Automated Morphosyntactic Analysis for Low-Resource Languages: A Comparative Study of Large Language Models

Anonymous EMNLP submission

Abstract

This comparative study investigates the use of LLMs for automatic morphosyntactic analysis of low-resource languages, such as Čakavian (Slavic) and Istro-Venetian (Romance). We assess three leading models, Gemini 3.1 Pro, Claude 4.6 Sonnet, and GPT 5.3. Using samples of audio interviews that have been manually transcribed and annotated as the ground truth, we provided text transcriptions to the LLMs with instructions to perform the tasks of lemmatization, part-of-speech tagging, morphosyntactic feature annotation, and translation. Our results show that dividing the transcription into short segments (segment-wise prompting) consistently performs better than zero-shot prompting. Our detailed analysis reveals that LLMs leverage similar high-resource languages to bridge data gaps, but they struggle with forms specific to the related low-resource varieties. However, using LLMs for automatic annotation with human-in-the-loop reduces manual labor while maintaining the necessary accuracy for language documentation. Furthermore, our comparative testing methodology can be scaled to evaluate and map LLM capabilities across many low resource languages.

1 Introduction

Despite rapid advances in the capabilities of Large Language Models (LLMs), low-resource languages continue to pose technical obstacles (Cahyawijaya et al., 2024), due to the scarcity of data. They also represent a sociolinguistic and cultural problem that requires urgent attention. Many of these languages are endangered and are facing the possibility of language death in the near future (Li et al.; Nguyen et al., 2024). They represent a source of cultural and linguistic knowledge that should be documented. Low resource languages also pose a problem of fairness in large language models: speakers of these languages do not have equal access to AI tools and may be disadvantaged in various other ways by the fact that their varieties are drastically under-represented or entirely absent from the training data used to develop these tools (Sindhujan et al., 2025; Robinson et al., 2023).

Audio or video data collected as part of language documentation projects must be transcribed, which can be extremely time-consuming. This is a well-known problem, commonly referred to as the transcription bottleneck, and there is a noticeable body of literature describing how AI tools can be used to mitigate this issue. However, for detailed linguistic analysis or other specialized purposes, additional annotations are also required, which include lemmatization, part of speech tagging, morphosyntactic feature tagging, word- and utterance-level translations, and/or other information. Transcription alone can take 40 or more hours per hour of audio (Himmelmann, 2018; Seifart et al., 2018), and manually adding multiple layers of annotation is even more time-consuming and requires specialist knowledge.

A case in point is the ELIC Corpus (Langston et al., 2026), consisting of over 60 hours of sociolinguistic interviews in four endangered languages spoken in the Istria and Kvarner regions of Croatia: Čakavian, Istriot, Istro-Romanian, and Istro-Venetian. Data are transcribed and time-aligned with the audio at the utterance and word levels, then annotated as described above. The annotation process has proven to be a major roadblock. Finding and training annotators for minority varieties is extremely difficult; the process of manual annotation is error-prone, requiring extensive review and correction; and the time required and costs are prohibitive.

Automated Morphosyntactic Analysis is the core of computational linguistics, yet for low resource languages,it remains under explored.Evaluating LLM models on low resource languages exposes the existing gaps of current cross-lingual transfer

methods as whether the patterns learned from high resource languages are generalized towards low resource languages. Developing benchmarks helps researchers annotate and analyze large amounts of unannotated data.This study bridges the gap between LLM evaluation for low resource languages and in preserving the language digitally by accelerating the analysis and documentation for low resource languages.

The goal of this study is to evaluate whether contemporary LLMs such as Gemini 3.1 Pro (Google DeepMind, 2026), GPT 5.3 (OpenAI, 2026), Claude 4.6 Sonnet (Anthropic, 2026) can be used to automate the morphosyntactic analysis pipeline for low-resource varieties such as those in the ELIC Corpus (Alam et al., 2024). We assess their performance based on:

- Lemmatization
- Part-of-speech tagging
- Morphosyntactic feature specification
- English glossing of individual lemmas

We focus here on samples from two languages, Čakavian and Istro-Venetian, which are closely related to higher resource languages included in the model training data: Croatian and Italian. However, these models have not been trained on the specific task of morphosyntactic annotation. Including an LLM in the pipeline for automation would allow researchers more time for linguistic analysis (Li et al., 2025; Vidal-Gorène et al., 2025; Ginn et al., 2024; Elsner and Liu, 2025; Zhong et al., 2024).

2 Background

2.1 LLMs and Machine Translation for Low-Resource Languages

Recent work shows that LLMs can produce fluent translations and can be competitive for high-resource pairs, but performance drops sharply in genuinely low-resource conditions (Court and Elsner, 2024; Moslem et al., 2023; Zhu et al., 2024; Zhong et al., 2024). Multiple studies argue for hybrid pipelines that combine LLMs with strong encoder–decoder NMT components, retrieval/fuzzy matching, or task-specific adapters to reduce data-scarcity failures (Moslem et al., 2023; Shu et al., 2024; AmericasNLP Organizers, 2025). Even when multilingual transfer enables partial translation in very-low or zero-resource settings (Zhu et al., 2024; Nguyen et al., 2024), persistent weaknesses remain in idiomatic and figurative language (Donthi et al., 2025; Ahmed, 2025) as well as systematic artifacts when comparing LLM/NMT outputs to human translations (Sizov et al., 2024). Safety filters can also lead to refusal, omission, or distortion in sensitive contexts (Tekgürler, 2025). Overall, the MT literature suggests LLMs are best treated as powerful components (post-editors, rerankers, data augmenters) rather than universal replacements for MT systems in low-resource settings (AmericasNLP Organizers, 2025; Shu et al., 2024).

