Low-Resource Morphological Inflection for Kashubian

Lucy Yang[1] **and Shu Okabe**[1,2] **and Alexander Fraser**[1,2,3]
[1]School of Computation, Information and Technology, Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
Corresponding authors: {lucy.yang,shu.okabe}@tum.de

Abstract

Generating inflected word forms requires substantial labelled datasets. Kashubian, a low-resource Slavic language, lacks sufficient annotated pairs for this task. To address this data scarcity, we fine-tune the byte-level transformer ByT5 on datasets from 7 related languages. We also continue pre-training on Kashubian and Polish monolingual corpora, and generate synthetic Kashubian word forms using LLM annotations. Multilingual fine-tuning and continued pre-training improve the monolingual ByT5 accuracy by 4–5 points, while incorporating synthetic data exceeds the previous strongest zero-shot baseline by 10 points. Training models on multilingual task datasets, monolingual sentences with self-supervision, and LLM-annotated data yields strong relative accuracy gains for the low-resource morphological inflection of Kashubian.

1 Introduction

Morphological inflection aims to modify the root form of a word (lemma) to agree with the sentence structure, such as in case and number in Figure 1. Supervised approaches struggle to inflect low-resource languages on limited labelled data (Cotterell et al., 2017). Similarly, Large Language Models (LLMs) notably show poorer performance on low-resource languages in the related morphological agreement task (Jumelet et al., 2026). Kashubian (ISO: csb, Glottocode: kash1274) is a Slavic language spoken in Northern Poland, with little annotated data (class 1, 'scraping-by', in the taxonomy of Joshi et al. (2020) on language resource status).

This article improves morphological inflection in Kashubian through cross-lingual transfer from Slavic languages, self-supervision on monolingual sentences, and LLM-based data generation. We fine-tune the byte-level transformer ByT5 (Xue et al., 2022) on Polish, its closest related high-resource language; three other West Slavic languages, Czech, Slovak, and Lower Sorbian; two East Slavic languages, Belarusian and Russian; and one South Slavic language, Macedonian.

Lemma	Features	Target
kòt	N;GEN(SG)	kòta

Figure 1: Inflecting the Kashubian noun lemma 'cat' for genitive (GEN) singular (SG).

Alongside multilingual fine-tuning with related languages, we continue pre-training on monolingual Polish and Kashubian sentences using a byte-level span-corruption objective. Finally, we perform zero-shot inflection with two LLMs, Gemini and GPT, and generate Kashubian task data with the better-performing Gemini.

We surpass the baselines through joint training on related languages and continued pre-training on monolingual sentences. However, our best model uses LLM-augmented Kashubian data, reaching 40 points in accuracy. We publish the code[1] and the trained models.[2]

2 Related work

The SIGMORPHON Shared Tasks, first introduced in Cotterell et al. (2016), have continuously addressed the morphological (re)inflection task, including low-resource settings. Later SIGMORPHON Shared Tasks have focused heavily on unseen lemmas (Kodner et al., 2022). The 2019 task evaluated cross-lingual transfer (McCarthy et al., 2019) with an additional 100 training samples of the low-resource language. There, an encoder-decoder architecture achieved the highest performance of 78% transferring from Polish to Kashu-

[1]https://github.com/TUM-NLP/kashubian-morphological-inflection
[2]https://huggingface.co/collections/livles/slavic-byt5-for-kashubian-inflection

bian. This model benefited from multi-source training across related languages (a strategy shown to be effective by Anastasopoulos and Neubig (2019)) and from augmented data that encourages character copying and common affixes. Based on the 2020 data (Vylomova et al., 2020), ByT5 (Xue et al., 2022), a byte-level transformer, reached strong performance on inflection and other word-internal tasks. Following the tasks, we train ByT5 on lemma-split data, augmenting it with related languages, particularly Polish. In contrast to the 2019 task, we include only samples in the high-resource language in our main setting to evaluate zero-shot cross-lingual transfer. Since the data split is also different, our results are not directly comparable.

Several studies have proposed additional methods to overcome data scarcity in low-resource NLP tasks. Murikinati et al. (2020) enhanced cross-lingual transfer for inflection through transliteration between different scripts. Morphological inflection also benefits from self-supervised character-level masking on larger datasets (Wiemerslage and Wense, 2025). Finally, Yoo et al. (2021) generated labelled training data through few-shot LLM prompting and improved text classification performance. Building on these findings, we examine self-supervised and LLM-based data augmentation approaches for morphological inflection in Kashubian.

