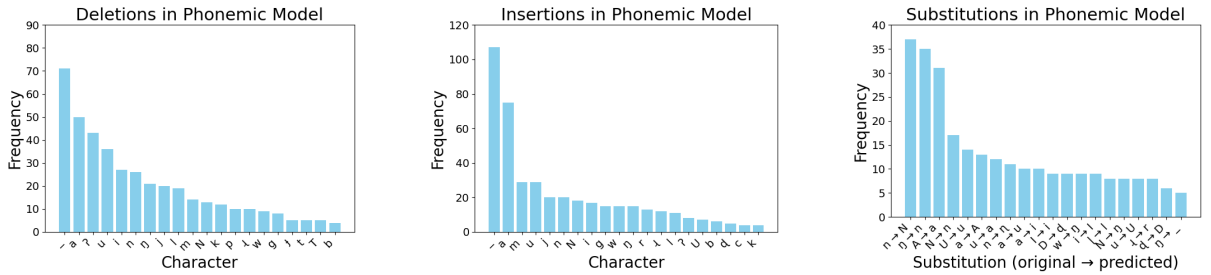
with vowel elision that occurs in Yan-nhangu, especially with *a*. Interestingly, the orthographic model frequently deletes tokens *n* and *h*, both of which are present in multiple digraphs. Across both models, sonorants (nasals, liquids, and vowels) are deleted much more frequently than stops.

Figure 3b and Figure 3e show the most frequently inserted characters, which mirror the most frequently deleted characters across both models. As such, whitespaces and *a* are the most frequent insertions, with *h* being the third most commonly inserted character for the orthographic model. Like with deletions, sonorants are inserted more frequently than stops.

Lastly, Figure 3c and Figure 3f show the most common token substitutions for the phonemic and orthographic models. This is the only setting in which we see a greater rate of errors for the phonemic model due to phonemic substitutions appearing as insertions or deletions for the orthographic model (e.g. $n \rightarrow N$, corresponding to $n \rightarrow nh$: in the orthographic model, this appears as an insertion). Another instance of ambiguous phones showing up as substitutions of phonologically similar phones include long vowels being substituted for their short counterparts, which occurs with an approximately equal and high frequency in both models. However, a stark asymmetry in substitution direction occurs with *n* and *ŋ*, where $ŋ \rightarrow n$ is the second most common substitution in the phonemic model, while $n \rightarrow ŋ$ is the second most common substitution in the orthographic model.

3.3 Further Comments on Errors

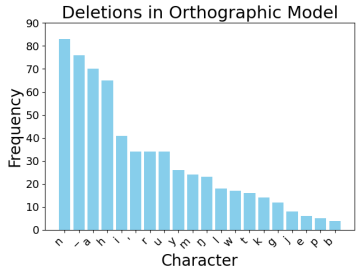
The last author, who is very familiar with Yan-nhangu, manually corrected 4 minutes of ASRed Yan-nhangu speech. It took 20 minutes to review each transcribed utterance to check, correct, and categorize the errors. At five minutes per minute of transcript, this is equal to the fastest unassisted transcription rate. Without using the ASR technology, it would take the last author approx. 15 minutes to transcribe 1 minute of Yan-nhangu, meaning that the introduction of ASR technology in



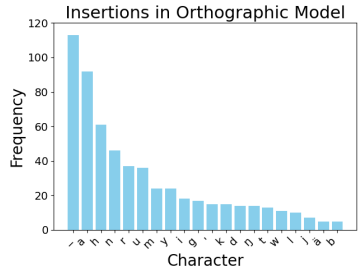
(a) Deletions for phonemic tokenization

(b) Insertions for phonemic tokenization

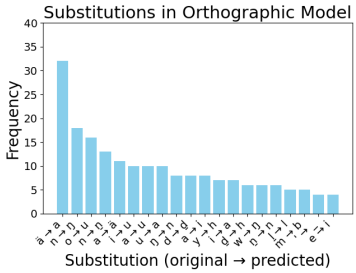
(c) Substitutions for phonemic tokenization



(d) Deletions for orthographic tokenization



(e) Insertions for orthographic tokenization



(f) Substitutions for orthographic tokenization

Figure 3: Counts for deletions, insertions, and substitutions for the best ASR models using phonemic and orthographic tokenization.

the transcription pipeline resulted in a three times speedup.

In the stretch of recording considered, errors fell into three (quantitative) categories. The first was grapheme substitutions, such as the word *d̲iltji* 'bush' was transcribed as *diltji*. The second set of "errors" involved items where ASR correctly rendered the material in the recording, but the transcription did not adhere to Yan-nhangu orthographic norms, such as word break errors or vowels in words not being transcribed due to them not being pronounced. Finally, there were 3 cases where, to the transcriber's ear, both the ASR word and the label word were possible representations of the recorded speech. The original transcripts were made in 2004–2008 and discussed with native speakers of Yan-nhangu, but there are several near homophone words that were, in context, also semantically plausible.

4 Discussion and Conclusions

This paper investigates whether phonologically informed tokenization improves ASR accuracy for the low-resource Australian language Yan-nhangu, providing the first comparative analysis of tokenization strategies for any Aboriginal language. Results show that phone-level tokenization improves WER and CER over a grapheme-level baseline for models trained on at least 30 minutes of data and improves CER across all models, suggesting that phonemic tokenization provides a more linguistically transparent representation of the language, reducing ambiguity and supporting more effective generalization in low-resource conditions. Models approach peak performance with approx. 90 minutes of training data, with the orthographic model showing more room for improvement. These findings align with prior research demonstrating the benefits of linguistically informed tokenization in low-resource NLP (Atuhurra et al., 2024).

We find that changing the tokenization scheme only results in meaningful differences in substitutions, insertions, and deletions if a speech sound is represented differently by the tokenizers. In Yan-nhangu, this occurs for single phonemes that are orthographic digraphs. We hypothesize that the substitution asymmetry involving *ŋ* and *n* arises because the orthographic model represents /n̪/ and /ɲ/ using the digraphs *nh* and *ny*, resuling in a single phoneme being encoded as two tokens. As a result, the probability of *n* becomes correlated with h or y. When the model encounters the velar nasal ŋ, the likelihoods of h and y (which cue apicals and palatals) decrease, narrowing down the places of articulation for the nasal leading to more frequent

substitutions of ŋ with *n*. This effect does not arise in the phonemic model, which lacks digraphs.

Speech transcription is essential but time-consuming in language documentation. While many factors affect transcription time, this experiment shows that correcting high-quality ASR transcriptions is about three times faster than manual transcription for someone familiar with the language. Based on the authors transcription rate, ASR-assisted transcription could provide a 10-hour speedup per hour of unannotated speech. This efficiency is crucial for developing resources for the world's most under-documented languages.

Limitations

This study is limited by its focus on a single language, Yan-nhangu. While this reflects realistic conditions in language documentation, the results may not directly generalize to languages with different phonological or orthographic properties. Similarly, by only testing on wav2vec2, we are limited in our ability to generalize these results to other model architectures. Additionally, the qualitative transcription speedup analysis is based on a single experienced annotator, and results may vary with regards to numerous factors which may lead to faster or slower transcription rates, including familiarity with the language, number of speakers, speech rate of participants, or complexity of the subject matter. Finally, our tokenization comparison is restricted to phonemic and orthographic representations, and does not explore alternative subword approaches (Bayram et al., 2025; Si et al., 2023), which remain an important direction for future work.

Ethical Considerations

While ASR-assisted transcription and annotation may increase the speed of language documentation workflows, its use raises important ethical considerations. Language documentation and fieldwork are often embedded within broader language revitalization efforts, where goals extend beyond corpus creation to include community engagement, skill development, and cultural preservation. In such contexts, the appropriateness of automated methods may depend on both the intent of the project and the nature of the materials being transcribed.

Prior work has shown that community members may prefer manual transcription even when ASR systems offer a faster alternative. For example, Prud'hommeaux et al. (2021) report that Indigenous participants preferred to transcribe from scratch, despite recognizing the efficiency of ASR-assisted transcription. This community emphasized the role of transcription in strengthening language proficiency through close engagement with the language. Participants also differentiated between content, remarking that ceremonial recordings should be transcribed manually, while less sensitive materials, such as childrens stories, may be more appropriate for ASR-assisted workflows.

In light of these considerations, linguists must work with language community members to identify the goals of the documentation project and how the methodology involved works to achieve them.

References

Kristine Mae M. Adlaon and Nelson Marcos. 2024. Finding the optimal byte-pair encoding merge operations for neural machine translation in a low-resource setting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14673–14682, Miami, Florida, USA. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Jesse Atuhurra, Hiroyuki Shindo, Hidetaka Kamigaito, and Taro Watanabe. 2024. Introducing syllable tokenization for low-resource languages: A case study with swahili. *Preprint*, arXiv:2406.15358.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Thibault Bañeras-Roux, Mickael Rouvier, Jane Wottawa, and Richard Dufour. 2024. A Comprehensive Analysis of Tokenization and Self-Supervised Learning in End-to-End Automatic Speech Recognition applied on French Language. In *32th European Signal Processing Conference (EUSIPCO)*, Lyon, France.

Laurie Baymarrwaŋa, Rita Gularrbanga, Laurie Milinditj, Rayba Nyaŋbal, Margaret Nyuŋunyuŋu, Allison Warrŋayun, and Claire Bowern. 2006. A learners guide to yan-nhaŋu.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümü, Sercan Karaka, Banu Diri, Sava Yldrm, and

Demircan Çelik. 2025. Tokens with meaning: A hybrid tokenization approach for nlp. *Preprint*, arXiv:2508.14292.

Shobhana L Chelliah. 2001. *The role of text collection and elicitation in linguistic fieldwork*, pages 152–165. Cambridge University Press.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Arienne M Dwyer. 2006. *Ethics and practicalities of cooperative fieldwork and analysis*, page Chapter 2. Mouton.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37. 1st International Conference on Natural Language and Speech Processing.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Disen Liao and Freda Shi. 2026. How tokenization limits phonological knowledge representation in language models and how to improve them. In *Tokenization Workshop*.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Kelly Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513. University of Hawaii at Mānoa.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Sub-character tokenization for chinese pretrained language models. *Preprint*, arXiv:2106.00400.

P Wittenburg, H Brugman, A Russel, A Klassman, and H Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*, pages 1556–1559. 00216.

R. David Paul Zorc. 1996. Yolngu-matha dictionary / r. david zorc.

A Additional Data Information

A.1 Speaker Demographics

All the data used in this experiment come from 5 women recognised by their community as authorities to speak about the language. At the time of recording they ranged in age from early 40s to early 70s. Four of the five were from Nhaŋu clans; the fifth was Warrawarra. The traditional lands of Yan-nhangu people are the Crocodile Islands of Northeast Arnhem Land (Northern Territory, Australia).

A.2 Recording and Archiving

Recordings were made with an Edirol R-01 solid state recorder and external microphone. They were recorded in a range of locations in Milingimbi Aboriginal Community and on Murruŋga Island. Language tasks included a range of structured and semi-structured elicitation tasks, including wordlist and sentence translation, descriptions of pictures and video clips, and discussion prompts around cultural concepts and practices.

Materials were transcribed in Elan (Wittenburg et al., 2006) by the last author and checked with speakers. Recordings, transcripts, and field notes have been deposited with the ELAR digital archive and the AIATSIS library. These materials are not publicly available due to usage agreements that respect Indigenous intellectual property; however, this work was conducted under agreements that permit the use of the recordings to support Yan-nhangu and other Indigenous language research.

A.3 Preprocessing

Incomplete transcriptions and audio containing English words were excluded, along with segments without transcriptions, leaving about 156 minutes of training data. Punctuation was removed, except for apostrophes, which denote glottal stops in Yan-nhangu orthography. The original transcriptions were all produced in Yan-nhangu orthography. Since the phonemic representation of Yan-nhangu words is predictable from the orthography, an automated script was used to generate phonemic transcriptions for all annotations used in training the phonemic models

p	t	ʈ (t̲)	t̪ (th)	c (tj)	k	ʔ (’)
b	d	ɖ (d̲)	d̪ (dh)	ɟ (dj)	g	
m	n	ɳ (n̲)	n̪ (nh)	ɲ (ny)	ŋ	
	l	ɭ (l̲)				
	r (rr)	ɹ (r)				
w				j (y)		

Table 2: Consonant inventory of Yan-nhangu

i	u	iː (e)	uː (o)
a		aː (ä)	

Table 3: Vowel inventory of Yan-nhangu

B Yan-nhangu Consonant and Vowel Inventory

Consonant and vowel charts are given below in Table 2 and Table 3 respectively in IPA, with orthography in parenthesis where different. Note that Yan-nhangu's orthography is the same as that used by the other Yolŋu languages of Arnhem Land.

Short-form verbal arts as a speech data resource in the field

Matthew Faytak[1], Tianle Yang[1], Pius W. Akumbu[2],
Ivo Forghema Njuasi[3], Éric Le Ferrand[1]
[1]University at Buffalo, USA,
[2]LLACAN - CNRS, France
[3]University of Buea, Cameroon

Abstract

We propose a method for efficient field data collection of speech resource data which leverages short-form verbal arts, namely riddles and proverbs, which permit a predictable transcript to be assigned to naturalistic but conventionalized utterances. As a proof of concept, we describe a 5.25 hour corpus of proverbs and riddles collected for Kom, a low-resource language of Cameroon, and conduct ASR modeling experiments on the corpus. Results suggest that the proposed method yields high quality speech data, albeit with relatively low lexical diversity. We highlight the alignment of the collected data with community priorities for cultural education and preservation in the Cameroonian context.

1 Introduction

Recent advances in natural language processing, particularly the development of foundation models and transfer learning techniques, have made language technologies more accessible to a wider range of languages. The amount of data required to train state-of-the-art architectures has greatly decreased, but *some* annotated data remains necessary for model training. While high-resource languages benefit from abundant, diverse online data, many low-resource languages studied by field linguists have no available online data sets. As a result, field collection of data resources remains essential to the development of automatic speech recognition (ASR) systems, which are increasingly used to facilitate further linguistic data collection (Seifart et al., 2018; Michaud et al., 2018).