2.2 Pre-LLM Morphosyntactic Tagging and Lemmatization

South Slavic has comparatively strong pre-LLM infrastructure for tokenization, POS tagging, morphological feature annotation, lemmatization, and parsing. The CLASSLA-Stanza ecosystem extends Stanza-style pipelines to multiple South Slavic languages and supports both standard and web/noisy varieties (Terčon et al., 2025; Terčon and Ljubešić, 2023; Ljubešić and Kuzman, 2024a; Qi et al., 2020). Earlier feature-rich CRF-based systems demonstrated strong joint tagging and lemmatization, particularly when combining annotated corpora with morphosyntactic lexicons (Ljubešić and Erjavec, 2016; Ljubešić et al., 2016). Work on Slovene user-generated content (Janes) shows that mixing standard and non-standard data and leveraging lexicon/cluster/normalization features significantly improves tagging robustness (Fišer et al., 2018; Ljubešić et al., 2017a; Zupan et al., 2019). In shared-task and social-media settings, comparisons of CRF vs BiLSTM-style neural taggers indicate consistent gains from character-aware neural models on morphology-heavy distinctions (Ljubešić, 2018; Meftah et al., 2018; Zampieri et al., 2018). Language-specific transformer pretraining further improves tagging accuracy; for example, BERTić-style models outperform multilingual baselines on South Slavic morphosyntax benchmarks (Ljubešić and Lauc, 2021; Ljubešić et al., 2017b; Ljubešić and Dobrovoljc, 2019).

For Romance languages such as Italian relied on the combination of rule based systems and statistical classifiers. Italian as a language has structural and lexical challenges, with regards to structural ambiguity where identical wordforms map to distinct parts of speech(POS) or lemma (Favaro et al., 2023) .To mitigate these challenges, early architec-

tures used Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) for tokenization and lemmatization (Alzetta and Montemagni, 2025). Datasets were introduced which are built on Itailian dependency tree banks (Favaro et al., 2023; Baffelli, 2016; Korkiakangas, 2020; Marotta et al., 2020). Preprocessing pipelines as TextPro (Pianta et al., 2008),Tanl (Attardi et al., 2010) combined language specific vocabularies along with supervised learning methods to map the syntactical tagging (Baffelli, 2016; Favaro et al., 2023) .While these helped in having interpretable and computationally efficient syntactic annotations, handling non-orthography, dialectical variations remained a bottleneck (Favaro et al., 2023).

2.3 Dialects, Minority Varieties, and Cross-Lingual Transfer

A major thread of morphosyntactic research addresses dialects and minority lects with scarce labeled data. Multi-source transfer and lexicon induction for spoken Rusyn demonstrates that combining related-language supervision (e.g., Russian/Ukrainian/Slovak/Polish) with induced lexicons can yield substantial gains for POS and full morphology (Scherrer and Rabus, 2019, 2017; Rabus and Scherrer, 2017). Similar lessons appear in border lect and dialect scenarios, where tagset alignment and lect-aware transcription choices can produce sizeable accuracy improvements over generic multilingual models (Afanasev, 2023). For South Slavic dialect speech, Torlak corpus and tool-adaptation work supports the view that small dialectal training samples can improve downstream tagging when integrated with larger standard-language resources (Vuković, 2021; Ljubešić and Kuzman, 2024b). Outside Slavic, results on dialectal Romance and endangered languages reinforce that lexicon expansion, active learning, and cross-lingual projection can be more cost-effective than large-scale manual annotation in low-resource settings (Hopton and Aepli, 2024; Vergez-Couret and Urieli, 2014; Anastasopoulos et al., 2018; Plank et al., 2018).

2.4 General Cross-Lingual Morphological Tagging Methods

Beyond dialect-specific adaptation, several general frameworks support morphosyntactic transfer to low-resource targets. Projection-based approaches map tags from high-resource languages via aligned bitext, enabling weak supervision at scale (Buys and Botha, 2016; Agić et al., 2015). Iterative pipelines for analyzer induction extend these ideas toward fuller morphosyntactic analyzers and lemmatizers (Nicolai and Yarowsky, 2019). Neural models that explicitly represent dependencies among morphological features (e.g., factorial CRF formulations with neural potentials) help predict unseen feature bundles in low-resource settings (Malaviya et al., 2018). Character-level multilingual transfer also improves robustness for morphologically rich languages (Cotterell and Heigold, 2017). These methods are relevant to Istrian varieties because they offer routes to bootstrap morphological annotation even when direct training data is limited.

2.5 LLMs for Lemmatization, POS Tagging, and Glossing

Recent studies suggest LLMs can serve as practical annotators in low-resource and historical contexts, especially with carefully designed prompting. Unsupervised prompting strategies that exploit English-dominant capabilities can improve low-resource performance without explicit fine-tuning (Nguyen et al., 2024). For joint POS tagging and lemmatization in under-resourced historical languages, few-shot prompting can rival strong supervised baselines, although complex morphology and non-Latin scripts remain challenging (Vidal-Gorène et al., 2025). For documentary-style outputs such as interlinear glossed text, multiple pipelines combine retrieval and structured prompting to improve morpheme-level glossing and reduce systematic errors (Shandilya and Palmer, 2024; Ginn et al., 2024; Elsner and Liu, 2025; Shandilya and Palmer, 2023; Yang and Nicolai, 2025). Finally, LLM-centered annotation pipelines emphasize phased prompt design, pre-hoc evaluation, batch annotation, and post-hoc validation as a general recipe for building reliable LLM-driven corpora (Morin and Larsson, 2025).