3 Language material

3.1 Languages of study

Our main language of study is Kashubian (csb), a West Slavic language, spoken in Northern Poland. As closely related languages, we consider four West Slavic languages which are written in the Latin script: Czech (ces), Lower Sorbian (dsb), Polish (pol), and Slovak (slk). We further include two East Slavic languages, Belarusian (bel) and Russian (rus), as well as a South Slavic language, Macedonian (mkd), in our analysis. All three are written in the Cyrillic script. Figure 4 in Appendix A shows the classification of these languages according to Glottolog (Hammarström et al., 2026).

3.2 Morphological inflection dataset

We rely on the UniMorph corpus, which provides language-independent morphosyntactic features for 169 languages with annotations detailed in Sylak-Glassman (2016). UniMorph 4.0 (Batsuren et al., 2022) introduces hierarchical feature annotation to represent case stacking (e.g., `SG;NOM` → `NOM(SG)`) and polypersonal agreement (e.g., `ARGNO1P;ARGAC2S` → `NOM(1,PL);ACC(2,SG)`). We hence adopt the recent UniMorph 4.0 format, also employed in the 2023 Shared Task (Goldman et al., 2023), and convert data from previous releases to the new schema.

Format conversion For all languages except Russian and Belarusian, we manually convert the sequential features to conform to the UniMorph 4.0 convention, as further detailed in Appendix B.

Dataset preparation Following the SIGMORPHON 2023 Shared Task (Goldman et al., 2023), we use 10,000 training samples and 1,000 validation and test samples each, avoiding lemma overlap across the sets. We downsample all language datasets except Belarusian, which is already at the right size in the 2023 Shared Task. We present them in Table 1. We note that Kashubian only features a *test* set (solely of nouns) with fewer than 1,000 instances.

Family	Language	train	dev	test
West Slavic	csb	-	-	509
	pol, ces, dsb, slk	10,000	1,000	1,000
East & South Slavic	bel, rus, mkd	10,000	1,000	1,000

Table 1: Sample size per train, development, and test set in all 8 languages.

4 Experimental methodology

4.1 Morphological inflection model

We use the byte-level transformer ByT5 (Xue et al., 2022) of small parameter size (300M) for our experiments. We choose this model because it is pretrained on 101 languages and has achieved strong performance in inflection.

Input Prompt	Target
Inflect **kòt** using **N;GEN(SG)**	**kòta**

Figure 2: ByT5 Prompt for Kashubian. The lemma *kòt* (English: "cat"), features, and target form are in bold.

Fine-tuning for inflection Given the absence of training data for Kashubian, we first fine-tune ByT5

on *Polish* in a monolingual setting. We then extend this to multilingual settings, incorporating various combinations of West, East, and South Slavic languages. We aggregate 10,000 training samples per language if available. A ByT5 input and target prompt, as presented in Figure 2, incorporates one (lemma, features, target) data triplet. Table 8 (Appendix C) lists the fine-tuning hyperparameters.

Prefix	Prompt
-	Inflect kòt using N;GEN(SG)
Language	**Kashubian**: Inflect kòt using N;GEN(SG)
Lang. family	**Slavic**: Inflect kòt using N;GEN(SG)

Figure 3: The ByT5 input prompt is either enriched with the language name, language family, or not at all.

Language name/family To evaluate whether explicit information about the language of study enhances performance, following Xue et al. (2022), we prepend the language name (`L`) or family (`F`) or an equal (50:50) mix of both (`L/F`) to the input prompt, as in Figure 3. While the language-family tag should incorporate shared intra-family morphological concepts, the language-name tag targets language-specific properties. Combining them in equal proportion allows the model to benefit from both.

4.2 Experimental setup

Baseline models We evaluate the two monolingual baseline models from the SIGMORPHON 2023 Shared Task (Goldman et al., 2023). The non-neural statistical system (Cotterell et al., 2017) predicts the inflected form according to longest-match affixation rules. The neural system (Wu et al., 2021) is based on a character-level transformer with an invariant encoding of morphological features.

Evaluation In line with previous SIGMORPHON Shared Tasks, we evaluate model predictions using exact-match per-prediction accuracy.