Naturalistic, spontaneous speech data is commonly collected in the field, given the difficulty of implementing controlled tasks and engaging speakers. However, these generate less predictable speech content which can only be deciphered with extensive work from expert transcribers (Himmelmann, 2018). In this paper, we introduce a novel method for efficient field collection of speech resource materials based around short-form verbal arts: namely, proverbs and riddles, which are not only easily obtained and efficiently transcribed, but also better aligned with community interests and priorities. We present as proof of concept a speech corpus created for Kom (Grassfields Bantu, Cameroon; ISO 639-3: bkm) using this approach.

In this paper, we first review the literature on challenges related to field data collection and ASR for field linguistics. We then describe in detail our use-case language, Kom, our data collection method, and aspects of the short-form verbal arts materials at issue. Finally, we present ASR experiments conducted on the collected corpus, which suggest generally good quality of the collected data but relatively low word character diversity.

2 Related work

Datasets collected for ASR in well-resourced languages often rely on speech or text data generated by institutions or media such as parliamentary discussions, audiobooks, or radio broadcasts (Wang et al., 2021; Panayotov et al., 2015; Kocabiyikoglu et al., 2018; Gelas et al., 2012). For low-resource languages, the range of available speech data sources is drastically reduced and is often limited to religious texts (Black, 2019; Zanon-Boito et al., 2020; Pratap et al., 2024), models trained on which may not extend well to other domains (Le Ferrand et al., 2025). While training models in low-resource contexts has become easier with the advent of transformers-based architectures – foundation models fine-tuned on minimal data have been reported to have good performances across domains (Havard et al., 2025; Billings and McDonnell, 2025; Geng et al., 2025) – collecting new datasets for fine-tuning is still a challenge for many languages and a popular research topic (Taguchi et al., 2024; Ngue Um et al.,

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 38–45
March 26, 2026.

2025).

Linguistic fieldwork is often the only source of materials for speech resource development (Michailovsky et al., 2014; Paschen et al., 2020; Tapo et al., 2024). Corpora of naturalistic, spontaneous speech are typically the major targets of this work. This reflects the difficulty of controlling the field recording environment, as well as the fact that speakers may be difficult to engage in repetitive or highly controlled tasks (Bowern, 2015; Le Ferrand et al., 2022). It also reflects the priority of naturalistic speech events in language documentation (Lüpke, 2010; Woodbury, 2011). However, development of annotated speech corpora is greatly facilitated by predictable transcripts, which spontaneous speech lacks. The resulting *transcription bottleneck* requires time-consuming manual transcription before further work with the resource can be done (Himmelmann, 2018; Seifart et al., 2018).

At issue in this paper are short-form verbal arts such as proverbs and riddles, specifically in the African context. In our view, they present a happy medium of naturalistic speech with consistent, manageably-sized transcripts. They also offer a crucial additional benefit: work centered on culturally significant verbal arts is more engaging for many participating speakers, compared to the arbitrary lists of text sentences often used for speech resource development (Gutkin et al., 2020; Gelas et al., 2012; Godard et al., 2018). Below, we consider the affordances of proverbs and riddles for efficient collection of a high-quality speech corpus in a field context, as well as their benefits for engaged speaker communities.

3 Background

3.1 Use-case: Kom

Kom (Itaŋikom) is a Grassfields Bantu language of Cameroon spoken by about 300,000 people. An official orthography is used in childhood education and small-press publications (Chia and Kimbi, 1992; Chuo, 2022). However, Kom totally lacks speech technology resources to our knowledge. The inventory of segments and surface tones is given in Table 1. The segmental inventory is notable for possessing front rounded vowels and *fricative vowels* /ᶻi, ᵛi/, which are associated with frication or affrication of preceding consonants (Connell, 2007; Faytak, 2017). Kom also has a complex tonal inventory, with at least eight surface tone contours and considerable postlexical changes to lexical tone patterns (Hyman, 2005).

Vowels			
ᶻi ⟨zi, si⟩			ᵛi ⟨vi, fi⟩
i	y ⟨ue⟩	ɨ	u
e	ø ⟨oe⟩		o
ɛ ⟨ae⟩			
	a		

Consonants			
b	t d	tʃ ⟨ch⟩ dʒ ⟨j⟩	k ⟨k, ’⟩ g
f v	s z		
m	n	ɲ ⟨ny⟩	ŋ
w	l	j ⟨y⟩	ɰ ⟨gh⟩

Surface tones	
⟨V⟩	H, HM, M, MH
⟨V̀⟩	L, LM
⟨V̂⟩	HL, ML

Table 1: Kom phonological inventory (Shultz, 1993; Hyman, 2005; Faytak, 2017); graphemes given in ⟨…⟩.

3.2 Proverbs and riddles as genres

Here, we highlight aspects of proverbs and riddles as the most conventionalized short-form verbal arts, with a particular focus on Kom as our use-case language. Table 2 shows a range of Kom proverbs and riddles.

Proverbs are fixed sayings, transmitted orally within the speech community, which have educational, critical, or advisory functions depending on the situation, both generally (Anchimbe, 2011; Etta and Mogu, 2012; Yankah, 1989) and in Kom specifically (Nkwi, 1987; Njwe, 2015). Proverb use demonstrates a speaker’s linguistic competence and higher-order thinking skills and is associated with wisdom and advanced age (Nkwi, 1987; Finnegan, 2012) . Proverbs generally convey moral lessons and emphasize pro-social behavior. They usually indirectly hint at their deeper meanings through figurative language, irony, and exaggeration. Because they are seen as indirect and conveying collective cultural wisdom rather than personal opinion, proverbs may be used to manage interpersonal conflict or negotiations (Finnegan, 2012; Fonkem, 2014).

A riddle is a verbal puzzle consisting of an short figurative description which guessers must identify with a scenario or object. Riddles are traditionally used in the more informal context of evening entertainment (Okpewho, 1992; Jick and Ngam, 2016; Akumbu et al., 2025). In Kom

Orthography	Literal translation
Awu à mò' a nɨ n-kulɨ wi ibu'.	One hand does not tie a bundle.
Gheli nɨ n-kôŋ sà chà' kɨ ibi zɨ̀ a yi n-chem.	People like to pick up kola that has dropped.
Wà tɨm àvɨ a kɨa, a kfɨ kɨ iti.	If you hit your foot, may only the stone break.
Chɨ̀sɨ nɨ̀ n-ku wi ɨlvâ ɨ yum	Charms don't catch an empty belly.
Afo kì a fòyn làlì si achi, a kɨ du'i.	A thing that sits when the chief stands up. Answer: fɨnjâenjàe (a fly)
Ghɨ se' ìghoŋ, a yɨ woynda.	They go to war, the children win. Answers: mɨlvɨ (soldier ants), ayôyn (speargrass)

Table 2: Representative Kom proverbs (top) and riddles (bottom) drawn from the collected corpus.

and other nearby societies, if the riddler outsmarts all guessers, they may demand symbolic payment to reveal the answer, in the form of titles of local chiefs (Jick and Ngam, 2016; Akumbu et al., 2025). Unlike proverbs, riddles usually do not convey moral lessons, but are seen as a sort of beneficial cognitive exercise.

3.3 Affordances of proverbs and riddles

Because proverbs and riddles are ubiquitous – the "palm oil with which words are eaten" (Achebe, 1959/1994, 14) – they are broadly known to speakers, and consistent transcripts can be generated for them. Because verbal arts are an oral tradition, they can be recalled by speakers without literacy in the target language. In our experience, short-form verbal arts are very effective at engaging participants in data collection. They also align better with the priorities of partner communities for cultural preservation than arbitrarily chosen materials, as proverbs are a repository of a community's worldview and epistemology. Recordings of short-form verbal arts may also have applications in childhood formal education, potentially filling gaps in availability of pedagogical material in local languages (Echu, 2004; Chiatoh, 2013) and contributing to decolonial practice in the classroom (Wolff, 2016; Akumbu et al., 2025).

4 Corpus collection and characteristics

For the verbal arts corpus, eighteen Kom speakers (7F, 11M) participated in recording near Douala, Cameroon, with a Zoom H4n digital recorder and Shure SM10A head-mounted cardioid dynamic microphones. Data were recorded as 16-bit mono WAV at a sampling rate of 44.1 kHz. For the out-of-domain evaluation described in Section 5.2, two speakers (1F, 1M) participated in similar recording sessions to collect spontaneous-speech narratives using different equipment: a Zoom H6Essential digital recorder with Shure WH20 head-mounted cardioid dynamic microphones, recording as 32-bit mono WAV at a rate of 44.1 kHz. The male participant was also a participant in the collection of the verbal arts corpus. The recording setting was a quiet but reverberant room in a concrete building without sound treatment. Because the data were originally collected for phonetic research on connected speech, the microphone hardware was chosen for its directional response and rejection of background noise. Speakers were recorded alone or in pairs, with the directional microphones ensuring isolation of one speaker per audio channel. Due to logistical limitations, not all proverbs or riddles are recorded for all speakers, and the total amount of material per speaker varies according to their availability.

Verbal arts data were collected in interaction with the first author; for full details on data collection see (Faytak et al., 2026). Speakers were prompted to re-speak proverbs read aloud from published sources by the first author (Loh, 1997; Lo-ah, 2018; Njwe, 2015). Speakers frequently volunteered conceptually related proverbs not attested in published sources; all riddles were volunteered, since riddles do not appear in published sources to our knowledge. Transcripts in Kom orthography were drawn from published resources or generated in consultation with participants if the proverb or riddle was volunteered. Due to the high degree of conventionalization of both proverbs and riddles, variation among speakers in the lexical or structural characteristics of a given item was quite low, and once a transcript was established it typically applied with few or no modifications to the versions provided by other speakers.

Once a proverb or riddle was successfully recalled, participants repeated it four to five times.

Repetition of this sort is common in phonetics research, as it increases the number of tokens per type and improves the statistical power of later analysis (Maddieson, 2001; Ladefoged, 2003; Bowern, 2015). Because speakers repeated the same utterance multiple times, the resulting corpus has reduced lexical diversity, which partly explains the low WER described in Section 5.

The resulting publicly available corpus[1] contains 315 minutes of Kom speech, with 55,092 word tokens covering 782 word types. This yields a very low token-type ratio of 0.0142. The out-of-vocabulary (OOV) rate on a random 80/20 split of the full corpus is 0.0243, and even across speakers (leave-one-out), the average OOV rate is 0.07%±0.04. This suggests that lexical variation is constrained, as may be expected given the repetitive data collection method. The token-type ratio remains low even if repetition is taken into account: taking *unique* utterances as the basis for calculations yields 14,469 word tokens and raises the token-type ratio only to 0.054.

5 Modeling

5.1 Experimental setup

As a proof of concept for the method, we trained two standard ASR architectures on the corpus under two evaluation strategies. Transcripts were preprocessed to remove tone diacritics due to inconsistent application across the corpus. The first evaluation setup was a regular random split where the utterances were shuffled and 80% were used for the training and the remaining 20% for testing. The second evaluation is a cross validation where the data collected from a given speaker is left out for evaluation for each speaker in the dataset.

The two ASR architectures used are XLSR, the multilingual version of wav2vec (Conneau et al., 2021; Baevski et al., 2020), and MMS (Pratap et al., 2024). For both architectures we used hyperparameters provided by the main HuggingFace tutorial with some modifications[2][3]. Models were trained for 20 epochs with a batch of size 16; acoustic features were kept unfrozen and CTC loss set to zero infinity. Since the classic MMS framework does not allow this, we evaluated only the XLSR models using trigram language models trained on the training sets of each model.

[1] https://huggingface.co/datasets/PhonLab-Buffalo/Kom
[2] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2
[3] https://huggingface.co/blog/mms_adapters

5.2 Results

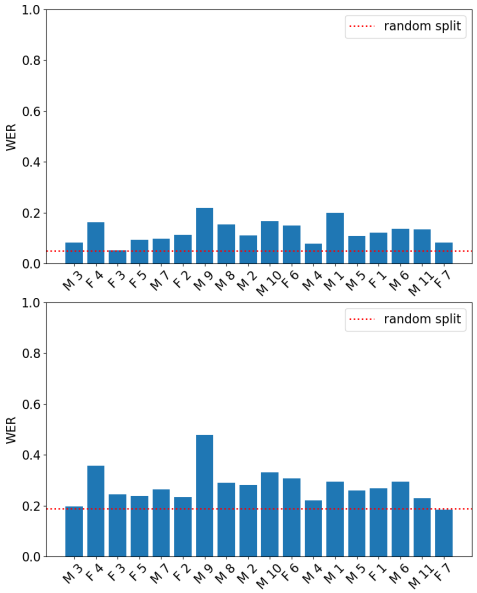


Figure 1: Word Error Rate by speaker for the cross validation (XLSR - top, MMS - bottom) compared with random split evaluation.

Word Error Rates for XLSR and MMS can be found in Figure 1. While WER is higher for MMS than XLSR, likely due to the former lacking a language model during decoding, all scores are below 0.2 for XLSR and 0.5 for MMS, which is within the range or below what is usually expected for a field-collected dataset (Jimerson et al., 2023). This confirms that the orthography of the corpus is consistent and the speech consistently intelligible.

To better inform future modeling decisions—particularly around use of the standard orthography for transcription of training data and around tokenization using characters rather than graphemes—we also conducted an analysis of the character recognition, using the Needleman-Wunsch algorithm to align the predictions to the gold standard at the character level. For both architectures, errors involving ⟨h⟩ are in the top three character recognition errors (Table 3). Because ⟨h⟩ appears in two digraphs ⟨ch⟩ and ⟨gh⟩ and is also written in some interjections (e.g. *mmhmm*, *eehee*), it may be hard to interpret due to its association with several phones' acoustic features.

To assess the quality of the data collected as ASR training data, we evaluated our baseline XLSR model on an additional 10 minutes of spontaneous speech. This additional data set consists of two unprompted monologues, one by a female speaker and one by a male speaker, concerning the

XLSR	CER	**MMS**	CER
h	0.05495	<space>	0.10946
<space>	0.05253	h	0.09041
y	0.03628	e	0.08239

Table 3: Top 3 CER per character by architecture.

recent history of their family and the traditional political structure of the Kom chiefdom, respectively. To obtain the gold standard transcriptions, raw ASR outputs were used as a starting point for corrections completed by the fourth author (a first-language speaker of Kom).