3 Dataset and Annotation Framework

3.1 Dataset Composition

The data used for evaluation consist of two sample interview segments from the ELIC corpus, one in Čakavian (duration 2m 55s, recorded in Trviž) and one in Istro-Venetian (duration 9m 08s, recorded in Pula). Each recording has two speakers, an interviewer and a language consultant, both of whom are native speakers of the local varieties.

3.2 ELAN Annotation Structure

The interview data have been transcribed and time-aligned at the utterance and word levels in Praat (Boersma, Paul and Weenink, David, 2026) and manually annotated using ELAN (Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive, 2026). The corrected manual annotations are treated as the ground truth for evaluation of LLM performance. ELAN files (.eaf) are XML files with an unlimited number of user-defined, hierarchically organized tiers. The tiers used in the ELIC project are described below.

Independent tier

- `Text@speakerID`: Time-aligned orthographic transcription of individual utterances

Tiers dependent on Text tier

- `engTrans@speakerID`: Free English translation
- `hrvTrans@speakerID`: Croatian translation (for non-Slavic varieties)
- `Words@speakerID`: Word-level tokenization, time-aligned subdivision of Text tier

Tiers dependent on Words tier

- `Lemma@speakerID`: Citation form of each word
- `PoS@speakerID`: Part-of-speech category for each word, using Universal Dependencies (De Marneffe et al., 2021) POS tags
- `Feats@speakerID`: Morphosyntactic features, using an adapted version of the Universal Dependencies inflectional feature values
- `Gloss@speakerID`: Literal English translation of the lemma

4 Experiment

The manually transcribed and corrected Text tier was exported from ELAN as a plain text file (.txt) with speaker labels identifying the utterances by the interviewer and language consultant. This text file was given as input to the LLMs Gemini 3.1 Pro, GPT 5.3, and Claude 4.6 Sonnet along with the prompt. Due to the timeline for the experimental evaluation and data processing, recent models such as GPT 5.5 ,Gemini 3.5 Flash and Opus 4.7 are not available during our primary evaluation phase, we will consider benchmarking these models for our future work. For the experiment, we considered every 10 lines as a **segment**. We performed both Zero-shot and Segment-wise prompting as described below.

- For **Zero-shot prompting**, we attached the transcript of the entire interview portion along with the prompt. For the sake of easier comparison, the results were then arranged into segments as described below.
- For **Segment-wise prompting**, the plain text file containing the transcription was divided into short segments consisting of 10 lines each. Each segment was given separately to the LLM after the prompt.

In order to empirically evaluate the data, we calculate an Annotation Error Rate (AnnER) for the tiers Words, Lemma, POS, and Feats, using the same formula as for WER. We categorize errors as:

- **Deletion** - annotation is missing
- **Substitution** - annotation is partly/wholly substituted by another value
- **Insertion** - an annotation is provided that is not present in the corresponding ground truth manual annotation

Because there can be legitimate differences in annotations (e.g., Čakavian *i* meaning 'even' (indicating emphasis) is described variously as a particle (POS=PART) or adverb (POS=ADV) and the order of morphosyntactic features can also vary (e.g., Masc.Nom.Sing vs. Nom.Sing.Masc), comparison of LLM annotations with the ground truth was performed manually by one of the authors. Different analytical choices or equivalent labels were not treated as errors. For morphosyntactic feature values, there can also be different levels of specificity. LLM-supplied values that were correct, but incomplete when compared to the corresponding ground truth annotation, were also not penalized as errors. Ambiguous forms, whose interpretation depends on the morphosyntactic context, were treated as errors when the LLM-provided annotation differed from the ground truth.

Prompt

Role: You are the Lead Annotator and Slavic Dialectologist for an Istrian Linguistic Corpus project. Your expertise covers Istrian Chakavian, Istro-Venetian, their historical development (jat reflexes and other phonological changes, infinitival reduction, language contact), and other characteristics of the languages spoken in the northern Adriatic region.
Task: Perform a multi-tiered interlinear linguistic analysis of the provided transcript segments. 1. Mandatory Data Pipeline For every input segment, follow this precise workflow:
Normalization: Strip non-linguistic markers (((noise)), @, etc.). Lemmatization: Provide dialectal lemmas, not standard language equivalents. Follow standard conventions for citation forms in Slavic and Romance languages Suppletion → list both forms as the lemma (e.g., dobar|bolji). POS Tagging (Rule 5.3): Use modified UD tags (CONJ for all conjunctions; NAME for personal names; DET for pronominal modifiers).
Morphosyntactic features: Use the UD universal features
2. Output Schema Each segment must be presented as a Markdown table with the following strict columns:
Word: Surface form. Lemma: Dictionary form (Slavic dialectal or Romance dialectal). POS: (NOUN, VERB, ADP, AUX, NAME, PROPN, PRON, ADJ, ADV, CONJ, NUM, DET, INTJ, PART). Feats: Morphological features (e.g., Masc.Nom.Sing, Pfv.Inf, Pres.3Sing, Past.Ptcp.Fem.Sing). Gloss: English translation of the corresponding Lemma Notes: Etymological origin (e.g., "Venetian"), notes on the reflex of jat or other dialectal differences
Eng Translation: At the end of the segment, English translation describing the whole segment.