5 Cross-lingual transfer results

The in-language evaluations benchmark the model's general inflection capabilities and provide a reference for target-language accuracy. Our main focus is on Kashubian; for consistency, we also report performance for the other Slavic languages to assess how the training settings affect their performance.

5.1 Transfer from Polish to Kashubian

As Kashubian does not have a training dataset on UniMorph, we perform zero-shot transfer from its closest related language, Polish. We compare ByT5 with the two baseline models; all three are trained on Polish inflection data and evaluated on the Kashubian and Polish datasets (for reference).

	NN-BL	N-BL	ByT5
csb	**29.1**	11.2	25.0
pol	**89.8**	**89.8**	77.2

Table 2: Accuracy for three models: the non-neural baseline (`NN-BL`), the neural baseline (`N-BL`), and ByT5. All models are fine-tuned on the Polish *training* set and evaluated on Kashubian and Polish test data.

In this monolingual zero-shot transfer to Kashubian, the non-neural model performs best, followed by the ByT5 model. The neural baseline performs comparably weakly. We note that data unavailability poses a significant challenge, as the Polish score is substantially higher.

5.2 Transfer from Slavic languages

As monolingual zero-shot transfer remains low (below 30%), we fine-tune ByT5 jointly on related Slavic languages. As detailed in Section 4.1, we include explicit language information, such as the language *name* (`L`), the language *family* (`F`), or a balanced set of both (`L/F`), for each language. The assigned prefixes can differ not only between the training and test sets, but also across individual samples.

train prefix	-	L	L/F	L/F	L/F
test prefix	-	L	F	L	L/F
csb	28.9	25.0	**29.5**	26.5	28.5
pol	81.9	**86.9**	83.4	85.9	84.5
ces	82.7	**94.6**	83.5	93.4	88.7
slk	88.9	**95.7**	87.3	**95.7**	88.6
dsb	76.6	**86.4**	73.2	85.0	78.5
avg	71.8	**77.7**	71.4	77.3	73.8

Table 3: ByT5 fine-tuned jointly on four **West Slavic** languages: Polish, Czech, Slovak, and Lower Sorbian (Latin script). The train or test set includes the language name (`L`), the language family (`F`), a balanced set of either (`L/F`), or no prefix (`-`).

West Slavic We fine-tune ByT5 on the four West Slavic languages, Polish, Czech, Slovak, and Lower Sorbian (i.e., 40,000 training instances), and report the results in Table 3. The language-name-family (L/F) model, tested with language-family tags (F), is the best setting for Kashubian, improving the monolingual score of Table 2 (25.0) by nearly 5 points. Since it also outperforms the setting without a prefix, the results suggest that language-family tags help to capture more intra-family morphological phenomena. Conversely, pure language-names (L→L setting) weaken zero-shot transfer because the training data lacks target language tags, as expected. Still, language-family tags yield weaker performance for the trained languages (pol, ces, slk, dsb), where language-name tags (L model) achieve the highest accuracies. The best setting overall is not optimal for Kashubian, due to its lack of training data.

trn	**tst**	csb	pol	ces	slk	dsb	bel	rus	mkd	**avg**
L	L	24.6	**89.0**	**96.4**	**96.2**	**89.1**	**75.5**	**93.2**	**94.9**	**82.4**
L/F	L	**24.8**	88.1	96.0	95.8	88.0	73.4	91.9	94.4	81.6

Table 4: ByT5 is fine-tuned jointly on **West, East, and South Slavic** languages: Polish (pol), Czech (ces), Slovak (slk), and Lower Sorbian (dsb) (all in Latin script), Belarusian (bel), Russian (rus), and Macedonian (mkd) (in Cyrillic script). The train or test set includes the language name (L), or a balanced set of language name/family (L/F) in the instruction prefix.

West, East, and South Slavic Furthermore, we incorporate three East and South Slavic languages, Belarusian, Russian, and Macedonian, all written in Cyrillic script, in addition to the previous four West Slavic languages. In Table 4, we report the scores of the two highest-averaging West Slavic settings (L->L and L/F->L) to assess transfer under optimal general inflection performance. The Kashubian scores decrease below the monolingual and the worst West Slavic score (both 25.0). This highlights that further language differences (such as script) restrict transfer, matching prior findings by Murikinati et al. (2020) and Anastasopoulos and Neubig (2019).