Using the baseline XLSR model, we obtained a WER of 63.9 and a CER of 28.6. The overall low level of performance is expected given low lexical diversity in the training data and the general lack of overlap between the domains at issue. In particular, the spontaneous speech data contained a number of English loanwords and proper nouns from both English and Kom; these categories of words are systematically absent from proverbs and riddles, reflecting the latter's concern with timeless generalities and normative values. However, the WER and CER obtained are surprisingly low for out-of-domain testing considering the low lexical diversity, and well within the range of results one should expect for this kind of dataset (Le Ferrand et al., 2025; Liang and Levow, 2025; Geng et al., 2025). The WER is also low relative to the size of the training data, in a similar range to rates obtained from models trained on Bible recordings (Le Ferrand et al., 2025), which typically provide five to ten times as much data as the present corpus (Meyer et al., 2022; Black, 2019).

6 Discussion and future work

The method presented here leverages short-form verbal arts to collect high-quality data sets in the field, improve transcriptional efficiency and consistency, and better align field data collection with speaker-community goals in cultural education and preservation. As a source of data for ASR model development, the data collected using this method appear to be very consistent for in-domain testing and offer a solid baseline for transcription of speech in other domains.

The proposed method shows no clear *degradation* of performance relative to other approaches to gathering ASR training data in low-resource contexts. Due to the many factors which affect ASR performance, we hesitate to claim an *improvement* at this stage, with such determinations requiring careful evaluation in future work. For instance, better-than-expected performance in OOD evaluation on spontaneous speech may be due to closer-than-expected vocabulary overlap with proverbs and riddles, due in part to the latter being embedded within everyday speech to a greater extent than domains involving a distinct performance style (e.g. traditional narratives, Bible recordings).

While the current low lexical diversity of the dataset limits its performance, materials collected using the proposed method offer a starting point for future model development, especially if augmented with additional training data. As such, future work aims to increase the lexical and domain diversity of the collected data. For instance, spontaneous explanations of the deeper significance of proverbs and riddles can be elicited, which may help to collect material complementary in style and more diverse in lexical content compared to the verbal arts items themselves. Refinements to transcription and tokenization may also yield incremental improvements in model performance: for instance, future models may tokenize by grapheme rather than character, avoiding issues related to recognition of ⟨h⟩ in the digraphs ⟨ch, gh⟩. Removal of tone diacritics for model training also likely reduced somewhat the number of lexical types in the corpus, and their reintroduction in future corpora may increase lexical diversity.

Limitations

Low WER and CER, as well as high recognition scores, suggest the present corpus has high recording quality and transcriptional consistency. However, we are aware that WER is also likely reduced due to low word character diversity, as may be expected given the repetition of the corpus materials. We also acknowledge that verbal arts, particularly proverbs, are not necessarily used by all speakers in a community, a limitation shared by other specialized genres such as traditional narratives. Performance in out-of-domain testing is rather low: WER and CER obtained in out-of-domain testing are fairly high, as may be expected, and the corpus needs to be augmented with more diverse data in order to build a robust ASR system for Kom.

We view these shortcomings as acceptable in light of the ease of use of short-form verbal arts materials and the efficiency with which they can

be used to collect a large amount of data in a range of field situations. The data which results is easy to consistently transcribe, and also has additional cultural significance for participating communities. The models presented here should be viewed as a starting point for further speech resource development, which is more viable if efficiency gains from imperfect ASR can be used to speed the annotation of additional training data.

Acknowledgments

We thank the members and executive of the Kom Cultural and Development Organization (Douala branch), Ikom Christopher Achuo, Pierpaolo Di Carlo, and Jeff Good for their logistical support. This work was supported by United States National Science Foundation award #2438916 to Matthew Faytak and Pius W. Akumbu.

References

Chinua Achebe. 1959/1994. Things Fall Apart.

Pius W Akumbu, Roland Kießling, Constantine Kouankem, Justine G Nzweundji, and Cornelius W Wuchu. 2025. Integrating traditional stories in formal education in the Cameroonian Grassfields. *Journal of the Cameroon Academy of Sciences*, 21(3):205–228.

Eric A. Anchimbe. 2011. *Language policy and identity construction: The dynamics of Cameroonian multilingualism*. Peter Lang.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Blaine Billings and Bradley McDonnell. 2025. Connecting automated speech recognition to transcription practices. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 128–132, Honolulu. Association for Computational Linguistics.

Alan W Black. 2019. CMU Wilderness multilingual speech dataset. In *Proc of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, pages 5971–5975. IEEE.

Claire Bowern. 2015. *Linguistic fieldwork: A practical guide*. Springer.

Emmanuel N. Chia and Joseph C. Kimbi. 1992. *Guide to the Kom alphabet*. SIL Cameroon.

Blasius A Chiatoh. 2013. Cameroonian languages in education: enabling or disenabling policies and practices. In *Language policy in Africa: perspectives for Cameroon*, pages 32–51. Miraclaire Academic Publications.

Godfrey Kain Chuo. 2022. *Achievements and Challenges of the Kom Multilingual Education Longitudinal Experience and the Impact on Cameroons Educational System.* SIL Cameroon.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Bruce Connell. 2007. Mambila fricative vowels and Bantu spirantisation. *Africana Linguistica*, 13(1):7–31.

George Echu. 2004. The language question in Cameroon. *Linguistik Online*, 18(1).

Emmanuel Efem Etta and Francis Ibe Mogu. 2012. The relevance of proverbs in African epistemology. *Lwati: A Journal of Contemporary Research*, 9(1).

Matthew Faytak. 2017. Sonority in some languages of the Cameroon Grassfields. In Martin J. Ball and Nicole Müller, editors, *Challenging Sonority: Cross-Linguistic Evidence*, pages 77–97. Equinox.

Matthew Faytak, Ivo Forghema Njuasi, Nicholas Mori, and Angelique Griffith. 2026. A speech corpus of Kom verbal arts and its applications. In Vicki Carstens, Katherine Russell, Olawale Akingbade, Deborah Morton, and Michael Diercks, editors, *Pamoja tena 'Together again': African linguistics after COVID*, pages 441–465. Language Science Press.

Ruth Finnegan. 2012. *Oral literature in Africa*. Open Book Publishers.

Achankeng Fonkem. 2014. Bekem in Peacemaking in Nweh Society. *Indigenous Conflict Management Strategies in West Africa: Beyond Right and Wrong*, page 307.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape Town.

Mengzhe Geng, Patrick Littell, PENÁĆ, Aidan Pine, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOŦEN Language Documentation Efforts with Automatic Speech Recognition. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39.

P. Godard, G Adda, Martine Adda-Decker, J Benjumea, Laurent Besacier, J Cooper-Leavitt, G-N

Kouarata, L Lamel, H Maynard, M. Müller, A Rialland, S. Stüker, F. Yvon, and M Zanon-Boito. 2018. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki.

Alexander Gutkin, Işın Demirşahin, Oddur Kjartansson, Clara Rivera, and Kọ́lá Túbọ̀sún. 2020. Developing an Open-Source Corpus of Yoruba Speech. In *Interspeech 2020*, pages 404–408.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Speech Technologies with Fieldwork Recordings: the Case of Haitian Creole. In *Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 40.

Nikolaus P. Himmelmann. 2018. Meeting the transcription challenge. In Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors, *Reflections on Language Documentation 20 Years after Himmelmann 1998*, pages 39–55. University of Hawai'i Press.

Larry M Hyman. 2005. Initial vowel and prefix tone in Kom: Related to the Bantu Augment. In Koen Bostoen and Jacky Maniacky, editors, *Studies in African comparative linguistics with special focus on Bantu and Mande: Essays in honour of Y. Bastin and C. Grégoire*, pages 313–341. Rüdiger Köppe Verlag Cologne.

Henry K. Jick and Gilead N. Ngam. 2016. Generic varieties and performance principles in Kom oral literature. *International Journal of Liberal Arts and Social Science*, 4(5):14–25.

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

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proc of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Peter Ladefoged. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Blackwell.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an Australian aboriginal community. In *Proc of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4988–4998.

Éric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. Faithful transcription: Leveraging Bible recordings to improve ASR for endangered languages. In *Proc of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, Mumbai. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In *Proc of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37.

F. Lo-ah. 2018. *The Wisdom of Fon Vincent Yuh II of Kom*. Bookman Communications.

Pius Loh. 1997. *Itaŋikom i tɨmlɨnɨ-i*. SIL Cameroon.

Friederike Lüpke. 2010. Research methods in language documentation. *Language Documentation and Description*, 7:55–104.

Ian Maddieson. 2001. Phonetic fieldwork. In *Linguistic Fieldwork*. Cambridge University Press.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. 2022. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. In *Interspeech 2022*, pages 2383–2387.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: The pangloss collection. *Language Documentation & Conservation*, 8:119–135.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating Automatic Transcription into the Language Documentation Workflow: Experiments with Na Data and the Persephone Toolkit. *Language Documentation & Conservation*, 12.

Emmanuel Ngue Um, Francis Tyers, Eliette-Caroline Emilie Ngo Tjomb, Florus Landry Dibengue, Blaise-Mathieu Banoum Manguele, Blaise Abbo Djoulde, Mathilde Nyambe, Brice Martial Atangana Eloundou, Jeff Sterling Ngami Kamagoua, José Mpouda Avom, Zacharie Nyobe, Emmanuel Giovanni Eloundou Eyenga, and André Likwai. 2025. Speech technologies datasets for african under-served languages. In *Proc of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 82–90.

Eyovi Njwe. 2015. "The palm oil with which words are eaten": Proverbs from Cameroons endangered indigenous languages. In Elizabeth C. Zsiga, One Tlale Boyer, and Ruth Kramer, editors, *Languages in Africa: Multilingualism, language policy, and education*, pages 118–126. Georgetown University Press.

Paul Nchoji Nkwi. 1987. *Traditional Government and Social Change: a study of the political institutions among the Kom of the Cameroon Grassfields*. Fribourg University Press.

Isidore Okpewho. 1992. *African oral literature: Backgrounds, character, and continuity*, volume 710. Indiana University Press.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proc 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

George Shultz. 1993. *Notes on the phonology of the Kom language*. SIL Cameroon.

Chihiro Taguchi, Jefferson Saransig, Dayana Velásquez, and David Chiang. 2024. Killkan: The Automatic Speech Recognition Dataset for Kichwa with Morphosyntactic Information". In *Proc of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9753–9763, Torino. ELRA and ICCL.

Allahsera Tapo, Éric Le Ferrand, Zoey Liu, Christopher Homan, and Emily Prud'hommeaux. 2024. Leveraging Speech Data Diversity to Document Indigenous Heritage and Culture. In *Interspeech 2024*, pages 5088–5092.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 993–1003.

H. Ekkehard Wolff. 2016. *Language and development in Africa: Perceptions, ideologies and challenges*. Cambridge University Press.

Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.

Kwesi Yankah. 1989. Proverbs: The Aesthetics of Traditional Communication. *Research in African Literatures*, 20(3):325–346.

Marcely Zanon-Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. In *Proc of the twelfth language resources and evaluation conference*, pages 6486–6493.

A Appendix

	XLSR		MMS	
Spkr.	WER	CER	WER	CER
M3	0.0823	0.0371	0.1979	0.0817
F4	0.1617	0.0632	0.3562	0.1110
F3	0.0519	0.0208	0.2448	0.1022
F5	0.0930	0.0408	0.2391	0.0762
M7	0.0975	0.0396	0.2650	0.0767
F2	0.1118	0.0479	0.2344	0.0871
M9	0.2181	0.0893	0.4781	0.1485
M8	0.1541	0.0632	0.2900	0.0895
M2	0.1112	0.0460	0.2807	0.1083
M10	0.1675	0.0849	0.3318	0.1265
F6	0.1494	0.0593	0.3083	0.1066
M4	0.0779	0.0351	0.2202	0.0924
M1	0.1993	0.0902	0.2947	0.0891
M5	0.1075	0.0462	0.2592	0.0901
F1	0.1220	0.0475	0.2688	0.1014
M6	0.1357	0.0664	0.2942	0.1049
M11	0.1341	0.0581	0.2299	0.0815
F7	0.0823	0.0333	0.1845	0.0605
Rand.	0.0600	0.0265	0.1880	0.0623

Table 4: Word and Character Error Rate by speaker and random split, for XLSR (left) and MMS (right).

Quantitative Lect Description: A Case Study of Lemko from the Field Data of 1920s-1930s

Ilia Afanasev
University of Vienna
ilia.afanasev.1997@gmail.com

Abstract

While qualitative descriptions (in the form of reference grammars) and benchmarks for low-resource languages are becoming increasingly widespread, computational linguists do not often use quantitative methods to describe a new lect rather than a new model. This paper intends to close this lacuna.

The case study is a Lemko text transcribed at the beginning of the twentieth century. Using morphosyntactic tagging and topic modelling, the study demonstrates areal influences and archaic features of the lect. Fine-grained evaluation significantly assists in identifying subtle patterns that are not readily apparent through traditional metrics such as accuracy score.

The results highlight the necessity of a more detailed analysis of model performance, which may yield more linguistically significant results than a purely manual check. This information is present in the resulting dataset, which can be used for further investigation into the structural features of the Lemko lect.

1 Introduction

The goal of this study is to provide an example of a quantitative characterisation of a low-resource lect that can enhance (but certainly not substitute for) the work of a field researcher through distributional analysis based on neighbouring lects. Quantitative characterisation is something that current studies of low-resource lects lack: while toolkits and benchmarks are crucial for language preservation, they do not in themselves provide a means of scientific description.

This often leaves researchers relying on qualitative methods for interpreting the errata produced by toolkits for benchmarks, and even more so in the age of Large Language Models (LLMs), which often become a single tool in a toolkit and remain barely explainable from any perspective, including a linguistic one (Rakotonirina et al., 2025). A more detailed investigation of this problem is presented in Section 2.

The distributions of particular features in raw texts of the lects remain in a blind spot between quantitative metrics that assess the number of correct outputs, qualitative interpretations of inconsistencies, and preprocessed datasets in variationist sociolinguistics. Section 4 proposes a way to close this gap, while Section 5 demonstrates a practical implementation of the algorithm. Section 6 provides an overview of the study and outlines its prospects.