5 Results

The Performance Table 1 of the three LLMs, Claude 4.6 Sonnet, GPT 5.3, and Gemini 3.1 Pro, are evaluated on the two datasets Čakavian and Istro-Venetian with the settings as Zero-Shot and Segment-Wise Prompting. Performance is evaluated using the metric of Annotation Error Rate (AnnER), calculated as the sum of the error types Deletion, Substitution, Insertion divided by the total number of annotations of the given type (Words, Lemma, POS, Feats). Figure 1 shows the improvements(ΔAnnER) of Segment-wise prompting over zero-shot prompting.

5.1 Dataset differences

The improvement in Annotation Error Rates differs between the two datasets 1. With Čakavian, the Zero-shot AnnER is relatively low (e.g., 4.11% for words and 4.59% for Feats for GPT 5.3). With segment-wise prompting, the AnnER further drops to as low as 0.48% for Gemini 3.1 Pro. The improvements on average range from 3%-10% AnnER. The Istro-Venetian dataset has high zero-shot AnnER rates (e.g., 96.88% for words, 97.01% for Lemma for Gemini 3.1 Pro) due to the very high number of deletions (e.g., 1426 for words and POS). Segment-wise prompting reduces these errors noticeably, with AnnERs decreasing from 96.88% to 7.07% for Gemini 3.1 Pro in words and from 97.01% to 8.02% in lemma for Gemini 3.1 Pro. noticeable improvement is identified for words in GPT 5.3 and claude4.6 Sonnet as 0.95% and 0.68%. Although the datasets involve two different languages, they are similarly noisy and we see across-the-board improvements with segment-wise prompting. A major difference between the two is the total number of words in the two transcripts: 414 words for Čakavian vs 1472 for Istro-Venetian. This indicates that the size of the dataset plays a major role. Performance improves as the number of words in the sample to be annotated decreases.

5.2 Linguistic level variation

Improvements are consistent across all levels (Words, Lemma, POS and Feats) Figure 1. For both datasets, major improvement is observed at the word and lemma levels. For example, for the Istro-Venetian dataset, the word level AnnER drops from 32.54% to 0.95% for GPT 5.3 and from 28.87% to 0.68% for Claude 4.6 Sonnet. Improvements for POS and Feats are smaller in comparision to Words and Lemma but they are still noticeable (e.g., POS AnnER Reduces from 35.87% to 7.27% for GPT5.3 on the Istro-Venetian dataset.

Dataset	Level	Metric	Zero-shot			Segment-wise		
			Gemini 3.1 Pro	GPT 5.3	Claude 4.6 Sonnet	Gemini 3.1 Pro	GPT 5.3	Claude 4.6 Sonnet
Čakavian	Words	Deletion	24	16	36	2	8	11
		Substitution	0	1	1	0	0	1
		Insertion	0	0	0	0	0	0
		AnnER ↓	5.80%	**4.11%**	8.94%	**0.48%**	1.93%	2.90%
	Lemma	Deletion	24	16	36	4	6	9
		Substitution	14	101	102	7	102	62
		Insertion	1	1	1	1	1	1
		AnNER ↓	**9.42%**	28.50%	33.57%	**2.90%**	26.33%	17.39%
	POS	Deletion	24	16	36	3	4	9
		Substitution	19	17	21	26	16	17
		Insertion	1	1	1	1	1	1
		AnNER% ↓	10.63%	**8.21%**	14.01%	7.25%	**5.07%**	6.52%
	Feats	Deletion	14	10	23	2	3	2
		Substitution	6	7	11	4	27	13
		Insertion	44	2	1	46	2	2
		AnNER ↓	15.46%	**4.59%**	8.45%	12.56%	7.73%	**4.11%**
Istro-Venetian	Words	Deletion	1426	479	424	101	13	8
		Substitution	0	0	0	2	0	0
		Insertion	0	0	1	1	1	2
		AnNER ↓	96.88%	32.54%	**28.87%**	7.07%	0.95%	**0.68%**
	Lemma	Deletion	1426	488	428	106	11	12
		Substitution	2	19	25	11	47	14
		Insertion	0	0	1	1	1	3
		AnNER ↓	97.01%	34.44%	**30.84%**	8.02%	4.01%	**1.97%**
	POS	Deletion	1426	483	422	108	17	11
		Substitution	1	45	33	30	88	26
		Insertion	0	0	0	1	2	1
		AnNER ↓	96.94%	35.87%	**30.91%**	9.44%	7.27%	**2.58%**
	Feats	Deletion	849	291	210	52	61	51
		Substitution	2	7	4	3	11	4
		Insertion	0	0	1	18	1	1
		AnNER ↓	57.81%	20.24%	**14.61%**	4.96%	4.96%	**3.80%**

Table 1: Annotation Error Rate (AnnER) where ↓ indicates that lower is better and error count across datasets and linguistic parameters. The error rate was counted across the metrics - Deletion, Substitution and Insertion.

