This paper was published in **Text, Speech, and Dialogue (TSD) 2025** – please cite the published version https://doi.org/10.1007/978-3-032-02551-7_5.

Flexing in 73 Languages: A Single Small Model for Multilingual Inflection

Tomáš Sourada[0009-0003-6792-825X] and Jana Straková[0000-0003-0075-2408]

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic
{sourada,strakova}@ufal.mff.cuni.cz

Abstract. We present a compact, single-model approach to multilingual inflection, the task of generating inflected word forms from base lemmas to express grammatical categories. Our model, trained jointly on data from 73 languages, is lightweight, robust to unseen words, and outperforms monolingual baselines in most languages. This demonstrates the effectiveness of multilingual modeling for inflection and highlights its practical benefits: simplifying deployment by eliminating the need to manage and retrain dozens of separate monolingual models.
In addition to the standard SIGMORPHON shared task benchmarks, we evaluate our monolingual and multilingual models on 73 Universal Dependencies (UD) treebanks, extracting lemma-tag-form triples and their frequency counts. To ensure realistic data splits, we introduce a novel frequency-weighted, lemma-disjoint train-dev-test resampling procedure. Our work addresses the lack of an open-source, general-purpose, multilingual morphological inflection system capable of handling unseen words across a wide range of languages, including Czech. All code is publicly released at: https://github.com/tomsouri/multilingual-inflection.

Keywords: inflection · multilingual inflection

1 Introduction

Morphological inflection, the process of modifying a base word form (lemma) to express grammatical categories (see example in Fig. 1), is a well-established and practically significant task.

Despite the field's high level of activity, driven mainly by the SIGMORPHON shared tasks [2,3,4,7,12,18,24], there is still no open-source, lightweight, neural morphological inflection generator for unknown (out-of-vocabulary, OOV) words in Czech, let alone one supporting multiple languages, available within the Czech academic environment. MorphoDiTa [23] relies on a morphological dictionary for Czech and lacks an out-of-vocabulary (OOV) guesser for generation.[1] UDPipe [22] does not support inflection generation at all. And although Sourada et al. (2024) [20] provide an OOV guesser, it is limited to Czech nouns.

[1] Although MorphoDiTa includes OOV heuristics for morphological analysis, the inverse process of generation, it does not support OOV generation.

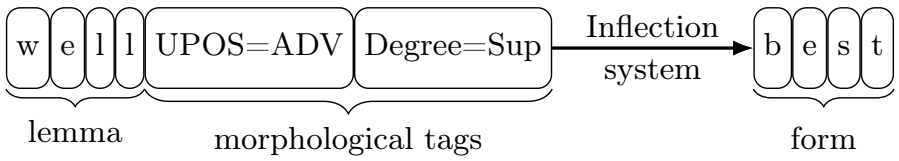

Fig. 1. The morphological inflection task: an input-output example, inflection of English lemma "well" to superlative, inflected form "best".

We address these shortcomings by experimenting with supervised deep learning models for morphological inflection, based on the Transformer architecture, with a focus on potential future production deployment and multilingual applicability.

Our multilingual system, trained jointly on 73 languages, is extremely lightweight, consisting of a single model with 5.69M trainable parameters, capable of performing inflection across all 73 languages. The model handles all parts of speech, inflects unknown words (with evaluation conducted entirely on words unseen during training), and crucially, outperforms both the baseline systems and separately trained monolingual models in most languages. Furthermore, it offers practical advantages for deployment: a single multilingual model eliminates the need to manage dozens of individual models in memory and significantly simplifies maintenance, avoiding the overhead associated with retraining 73 separate models.

We make all our work open-source by publicly releasing our code at GitHub: https://github.com/tomsouri/multilingual-inflection.

2 Related Work

Morphological inflection has seen considerable research interest in the recent years, mainly thanks to the SIGMORPHON shared tasks held from 2016 to 2023 [2,3,4,7,12,18,24]. This period was marked by major advancements in model architectures, particularly recurrent neural networks and Transformers.

The field has also undergone significant methodological changes, such as the shift from naïve splitting by form to a *lemma-disjoint* split [8], and has identified weaknesses in previously established practices. Notable examples include the deficiencies of the original uniform train-dev-test split, as compared to the *frequency-weighted* split proposed by Kodner et al. (2023) [13] to achieve a more realistic train-test distribution. However, these challenges have not yet been fully addressed.