The case study for this research is the Lemko[1] lects of the western part of the Transcarpathian region. Section 3 describes this dataset in detail. Using it, the paper demonstrates the benefits of taking a distant, *a posteriori* perspective on a new lect that is not heavily influenced by presupposed categories.

2 Related Work

The following section begins with a presentation of existing frameworks for quantitative and qualitative description. The next part focuses on quantitative studies of variation within corpora. The section ends with an overview of existing NLP studies of Lemko.

2.1 Frameworks of language description

Language documentation and description were among the earliest linguistic endeavours, especially in the study of newly documented lects. Prominent examples include the first grammars of South American languages (Domingo de Santo Tomás, 1560; Ruiz de Montoya, 1640; Adam and Henry, 1880). However, for a long time a much more widespread practice (illustrated below with examples from indigenous languages of Asia and

[1] Lemko is a denotation preferred by native speakers compared to Lemkian: https://uwr.edu.pl/en/lemkos-who-are-they/ (date of access: February 12, 2026)

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 46–59
March 26, 2026.

Africa) was the collection of basic vocabulary lists (Dobrotvorskii, 1875; Munkácsi, 1894; Zukowsky, 1924) or text corpora (Bleek and Lloyd, 1911; Bleek, 1929). Grammars, enhanced by collections of texts, have become a standard means of describing a language only relatively recently; see, among others, Wiedemann (1884); Nakonečna and Rudnyc'kyj (1940).

In modern linguistics, the most traditional form of language description is the production of a reference grammar; some publishers dedicate entire series to this purpose (Lehman, 1989; Zigmond et al., 1991; Kimball, 1991; Osada, 1992; Brindle, 2017; Daniel et al., 2019; Namyalo et al., 2021). This is an extremely valuable and well-developed genre of linguistic literature that covers the historical development, phonetics, morphology, and syntax of a given lect. It recognises and critically assesses the *a priori* nature of many linguistic categories with which it must operate (Terhart, 2024, pp.113–114). Still, it rarely focuses on quantitative characteristics and distributions within corpora that represent actual linguistic practices of communities.

Natural Language Processing (NLP) studies, in contrast, focus on developing toolkits (Bolt et al., 2019; Tolmachev et al., 2018; Pauli et al., 2021; Rennes et al., 2022) and designing benchmarks (Shavrina et al., 2020; Aparovich et al., 2025; Chirkin et al., 2025; Umbet et al., 2025). These toolkits and benchmarks, while indispensable for increasing the presence of low-resource lects, are almost always based on existing descriptions and understandings of a given lect, employing either expert knowledge or reference grammars. Their use in model assessment is vital for understanding linguistic capabilities, but it reveals little about the language itself. Some works attempt to bring evaluation closer to linguistic exploration (Bindi, 2025; Neumann, 2025), but they still primarily guide researchers through tools, concentrating attention on the qualitative evaluation of quantitative methods.

2.2 Variation in corpora

Quantitative studies of variation are almost impossible without corpora (for an overview, see Tagliamonte (2025, pp. 1–15)). However, most rely on extracting units belonging to pre-determined categories from the material. The closest analogue to a quantitative description would be *a posteriori* studies that do not assume pre-given language categories (Otheguy, 2002) and instead operate with distributional skewings that mark the presence of semantic substance expressed through specific signals (García, 1989; Diver, 2012). When fully integrated into quantitative linguistics, this approach allows for the identification and explanation of less trivial regularities and irregularities in the data.

2.3 NLP and Transcarpathian lects

Despite its significance for studies of the historical development of the Slavic clade (Ševel'ov, 1979, p. 37), the Transcarpathian group remains underrepresented in NLP, as do most Slavic territorial lects. There has been an effort to create a corpus representing their modern state[2], which resulted in the development of some language-specific tools (Scherrer and Rabus, 2019), as well as computational research (Rabus and Scherrer, 2017; Rabus, 2018; Lahjouji-Seppälä and Rabus, 2021). However, these studies operate at a regional scale, and smaller lects, such as Lemko, have not received full-scale attention.

3 Data

The description of the research data contains two subsections. The first delves into the overall characterisation of the lect under consideration; the second reports on the subcorpus used in the current study.

3.1 Lemko (Lemkian)

Lemko is a group of small territorial East Slavic lects historically spread across the territories of southern Poland and north-western Slovakia, as Figure 1 shows. It is in this territory that scholars collected the material constituting the dataset (Nakonečna and Rudnyc'kyj, 1940, p. 15). Unfortunately, after the Second World War most Lemkos living in this territory lost their homes due to discriminatory policies (Magocsi, 2015, pp. 336–338), so the material gathered at the beginning of the twentieth century may well be the latest available fragment (Baglioni and Rigobianco, 2024, pp. 1–9) of this lect.

Lemko is part of the Carpathian group (Del Gaudio, 2017, p. 78), which itself is part of the Transcarpathian area (Ševel'ov, 1979, p. 37), incorporating Hutsul, Bojko, and Central and South Carpathian lects (Del Gaudio, 2017, p. 78). There are at least three standards that roof these lects: Polish, Slovak, and Ukrainian. Modern standard Ukrainian, which stems from the Central

[2] http://www.russinisch.de/VarchoLatin2/login.php (date of access: February 12, 2026)

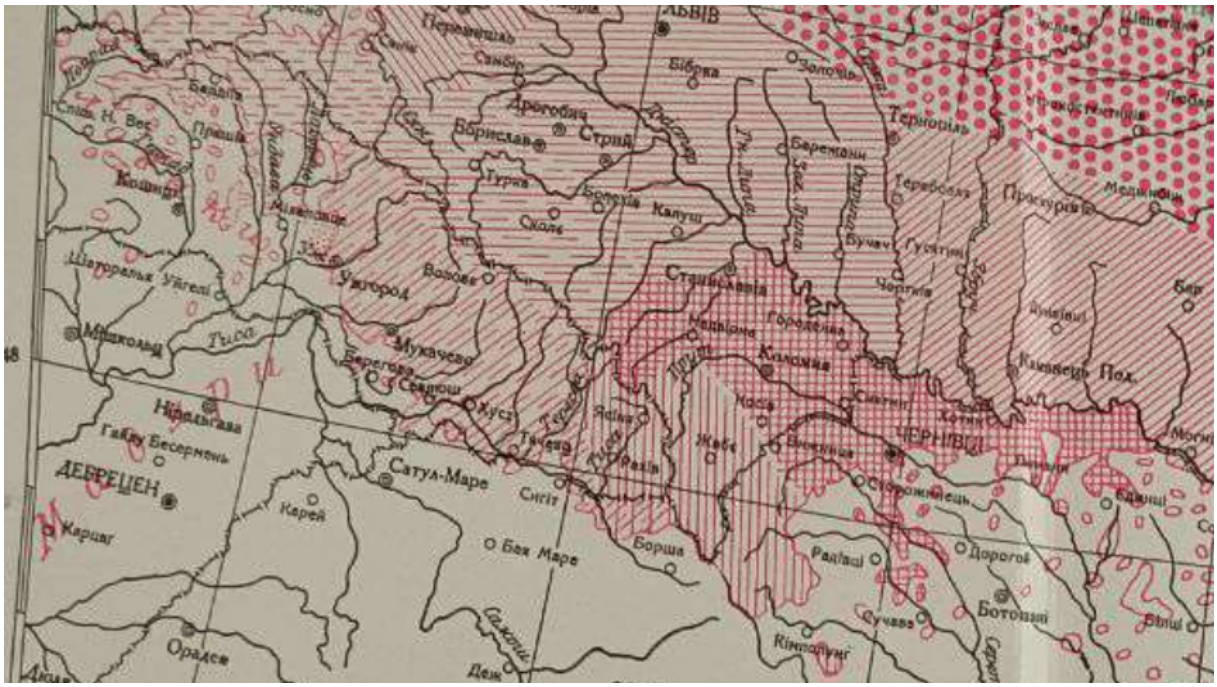


Figure 1: The territory of the Transcarpathian lects spread at the beginning of the twentieth century (Zilyns'kyj, 1933). Lemko is in the left corner, marked by rare horizontal strikes.

Dnipro group (Del Gaudio, 2017, p. 92), is the most closely related to Lemko among these standards. Nowadays, the Rusyn (micro)standard is also emerging (Dulichenko, 1981, p. 14; Magocsi, 2004). There has also been an attempt to standardise Lemko itself, resulting in the appearance of a grammar (Fontański and Chomiak, 2000). Most written or printed material in Lemko is in the Cyrillic script.

There are distinct characteristics that differentiate Lemko from neighbouring lects at all levels of the language system. For this paper, the most relevant are grammatical and lexical features.

3.1.1 Morphology

The inflection of nominal and pronominal forms in Lemko has significant peculiarities which, in the case of transfer learning, are likely to influence system performance. Among these are unique declension paradigms. For instance, the word denoting 'young girl' is *дівча* (*divča*, see Fontański and Chomiak (2000, pp. 79–80)), unlike Polish *dziewczyna* or Ukrainian *дівчина* (*divčyna*).

The instrumental singular feminine form of adjectives (and words undergoing similar inflection, such as some relative pronouns and nouns) has the *-ом* (*-om*) ending, cf. *к'отром* (*k'otr-om* 'which-FEM.INS.SG'[3]). For comparison, Polish has *-ą* (*ma l-ą* 'small-FEM.INS.SG' (Wróblewska, 2018)[4]), Ukrainian has *-ою* (*скороченою* (Kopp et al., 2023) (*skoročen-oju* 'shortened-FEM.INS.SG')), and Slovak has *-ou* (*z'ahrebsk-ou* (Zeman, 2017) 'Zagrebian-FEM.INS.SG').

The instrumental plural form is often *-ма* (*-ma*): *н'іма* (*n'i-ma* 'PRON.3PL-INS'). Slovak, Polish, and Ukrainian possess different forms: *ni-mi* (Slovak (Zeman, 2017); Polish (Wróblewska, 2018); Ukrainian *ними* (*ny-my* (Kopp et al., 2023))).

Another noteworthy feature of Lemko is the reduplicated form of determinative pronouns: *momo* (*toto*) instead of Ukrainian *це* (*ce*) (Kopp et al., 2023), Polish *tamto* (Brooks, 1975, p. 306), and Slovak *to* (Zeman, 2017).

3.1.2 Lexis

As the lect is part of an intense contact area in the Carpathian mountains, Lemko material demonstrates a significant presence of borrowings, mostly from Slovak and Polish. Slovak borrowings include words such as *вельо* (*vel'o* 'many') and *кед* (*ked* 'when, if') (Nakonečna and Rudnyc'kyj, 1940, p. 30). Material borrowed from Polish includes, among other items, *барз* (*barz* 'very') and *мераз* (*teraz* 'now') (Nakonečna and Rudnyc'kyj, 1940, p.

[3]Glosses given according to Comrie et al. (2008)

[4]See Appendix A for more detailed information on the sources of examples.

30). There are also Hungarian and German borrowings, for instance *кіральфій* (*kiral'fij* 'prince, son of king' (< Hungarian *kir'alyfi* 'id.' (Nakonečna and Rudnyc'kyj, 1940, p. 30))), but these are mostly long-integrated nouns, and any incorrect description by the model is more likely to be due to morphological reasons than to out-of-vocabulary status.

3.1.3 Syntax

Unlike Ukrainian, and like Polish and Slovak, Lemko tends to make heavy use of copular clauses of a very specific type. The subject is a determinative pronoun (in all of these languages in neuter gender form), the predicate is a noun, and the link between them is an auxiliary verb 'to be'. Examples from all of these languages are given below; the relevant parts of the sentences are in bold.

(1) Lemko (Nakonečna and Rudnyc'kyj, 1940, p. 31)

То				**сут**
T-o				**sut**
DET-NEUT.NOM.SG				**be.PRES.3PL**
вши́тки				лемки́вски
všýtk-y				lemkývsk-y
all-NOM.PL				Lemko-NOM.PL
се́ла	,	де	по	лемки́вски
sél-a	,	de	po	lemkývsky
village-NOM.PL	,	where	in	Lemko
гва́рят	.			
hvárj-at	.			
speak-PRES.3PL	.			

'**These are** all Lemko **villages**, where one speaks Lemko.'

(2) Polish (Wróblewska, 2018)

Co	**t-o**		**są**
What	**DET-NEUT.NOM.SG**		**be.PRES.3PL**
mechanizm-y		obronn-e	?
mechanism-NOM.PL		defence-NOM.PL	?

'**Are these** defence **mechanisms**?'

(3) Slovak (Zeman, 2017)

Je-∅			**t-o**
be.PRES-3SG			**DET-NEUT.NOM.SG**
naj-bežn-ějš-ia			
SUP-common-CMPR-FEM.NOM.SG			
kukuric-a	pre	priam-u	
corn-NOM.SG	for	direct-FEM.ACC.SG	
ľudsk-ú		spotreb-u	
human-FEM.ACC.SG		consumption-ACC.SG	
.			
.			

'**This is** the most common **corn** for direct human consumption.'

(4) Ukrainian (Kopp et al., 2023)

це	дуже
c-e	duže
DET-NEUT.NOM.SG	very
важлива	
važlyv-a	
important-FEM.NOM.SG	
поправка	
popravk-a	
amendment-NOM.SG	

'This is a very important amendment'

As can be seen from the examples, Lemko and Slovak use the auxiliary 'to be' (in this case in the present tense, third-person singular or plural form), whereas modern standard Ukrainian does not. Given the relatively high degree of descriptiveness in Nakonečna and Rudnyc'kyj (1940), which entails a high frequency of identificational constructions (roughly translated into English as *This is*), this discrepancy may create severe issues for the syntactic parser module, which is not trained for this type of sentence. Apart from this, however, there are no syntactic features that would characterise Lemko as a specific part of the Carpathian area.

3.2 Dataset

The case study is *LA1407*, a set of three texts in Lemko written sometime between the 1930s and the 1940s by a person who learned the lect in the settlement of Kamienka (Lemko *Камюнка*, modern Prešov Region of northern Slovakia) during childhood. The first text provides a general metacharacterisation of the lect, the second discusses traditional social gatherings, and the third is a folklore story about a devil who intended to destroy Stará Ľubovňa Castle.