5.3 Comparision of Models

While all the models benefit from segment-wise prompting, there are noticeable differences among them. For Zero-Shot, GPT 5.3 has a lower AnnER than Gemini 3.1 Pro and Claude 4.6 Sonnet on Čakavian, whereas Claude 4.6 Sonnet performs best on the Istro-Venetian dataset. For example, on Istro-Venetian dataset, for Features, the AnnER converges to 3.80% (Claude 4.6 Sonnet) and 4.96% for Gemini 3.1 Pro and GPT 5.3, compared to a larger range in zero-shot setting, ranging from 14.61% to 57.81%. This shows that segment-wise prompting reduces variability across models.

5.4 Error Analysis

The Reduction in AnnER is primarily due to a decrease in deletion errors, which are higher for the larger dataset (Istro-Venetian) and for zero-shot vs. segment-wise prompting, and which propagate across all categories. Deletions of words from the original transcript result in at least the same number of deletion errors for the categories Lemma and POS (not all words are marked for morphosyntactic features, so the numbers are lower here). In contrast, the substitution and insertion errors remain relatively small across all settings, but may increase with segment-wise prompting because the increase in the number of words from the original transcript that are included in the Words tier increases the number of possible errors of these other two types. This shows that zero-shot prompting struggles with covering all the data, by not aligning the elements, whereas segment-wise prompting improves completeness by having consistent alignment across segments.

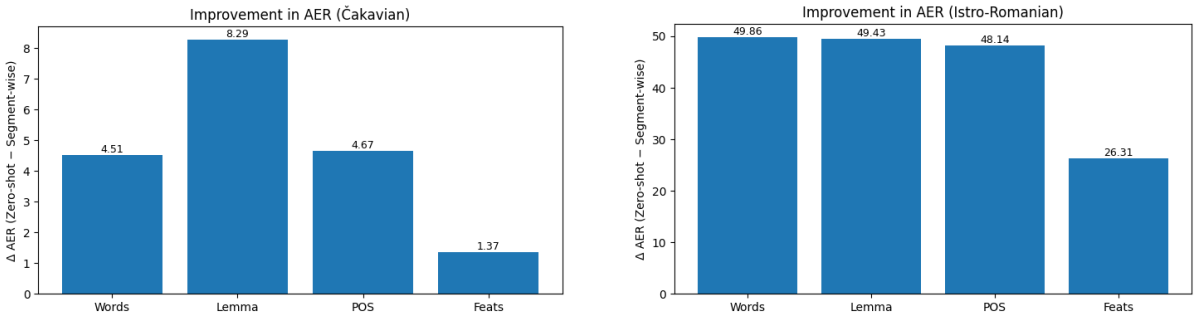

Figure 1: The noticeable improvement in Annotation Error Rate(AnnER) across the datasets from zero-shot to segment-wise results. Bars were calculated as ΔAnnER (zero-shot − segment-wise), where higher values indicate major improvements. Segment-wise prompting reduces error across all linguistic levels consistently, with noticeable changes across Istro-Venetian dataset.

5.5 Summary

Segment-wise prompting gives more consistent results and noticeable reduction in error count across datasets, linguistic metrics, and models. This indicates that segment-wise prompting is effective when zero-shot performance degrades. The Figure 1 analysis further highlights the fact that the improvements are substantial across the datasets.

6 Detailed Analysis

All models delete words from the given transcripts. The ELIC project uses literal transcriptions of interviews, since the utterances and individual words are time-aligned with the audio. With zero-shot prompting, for the shorter Čakavian transcript the deletions are almost all repetitions of words and incomplete utterances, which are frequent in speech. Essentially, the models are normalizing the output to make it more consistent with written language. For the longer Istro-Venetian transcript, model performance degrades substantially, omitting 29% to 97% of the given text, with Gemini 3.1 Pro being especially sensitive to the size of the dataset. With segment-wise prompting, performance improves across all models. However, Gemini 3.1 Pro still omits 7% of the words in the original text for the longer Istro-Venetian sample.

When words from the original transcript are deleted, the corresponding annotations for lemma, POS, and morphological features are consequently also absent. Deletions therefore make up the majority of errors for these categories as well with zero-shot prompting. With segment-wise prompting, the proportion of errors in lemmatization and POS/-morphosyntactic feature-tagging increases. Under both conditions we observe some differences in model performance.

All models generally provide the correct lemmas for words, in the expected Čakavian and Istro-Venetian spellings. GPT 5.3 and Claude 4.6 Sonnet have the largest number of lemmatization errors for Čakavian, which almost all represent the substitution of standard Croatian infinitive endings for Čakavian forms (the infinitive form serves as the lemma for verbs). With segment-wise prompting, Claude corrects most of these errors, providing the correct Čakavian lemmas except for the high-frequency auxiliary verb *biti* 'to be' (Čakavian *bit*). In the Istro-Venetian sample, substitution errors are less frequent overall, and often represent repeated substitutions for a few high-frequency lemmas; e.g., GPT 5.3 consistently substitutes *eser* for *esser* 'to be', and the Croatian form *čakavski* 'Čakavian' for the Istro-Venetian equivalent *ciacavo*.

POS substitution errors are frequent only for forms whose interpretation depends on the context; e.g., Čakavian *onda* 'then (at some time in the past)' as an adverb vs. *onda* 'then' as a conjunction. Gemini 3.1 Pro frequently provides both possible labels, which was not counted as an error; e.g., Čakavian *blizu* 'nearby, near' is labeled as ADV/ADP (adverb or adposition) in a context where it is clearly functioning as the latter: *blizu Pazina* 'near Pazin'.