Furthermore, many previous studies were experimental and/or trained on small datasets, making the transition from academic research to production deployment challenging.

3 Data

We use the SIGMORPHON 2022 [12] and the SIGMORPHON 2023 [7] shared task benchmarks for evaluating the strength of our models.[2] Furthermore, with the goal of future deployment and multilingual applicability, we select 73 languages of the Universal Dependencies [16] corpora[3] and lexicalize them by extracting unique lemma-tag-form triples along with their occurrence counts for training and evaluation.[4]

4 Methodology

Data Splitting Strategies for data splitting in the inflection field differ in the degree of overlap control between train-test splits, from an uncontrolled overlap (where lemma-tag-form may overlap randomly) to varying degrees of control over lemma and/or morphological feature overlap. The former has been the traditional approach for decades but has recently come under critique for hindering evaluation of models' generalization abilities.

At the very least, a *lemma-disjoint* option [8] is recently advisable, and SIGMORPHON 2022 [12] probed this setting in its *feature overlap* setting: A test pair's feature set is attested in training, but its lemma is novel. The latest SIGMORPHON shared task installation completely moved to the *lemma-disjoint* setting in 2023 [7]. We use the official train-dev-test split for both the SIGMORPHON benchmarks.

Another axis for consideration is the manipulation with the lemma-tag-form corpus frequencies during sampling into the train-test splits. Traditionally, the corpus frequencies are neglected (or not even present in the mostly used data source, UniMorph [11]) and the lemma-tag-form triples are sampled uniformly. Kodner et al. (2023) [13] argued that uniform sampling leads to an unrealistic train-test split, with train set unnaturally biased towards low-frequency types. They introduced a *frequency-weighted sampling* method to produce a more realistic train-test distribution. However, the suggested split is not lemma-disjoint. Also, no further work has experimented with that split.

Due to the lack of standardized practice for sampling with corpus frequencies, we design a new train-dev-test split technique for the UD datasets that satisfies

[2] Hebrew was excluded from SIGMORPHON 2023 due to filename inconsistencies that prevented reliable comparison.

[3] The selection contains 32 Indo-European languages written in the Latin script and 41 languages without restriction on language family or script. For languages with multiple corpora in UD, we chose one corpus per language based on historical preferences, practical considerations, and the defaults in the widely adopted UDPipe tool [21]. We prioritized large, canonical corpora, excluded those with narrow or specialized domains, and favored manually over semi-automatically annotated corpora. Corpora lacking lemma annotations were excluded.

[4] Unlike UniMorph, the Universal Dependencies data can be more prone to noise. We rely on a neural network to mitigate its impact.

Table 1. SIGMORPHON 2022 (Feature Overlap): Comparison across all 16 development languages that provided both large training dataset (7k examples) and test data under the feature overlap evaluation condition. Reported exact-match accuracy (%). ‡ marks datasets with less than 100 test items. The systems are: CLUZH [26], Flexica [19], OSU [5], TüM [15], UBC [27], and baselines [12], Sou is Sourada et al. (2024) [20].