The overall size of *LA1407* is 609 tokens, split into 34 predications[5], but the metainformation for each sentence is quite extensive. Nakonečna and Rudnyc'kyj (1940) provides detailed phonetic transcription, standardised transcription, and a German translation, which greatly aid modern scholarship. The format of the existing digitised version is CoNLL-U. The phonetic transcription (converted to IPA) and the German translation are provided as metadata fields for each sentence. Table 1 shows an example.

[5]For some utterances, Nakonečna and Rudnyc'kyj (1940) preserved them as sentences; others were chunked into smaller units, likely due to the need to include as much information as possible about each of them on a given page.

sent_id = LA1407.3.3
IPA_transcription = tɕʲ’ort maw prïkaz’ano
do dvan’atsʲatoj fiodïnɨ̈ v‿n’otɕʲɨ̈ rozb’ïtı z’amök //
standard_text = Чорт мав пріказа́но до дванаця́той годи́ни в но́чи розби́ті за́мок.
german_text = Es wurde dem Teufel befohlen, bis zwölf Uhr nachts das Schloß zu zerstören.

1 Чорт _ _ _ _ _ _ _ wf="Чорт"|tf="tɕʲ’ort"

(...)

Table 1: The initial digitisation of LA1407.3.3. Underscores denote the fields, obligatory for CoNLL-U format, but not yet filled with morphosyntactic tags. As the table is an illustration, it shows (for brevity considerations) only the first token. The translation of the example is *The devil had an order: before the clock strikes midnight, he should destroy the castle*.

During this study, the existing digitisation (Nakonetschna et al., 2025), which also contains information on named entities and basic vocabulary items, underwent an additional round of checks for consistency and correctness. While the tagging remained mostly intact, some normalisation was necessary. In this process, both instances of *староlюбов’ентскій* (*staroljubov’entskij* ‘of Stará Ľubovňa-ADJ’ in the standardised transcription) received the proper representation of the epenthetic *m* (*t*): in the original rendering, one instance lacked it, while the other represented *mc* (*ts*) as *ц* (*c*). This adjustment was required for an accurate visual representation of the linguistic variation.

For this study, the dataset underwent additional manual preprocessing by a linguist specialising in East Slavic languages. Some graphic variation in the standardised transcription, such as *хт’овди/хд’овди* (*ht’ovdy/hd’ovdy* ‘then’), was normalised to a single form corresponding to the phonetic transcription. Predications that were originally part of a single sentence were merged to restore the original structure and ensure correct dependency parsing. The sentences were provided with an English translation in addition to the original German one.

The final step consisted of tagging linguistic variation using a schema similar to UA-GEC (Syvokon et al., 2023). If no variation is present, the schema reproduces the sentence unchanged. Where variation occurs, the relevant part of the token receives the following tag: `{SOURCE=>INVARIANT:::variation_type=GROUP~TYPE}`, where `SOURCE` is the original segment of the token, `INVARIANT` is its normalised standard equivalent, and `GROUP/TYPE` is a variation label indicating the broader category (phonetic/morphological) and the specific subtype.

For instance, the sentence in Table 2 shows three distinct types of phonetic (`Phon`) variation: `G` (fricative/plosive velar), `NTSK` (presence or absence of the epenthetic *m* (*t*) between *н* (*n*) and *ск* (*sk*)), and `Edn` (presence or absence of prothetic *v* before *єдн* (*jedn-* ‘one-’)). This tagging is stored as a separate metadata field for each sentence in the dataset..

4 Method

This section consists of two parts. The first outlines the general methodological considerations of the current study, forming its theoretical backbone. The second provides the workflow for data annotation and the subsequent experiments that constitute the application of the theoretical principles developed in the first subsection.

4.1 How to describe a lect quantitatively?

The goal of quantitative description is to produce a model or a set of models that describe a lect as accurately as possible. The most appropriate strategy for performing this description for low-resource lects is to utilise models (pre-)trained on neighbouring higher-resource lects. In this type of exploration, it is crucial to treat the studied lect as an independent system rather than as an offshoot of a neighbouring higher-resource lect. Dialectologists have identified such biases even in qualitative research (Saenko, 2018) and have cautioned against them, arguing for describing a smaller lect as a self-contained system (*integral approach*) rather than as a deviation from a roofing standard (*differentiating approach*) (Goldin and Kryuchkova, 2011; Otheguy and Stern, 2011; Hromko, 2020).

The purpose of the subsequent analysis is to explain how the distribution of the grammatical features within a lect affects model performance and

За́мок старолюбове́{нтскі=>нскі:::variation_type=Phon~NTSK}й сто́їт ко́ло Попра́да бли́зко місто́чка Старолюбо́вні і Камю́нки, а є ма́етком {єдн=>єдн:::variation_type=Phon~Edn}о́{г=>г:::variation_type=Phon~G}о польско́{г=>г:::variation_type=Phon~G}о {ґ=>г:::variation_type=Phon~G}ро́фа.

Table 2: The example of tagging of grammatical variation in LA1407. The English translation is *The castle of Stará Ľubovňa stands on the river Poprad, near the Stará Ľubovňa city and the Kamienka village; it is a property of a Polish count.*

what this reveals about the distributional properties of this lect. A necessary component of the analysis is the combination of close reading, which selectively and thoroughly examines sections of the material to capture the full scale of variation, with distant reading, which traces a single feature across a larger body of data.

4.2 Experiment outline

As *LA1407* is a single digitised Lemko text, the current study primarily focuses on providing baseline morphosyntactic tagging. The analysis section discusses both its automatic and manual evaluation.

4.2.1 Tagging

For morphosyntactic tagging, the paper uses a modern standard Ukrainian-trained model within Stanza (Qi et al., 2020), an NLP toolkit covering a wide range of languages. The study implements only one model, as its focus is not on comparing the performance of different tools on a single dataset, but rather on examining what this specific model reveals. This is also why the paper does not employ modern generative AI models as tools, as their use introduces additional methodological variables that fall outside the scope of the present study.

The tagging process consists of three stages (PoS/morphological tagging, lemmatisation, dependency parsing), applied sequentially with intervening manual correction. This workflow reduces the number of errors propagated to subsequent stages. After preprocessing is complete, the study uses Latent Dirichlet Allocation (Blei et al., 2003) to generate topics for the text set, thereby adding a lexical layer to the analysis.

4.2.2 Evaluation

The study relies heavily on fine-grained evaluation. For part-of-speech and morphological tagging, it reports the number of errors for each tag. During the lemmatisation stage, in addition to the traditional accuracy score, the article implements string similarity measures: Levenshtein distance (Levenshtein, 1966) and Jaro–Winkler distance (Winkler, 1990). In addition to Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS), dependency parsing utilises Unlabelled Complete Predication (UCP) and Labelled Complete Predication (LCP), which indicate whether the model identified all dependencies of the root verb and whether these dependencies received correct labels (Plank et al., 2015, p. 315).

This study slightly modifies UCP and LCP to account for cases in which the model assigns more labels than required (the original metric tests only the presence or absence of gold labels among the predicted ones). The modified version relaxes the metrics by scoring the proportion of correctly identified dependencies for each verb, rather than using a binary judgement of complete success or failure. An additional metric is the average tree edit distance (TED), which measures the number of edits required to transform a predicted tree into the gold tree (Plank et al., 2015, p. 315).

4.2.3 Qualitative exploration

Fine-grained evaluation highlights distributional skews present in the data itself or in comparison with other datasets (for instance, the initial input data). Close reading, by contrast, examines the tagging of specific forms. Its purpose is to illustrate the structural properties of Lemko, showing how different linguistic levels interact to create a distinctive combination of Transcarpathian features.

5 Experiments and Analysis

This section briefly discusses the tagging experiment results, presents the quantitative evaluation, and then explores the identified patterns. The tagging code is available in open access (Afanasev, 2026).

5.1 Tagged dataset

After the combination of manual (Section 3.2) and automatic (Section 4.2.1) tagging, the dataset

becomes substantially richer in linguistic annotation. The representation of sentence LA1407.3.3 is shown in Table 3.

The tokens also contain information on the errors that occurred during the tagging phases. The miscellaneous field `PosRapidity` denotes the quantitative measurement of errors made by the model for both part-of-speech and morphological tags (in this case, 0), which in subsequent visualisation studies is converted into a heat rate. The miscellaneous field `LemmaErrorSpots` indicates (in this case, absent) the differences between the gold lemma and the predicted lemma, while `TaggedLemma` preserves the predicted lemma.

5.2 Evaluation results

5.2.1 Morphological tagging

In evaluation terms, the model shows a moderate decline relative to the results it achieved on the test subset of the standard Ukrainian corpus on which it was trained [6]. For part-of-speech tagging, the macro F1-score is 66.78%, with recall of 68.71% and relatively higher precision of 75.88%. The exact match rate for morphological tags is 55.99%.

Table 4 shows the fine-grained evaluation results, excluding punctuation, interjections, and coordinating conjunctions (as there were no errors in these categories).

As can be seen, the categories that affect the model most are adverbs (`ADV`, which account for almost ten percent of the dataset), subordinate conjunctions (`SCONJ`, almost five percent), and verbs (`VERB` and `AUX`, more than ten percent). These errors concentrate around Satellite Cluster B (Reid, 2011, pp. 1107–1108), the verb, while the effect on the nominal group, Satellite Cluster A (adjectives `ADJ`, nouns `NOUN`, proper nouns `PROPN`, and adpositions `ADP`), is significantly smaller. The key factors here are lexical and syntactic differences. Many verbs, such as гварят (*hv'ar-jat* 'speak-`IPFV.PRES.3PL`'), are unknown to the model (standard Ukrainian would yield говорят (*hovor-jat*)), which leads to incorrect tagging. This, in turn, propagates errors down the syntactic tree, as the model misinterprets the remaining signals. This suggests that the model is unable to rely on character-level sequences within tokens that might otherwise signal a particular meaning.

The very low score for `AUX` is likely due to the absence of copular constructions in the training dataset, as mentioned in Section 3.1.3. Because the model was not adjusted to this type of signal grouping, it fails to distinguish its constituents, especially auxiliary verbs.

One of the greatest difficulties for the model is identifying tense (32.50%) and aspect (39.56%) in the verbs. Explaining these failures is more complex, but it is clear that the stanza-uk model does not adequately capture the semantic distinctions encoded in the verbal inflectional system. For instance, *жúют* (*žý-jut*'live-`IPFV.PRES.3SG`') receives the tag `Aspect=Perf`.

5.2.2 Lemmatisation

The model performs better in lemmatisation: the accuracy score is 75.69%. The Levenshtein (1.44) and Jaro-Winkler (87.40%) distances, however, indicate even stronger performance. This suggests that the concept of lemmatisation does not differ substantially cross-linguistically, or at least not between standard Ukrainian and Lemko. When the model makes an error, it rarely exceeds two characters and typically reflects an incorrect inflection rather than a random substitution.

The most problematic categories are, once again, verbs and auxiliaries. The accuracy score for verbs is 4.11% and for auxiliaries is 15.79%. It is noteworthy, however, that string similarity metrics for verbs are much better: a Levenshtein distance of 1.15 and a Jaro-Winkler distance of 92.04%. Errors are almost absent; the only significant issue is that the model predicts the infinitive ending as *-mu* (*-ty*), whereas in Lemko it is generally -ти (*-ti*) with some exceptions.

For auxiliaries, by contrast, all metrics indicate poor performance: the Levenshtein distance is 2.88, and the Jaro–Winkler distance is 56.19%. The model handles this class very poorly, which reflects its substantially different usage in the training dataset (see Section 3.1.3).

5.2.3 Dependency parsing

The syntax is seemingly the weakest spot of the model, as shown in Table 5.

On the surface level, the performance is acceptable. UAS and TED could have been considerably worse (given that there are sentences of length 60, an average error of 9 is substantial, but not critical). LAS and both CP metrics, however, indicate

[6] See the results of the model on the test subset here (date of access: February 12, 2026). The work uses results from the test subset as comparative material in accordance with common practice in NLP for closely related varieties (Bhatia et al., 2021; Blaschke et al., 2023; Pugh and Tyers, 2024).

sent_id = LA1407.3.3
IPA_transcription = tɕʲ’ort maw prïkaz’ano
do dvan’atsʲatoj fiodïnɨ̈ vˍn’otɕʲɨ̈ rozb’ïtɪ z’amök //
variation_text = Чорт мав пріказа́но до дванаця́той {г=>г:::variation_type=Phon~G}оди́ни в но́чи розби́{ті=>ті:::variation_type=Morph~Infinitive} за́мок.
standard_text = Чорт мав пріказа́но до дванаця́той годи́ни в но́чи розби́ті за́мок.
german_text = Es wurde dem Teufel befohlen, bis zwölf Uhr nachts das Schloß zu zerstören.
english_text = The devil had an order to destroy the castle before midnight.

1 Чорт чорт NOUN _ Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
3 nsubj:pass _ wf="Чорт"|tf="tɕʲ’ort"|PosRapidity=0|LemmaErrorSpots=____|TaggedLemma=так

(...)

Table 3: The final digitisation of LA1407.3.3. The underscore denotes an empty field of language-specific morphosyntactic tagging (XPOS): such tagging requires additional effort that is out of the scope for this research. As the table is an illustration, it shows (for brevity considerations) only the first token.

PoS	ADV	PRON	ADJ	SCONJ	PROPN	ADP	NOUN	DET	VERB	PART	AUX
A	54.17	60.00	67.65	34.62	97.22	80.00	97.06	56.52	67.12	90.00	26.32
S	7.88	5.75	5.58	4.27	5.91	10.67	16.75	3.78	11.99	1.64	3.12

Table 4: The accuracy scores (A), % and shares (S), % (rounded to the second digit) of PoS within the dataset (excluding 100% results).

that the model does not infer syntactic categories reliably. One of the most problematic categories is reflexivity (`expl:pv`); the model never predicts it correctly. This is due to the reflexive short pronoun *ся* (*sja* ‘REFL’) behaving more freely than, for example, in standard Ukrainian, in a manner closer to Slovak or Polish.

(5) Lemko (Nakonečna and Rudnyc’kyj, 1940)
A на́ них **ся**
A ná n-ych **sja**
And on PRON.3-ACC.PL **REFL**
призіра́ют ня́ньо ,
pryzirá-jut nján’-o ,
watch-PRES.3PL father-NOM.SG ,
ма́ма , ді́до
mám-a , díd-o
mother-NOM.SG , grandfather-NOM.SG
, ба́ба ,
, báb-a ,
, grandmother-NOM.SG ,
невíсти і дру́ги
nevíst-y i drúh-y
daughter.in.law-NOM.PL and other-NOM.PL
з роди́ни .
z rodín-y .
from family-GEN.SG .
’And watching them are: the father, the mother, the grandfather, the grandmother, the daughters-in-law and other family members.’