The same is true for morphosyntactic features, although the models also often identify ambiguous forms correctly, based on the context. In many instances, the models omit some possible feature specifications; e.g., the Istro-Venetian feminine singular definite article *la* 'the' is tagged as Fem.Sing rather than Art.Def.Fem.Sing. Since the prompt did not stipulate which features needed to be specified for every part of speech (which is also language-dependent), these differences were not considered

errors. In general, Gemini 3.1 Pro provided more feature specifications than the other models, and all models specified a somewhat larger number of feature values with segment-wise than with zero-shot prompting. Although the prompt specifies that feature values should be separated by periods, the Gemini 3.1 Pro output was consistently in the feature name=feature value format (e.g., Case=Ins|Gender=Masc|Number=Sing), except in the Istro-Venetian zero-shot condition.

Error rates are not reported for the literal English glosses for individual lemmas. Since multilingual LLMs are known to perform well on translation tasks, it is not surprising that the word-by-word and free translations of utterances provided by all models were highly accurate, even for many Čakavian and Istro-Venetian words that do not occur in the closely related higher-resource languages Croatian and Italian. The Istro-Venetian sample contained a few instances of code-switching to Croatian, which all models annotated correctly.

Errors that could be categorized as hallucinations are rare, and there is not enough data to determine whether zero-shot prompting reduces hallucinations. For example, in the zero-shot context, GPT 5.3 misidentifies the form *preteli* 'friends' as a verb, tags it as a masculine plural past participle, and creates an infinitive to serve as the lemma, **preteliti*, with the gloss 'become friends'. This is at least semantically and morphologically plausible (if accompanied by the reflexive pronoun *se*), but does not fit the syntactic context. With segment-wise prompting, GPT 5.3 correctly identifies this form as a noun with the features Masc.Nom.Plur, but provides the standard Croatian form *prijatelj* as the lemma. All of the models hallucinate a translation for the form *kalotine* 'logs' (defining it as 'blocks', 'pieces', 'shells', 'bowls', 'small bowls', 'stumps', 'stone cap-stones', 'flat stones') with both zero-shot and segment-wise prompting. The most obvious examples of hallucinations would be insertions, which again are relatively rare in our data. Even here, there is often an understandable connection to the actual linguistic structure. For example, prepositions do not have any inherent morphosyntactic features, but Gemini 3.1 Pro assigns case features to them. However, these actually represent the case required for the noun phrase that is the object of the given preposition.

7 Conclusion

This comparative study indicates the potential for a pivotal shift in linguistic documentation. While LLMs cannot provide fully accurate automated morphosyntactic annotations, they can be transformative as assistive technologies for language documentation. Our analysis and findings for the low-resource varieties Čakavian and Istro-Venetian three major LLM Models (Gemini 3.1 Pro, GPT 5.3 and Claude 4.6 Sonnet) demonstrate that they have the strong potential to accelerate corpus development with LLM Assisted Annotation. As our study focuses on Čakavian and Istro-Venetian Languages, the testing methodology can be scaled to analyze model performance for under-resourced and other low resource languages as well. Expanding the comparative benchmarks across diverse languages is important for the NLP community to identify model deficiencies and track our progress towards multilingual LLMs.

While the immediate goal here is to facilitate language documentation by helping overcome the bottleneck of manual annotation, which can help in the preservation and revitalization of endangered languages, this type of detailed linguistic analysis can also assist in the development of better AI models for low-resource varieties more generally. However, it is essential to recognize the strengths and weaknesses of these tools and to use them accordingly, retaining human intervention for quality assurance. Our comparative approach has a strong potential for future applications by having an evaluation for model accuracy, this work enables teams to deploy LLMs for culture preservation.

8 Limitations

The study focuses on the potential of LLMs to help automate the tedious and extremely time-consuming work of morphosyntactic analysis. While they are good for high resource languages, they struggle with low resource languages. Newer Models which were released while doing the experiments such as GPT 5.5, Gemini 3.5 Flash and Opus 4.7 are not included in our active experiments, which can be taken as future direction to benchmark these models as well. Major limitations include:

- **Lack of Training Data :-** Even the most advanced models struggle with features that are missing or inadequately represented in their

training data. This indicates a need for finetuning (Sindhujan et al., 2025; Joshi et al., 2024; Zhong et al., 2024) and the need to look for new methods beyond finetuning for processing low resource languages.

- **Error Patterns:-** Although LLMs can facilitate morphosyntactic analysis, they may provide incorrect or incomplete annotations, which can easily be overlooked if not reviewed carefully. As we are manually annotating and reviewing for corpus development, for the LLM generated output we need manual review.

- **Transition from model performance to data scarcity as a barrier:-**For low resource languages/ endangered languages with small communities of users and less documentation, the problem transitions from model performance towards data scarcity as a barrier. The languages in our sample are closely related to higher resource languages, and the good performance of the LLMs in our test is likely due to similarities between our data and data included in the models' training sets.

9 Ethical Considerations

Ethical concerns related to privacy and consent must be taken into account when using LLMs for the analysis of linguistic data collected from human subjects, since the creators of these models may store data and use them for model training or other purposes that individuals may not consider when they agree to participate in a research project. Minority language communities are often especially sensitive to the appropriation of their language data by outsiders. It is essential to explain possible implications of data sharing and to obtain informed consent in a manner that is appropriate for the specific cultural context. Errors in automated analysis may also misrepresent the language of individual speakers or of the community as a whole, so review and correction of the output is important. There is also the possibility of misuse of shared audio and textual data for malicious purposes. Transparency about the use of LLMs and their potential risks is essential to protect the rights of participants.