	Submitted systems					Baselines		Sou		OURS	
lang	CLUZH	Flexica	OSU	TüM	UBC	Neural	NonNeur	LSTM	TRM	MONO	MULTI
ang	76.6	64.4	73.7	71.9	74.1	73.4	68.7	76.3	75.5	73.4	**77.3**
ara	81.7	65.5	78.7	78.5	65.5	81.9	50.8	79.2	**82.6**	81.0	79.6
asm‡	83.3	75.0	75.0	**91.7**	83.3	83.3	83.3	83.3	83.3	83.3	83.3
got	92.9	41.4	**94.1**	91.7	91.7	93.5	87.6	92.3	92.3	91.1	91.7
hun	93.5	62.9	93.1	92.8	91.5	**94.4**	73.1	92.8	**94.4**	93.2	93.2
kat	96.7	95.7	96.7	96.7	96.7	97.3	96.7	97.3	**97.8**	96.7	97.3
khk‡	94.1	47.1	94.1	94.1	88.2	94.1	88.2	**100.0**	94.1	94.1	94.1
kor‡	**71.1**	55.4	50.6	56.6	60.2	62.7	59.0	49.4	62.7	63.9	62.7
krl	87.5	69.8	85.9	57.8	85.4	57.8	20.8	**89.1**	85.9	85.9	86.5
lud	87.3	92.0	92.9	93.4	88.2	**94.3**	93.4	89.2	92.0	91.5	93.4
non‡	85.2	77.0	85.2	80.3	**90.2**	88.5	80.3	83.6	88.5	**90.2**	86.9
pol	**96.1**	85.9	94.9	74.0	95.7	74.4	86.3	**96.1**	95.6	95.7	95.5
poma	76.1	54.5	70.1	69.4	73.3	74.1	47.8	75.2	**76.3**	73.9	72.8
slk	93.5	90.0	92.2	70.4	**95.7**	71.1	92.4	95.2	**95.7**	94.4	95.2
tur	93.7	57.9	**95.2**	80.2	92.9	79.4	66.7	**95.2**	92.9	91.3	**95.2**
vep	**71.5**	58.8	70.0	57.5	68.8	59.2	60.4	70.7	68.8	67.8	70.2
avg	**86.3**	68.3	83.9	78.6	83.8	80.0	72.2	85.3	86.2	85.5	85.9

both recent methodological standards (being *lemma-disjoint* [7,8]) and practical requirements for real-world deployment by ensuring a realistic distribution of items with varying corpus frequencies through *frequency-weighted sampling* [13].

Model We employ a Transformer architecture of the encoder-decoder, sequence-to-sequence type, characterized by a small capacity according to state-of-the-art standards in the inflection field. This model is trained from scratch on inflection data, using characters as tokens, with the input consisting of lemma-tag pair and the output being the corresponding inflected form (see Fig. 1).

Multilingual Model We use a special language ID token as part of the input sequence [10,17] by prepending it to the sequence of morphological tags. This should help the model disambiguate identical lemma strings corresponding to different languages.

Unlike previous work on multilingual inflection which usually used the SIGMORPHON shared task data where the train set for all languages was of the same size, we need to deal with the problem of mixing corpora of substantially different sizes. If we naïvely pooled the training datasets, the model may be

Table 2. SIGMORPHON 2023 (Lemma Disjoint): exact-match accuracy (%), with systems AZ1-4 [14], TÜB [6], IL1-4 [1], baselines [7]. As multilingual training wasn't clearly defined in 2023, we highlight (*) the best systems excluding MULTI (†).