6 Continued Pre-Training (CPT)

6.1 Methodology

To overcome the low accuracy plateau (below 30.0) reached by multilingual fine-tuning, we continue pre-training ByT5 with the byte-level span-corruption objective on monolingual Kashubian (and Polish) sentences. This helps to adapt the model to Kashubian, as the language is not featured in ByT5's pre-training dataset. We also explore incorporating Polish text to better ground the cross-lingual transfer. After monolingual pre-training, we fine-tune the model on inflection, using data from its most closely related language, Polish, as in Section 5.1.

To continue pre-training ByT5, we use 10,000 monolingual sentences of the 2021 Wikipedia data from the Leipzig Corpora Collection (Goldhahn et al., 2012). We continue the pre-training with 9,000 pre-training and 1,000 development sentences. The sentences are either in Kashubian (and the model denoted csb-CPT) or in Polish and Kashubian (csb-pol-CPT). The Polish-Kashubian mix has an equal proportion of sentences in both languages. This aims to capture similarities between the two languages for transfer and to maintain Polish knowledge. We use the same hyperparameters for pre-training and fine-tuning, as shown in Table 8. We mask 25% of the bytes per sequence, as detailed in Appendix C, as it led to the highest validation score in our experiments.

6.2 Results

model	pol	csb-CPT + pol	csb-pol-CPT + pol
csb	25.0	30.5	**31.8**
pol	77.2	79.1	**79.7**

Table 5: ByT5 accuracy for Kashubian and Polish inflection after monolingual continued pre-training (CPT) and fine-tuning on Polish data (pol).

We compare continued pre-training and fine-tuning (denoted CPT name + pol model) with fine-tuning only (pol model) in Table 5. Pre-training on Kashubian improves the score by 5 points, while mixed pre-training enhances it further. Our findings indicate that both intra-family language transfer and target-language context benefit morphological inflection accuracy. We also outperform the best setting for Kashubian using only cross-lingual transfer (in Section 5).

7 Experiments with LLMs

7.1 Zero-shot LLMs vs our approach

Methodology To investigate how well massive pre-training and large-scale models solve this task without additional training, we employ Gemini 2.5

Flash (Comanici et al., 2025), which is already pre-trained on 400 languages, and the large, open-weight GPT OSS-120B (OpenAI et al., 2025). To generate the inflected form, each held-out (lemma, features) pair is provided with task instructions as presented in Figure 5 (Appendix D).

	CPT+pol	GPT	Gemini
csb	31.8	32.2	**36.1**

Table 6: Both zero-shot LLMs (GPT and Gemini) outperform our best approach of the Polish ByT5 with continued pre-training. CPT+pol refers to csb-pol-CPT + pol, the best setting in Table 5.

Results Both LLMs outperform the previous models, showing strong zero-shot inflection capability. We cannot, however, rule out partial data contamination for this performance. Besides, we recall that our model, although fine-tuned, is notably smaller in size (300M parameters).

7.2 LLM data generation experiment

Given the strong results of LLMs on this task and ByT5's competitive in-language performance, we combine fine-tuning ByT5 with data generated by LLMs.

Methodology As Gemini achieved better performance than GPT, we generate training samples with Gemini from words in the Kashubian corpus of Section 6. Appendix D further details the prompt design for data generation. Using the prompt given in Figure 6, the LLM outputs the feature sequence and lemma for ~20,000 inflected forms. After noise filtering, we downsample to the same amount of validation and training samples as for monolingual fine-tuning. We then fine-tune ByT5 on synthetic data alone (denoted syn-csb, with 10,000 samples) and on synthetic data combined with the complete Polish training data (denoted syn-csb+pol, with 20,000 samples). We report the results in Table 7.

	CPT+pol	Gemini	syn-csb	syn-csb+pol
csb	31.8	36.1	35.8	**40.1**

Table 7: Fine-tuning ByT5 on inflection data generated by Gemini, either on Kashubian alone (syn-csb) or with Polish (syn-csb+pol). The latter outperforms the best continued pre-trained ByT5 model (CPT+pol) and zero-shot Gemini.