The freer behaviour of the reflexive in Lemko results from its status as a clitic that most frequently occupies the position of the second phonetic word in a phrase (Kolaković et al., 2022, pp. 22–32). As Polish and Slovak, alongside other Slavic languages (Kolaković et al., 2022, pp. 22–32), preserve this pattern, the most likely explanation is a shared archaism inherited from Proto-Slavic. While Central Dnipro lects (and, subsequently, modern standard Ukrainian) innovated toward a tighter attachment of the reflexive to the verb, Lemko and other lects of the Transcarpathian area retained the older structure. In this respect, they are closer to historical Slavic lects such as Old Church Slavonic (Polivanova, 2013, pp. 462–464).

5.2.4 Thematic modelling

The hyperparameters for the LDA model are given in Appendix B. The extracted topics are *под* (*pod* ’under’), *познати* (*poznati* ’know’), *польский* (*pol’skyj* ’Polish’), *полюбити* (*poljubyti* ’love’), *посмотрiти* (*posmotriti* ’take a look’), *пост* (*post* ’fasting-NOM.SG’), *походити* (*pohodyti* ’originate’), *през* (*prez* ’through’), *премінiти* (*preminiti* ’change’), *призірати* (*pryzirati* ’watch’). Together,

Metrics	UAS	LAS	TED	UCP	LCP
Value	67.65%	52.71%	9.00	47.18%	35.57%

Table 5: UAS, LAS, UCP, LCP % and tree-edit distance for dependency parsing of LA1407 with Stanza

they summarise the three texts with considerable clarity. For instance, *под* (*pod* 'under'), *польский* (*pol'skyj* 'Polish'), *походити* (*pohodyti* 'originate') and *през* (*prez* 'through') characterise the first text, which describes the speaker and the geographical distribution of the lect. Even the functional words are thematic here, as they emphasise spatial relations between the entities expressed through the content words. The item *польский* (*pol'skyj* 'Polish') highlights the role of neighbouring Slavic languages in the development of Lemko, already visible in its syntax.

The second group of words evokes the atmosphere of a traditional gathering: *познати* (*poznati* 'know'), *полюбити* (*poljubiti* 'love'), *посмотріті* (*posmotriti* 'take a look'), *пост* (*post* 'fasting'), *призірати* (*pryzirati* 'watch'). These items relate either to the social purpose of such meetings (courtship and marriage) or to their temporal setting (summer fasting). The word *преміниті* (*preminiti* 'change') criginates from the third text and does not characterise it as directly. However, it may function as a predicate summarising the story of the devil transforming into scale grease after failing to fulfil an order. In that sense, it remains thematically appropriate.

5.3 Discussion

Through its errors, the model facilitates a description of Satellite Cluster B (the lexical verb and its adjacent elements, excluding the subject and its adjacent morphemes, Satellite Cluster A), whose structure differs substantially between Lemko and modern standard Ukrainian. Some predication types are entirely unknown to the model, while those it recognises exhibit a divergent distribution (see paragraph 5.2.3).

The analysis not only supports the existing body of qualitative research, but also highlights the Transcarpathian features of Lemko by introducing a quantitative perspective. The distributional skewings present in data affect the clause types and, consequently, roots and affixes that structure semantic space through formal markers. This organisation diverges markedly from that of modern standard Ukrainian.

One of the key features is the copular construction, widely distributed across the Transcarpathian area but absent in modern standard Ukrainian. In this respect, the Lemko lect, as part of the Transcarpathian area, places relatively strong emphasis on temporal anchoredness of identificational constructions, expressed through the explicit marking of past, present, and future forms of the verb 'to be'. In addition, aspectual marking appears less grammaticalised, as evidenced by persistent tense/aspect tagging issues, and more semantically diffuse. This may point to a grammatical system that differs not only from modern standard Ukrainian but potentially also from other Transcarpathian lects.

6 Conclusion

The paper applies Stanza to the quantitative linguistic description of the Lemko lect (East Slavic, Transcarpathian region). It identifies areal distributions in syntactic structures and lexical items that complicate tagging by models trained on neighbouring lects, most notably copular predications of the type *short determinative + auxiliary verb + noun*. The research produces a morphologically tagged dataset of Lemko from the 1920s–1930s, enriched with variation annotation and topic modelling. The tagging underwent manual verification by a linguist specialising in East Slavic languages and can therefore be considered reliable.

Future work should prioritise extending the dataset to additional Lemko texts and, more broadly, to other Transcarpathian materials from the same period. At present, much of the description remains contrastive; only the topic modelling component allows the Kamienka lect to be examined on its own terms.

For a fully adequate quantitative description, it would be preferable to develop a model designed specifically for Lemko morphosyntax. Any future topic-modelling study should incorporate stopword removal in order to obtain a clearer thematic profile of the texts. A crucial long-term objective is the development of a model that represents the lect in a more transparent and linguistically interpretable manner (Chung and Chou, 2025).

Limitations

LA1407 (Nakonečna and Rudnyc'kyj, 1940, 31–37), the primary material of this study, does not represent Lemko (or Transcarpathian lects more generally) in their entirety; a large portion of the available material is still undergoing digitisation. Moreover, LA1407 reflects the speech of a single Lemko speaker, which may influence the observed distributions.

Ethical Considerations

The data were published in printed form and have been available for research purposes for fifty to ninety years at the time of writing. Nevertheless, the metatagging has been anonymised where possible, masking speaker names in order to mitigate potential ethical issues related to historical data collection practices.

The texts contain occasional references to xenophobic behaviour and religious (primarily Christian) imagery. Reader discretion is advised.

Disclosure of Generative AI use

This study does not employ generative AI in the research process. While Stanza (Qi et al., 2020) technically belongs to the broader class of generative models in the modern colloquial meaning (the decoder models with more than a billion parameters trained on high-resource corpora), it operates at a much smaller scale and is locally reproducible. During the editing stage, the author used generative AI tools (Grammarly and OpenAI) solely for language polishing where non-native proficiency might otherwise have limited grammatical or stylistic clarity. The intellectual content of the article is entirely human-authored.

Acknowledgements

I thank the anonymous reviewers for their insightful feedback, which substantially improved the article. I am also grateful to the speakers whose recorded speech preserves the studied lects, to the scholars responsible for the original transcriptions, and to the research teams who produced the revised versions. Special thanks are due to Olha Fedorivna Mygolynets (ukr. Ольга Федорівна Миголинець, University of Uzhhorod) for her invaluable assistance with transcription systems and the phonetics of the analysed lects.

References

Lucien Adam and Victor Henry. 1880. *Arte y vocabulario de la lengua chiquita. Con algunos textos traducidos y explicados compuestos sobre manuscritos inéditos del XVIII siglo*. Maisonneuve y Cia., Paris.

Ilia Afanasev. 2026. Quantitative lect description: A case study of lemko from the field data of 1920s-1930s - supplementary material.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. BelarusianGLUE: Towards a natural language understanding benchmark for Belarusian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.

Daniele Baglioni and Luca Rigobianco. 2024. *Chapter 1 Rethinking Fragmentariness and Reconstruction: An Introduction*, pages 1 – 25. Brill, Leiden, The Netherlands.

Kushagra Bhatia, Divyanshu Aggarwal, and Ashwini Vaidya. 2021. Fine-tuning distributional semantic models for closely-related languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 60–66, Kiyv, Ukraine. Association for Computational Linguistics.

Beatrice Bindi. 2025. Evaluating stanza and udpipe for morphosyntactic annotation of old russian: A case study on maximus the greek. *Scripta & e-Scripta*, 25:39–60. Pages: 22. Language: English. Published by: Institute for Literature, Bulgarian Academy of Sciences.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Dorothea Frances Bleek. 1929. *Bushman Folklore*. The African Review.

Wilhelm Heinrich Immanuel Bleek and Lucy Catherine Lloyd. 1911. *Specimens of Bushman Folklore*. George Allen & Company.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Thomas J. Bolt, Jeffrey H. Flynt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. A stylometry toolkit for Latin literature. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 205–210, Hong Kong, China. Association for Computational Linguistics.

Jonathan Brindle. 2017. *A dictionary and grammatical outline of Chakali*. Number 2 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Maria Z. Brooks. 1975. *Polish Reference Grammar*. De Gruyter Mouton, Berlin, Boston.

Andrey Chirkin, Svetlana Kuznetsova, Maria Volina, and Anna Dengina. 2025. RusConText benchmark: A Russian language evaluation benchmark for understanding context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1158–1170, Vienna, Austria. Association for Computational Linguistics.

Meng-Hsuan Chung and Chao-Ting Tim Chou. 2025. Climbing towards the nlu of the universal reading of shei 'who'. *Concentric*, 51(2):303–348.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. `https://www.eva.mpg.de/lingua/resources/glossing-rules.php`. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Michael Daniel, Nina Dobrushina, and Dmitry Ganenkov, editors. 2019. *The Mehweb language*. Number 1 in Languages of the Caucasus. Language Science Press, Berlin.

Salvatore Del Gaudio. 2017. *An Introduction to Ukrainian Dialectology*. Peter Lang Verlag, Berlin, Germany.

Salvatore Del Gaudio. 2017. *An introduction to Ukrainian dialectology*. Wiener slawistischer Almanach. Linguistische Reihe Sonderband 94. Peter Lang, Frankfurt am Main Bern Wien.

William Diver. 2012. Theory. meaning as explanation: Advances in linguistic sign theory. In Alan Huffman and Joseph Davis, editors, *Language: Communication and Human Behavior: The Linguistic Essays of William Diver*, pages 445–519. Brill, Leiden/Boston. Revised and reprinted from the 1995 original: Contini-Morava, E., & Sussman-Goldberg, B. (Eds.). (1995). *Meaning as Explanation: Advances in Linguistic Sign Theory* (pp. 43–114). Mouton de Gruyter.

M. M. Dobrotvorskii. 1875. *Ainsko-russkij slovar' [Ainu-Russian Dictionary]*. Universitetskaya tipografiya[Kazan' University Typography], Kazan'.

Domingo de Santo Tomás. 1560. *Grammatica o Arte de la lengua general de los Indios de los Reynos del Peru*. Valladolid. First grammar of the Quechua language, printed in Valladolid. Often attributed to the Dominican friar Domingo de Santo Tomás.

A. D. Dulichenko. 1981. *Slavjanskie literaturnye mikrojazyki: voprosy formirovanija i razvitija [Slavic Literary Microlanguages: Questions of Formation and Development]*. Valgus, Tallinn.

Henryk Fontański and Mirosława Chomiak. 2000. *Gramatyka języka łemkowskiego*. Śląsk, Katowice.

Erica C. García. 1989. Quantitative aspects of diachronic evolution:: The synchronic alternation between o.sp. y, alli' there'. *Lingua*, 77(2):129–149.

V. E. Goldin and O. Yu. Kryuchkova. 2011. Korpus russkoi dialektnoi rechi: kontseptsiia i parametry otsenki [Corpus of Russian Dialectal Speech: Concept and Evaluation Parameters]. In *Komp'uternaia lingvistika i intellektual'nye tekhnologii : Materialy ezhegodnoi Mezhdunarodnoi konferentsii, Bekasovo, 25–29 maia 2011 goda [Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference, Bekasovo, May 25–29, 2011]*, volume 10, pages 359–367, Moscow. Russian State University for the Humanities.

Tetiana Vasylivna Hromko. 2020. Stanovlennia monohovirkovoi deskryptsii u vitchyznianomu movoznavstvi (kinets' xix – 40-i roky xx st.) [formation of monographic description in domestic linguistics (end of the xix – 40s of the xx century)]. In M. Pantiuk, A. Dushnyi, and I. Zymomria, editors, *Aktual'ni pytannia humanitarnykh nauk: mizhvuzivs'kyj zbirnyk naukovykh prats' molodykh vchenykh Drogobyts'koho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka [Current Issues of the Humanities: Interuniversity Collection of Scientific Works of Young Scientists of the Drohobych Ivan Franko State Pedagogical University]*, Vypusk 34, Tom 2, pages 118–123. Vydavnychyi dim "Hel'vetyka", Drohobych.

Geoffrey D. Kimball. 1991. *Koasati grammar*. Brill, Lincoln.

Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher, Björn Hansen, Dušica Filipović Đurđević, and Nataša Fritz. 2022. *Clitics in the wild*. Number 7 in Open Slavic Linguistics. Language Science Press, Berlin.

Matyáš Kopp, Anna Kryvenko, and Andriana Rii. 2023. Ukrainian parliamentary corpus ParlaMint-UA 4.0.1. Slovenian language resource repository CLARIN.SI.

Z. Lahjouji-Seppälä and A. Rabus. 2021. A robust approach to variation in carpathian rusyn: Resampling-based methods for small data sets. *Jazykovedný časopis*, 72(2):603–617.

Thomas Lehman. 1989. *A grammar of modern Tamil*. Number 1 in Pondicherry Institute of Linguistics and Culture publications. Pondicherry Inst. of Ling. and Culture, Pondicherry.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Paul R Magocsi. 2015. *With their backs to the mountains : a history of Carpathian Rus' and Carpatho-Rusyns*. Central European University Press,, Budapest :.

Paul Robert Magocsi. 2004. Jazykovyj vopros [the language question]. In Paul Robert Magocsi, editor, *Rusyns'kyj jazyk [The Rusyn Language]*, Najnowsze dzieje języków słowiańskich, pages 39–66. Uniwersytet Opolski, Opole.

B. Munkácsi. 1894. *A vogul nyelvjárások szóragozásának ismertetve [Description of the Conjugation of Vogul Dialects]*. Budapest.

Hanna Nakonetschna, Jaroslau Rudnyćkyj, and Ilia Afanasev. 2025. Computer-assisted study of historical lemkian (transcarpathian ukraine) lects: basic vocabulary approach - supplementary material 1 (dataset).