Another ethical concern would be the potential for linguistic corruption. As all the models are trained on standardized, high resource languages, they tend to exhibit 'standardization bias'. When given the task of annotation, they can smooth over nuances unique to local varieties or eliminate rare language features by replacing them with more common forms of a dominant related language. There is also the risk of a digital feedback loop. If the outputs are not consistently audited by linguists or native speakers, this might lead to a sterilized version of the language where the very tools used to preserve it could lead to the loss of its unique linguistic identity, eventually resulting in a scenario in which the grammatical norms of high resource languages are imposed on low resource languages. Therefore, human-in-the-loop is not only a technical requirement but also a moral requirement to protect the authenticity of low resource languages.

References

I. Afanasev. 2023. The use of khislavichi lect morphological tagging to determine its position in the east slavic group. In *Proceedings of VarDial 2023*. . Doi:10.18653/v1/2023.vardial-1.18.

Zeljko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. . Doi:10.3115/v1/P15-2044.

Mahdi Aben Ahmed. 2025. From text to understanding the inner text: Llms and translation accuracy and fluency. *International Journal of Language and Literary Studies*, 7(2):139–156. Doi:10.36892/ijlls.v7i2.2072.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Chiara Alzetta and Simonetta Montemagni. 2025. Low- vs high-level lemmatization for historical languages: A case study on italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, Cagliari, Italy. CEUR Workshop Proceedings.

AmericasNLP Organizers. 2025. AmericasNLP 2025 Shared Task 1: Machine Translation Codebase. https://github.com/AmericasNLP/americasnlp2025/tree/main/ST1_MachineTranslation. GitHub Repository; Accessed: May 2026.

Antonios Anastasopoulos, M. Lekakou, J. Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang.

2018. Part-of-speech tagging on an endangered language: a parallel griko-italian resource. In *In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA. Association for Computational Linguistics*, pages 2529–2539.
. Https://aclanthology.org/C18-1214/. aclanthology.org/C18-1214/.

Anthropic. 2026. Claude (version 4.6 sonnet). https://www.anthropic.com/claude. Large language model; Accessed: May 2026.

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. The tanl pipeline. In *Proceedings of the LREC 2010 Workshop on Web Services and Processing Pipelines in HLT*, Valletta, Malta. European Language Resources Association (ELRA).

C. Baffelli. 2016. An annotation pipeline for italian based on dependency parsing. Master's thesis, University of Zurich. Multilinguale Textanalyse MLTA.

Boersma, Paul and Weenink, David. 2026. Praat: doing phonetics by computer [computer software]. Version 6.4.65.

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 (ACL 2016)*.
. Doi:10.18653/v1/P16-1184.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Ryan Cotterell and G. Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of EMNLP 2017*.
. Doi:10.18653/v1/D17-1078.

Sara Court and Micha Elsner. 2024. Low-resource language processing with large language models. Https://arxiv.org/html/2406.15625v1. arXiv preprint.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving llm abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, United Arab Emirates. Association for Computational Linguistics. Abu Dhabi.

Micha Elsner and David Y. Liu. 2025. Prompt and circumstance: A word-by-word llm prompting approach to interlinear glossing for low-resource languages. Doi:10.48550/arXiv.2502.09778.

Manuel Favaro, Marco Biffi, and Simonetta Montemagni. 2023. Pos tagging and lemmatization of historical varieties of languages. the challenge of old italian. *Italian Journal of Computational Linguistics*, 9.

Darja Fišer, Nikola Ljubešić, and T. Erjavec. 2018. *The Janes project: language resources and tools for Slovene user generated content*. Language Resources and Evaluation.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? Doi:10.48550/arXiv.2406.18895.

Google DeepMind. 2026. Gemini 3.1 pro. https://gemini.google.com. Large language model; Accessed: May 2026.

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 Hemmelmann 1998*, number 15 in Language Documentation & Conservation Special Publication, pages 33–40. University of Hawai'i Press.

Zachary Hopton and Noemi Aepli. 2024. Modeling orthographic variation in occitan's dialects. Doi:10.48550/arXiv.2404.19315.

Shreyas Joshi, Muhammad Shahnawaz Khan, Aditya Dafe, Kavita Singh, Vedant Zope, and Tanish Jhamtani. 2024. Fine tuning llms for low resource languages. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 511–519.

Timo Korkiakangas. 2020. Theoretical and pragmatic considerations on the lemmatization of non-standard early medieval latin charters. *Studi e Saggi Linguistici*, 58(1):—.

Keith Langston, Zvjezdana Vrzić, Margaret E. L. Renwick, and John T. Hale. 2026. Endangered languages in contact in istria and kvarner, croatia (elic).

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29544–29559, Suzhou, China. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: A metric for quantifying llm performance across high and low-resource languages.

N. Ljubešić and D. Lauc. 2021. *BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian*. Preprint.

Nikola Ljubešić. 2018. Comparing crf and lstm performance on the task of morphosyntactic tagging of non-standard varieties of south slavic languages. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of slovenian, croatian and serbian. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2019)*. . Doi:10.18653/v1/W19-3704.

Nikola Ljubešić and T. Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of slovene. In *Proceedings of LREC 2016*.