	Submitted systems and official baselines											OURS	
lang	AZ3	AZ1	non-neur	AZ2	AZ4	TÜB	neur	IL1	IL2	IL3	IL4	MONO	MULTI†
afb	34.5	30.8	30.8	52.7	52.7	75.8	80.1	80.7	82.2	84.1	**84.6**	80.3	79.6
amh	59.9	65.4	65.4	74.0	74.0	83.8	82.2	88.9	**90.6***	88.9	88.6	88.6	**90.8**
arz	75.7	77.2	77.9	80.8	80.8	87.6	**89.6***	89.2	88.7	89.1	88.7	88.5	**89.9**
bel	46.2	68.1	68.1	64.5	64.5	56.3	74.5	73.5	74.7	72.9	72.9	**75.4***	**75.9**
dan	64.8	**89.5**	**89.5**	87.4	87.4	85.7	88.8	88.8	**89.5**	86.5	87.5	86.0	89.4
deu	59.9	79.8	79.8	77.9	77.9	74.5	**83.7***	79.7	79.7	80.2	79.7	80.5	**84.8**
eng	67.0	**96.6***	**96.6***	96.2	96.2	96.0	95.1	95.6	95.9	94.6	95.0	94.5	**97.5**
fin	48.2	80.8	80.8	80.6	80.6	67.6	85.4	79.2	80.6	85.7	**86.1***	84.1	**88.0**
fra	76.7	**77.7***	**77.7***	76.3	76.3	67.9	73.3	69.3	74.7	71.7	72.9	77.2	**80.0**
grc	40.4	52.6	52.6	54.8	54.8	36.7	54.0	48.9	53.7	**56.0***	**56.0***	49.7	**61.7**
hun	45.9	74.7	74.7	74.7	74.7	75.9	80.5	76.3	79.8	84.3	**85.0***	84.6	**88.9**
hye	88.9	86.3	86.3	86.2	88.9	85.9	91.0	88.4	91.5	94.4	94.3	**95.1***	**98.4**
ita	78.0	75.0	75.0	63.6	78.0	84.7	94.1	95.8	**97.2***	92.1	92.2	95.8	**97.8**
jap	67.0	64.1	64.1	64.1	67.0	**95.3**	26.3	92.8	94.2	94.9	94.9	36.2	48.9
kat	71.7	82.0	82.0	82.1	82.1	70.5	84.5	84.1	84.7	81.3	82.9	**87.1**	85.4
klr	27.8	54.5	54.5	53.1	53.1	96.4	**99.5**	99.4	99.4	99.4	99.4	99.1	99.2
mkd	64.9	91.6	91.6	90.8	90.8	86.7	**93.8**	91.9	92.4	92.1	92.4	92.9	93.2
nav	23.7	35.8	35.8	41.8	41.8	53.6	52.1	54.0	55.1	55.1	**55.6***	52.6	**60.0**
rus	66.8	86.0	86.0	85.6	85.6	82.1	**90.5***	87.4	87.3	84.2	85.5	87.8	**91.9**
san	47.0	62.2	62.2	62.1	62.1	54.5	66.3	63.3	**69.1***	67.7	65.9	68.6	**73.4**
sme	30.1	56.0	56.0	49.7	49.7	58.5	**74.8***	69.9	71.8	67.4	67.3	73.0	**83.1**
spa	86.3	87.8	87.8	87.4	87.4	88.7	93.6	90.9	91.4	93.8	93.1	**94.9***	**95.2**
sqi	73.8	19.3	83.4	78.1	78.1	71.5	85.9	87.6	88.9	**92.0***	91.6	89.3	**93.0**
swa	56.2	60.5	60.5	65.0	65.0	94.7	93.7	93.1	93.1	96.6	96.6	**98.4***	**99.2**
tur	28.1	64.6	64.6	64.6	64.6	81.8	**95.0***	90.9	90.8	90.3	92.0	93.8	**95.8**
avg	57.2	68.8	71.4	71.8	72.6	76.5	81.1	82.4	83.9	83.8	**84.0***	82.2	**85.6**

overly influenced by the high-resource languages. To address this issue, we adapt a corpus upsampling method by Wang et al. (2020) [25]. For better control over sampling into training batches, we also adapt a temperature value $\tau \in [0, 1]$ (our $\tau = 0.5$) to smooth the distribution of training data from datasets of different sizes, see van der Goot et al. (2021) [9].

5 Results

SIGMORPHON Benchmarks Our monolingual and multilingual systems achieve competitive performance on a global inflection benchmark, specifically the SIGMORPHON Inflection Shared Tasks from 2022 [12] and 2023 [7], as

Table 3. UD test accuracy — part 1. The COPY baseline copies the input lemma to output. †The COPY baseline fails on lemma-to-form transfer in Ancient_Hebrew-PTNK due to special diacritics (e.g., cantillation marks) present in surface forms but absent in standardized lemmas.