Hanna Nakonečna and Jaroslav Bohdan Rudnyc'kyj. 1940. *Ukrainische Mundarten : Südkarpatoukrainisch ; (Lemkisch, Bojkisch und Huzulisch) [Ukrainian dialects: South Carpathian Ukrainian; Lemkian, Bojkian and Huzulian]*. Arbeiten aus dem Institut für Lautforschung an der Universität Berlin ; 9. Otto Harrassowitz, Berlin.

Saudah Namyalo, Alena Witzlack-Makarevich, Anatole Kiriggwajjo, Amos Atuhairwe, Zarina Molochieva, Ruth Gimbo Mukama, and Margaret Zellers. 2021. *A dictionary and grammatical sketch of Ruruuli-Lunyala*. Number 5 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Vladimir Neumann. 2025. Effektiver einsatz von nlp-methoden am beispiel des codex suprasliensis [effective use of nlp methods using the example of the codex suprasliensis]. *Scripta & e-Scripta*, 25:79–100. Pages: 22. Language: German. Published by: Institute for Literature, Bulgarian Academy of Sciences.

Toshiki Osada. 1992. *A reference grammar of Mundari*. Tokyo Univ. of Foreign Studies, Inst. for the Study of Languages and Cultures of Asia and Africa (ILCAA), Tokyo.

Ricardo Otheguy. 2002. *Saussurean Anti-Nomenclaturism in Grammatical Analysis: A Comparative Theoretical Perspective*, pages 373–403. John Benjamins Publishing Company.

Ricardo Otheguy and Nancy Stern. 2011. On so-called spanglish. *International Journal of Bilingualism*, 15(1):85–100.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.

A. K. Polivanova. 2013. *Staroslavjanskij jazyk: Grammatika. Slovari [Old Church Slavonic Language: Grammar. Dictionaries]*. Universitet Dmitrija Pozharskogo, Moscow.

Robert Pugh and Francis Tyers. 2024. Experiments in multi-variant natural language processing for Nahuatl. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 140–151, Mexico City, Mexico. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

A. Rabus. 2018. Obrazovanie prošedšego vremeni v raznovidnostjach karpatorusinskogo: kvantitativnyj analiz [the formation of the past tense in varieties of carpathian rusyn: A quantitative analysis]. In Kvetoslava Koporova, editor, *20 rokiv vyšokoškol'skoj rusynistikÿ na Slovakiji [20 Years of University Rusyn Studies in Slovakia]*, pages 139–151. Prešovská univerzita v Prešove, Prešov.

Achim Rabus and Yves Scherrer. 2017. Lexicon induction for spoken Rusyn – challenges and results. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 27–32, Valencia, Spain. Association for Computational Linguistics.

Nathanaël Carraz Rakotonirina, Corentin Kervadec, Francesca Franzon, and Marco Baroni. 2025. Evil twins are not that evil: Qualitative insights into machine-generated prompts. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 48–68, Suzhou, China. Association for Computational Linguistics.

Wallis Reid. 2011. The communicative function of English verb number. *Natural Language & Linguistic Theory*, 29(4):1087–1146.

Evelina Rennes, Marina Santini, and Arne Jönsson. 2022. The Swedish simplification toolkit: Designed with target audiences in mind. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 31–38, Marseille, France. European Language Resources Association.

Antonio Ruiz de Montoya. 1640. *Arte y vocabulario de la lengua guaraní*. Madrid, Madrid. Published as a quarto.

M. N. Saenko. 2018. Netochnosti v opisanii semantiki, vyzvannye vosprijatiem dialektnoj leksiki skvoz' prizmu literaturnogo jazyka: neskol'ko primerov [inaccuracies in the description of dialect lexis semantics, caused by literary language interference. a few examples from the east slavic dialect dictionaries]. In L. Ė. Kalnyn', editor, *Issledovanija po slavjanskoj dialektologii 19–20. Slavjanskie dialekty v sovremennoj jazykovoj situacii. Dialektnyj slovar' kak sposob issledovanija slavjanskix dialektov [Studies in Slavic dialectology 19–20. Slavic dialects in the modern language situation. Dialect dictionary as a method of studying Slavic dialects]*, pages 218–222. Institut slavjanovedenija RAN, Moscow.

Yves Scherrer and Achim Rabus. 2019. Neural morphosyntactic tagging for rusyn. *Natural Language Engineering*, 25(5):633–650.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Oleksiy Syvokon, Olena Nahorna, Pavlo Kuchmiichuk, and Nastasiia Osidach. 2023. UA-GEC: Grammatical error correction and fluency corpus for the Ukrainian language. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 96–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Sali A. Tagliamonte. 2025. *Analysing Sociolinguistic Variation*. Cambridge University Press.

Lena Terhart. 2024. *A grammar of Paunaka*. Number 7 in Comprehensive Grammar Library. Language Science Press, Berlin.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. KazBench-KK: A cultural-knowledge benchmark for Kazakh. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.

Ferdinand Johann Wiedemann. 1884. *Grammatik der Syrjänischen Sprache mit Berücksichtigung ihrer Dialekte und des Wotjakishen [Grammar of the Syrjän Language with Consideration of its Dialects and of the Wotjak]*. Russian Academy of Sciences, St. Petersburg.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.

Alina Wróblewska. 2018. Extended and enhanced polish dependency bank in universal dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics.

Daniel Zeman. 2017. Slovak Dependency Treebank in Universal Dependencies. *Jazykovedný časopis / Journal of Linguistics*, 68(2):385–395.

Maurice L. Zigmond, Curtis G. Booth, and Pamela Munro. 1991. *Kawaiisu : a grammar and dictionary, with texts*. Number 119 in Univ. of California publications. Linguistics. Univ. of California Press, Berkeley, CA.

Ivan M Zilyns'kyj. 1933. *Karta ukraïns'kych hovoriv : z pojasnennjamy ; mirylo 1:4.000.000*. Praci Ukräins'koho Naukovoho Institutu 14. Ukraïns'kyj Naukovyj Instytut, Warszawa.

Ludwig Zukowsky. 1924. Beitrag zur kenntnis der säugetiere der nördlichen teile deutsch-südwestafrikas [contribution to the knowledge of the mammals of the northern parts of german south west africa]. *Archiv für Naturgeschichte*, 90(A, 8):43–139.

Jurij Volodymyrovyč Ševel'ov. 1979. *A historical phonology of the Ukrainian language*. Historical phonology of the Slavic languages ; 4. Winter, Heidelberg.

A Example sources

The modern standard Ukrainian examples are from *dev* branch of (Kopp et al., 2023), date of access: February 12, 2026. The Polish examples are from *dev* branch of Wróblewska (2018), date of access: February 12, 2026. The Slovak examples are from *dev* branch of (Zeman, 2017), date of access: February 12, 2026.

B LDA hyperparameters

Parameter	Value
seed	1590
num_topics	10
alpha	auto
epochs	300
passes	500
random_state	0
taken topics	2 – 9

Table 6: LDA model hyperparameters. Parameter "taken topics" denotes the topics selected from the result of topic modelling in the order that the model yielded.

The order of subject, object, and verb in Tatyshly Udmurt

Daria Belova
Institute of Linguistics, RAS
HSE University
dd.belova@yandex.ru

Irina Khomchenkova
Institute of Linguistics, RAS
irina.khomchenkova@yandex.ru

Abstract

We conduct a preliminary study of the order of subject (S), object (O), and verb (V) in Tatyshly Udmurt (Finno-Ugric) on the basis of approximately 900 clauses from oral folklore and non-folklore narratives (including contemporary texts and texts recorded earlier) using a gradient approach. We show that the most frequent word orders are SOV, SV, and OV. In full clauses (with both S and O), in folklore texts SOV order (≈ 70%) is followed by OSV order (≈ 15%). In contemporary non-folklore texts, however, SOV order competes with SVO order (50% vs 30%), which may be explained by the influence of Russian. We note that full clauses may differ from clauses with only S or with only O: in contemporary folklore texts VS order is much more frequent in S-only clauses (≈ 23%) than in full ones (≈ 4%), and in contemporary non-folklore texts VO order is more frequent in full clauses (≈ 35%) than in O-only ones (≈ 12%). Moreover, we show that word order can depend on the type of clause. For example, in existential clauses the order is almost always SV, while clauses with verbs of speech are often VS.

1 Introduction

The study deals with word order in Tatyshly Udmurt, which is an understudied and low-resource subdialect of Udmurt (Permic < Finno-Ugric < Uralic). We will present quantitative data on the most important elements (S, O, V) of a clause.

Udmurt is mainly spoken in the Udmurt Republic (Russia); a significant number of speakers also live compactly in the Republics of Bashkortostan, Tatarstan and Mari El, in Perm Krai, and in Sverdlovsk and Kirov Oblasts (Russia), see, for example, (Edygarova, 2022, 507). The speakers of the Tatyshly subdialect live in the Tatyshly district of the Republic of Bashkortostan.

The vast majority of Udmurt speakers are bilingual in Udmurt and Russian (Salánki, 2007). Apart from this, Tatyshly Udmurt is significantly influenced by contact Turkic idioms (dialects of Tatar and Bashkir), see e.g. (Baidoullina, 2003, 5–7).

Udmurt is often considered a non-rigid SOV language: SOV order is the most common, while other orders are possible depending on the information structure (Bulyčov, 1947; Gavrilova, 1970; Timerkhanova, 2011; Karpova, 2015, *inter alia*). However, some researchers claim that VO order can be pragmatically unmarked as well as OV order. For example, Asztalos (2021) shows that Udmurt is undergoing a typological change from OV to VO based on her elicitation data: younger speakers tend to use VO order more often than older ones, which she explains by the influence of Russian basic SVO (Bailyn, 2012, 239–244). She also reports an areal difference: Udmurt speakers from Tatarstan used VO order less often than speakers from the Udmurt Republic. The author suggests that this may be explained by the additional influence of Tatar, which is also a (non-rigid) SOV language (Kashaeva, 2012, 77–78).

Tatyshly Udmurt may serve as another curious example of a "double-influenced" SOV language variety, since it is spoken in the area where both Russian and Turkic languages are spread. To show whether word order serves as a parameter of variation between the Udmurt dialects, one should analyze it in Tatyshly Udmurt and compare it with Tatar and Bashkir data on the one hand and the data of other Udmurt varieties on the other. However, existing research on Udmurt, Tatar, and Bashkir provides only qualitative analysis with detailed descriptions and illustrations of the possible arrangement of various sentence elements, and no statistical analysis is conducted. This hinders our ability to directly compare these data.

This paper focuses on Tatyshly Udmurt with the aim of getting reliable results and opening the Udmurt data to typological comparison by using statistical corpus methods.

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 60–65

2 Materials and methods

In linguistic typology, word order is an essential variation parameter. Some authors use a type-based approach (in terms of Levshina, 2019) and classify languages as SVO, SOV, etc based on the dominant order. Such word order types are listed for many languages in databases such as WALS (Dryer and Haspelmath, 2013) or Grambank (Skirgård et al., 2023).

There is another approach, the token-based (Levshina, 2019) or gradient (Levshina et al., 2023) approach. Within this framework, the labelling of languages as SOV, SVO etc. or as having fixed, flexible, or free word order is substituted for (or supplemented by) the proportion of different orders and the degree of word order variability (Levshina et al., 2023).

The degree of variability in word order is evaluated using Shannon entropy (Levshina et al., 2023, 850). The Shannon entropy is computed as follows (Shannon, 1948):[1]

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i))$$

Shannon entropy quantifies the average level of uncertainty associated with a set of possible outcomes. Low entropy indicates high predictability. For example, the entropy is zero if there is only one outcome. The entropy is maximal if there is a uniform distribution; in this case, the outcome is the hardest to predict. Thus, the higher the entropy, the more variable the word order.

In our study, we use this gradient approach and count the proportion and degree of variability of different word orders in Tatyshly Udmurt using corpus data. This approach is chosen because Udmurt in general is considered as a SOV language, but there is clear interdialectal variability, and it is important to have quantitative data.

The corpus of Tatyshly Udmurt[2] contains approximately 69.5 thousand tokens (as of November 2025). The corpus consists of two subcollections. The first one is sound-aligned transcriptions of recordings done in 2019, 2021–2025 during fieldwork. The second one is texts from previously published literature that were collected in 1963–2003. Further in the article we will call them the *newer* and the *older* texts. Note that both subcorpora are oral and not written texts. It is also important to note that since we do not have access to source materials, we do not know whether the published texts have undergone any editorial work. Nevertheless, this data is important to assess language change.

The main genres in the corpus are folklore narratives, non-folklore narratives, and non-folklore dialogues. For our preliminary analysis, we used the folklore narratives: 13 newer texts with 3338 tokens (from 9 speakers) and 12 older texts (which were mainly recorded in 1970) with 2899 tokens. The folklore sample was then compared with non-folklore narratives: 24 newer texts with 5288 tokens from 13 speakers and 11 older texts with 1387 tokens. In total, we analyzed texts from 15 speakers (9 women, 6 men) from 23 to 79 years old (the median age is 60).

From the sample, we manually selected and annotated 928 declarative clauses (both affirmative and negative) with noun phrases (including pronouns) as S and O, see Table 1. We did not include clauses without overt S and O and several other types of clauses (e.g., existential, with verbs of speech, with non-verbal predicates, imperative and interrogative sentences), since word order proportions in such clauses may differ (see some observations in Section 4).

	Newer		**Older**	
	Folklore	Other	Folklore	Other
S and O	70	48	44	11
Only S	179	155	122	55
Only O	61	107	29	47
Total	310	310	195	113

Table 1: Clause types under consideration.

Note that folklore and non-folklore subcorpora differ in the number of clauses with both S and O, which we will call "full clauses". It may be related to the fact that there are more nominal subjects in folklore texts than in other types of narratives. It also makes it difficult to balance the sample, especially with the older texts. In this article, we examined all the older texts and all the newer folklore texts available in the corpus to maximize the sample size. The number of newer non-folklore monologues was chosen so that it roughly corresponded to the newer folklore in number of clauses.

[1] The Shannon entropy can be calculated, e.g., here: https://www.mytimecalculators.com/tools/entropy-calculator/.

[2] The corpus is available at the following link: https://udmurt.web-corpora.net/tatyshly/index_en.html

Let us examine the types of clauses attested in our sample. The first group consists of full clauses (with both S and O), as in (1).