Results Fine-tuning ByT5 only on the generated Kashubian data for inflection (syn-csb) did not surpass Gemini's zero-shot performance. However, when we further include the actual Polish task data (syn-csb+pol), we reach the highest score of 40%. We note that, despite fine-tuning plain ByT5 on more data (20,000 samples) than the monolingual model (10,000 samples) in Section 5.1, Kashubian accuracy remains lower than that of Polish, leaving room for improvement.

8 Conclusion

This study enhances morphological inflection for Kashubian with the byte-level transformer ByT5, using transfer from related Slavic languages, continued pre-training, and LLM-generated data. We outperform monolingual zero-shot transfer scores using West Slavic languages or by continuing pre-training with monolingual sentences in Kashubian and Polish, which reaches over 30% accuracy. Finally, we fine-tune ByT5 on LLM-annotated synthetic data, reaching 40 points. Our experiments show the potential of the hybrid approach to generate more fine-tuning data and stress the benefit of related languages during training for Kashubian.

Future work will explore few-shot LLM prompting and transliteration of South and East Slavic data from Cyrillic to Latin. Furthermore, the quality of synthetic data can be improved through error analysis. Additionally, we will investigate whether our methods extend to other low-resource languages.

Limitations

Since our experiments focus exclusively on one West Slavic language, Kashubian, the findings may not generalize to other languages, particularly those with different typological properties or with fewer related mid- to high-resource languages. However, since linguistic relatedness has been shown to facilitate cross-lingual transfer (Anastasopoulos and Neubig, 2019), these results still support multilingual approaches as a means of improving inflection scores for other low-resource languages.

The Kashubian test data consists solely of nouns, which introduces a bias in the dataset. Nevertheless, our models are not restricted to nouns and are therefore compatible with other parts of speech; the synthetic data we generated contains verbs or adjectives as well. We note here, however, that the test set is smaller for Kashubian (cf. Table 1).

Ethics statement

LLM-based methods have poor sustainability in terms of material resources, economic cost, and future availability. Our best approach, therefore, integrates LLMs only for data generation, as a one-time step.

We acknowledge that data produced by LLMs may be inaccurate, unreliable, and irreproducible. Nevertheless, our models do not rely exclusively on such synthetic data, as we further pre-train the base model, ByT5, on Kashubian (continued pre-training) and on actual Polish inflection data in the best case. The LLM-generated data is based on words from a Kashubian corpus, and therefore not completely from scratch. Furthermore, evaluation is conducted on the same benchmark dataset, which ensures the reliability of the obtained scores, while we ensure no overlap between the fine-tuning and test sets.

Acknowledgments

We thank the anonymous reviewers for their comments. This work was partly funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. 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.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. Glottolog 5.3.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, and 11 others. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Nikitha Murikinati, Antonios Anastasopoulos, and Graham Neubig. 2020. Transliteration for cross-lingual morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–197, Online. Association for Computational Linguistics.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Hai-Biao Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sébastien Bubeck, Cheng Chang, and 106 others. 2025. gpt-oss-120b&gpt-oss-20b model card.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema).

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, and 9 others. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Adam Wiemerslage and Katharina von der Wense. 2025. Improving low-resource morphological inflection via self-supervised objectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24494–24510, Vienna, Austria. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Classification of Slavic languages

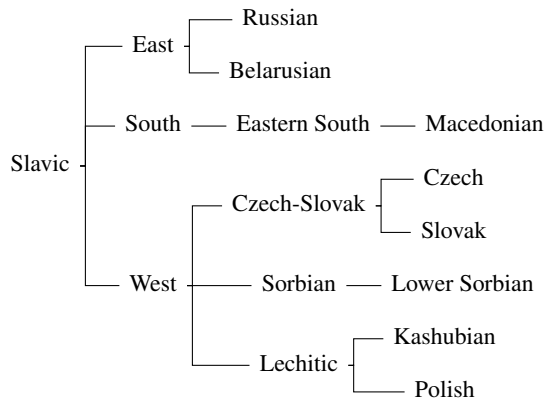

Figure 4: Classification of the Slavic languages in this study.

Figure 4 shows a traditional categorization of the 8 Slavic languages of interest, according to Glottolog (Hammarström et al., 2026).[3] All five West Slavic languages (Polish, Kashubian, Lower Sorbian, Czech, and Slovak) use the Latin script. The two East Slavic and one South Slavic languages (Russian, Belarusian, and Macedonian) use Cyrillic. As illustrated in the Figure, Polish and Kashubian are closely related.