	size (# lemma-tag-form)			accuracy (%)		
lang-corpus	train	dev	test	COPY	MONO	MULTI
Afrikaans-AfriBooms	1.9K	2.4K	2.3K	69.18	83.72	**89.41**
Ancient_Greek-PROIEL	12.8K	11.7K	11.2K	10.67	53.63	**56.14**
Ancient_Hebrew-PTNK	4.8K	2.8K	2.7K	0.00†	1.40	**1.66**
Arabic-PADT	14.7K	12.2K	11.9K	24.08	86.94	**87.17**
Armenian-ArmTDP	6.5K	3.7K	3.6K	35.69	88.06	**91.04**
Basque-BDT	13.0K	7.8K	7.7K	39.74	87.13	**88.42**
Belarusian-HSE	21.4K	19.4K	19.3K	40.88	88.02	**89.15**
Breton-KEB	1.2K	696	718	57.66	64.62	**75.91**
Bulgarian-BTB	10.0K	9.1K	9.0K	37.90	90.94	**92.00**
Catalan-AnCora	7.4K	15.6K	15.4K	64.73	93.62	**95.39**
Chinese-GSDSimp	7.2K	7.7K	7.7K	82.10	80.88	**86.45**
Classical_Armenian-CAVaL	2.6K	2.9K	2.6K	28.52	55.64	**72.20**
Classical_Chinese-Kyoto	3.9K	6.4K	6.4K	13.99	1.62	**46.27**
Coptic-Scriptorium	685	1.5K	1.3K	**83.18**	54.86	79.45
Croatian-SET	17.7K	12.6K	13.0K	36.10	92.10	**93.45**
Czech-PDT	47.9K	62.0K	62.3K	37.53	**97.44**	97.37
Danish-DDT	5.8K	6.6K	6.7K	55.61	89.92	**92.95**
Dutch-Alpino	7.0K	11.7K	11.8K	55.48	77.66	**79.16**
English-EWT	6.6K	9.5K	9.7K	76.67	93.63	**94.96**
Erzya-JR	4.3K	1.7K	1.6K	29.63	82.27	**86.10**
Estonian-EDT	31.7K	27.8K	28.2K	23.24	**88.83**	87.93
Finnish-TDT	26.3K	15.2K	15.0K	23.88	**90.58**	88.84
French-GSD	10.4K	20.0K	19.6K	72.55	95.36	**96.79**
Galician-TreeGal	2.4K	1.7K	1.7K	59.67	94.09	**97.36**
Georgian-GLC	988	229	224	44.64	68.30	**75.45**
German-GSD	25.5K	23.7K	23.8K	75.22	89.12	**89.37**
Gothic-PROIEL	4.1K	3.1K	2.9K	17.50	71.13	**79.30**
Greek-GDT	5.2K	4.0K	3.9K	34.45	82.69	**86.18**
Hebrew-HTB	6.7K	6.9K	7.0K	51.75	88.78	**90.47**
Hindi-HDTB	10.4K	12.3K	12.1K	79.64	92.31	**92.71**
Hungarian-Szeged	7.4K	3.4K	3.4K	51.34	92.55	**93.23**
Icelandic-Modern	4.6K	4.5K	4.5K	35.71	70.62	**77.57**
Indonesian-GSD	7.3K	7.4K	7.6K	88.57	89.89	**91.59**
Irish-IDT	6.8K	6.8K	6.8K	51.93	83.75	**87.74**
Italian-ISDT	7.8K	11.9K	11.7K	62.95	95.26	**96.02**
Japanese-GSDLUW	8.4K	11.6K	11.6K	74.77	75.31	**79.74**
Korean-Kaist	45.9K	28.3K	28.5K	10.06	95.90	**96.19**

Table 4. UD test accuracy — part 2. The COPY baseline copies the input lemma to output.