(1) *äd'žämi vedra kut-em*
person bucket grasp-PST2
'The man took a bucket.' (older)

Clauses with only S are mainly intransitive (2). Transitive clauses without an overt O, as in (3), were also classified as S-only clauses.

(2) *äǯ'ämi lə̑kt-e*
person come-PRS.3SG
'The man comes.' (newer)

(3) *tolez' žal'a-∅*
moon pity-PRS.3SG
'The moon pities [her].' (newer)

Finally, clauses with only O are transitive clauses without an overt subject, as in (4).

(4) *tolez'-ez a$^{?}$ǯ'-i-z*
moon-ACC see-PST-3SG
'[She] saw the moon.' (newer)

3 Results

3.1 SOV vs other word orders

The most frequent word order in both subcorpora is SOV, see Tables 2–3. In folklore texts, it is attested in about 70% of all clauses. The second most common word order is OSV, which was attested in about 15% of all folklore examples. Other orders are much less common. The entropy of the newer and the older folklore subcorpora is quite similar: $H_{newer} = 1.27$ and $H_{older} = 1.24$.

	Folklore	Other
SOV	**51 (72.85%)**	**24 (50%)**
OSV	10 (14.3%)	4 (8.3%)
SVO	6 (8.6%)	**15 (31.25%)**
VSO	2 (2.85%)	1 (2.1%)
OVS	1 (1.4%)	3 (6.25%)
VOS	0	1 (2.1%)

Table 2: Word order: newer subcorpus.

In non-folklore newer texts, the entropy is higher: $H = 1.8$, which means that word order is more variable. SOV order is still dominant (50%), but the second most common order is SVO (≈ 30%). In comparison, in newer folklore texts SVO order is rather rare (≈ 9%). The more common use of SVO order in non-folklore newer texts may be due to the influence of Russian, and its rarity in newer folklore texts may be explained by the genre: folklore texts tend to conserve more archaic features, including fossilized verbal constructions.

	Folklore	Other
SOV	**31 (70.4%)**	9 (81.8%)
OSV	8 (18.2%)	0
SVO	4 (9.1%)	1 (9.1%)
VSO	1 (2.3%)	0
OVS	0	1 (9.1%)
VOS	0	0

Table 3: Word order: older subcorpus.

The number of clauses with both S and O in non-folklore older texts is rather low, so it is difficult to draw any reliable conclusions based on these data.

3.2 SV vs VS

To analyze order of S and V, apart from the examples with full clauses (with both S and O) discussed above, we also used S-only clauses. We found that the difference between full clauses and S-clauses in the newer folklore subcorpus is statistically significant according to the Fisher exact test ($p = 0.0003$), see Table 4. The entropy differs as well: $H_{S,O} = 0.26$ and $H_S = 0.78$.

	S and O	Only S
SV	67 (95.7%)	138 (77.1%)
VS	3 (4.3%)	41 (22.9%)

Table 4: The order of S and V: newer folklore texts.

However, in newer non-folklore texts, there was no such correlation ($p = 0.6317$), see Table 5. The total entropy is $H_{S,O} = 0.57$.

	S and O	Only S	Total
SV	43	133	176 (86.3%)
VS	5	23	28 (13.7%)

Table 5: The order of S and V: newer non-folklore texts.

As for the older subcorpus (see Table 6), the difference between these two types of clauses is not statistically significant ($p = 0.2911$ and $p = 1$). Word order in the non-folklore subcorpus is a little more variable, cf. the entropy scores: $H_{folklore} = 0.35$ and $H_{other} = 0.55$.

	S and O	Only S	Total
Folklore			
SV	43	111	154 (93.3%)
VS	1	10	11 (6.7%)
Other			
SV	9	45	54 (87.1%)
VS	1	7	8 (12.9%)

Table 6: The order of S and V: older texts.

To sum up, the most frequent order is SV. In the newer folklore subcorpus, VS order in S-clauses is more common than in the other subcorpora.

It seems curious that one subcorpus out of four stands out in such a manner. We suppose that in this case we observe the influence of a specific speaker. Table 7 presents the distribution of SV and VS order in S-only clauses from two individual speakers with the most data in the folklore subcorpus (other 7 speakers in our dataset have less than 20 S-only clauses each). Both speakers under consideration are women, Speaker A was born in 1951, and Speaker B was born in 1978. The speakers also differ in profession and place of residence. Speaker B uses VS clauses much more frequently than Speaker A, and this difference is statistically significant ($p = 0.008702$).

	Speaker A	Speaker B
SV	41 (87.2%)	41 (65.1%)
VS	6 (12.8%)	22 (34.9%)

Table 7: Word order in S-only clauses in newer folklore texts of two speakers

Thus, the explanation through individual preferences seems plausible, but the existing data are insufficient to draw any strong conclusions based on sociolinguistic factors.

Apart from this, different clause types are worthy of separate consideration when analyzing SV/VS order. As an example, we will discuss clauses with existential verbs (Subsection 4.1) and with verbs of speech (Subsection 4.2).

3.3 OV vs VO

In addition to full clauses, we analyzed examples without an overt subject. We found that the difference between full clauses and O-clauses is statistically significant in the newer non-folklore subcorpus ($p = 0.0016$), but not in the newer folklore subcorpus ($p = 0.1017$), see Tables 8–9. The more frequent occurrence of VO order in newer non-folklore texts in full clauses with both S and O can also probably be explained by the Russian influence (cf. Table 2: we have shown that SVO order is frequent in newer non-folklore texts).[3]

	S and O	Only O
OV	31 (64.6%)	94 (87.9%)
VO	17 (35.4%)	13 (12.1%)

Table 8: The order of O and V: newer non-folklore texts.

	S and O	Only O	Total
OV	62	47	109 (83.2%)
VO	8	14	22 (16.8%)

Table 9: The order of O and V: newer folklore texts.

The difference between these two clause types is absent in both folklore and non-folklore texts from the older subcorpus, see Table 10.

	S and O	Only O	Total
Folklore			
OV	39	26	65 (89%)
VO	5	3	8 (11%)
Other			
OV	10	47	57 (98.3%)
VO	1	0	1 (1.7%)

Table 10: The order of O and V: older texts.

Overall, OV order is the most preferred in all of the corpora; however, it is the most common in older non-folklore texts. The entropy scores are as follows: $H_{newerfolklore} = 0.65$, $H_{olderfolklore} = 0.5$, $H_{olderother} = 0.13$.

4 Clause types

4.1 Existential clauses

Asztalos (2021) in her elicitation tasks analyzed SV/VS order in existential sentences separately. (In general Asztalos (2021) considered only existential and predicative possessive sentences when discussing SV vs VS order.) She showed that the speakers of the older generation demonstrated unanimity in their choice of the SV order, while the younger participants showed greater variability in their preferences: the majority of respondents

[3] Unfortunately, we are not able to check whether there is any individual influence, because we have even fewer data on each speaker than in SV/VS comparison (see Subsection 3.2).

from the Udmurt Republic and a third of respondents from the Republic of Tatarstan judged VS order at least as good as the SV order.

We annotated 29 existential clauses, such as (5), with the markers *van'* 'EXST', *jevâl* 'NEG' *val* 'be.PST', *vâlem* 'be.PST2', and the verb *luəńə* 'be'.

(5) *mânam tod-em-e van'*
1SG.GEN know-NMLZ-POSS.1SG EXST
'I have knowledge.' (newer)

The difference between subcorpora has not been attested, see Table 11. As we have shown in Subsection 3.2, in newer folklore texts S-only clauses have VS order more often than in clauses with S and O. Thus the distribution of SV/VS order in existential clauses and S-only clauses differs. However, this difference is not statistically significant.

	Newer	Older	Total
Folklore			
SV	17	9	26 (89.7%)
VS	3	0	3 (10.3%)
Other			
SV	28	19	47 (95.9%)
VS	1	1	2 (4.1%)

Table 11: The order of S and V: existential sentences.

4.2 Verbs of speech

In some cases, on the contrary, VS order is more widespread. This is typical for verbs of speech, see Table 12. We annotated these clauses in the folklore subcorpus, since they were almost absent in the non-folklore texts.

	Newer	Older	Total
SV	8	12	20 (27.4%)
VS	22	31	53 (72.6%)

Table 12: The order of S and V: verbs of speech.

VS order is attested when the reported speech precedes the verb (6) while SV order is used when it follows the verb (7). (This is also noted in (Vakhrushev, 1974, 130).)

(6) *d'źeź gəne iź-i-∅ no, šü-e*
good only sleep-PST-1SG ADD say-PRS.3SG
äd'źämi
person
'I slept pretty well, the man says.' (older)

(7) *pijosmurt šü-em: danag ȯj uža*
man say-PST2 plenty NEG.PST.1SG work
'The man said: "I haven't done much work."' (older)

Thus, if one uses automatically annotated texts to study the order of S and V and does not remove speech predicates from consideration, the overall proportion of word orders may change significantly.

5 Conclusion

We presented a preliminary study of the order of S, O and V in the corpus of Tatyshly Udmurt. We analyzed subcorpora that differ in genre (folklore, non-folklore) and in time period (2019–2025, 1963–2003). The most frequent orders are SOV, SV, and OV. In full sentences (with both S and O), SOV word order is the most common in all subcorpora. In the newer non-folklore subcorpus, the second most frequent word order is SVO. This fact distinguishes this subcorpus from both newer folklore and older non-folklore texts. Hence, it may be taken as a sign of Russian influence on contemporary Tatyshly Udmurt speech. Our findings show that folklore texts should be treated separately from non-folklore ones. An approach that combines these two text classes as a single sample (such as in (Karpova, 2015)) may be too simplistic. Moreover, word order proportions depend on argument structure and clause type. They may be different in full, S-only and O-only clauses. Clauses with verbs of speech demonstrate "deviating" word order distributions. Unlike other S-only clauses in all four subcorpora, in clauses with verbs of speech VS order is the most common.

Limitations

Although comparing contemporary speech with texts from 30 to 60 years ago gives us important insights into contact-induced language change, consideration of sociolinguistic parameters such as age and language skills would be beneficial for future research. These limitations are closely tied to the Tatyshly subdialect being a low resource language variety: the volume of the corpus, especially the older subcorpus, and different proportions of texts from individual speakers in different subcorpora impede our ability to properly assess whether and how sociolinguistic parameters influence our results. The use of other methods such as elicitation is needed to verify corpus findings, as well.

Abbreviations

1, 3 — 1st, 3rd person; ACC — accusative; ADD — additive particle; EXST — existential marker; GEN — genitive; NMLZ — nominalization; NEG — negation; POSS — possessive suffix; PRS — present tense; PST, PST2 — past tense; SG — singular.

Acknowledgments

This research is supported by Russian Science Foundation, RSF project No. 24-18-00199 "Clause structure and positional phenomena in SOV languages" carried out at the Institute of Linguistics, Russian Academy of Sciences.

References

Erika Asztalos. 2021. From head-final towards head-initial grammar. generational and areal differences concerning word order usage and judgement among Udmurt speakers. In *Language contact in the territory of the former Soviet Union*, pages 143–182, Amsterdam/Philadelphia. John Benjamins Publishing Company.

Anna Baidoullina. 2003. Tatyshlinskii govor udmurtskogo yazyka: fonetika i morfologiya [Tatyshly subdialect of Udmurt: phonetics and morphology]. Master's thesis, University of Tartu.

John Bailyn. 2012. *The syntax of Russian*. Cambridge University Press, Cambridge.

Mikhail Nikandrovich Bulyčov. 1947. *Porjadok slov v udmurtskom prostom predlozhenii [Word order in Udmurt simple clauses]*. Udmurtgosizdat, Izhevsk.

Matthew Dryer and Martin Haspelmath. 2013. Wals online (v2020.4). Dataset.

Svetlana Edygarova. 2022. Udmurt. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 507–522. Oxford University Press, Oxford.

Tatyana Gennadievna Gavrilova. 1970. Porjadok slov v udmurtskom prostom povestvovatel'nom predlozhenii [Word order in Udmurt simple declarative sentences]. In *Zapiski. Issue 21: Philology*, pages 107–118, Izhevsk. Udmurtskij NII istorii, ekonomiki, literatury i jazyka pri Sovete Ministrov Udmurtskoj ASSR.

Lyudmila Leonidovna Karpova. 2015. Osobennosti poryadka slov v udmurtskoi narodno-razgovornoi rechi (na materiale severnykh dialektov) [Peculiarities of word order in the Udmurt local colloquial speech (on the material of northern dialects)]. *Bulletin of Udmurt University. History and Philology Series*, 25(1):85–91.

Goljihan Kashaeva. 2012. The Tatar IP-field. In *Generative Grammar in Geneva*, volume 8, pages 77–94, Geneva. University of Geneva.

Natalia Levshina. 2019. Token-based typology and word order entropy. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Zsuzsanna Salánki. 2007. *The present-day situation of the Udmurt language*. Ph.D. thesis, Eötvös Loránd University.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, and 100 others. 2023. Grambank v1.0. Dataset.

Nadezhda Nikolaevna Timerkhanova. 2011. Osobennost' porjadka slov v prozaičeskix proizvedenijax G. E. Vereshchagina i v sovremennom udmurtskom jazyke [Word order in the prosaic works of G. E. vereshchagin and in contemporary Udmurt]. In *Tipologicheskie aspekty mnogojazychija v sovremennom obrazovatel'nom prostranstve [Typological aspects of multilingualism in the modern educational space]*, pages 180–185, Izhevsk. Udmurtskij Universitet.

Vasily Maksimovich Vakhrushev. 1974. *Grammatika sovremennogo udmurtskogo yazyka: sintaksis slozhnogo predlozheniya [The grammar of contemporary Udmurt: compound sentence syntax]*. Udmurtiya, Izhevsk.

Author Index

Afanasev, Ilia, 46
Akumbu, Pius Wuchu, 38

Belova, Daria, 60
Bowern, Claire, 31

Daul, Massimo Marie, 31

Faytak, Matthew, 38

K H, Manodyna, 8
Khomchenkova, Irina, 60

Le Ferrand, Éric, 38
Levow, Gina-Anne, 16
Liang, Siyu, 16

Mawkanuli, Talant, 16

Njuasi, Ivo Forghema, 38

Stephen, Abishek, 1

Tosolini, Alessio, 31

van Dam, Kellen Parker, 1

Yang, Tianle, 38