Nikola Ljubešić, T. Erjavec, and Darja Fišer. 2017a. Adapting a state-of-the-art tagger for south slavic languages to non-standard text. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*. . Doi:10.18653/v1/W17-1410.

Nikola Ljubešić, T. Erjavec, Maja Miličević, and T. Samardžić. 2017b. Croatian twitter training corpus reldi-normtagner-hr 2.0. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*.

Nikola Ljubešić, Filip Klubička, Željko Agić, and I. Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In *Proceedings of LREC 2016*.

Nikola Ljubešić and Taja Kuzman. 2024a. Classlaweb: Comparable web corpora of south slavic languages enriched with linguistic and genre annotation. Doi:10.48550/arXiv.2403.12721.

Nikola Ljubešić and Taja Kuzman. 2024b. Classlaweb: Comparable web corpora of south slavic languages enriched with linguistic and genre annotation. Doi:10.48550/arXiv.2403.12721.

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 (ACL 2018)*. . Doi:10.18653/v1/P18-1247.

Giovanna Marotta, Francesco Rovai, Irene De Felice, and Lucia Tamponi. 2020. Classes: Orthographic variation in non-literary latin. *Studi e Saggi Linguistici*, 58(1):—.

Sara Meftah, N. Semmar, F. Sadat, and S. Raaijmakers. 2018. Using neural transfer learning for morphosyntactic tagging of south-slavic languages tweets. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*.

Cameron Morin and Matti Marttinen Larsson. 2025. A large-scale pipeline for automatic corpus annotation using llms: variation and change in the english consider construction. *arXiv:2510.12306*. Doi:10.48550/arXiv.2510.12306.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Finland. European Association for Machine Translation. Tampere.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1*, pages 3501–3516, Thailand. Association for Computational Linguistics. Bangkok.

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. . Doi:10.18653/v1/P19-1172.

Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. 2026. Elan [computer software]. Version 7.1.

OpenAI. 2026. Chatgpt (version gpt-5.3). `https://chatgpt.com`. Large language model; Accessed: May 2026.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Barbara Plank, Sigrid Klerke, and Željko Agić. 2018. The best of both worlds: Lexical resources to improve low-resource part-of-speech tagging. Doi:10.48550/arXiv.1811.08757.

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*. . Doi:10.18653/v1/2020.acl-demos.14.

Achim Rabus and Yves Scherrer. 2017. Lexicon induction for spoken rusyn – challenges and results. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*. . Doi:10.18653/v1/W17-1405.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Yves Scherrer and A. Rabus. 2019. *Neural morphosyntactic tagging for Rusyn*. Natural Language Engineering.

Yves Scherrer and Achim Rabus. 2017. Multi-source morphosyntactic tagging for spoken rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*. . Doi:10.18653/v1/W17-1210.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language Documentation Twenty-Five Years On. *Language*, 94(4):e324–e345.

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 SIGMORPHON 2023*. . Doi:10.18653/v1/2023.sigmorphon-1.9.

Bhargav Shandilya and Alexis Palmer. 2024. Boosting the capabilities of compact models in low-data contexts with large language models and retrieval-augmented generation. Doi:10.48550/arXiv.2410.00387.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. ArXiv preprint.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of llms, nmts, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Florida, USA. Association for Computational Linguistics. Miami.

Merve Tekgürler. 2025. Llms for translation: Historical, low-resourced languages and contemporary ai models. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 227–237, New Mexico. Association for Computational Linguistics. Albuquerque.

Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić. 2025. *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*. Contributions to Contemporary History.

Luka Terčon and Nikola Ljubešić. 2023. Classla-stanza: The next step for linguistic processing of south slavic languages. Doi:10.48550/arXiv.2308.04255.

Marianne Vergez-Couret and A. Urieli. 2014. Pos-tagging different varieties of occitan with single-dialect resources. In *Proceedings of the First Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2014)*. . Doi:10.3115/v1/W14-5303.

Chahan Vidal-Gorène, Florian Cafiero, and Bastien Kindt. 2025. Under-resourced studies of under-resourced languages: lemmatization and pos-tagging with llm annotators for historical armenian, georgian, greek and syriac. *HAL Open Science*, h 485:al–05119. Https://hal.science/hal-05119485.

Teodora Vuković. 2021. *Representing variation in a spoken corpus of an endangered dialect: the case of Torlak*. Language Resources and Evaluation.

Changbing Yang and Garrett Nicolai. 2025. Learning beyond limits: Multitask learning and synthetic data for low-resource canonical morpheme segmentation. Doi:10.48550/arXiv.2505.16800.

Marcos Zampieri, S. Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James R. Glass, Yves Scherrer, T. Samardžić, Nikola Ljubešić, J. Tiedemann, Chris van der Lee, S. Grondelaers, Nelleke Oostdijk, D. Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second vardial evaluation campaign. In *Proceedings of VarDial 2018*.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong sky Zhang, Weihang You, Yiheng Liu, Ocean Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, Xiang Li, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Findings of the Association for Computational Linguistics: NAACL*, 2024, , Mexico City, Mexico. Association for Computational Linguistics:2765–2781.

Katja Zupan, Nikola Ljubešić, and T. Erjavec. 2019. *How to tag non-standard language: Normalisation versus domain adaptation for Slovene historical and user-generated texts*. Natural Language Engineering.

A Appendix

All the datasets and models used for comparision are publicly available. All the artifacts the datasets and LLM models were used strictly for linguistic analysis and academic benchmarking which is consistent with the intended use cases.