	size (# lemma-tag-form)			accuracy (%)		
lang-corpus	train	dev	test	COPY	MONO	MULTI
Kyrgyz-KTMU	2.9K	675	673	42.20	64.34	**66.57**
Latin-ITTB	5.8K	9.7K	9.7K	14.58	76.33	**88.94**
Latvian-LVTB	22.6K	18.3K	18.2K	27.81	**96.48**	96.15
Lithuanian-ALKSNIS	9.3K	5.0K	5.0K	27.79	92.74	**93.38**
Low_Saxon-LSDC	3.5K	1.9K	1.8K	54.42	54.14	**62.49**
Maghrebi_Arabic_French-Arabizi	5.9K	1.8K	1.8K	**17.46**	9.54	15.35
Manx-Cadhan	762	1.1K	1.2K	65.15	66.84	**82.03**
Marathi-UFAL	759	302	302	47.02	56.62	**74.17**
Naija-NSC	916	2.4K	2.3K	86.41	70.64	**96.04**
North_Sami-Giella	4.3K	2.1K	2.1K	33.00	68.50	**75.96**
Norwegian-Bokmaal	7.4K	14.8K	14.2K	53.62	93.08	**95.26**
Old_Church_Slavonic-PROIEL	24.7K	12.9K	13.2K	6.35	24.34	**27.56**
Old_East_Slavic-TOROT	28.6K	16.7K	16.2K	7.49	31.08	**33.82**
Old_French-PROFITEROLE	19.4K	2.8K	3.0K	**19.82**	0.13	18.05
Ottoman_Turkish-BOUN	3.0K	781	789	47.66	78.20	**84.03**
Persian-PerDT	12.4K	14.7K	14.2K	67.70	**72.96**	70.73
Polish-PDB	28.3K	22.6K	23.1K	28.42	**93.78**	93.64
Pomak-Philotis	3.0K	2.2K	2.2K	20.98	43.21	**54.99**
Portuguese-Bosque	9.7K	11.6K	11.7K	67.66	96.21	**97.63**
Romanian-RRT	11.1K	11.7K	11.6K	42.00	92.65	**93.60**
Russian-SynTagRus	44.4K	64.1K	63.2K	26.35	93.28	**93.57**
Sanskrit-Vedic	19.8K	11.3K	11.7K	9.43	**75.55**	75.22
Scottish_Gaelic-ARCOSG	2.7K	4.1K	4.0K	59.71	60.65	**68.46**
Slovak-SNK	14.1K	8.1K	7.8K	31.83	94.13	**94.45**
Slovenian-SSJ	22.3K	18.3K	18.2K	33.31	95.34	**96.06**
Spanish-AnCora	9.5K	17.7K	17.7K	61.38	95.16	**96.59**
Swedish-Talbanken	4.9K	5.7K	5.6K	48.63	89.63	**92.51**
Tamil-TTB	2.3K	727	734	50.54	68.66	**69.62**
Turkish-BOUN	23.4K	9.1K	9.2K	42.80	81.77	**83.00**
Ukrainian-IU	16.4K	9.7K	9.6K	32.11	92.56	**93.81**
Urdu-UDTB	9.0K	6.5K	6.7K	82.32	87.65	**89.05**
Uyghur-UDT	8.8K	2.1K	2.0K	44.03	90.36	**93.67**
Vietnamese-VTB	2.2K	2.7K	2.7K	92.96	97.74	**99.44**
Welsh-CCG	3.0K	2.8K	2.8K	60.11	84.91	**88.80**
Western_Armenian-ArmTDP	11.1K	7.5K	7.5K	36.33	91.63	**93.39**
Wolof-WTB	2.3K	2.3K	2.4K	78.34	87.35	**89.76**
total size / avg accuracy	809.6K	723.3K	720.2K	45.27	76.67	**81.36**

shown in Table 1 and Table 2, respectively: MONO ranks 4th in the 2022 data and 6th in the 2023 data, while MULTI attains 3rd place in 2022 and 1st place in 2023, based on the average across languages. For clarity, we note that the models were trained from scratch exclusively on the SIGMORPHON benchmark data.

Universal Dependencies The two-part Table 3 and 4 presents the results on 73 corpora from UD. The overall best-performing system is MULTI, which was trained jointly on all 73 languages. Among these languages, it is outperformed in 3 cases by the simple COPY baseline, and in 7 cases by the monolingual models (MONO), each trained separately for a single language. On macro average across all languages, the MULTI model improves over the MONO models by 4.69%.

6 Conclusions

We focused on the task of automatic morphological inflection in a multilingual setting, with the aim of enabling future deployment in an open-source tool or web service. Trained jointly on 73 languages, our multilingual model is remarkably lightweight (5.69M trainable parameters), handles unseen words well, works across parts of speech, and importantly, outperforms monolingual models in most languages. We designed and implemented a novel frequency-weighted, lemma-disjoint train-dev-test split method, which combines recent evaluation practices (lemma-disjoint splits) with a realistic frequency distribution. This work establishes a foundation for deploying a practical and multilingual inflection system.