B Data preparation

Feature conversion To transform the features of previous UniMorph releases into the schema of UniMorph 4.0 (Batsuren et al., 2022), we identify the feature categories to analyze the UniMorph 4.0 convention. The person, number, gender, and animacy features are in brackets, stacked under the case feature (e.g. `ADJ;INS;PL` → `ADJ;INS(PL)`). For all languages except Russian and Belarusian, which are already in the correct format, we manually identify the features corresponding to these dimensions, following Sylak-Glassman et al. (2015) and Sylak-Glassman (2016).

C ByT5 training hyperparameters

Fine-tuning We adopt standard fine-tuning hyperparameters used in the summarization task.[4] We disable FP16 to prevent numerical overflow during validation. Additionally, we increase the learning rate and introduce 500 warm-up steps to accelerate convergence and stabilize early training.

[3] https://glottolog.org/

[4] https://huggingface.co/docs/transformers/tasks/summarization

Hyperparameter	Configuration
Learning rate	5×10^{-5}
Number of epochs	4
Weight decay	0.01
Evaluation strategy	Per epoch
FP16 precision	Disabled
Warmup steps	500
Per-device batch size (train / eval)	16

Table 8: Hyperparameters for ByT5 fine-tuning and continued pre-training.

Continued pre-training We apply the same hyperparameters for continued pre-training as for fine-tuning. Raffel et al. (2020) report that the T5 framework, on which ByT5 is built, masks 15% of tokens. In our experiments, we randomly mask 15%, 20%, 25%, and 30% of the bytes per sequence, leaving out the full stop and eos bytes. A masking rate of 25% yielded the best validation performance.

Fine-tuning with LLM-generated data While previous fine-tuning runs achieved validation losses below 0.1, the same configuration but with synthetic task data yielded 0.3. Consequently, we increase the number of training epochs from 4 to 12 and select the best-performing model at each epoch. This configuration is also applied when fine-tuning with the combined synthetic Kashubian and existing Polish data. Fine-tuning the best continued pre-trained model under this configuration decreased performance.

D LLM prompts

We present the prompts used in the LLM experiments of Section 7.

System role:
`You are an expert documentary linguist specializing in morphological inflection in the language Kashubian. You are given a lemma in Kashubian and features in the hierarchical annotation schema UniMorph 4.0. Inflect the lemma according to the features and output only the inflected form.`
Prompt:
`Inflect lemma + features: <lemma> + <features>`

Figure 5: **Zero-shot LLM prompt.** The system role is provided before each input prompt. The test lemma `<lemma>`, and morphosyntactic features `<features>` are inserted in every prompt.

Zero-shot inflection We use an adapted version of the LLM prompt from Ginn et al. (2024), originally designed for interlinear glossing. Figure 5

presents our prompt. It details the task before providing the (lemma, features) bundle.

```
You are an expert documentary linguist specializing in morphological inflection in the Kashubian language. You are given the (inflected) word "<word>". If it is neither a verb, a noun, an adverb, nor an adjective, return <invalid>. Otherwise, return its lemma and morphosyntactic features in the feature schema of UniMorph, as follows: <lemma>\t<features>\t<word>. Here is an example of the morphosyntactic features: N;ACC;SG
```

Figure 6: **LLM synthetic data generation prompt.** From a given <word> of the Kashubian corpus, the LLM generates the morphosyntactic features and the lemma.

Synthesizing Kashubian data For generating Kashubian task data, we modify the previous prompt, as shown in Figure 6. Here, we find the (lemma, features) pair from the inflected form. We choose unique words longer than 4 characters that the LLM classifies as adjectives, adverbs, nouns, or verbs. We only accept lemmas that are absent from our Kashubian test set to avoid lemma overlap between the train and test sets. Outputs longer than expected (e.g., sentences) are discarded. Finally, we downsample to 10,000 training and 1,000 validation forms. To ensure consistent feature schemas, we include an example of the morphosyntactic features (N;ACC;SG). The separator characters ., :, (,), and - in the features are replaced by semicolons (e.g., MASC.NOM → MASC;NOM). These are later converted to the hierarchical feature schema of UniMorph 4.0, as described in Section 3.2.