Acknowledgments. This research was supported by the Johannes Amos Comenius Programme (P JAC) project No. CZ.02.01.01/00/22_008/0004605, Natural and anthropogenic georisks. Computational resources for this work were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work described herein uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (projects LM2018101 and LM2023062, supported by the Ministry of Education, Youth and Sports of the Czech Republic). The authors would like to thank Aleš Manuel Papáček and Milan Straka for their thoughtful comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

7 Appendix: Hyperparameters

For our MONO and MULTI model, hyperparameters tuned on UD development data are: layers (encoder/decoder) = 3/3 (MONO) and 4/4 (MULTI), layer dimension = 256, feed-forward dimension = 64, attention heads = 4, dropout = 0.15, attention dropout = 0.1, activation dropout = 0.35, and layer drop = 0.2. We

use gradient clipping (max norm 1.0), L2 regularization (0.01), and train with Adam using cosine learning rate decay, initial LR = 0.001, batch size = 512 (MONO) and 1024 (MULTI). Checkpoints are selected by dev set performance: on the target language for MONO, or macro-averaged across languages for MULTI, over up to 960 epochs. The MONO model has 3.12M-3.69M parameters,[5] while the MULTI model has 5.69M parameters.

References

1. Canby, M., Hockenmaier, J.: A framework for bidirectional decoding: Case study in morphological inflection. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 4485–4507. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.297, https://aclanthology.org/2023.findings-emnlp.297/
2. Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A.D., Kann, K., Mielke, S.J., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., Hulden, M.: The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In: Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. pp. 1–27. Association for Computational Linguistics, Brussels (Oct 2018). https://doi.org/10.18653/v1/K18-3001, https://aclanthology.org/K18-3001
3. Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M.: CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In: Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection. pp. 1–30. Association for Computational Linguistics, Vancouver (Aug 2017). https://doi.org/10.18653/v1/K17-2001, https://aclanthology.org/K17-2001
4. Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., Hulden, M.: The SIGMORPHON 2016 shared Task—Morphological reinflection. In: Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 10–22. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2002, https://aclanthology.org/W16-2002
5. Elsner, M., Court, S.: OSU at SigMorphon 2022: Analogical inflection with rule features. In: Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 220–225. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.22, https://aclanthology.org/2022.sigmorphon-1.22
6. Girrbach, L.: Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In: Nicolai, G., Chodroff, E., Mailhot, F., Çöltekin, Ç. (eds.) Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 151–165. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.sigmorphon-1.17, https://aclanthology.org/2023.sigmorphon-1.17/

[5] varying by vocabulary size

7. Goldman, O., Batsuren, K., Khalifa, S., Arora, A., Nicolai, G., Tsarfaty, R., Vylomova, E.: 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. pp. 117–125. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.sigmorphon-1.13, https://aclanthology.org/2023.sigmorphon-1.13
8. Goldman, O., Guriel, D., Tsarfaty, R.: (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 864–870. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-short.96, https://aclanthology.org/2022.acl-short.96/
9. van der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., Plank, B.: Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In: Gkatzia, D., Seddah, D. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pp. 176–197. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-demos.22, https://aclanthology.org/2021.eacl-demos.22/
10. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics **5**, 339–351 (2017). https://doi.org/10.1162/tacl_a_00065, https://aclanthology.org/Q17-1024/
11. Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S.J., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., Hulden, M.: UniMorph 2.0: Universal Morphology. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1293/
12. Kodner, J., Khalifa, S., Batsuren, K., Dolatian, H., Cotterell, R., Akkus, F., Anastasopoulos, A., Andrushko, T., Arora, A., Atanalov, N., Bella, G., Budianskaya, E., Ghanggo Ate, Y., Goldman, O., Guriel, D., Guriel, S., Guriel-Agiashvili, S., Kieraś, W., Krizhanovsky, A., Krizhanovsky, N., Marchenko, I., Markowska, M., Mashkovtseva, P., Nepomniashchaya, M., Rodionova, D., Scheifer, K., Sorova, A., Yemelina, A., Young, J., Vylomova, E.: 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. pp. 176–203. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.19, https://aclanthology.org/2022.sigmorphon-1.19
13. Kodner, J., Payne, S., Khalifa, S., Liu, Z.: Morphological inflection: A reality check. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6082–6101. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.335, https://aclanthology.org/2023.acl-long.335/

14. Kwak, A., Hammond, M., Wing, C.: Morphological reinflection with weighted finite-state transducers. In: Nicolai, G., Chodroff, E., Mailhot, F., Çöltekin, Ç. (eds.) Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 132–137. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.sigmorphon-1.15, https://aclanthology.org/2023.sigmorphon-1.15/
15. Merzhevich, T., Gbadegoye, N., Girrbach, L., Li, J., Shim, R.S.E.: SIGMORPHON 2022 task 0 submission description: Modelling morphological inflection with data-driven and rule-based approaches. In: Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 204–211. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.20, https://aclanthology.org/2022.sigmorphon-1.20
16. Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, C.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.497/
17. Peters, B., Dehdari, J., van Genabith, J.: Massively multilingual neural grapheme-to-phoneme conversion. In: Bender, E., Daumé III, H., Ettinger, A., Rao, S. (eds.) Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. pp. 19–26. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-5403, https://aclanthology.org/W17-5403/
18. Pimentel, T., Ryskina, M., Mielke, S.J., Wu, S., Chodroff, E., Leonard, B., Nicolai, G., Ghanggo Ate, Y., Khalifa, S., Habash, N., El-Khaissi, C., Goldman, O., Gasser, M., Lane, W., Coler, M., Oncevay, A., Montoya Samame, J.R., Silva Villegas, G.C., Ek, A., Bernardy, J.P., Shcherbakov, A., Bayyr-ool, A., Sheifer, K., Ganieva, S., Plugaryov, M., Klyachko, E., Salehi, A., Krizhanovsky, A., Krizhanovsky, N., Vania, C., Ivanova, S., Salchak, A., Straughn, C., Liu, Z., Washington, J.N., Ataman, D., Kieraś, W., Woliński, M., Suhardijanto, T., Stoehr, N., Nuriah, Z., Ratan, S., Tyers, F.M., Ponti, E.M., Aiton, G., Hatcher, R.J., Prud'hommeaux, E., Kumar, R., Hulden, M., Barta, B., Lakatos, D., Szolnok, G., Ács, J., Raj, M., Yarowsky, D., Cotterell, R., Ambridge, B., Vylomova, E.: SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In: Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 229–259. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.sigmorphon-1.25, https://aclanthology.org/2021.sigmorphon-1.25
19. Sherbakov, A., Vylomova, E.: Morphology is not just a naive Bayes – UniMelb submission to SIGMORPHON 2022 ST on morphological inflection. In: Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 240–246. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.25, https://aclanthology.org/2022.sigmorphon-1.25
20. Sourada, T., Straková, J., Rosa, R.: OOVs in the spotlight: How to inflect them? In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings

of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 12455–12466. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.1091/
21. Straka, M.: UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 197–207. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-2020, https://www.aclweb.org/anthology/K18-2020
22. Straka, M., Straková, J.: Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/K/K17/K17-3009.pdf
23. Straková, J., Straka, M., Hajič, J.: Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 13–18. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/P14-5003, https://aclanthology.org/P14-5003
24. Vylomova, E., White, J., Salesky, E., Mielke, S.J., Wu, S., Ponti, E.M., Maudslay, R.H., Zmigrod, R., Valvoda, J., Toldova, S., Tyers, F., Klyachko, E., Yegorov, I., Krizhanovsky, N., Czarnowska, P., Nikkarinen, I., Krizhanovsky, A., Pimentel, T., Torroba Hennigen, L., Kirov, C., Nicolai, G., Williams, A., Anastasopoulos, A., Cruz, H., Chodroff, E., Cotterell, R., Silfverberg, M., Hulden, M.: SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In: Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 1–39. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.sigmorphon-1.1, https://aclanthology.org/2020.sigmorphon-1.1
25. Wang, X., Tsvetkov, Y., Neubig, G.: Balancing training for multilingual neural machine translation. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8526–8537. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.754, https://aclanthology.org/2020.acl-main.754/
26. Wehrli, S., Clematide, S., Makarov, P.: CLUZH at SIGMORPHON 2022 shared tasks on morpheme segmentation and inflection generation. In: Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 212–219. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.21, https://aclanthology.org/2022.sigmorphon-1.21
27. Yang, C., Yang, R.R., Nicolai, G., Silfverberg, M.: Generalizing morphological inflection systems to unseen lemmas. In: Nicolai, G., Chodroff, E. (eds.) Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. pp. 226–235. Association for Computational Linguistics, Seattle, Washington (Jul 2022). https://doi.org/10.18653/v1/2022.sigmorphon-1.23, https://aclanthology.org/2022.sigmorphon-1.23/