OOVs in the Spotlight: How to Inflect them?

Tomáš Sourada, Jana Straková, Rudolf Rosa
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University, Czech Republic
{sourada,strakova,rosa}@ufal.mff.cuni.cz

Abstract

We focus on morphological inflection in out-of-vocabulary (OOV) conditions, an under-researched subtask in which state-of-the-art systems usually are less effective. We developed three systems: a retrograde model and two sequence-to-sequence (seq2seq) models based on LSTM and Transformer. For testing in OOV conditions, we automatically extracted a large dataset of nouns in the morphologically rich Czech language, with lemma-disjoint data splits, and we further manually annotated a real-world OOV dataset of neologisms.
In the standard OOV conditions, Transformer achieves the best results, with increasing performance in ensemble with LSTM, the retrograde model and SIGMORPHON baselines. On the real-world OOV dataset of neologisms, the retrograde model outperforms all neural models. Finally, our seq2seq models achieve state-of-the-art results in 9 out of 16 languages from SIGMORPHON 2022 shared task data in the OOV evaluation (feature overlap) in the large data condition.
We release the Czech OOV Inflection Dataset for rigorous evaluation in OOV conditions. Further, we release the inflection system with the seq2seq models as a ready-to-use Python library.

Keywords: morphological inflection, out-of-vocabulary words, OOV, retrograde, seq2seq, LSTM, Transformer

1. Introduction

Inflection is a process of word formation in which a base word form (lemma) is modified to express grammatical categories. Many natural language generation systems that have natural text on the output, such as dialogue systems, need to be able to correctly inflect words. However, it has been shown that the state-of-the-art systems achieve rather poor results when tested on previously unseen inputs (OOV words) (Liu and Hulden, 2021; Goldman et al., 2022). Despite an extensive exploration of the inflection task in recent years (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023) and outstanding results of the state-of-the-art systems, especially when the training data was plentiful (Wu et al., 2020; Pimentel et al., 2021), the poor performance on OOV words has not been fully realized until recently, because the results had been inflated by the presence of training lemmas in the test dataset (Liu and Hulden, 2021; Goldman et al., 2022).

To provide a consistent benchmark for inflection in OOV context, we release the Czech OOV Inflection Dataset[1] for rigorous evaluation, with a lemma-disjoint train-dev-test split of the pre-existing large morphological dictionary MorfFlex (Hajič et al., 2020). This benchmark is accompanied by a manually annotated small dataset of real-world OOV words (neologisms) in Czech. Unlike English, which has relatively simple morphology, e.g. just adding '-s' when forming the plural, the task of automatic inflection in morphologically rich languages such as Czech is quite difficult. An example of an inflection of a Czech neologism is shown in table 1.

LINGEBRA		
Case / Number	Singular	Plural
1. Nominative	lingebra	lingebry
2. Genitive	lingebry	lingeber
3. Dative	lingebře	lingebrám
4. Accusative	lingebru	lingebry
5. Vocative	lingebro	lingebry
6. Locative	lingebře	lingebrách
7. Instrumental	lingebrou	lingebrami

Table 1: An example of an inflection of a Czech neologism not covered by the Czech inflection dictionary MorfFlex (Hajič et al., 2020). "*Lingebra*" is a playful compound of words "*lineární*" (linear) and "*algebra*" (algebra) (both in vocabulary; the neologism is inflected in the same way as "*algebra*").

To our knowledge, this is the first dataset designed specifically for evaluation of inflection in the OOV conditions in Czech. In addition, Czech was not included in the 2022's iteration of the SIGMORPHON shared task (Kodner et al., 2022), which evaluated the performance of submitted systems on the implicit OOV subset of the shared task.

We develop three different systems, all data-driven, and compare them to several well-established systems, both Transformer-based and rule-based ones, in OOV conditions. We also build a state-of-the-art ready-to-use guesser for morphological inflection of Czech OOV nouns.

[1] http://hdl.handle.net/11234/1-5471

LREC-COLING 2024, pages 12455–12466
20-25 May, 2024.

Our first approach is a dictionary-based retrograde model: when given a lemma, search the database for a word that is most similar (has the longest common suffix), and inflect the lemma according to it.

The second and the third approach follow the standard neural approach using sequence-to-sequence architecture based on either LSTM (Hochreiter and Schmidhuber, 1997) or Transformer (Vaswani et al., 2017).

We adapt the systems to the OOV setting and tune them extensively. Then we evaluate them and compare them to one existing ready-to-use system, and to SIGMORPHON shared task baselines (Pimentel et al., 2021) on the Czech OOV Inflection Dataset. Our systems either outperform the other evaluated systems or perform comparably.

In addition, we train and evaluate our neural setups on SIGMORPHON 2022 shared task data (16 languages, Czech not included, all parts-of-speech) (Kodner et al., 2022) in the large training data condition, and in 9 languages we achieve state-of-the-art results in the OOV evaluation (feature overlap).

Finally, we address the lack of a reliable morphoguesser for generation in Czech by releasing a ready-to-use Python library with our seq2seq models.[2] We also release the Czech OOV Inflection Dataset.[3]

A more detailed description of our work, dataset creation, other exploratory experiments, as well as a profound summary of the related work, is provided in Sourada (2023).

We describe the new Czech OOV Inflection Dataset in section 3 and our methodology in section 4, with results and comparison of the three approaches in section 5 and an error analysis in section 6. Finally, we conclude in section 7.

2. Related Work

Earlier inflection systems were based on rules and dictionaries. For Czech language, the linguistic module of the ASIMUT system (Králíková and Panevová, 1990) determined inflection paradigms according to lemma endings based on the retrograde dictionary of Slavíčková (1975); the sklonuj.cz system[4] directly maps lemma endings to form endings based on hand-crafted rules. Such simple ways of paradigm assignment have limited precision, often selecting an incorrect paradigm. MorphoDiTa (Straková et al., 2014), on the other hand, only outputs inflections from the MorfFlex morphological dictionary (Hajič et al., 2020), leading to high precision but low recall, as no output is generated for OOV lemmas. Later systems tried to extract and apply string transformation rules based on learned models (Dušek and Jurčíček, 2013; Durrett and DeNero, 2013).

Since 2016, research has been considerably fueled by the annual SIGMORPHON shared task on morphological inflection (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023), including the release of datasets for up to 103 languages. The increasingly prevalent approach has been the employment of the sequence-to-sequence (seq2seq) neural network architectures (Sutskever et al., 2014), often inspired by machine translation approaches, with hyperparameters adapted and tuned for the morphological inflection task. The systems have been based on GRU or LSTM recurrent neural networks with attention (Faruqui et al., 2016; Kann and Schütze, 2016), and, since 2020 (Wu et al., 2020), also the Transformer (Vaswani et al., 2017). The models typically operate on sequences of individual characters and morphological labels, taking the lemma and the morphological information as input and producing the inflected form as the output.

Currently, the Transformer-based systems seem to have almost completely mastered the task, achieving outstanding results especially when the training data is plentiful; with low training data and for unseen inputs (OOV words), the accuracies often plummet (Liu and Hulden, 2021; Goldman et al., 2022). This can be partially alleviated by data augmentation techniques, such as data hallucination (Yang et al., 2022), and by employing multilingual approaches. Since 2021, the SIGMORPHON shared tasks include evaluation on unseen lemmas, but not for Czech language.

3. Czech OOV Inflection Dataset

To allow for consistent evaluation of inflection in OOV conditions and for development of inflection systems, we create lemma-disjoint splits of an existing morphological dictionary. In addition, we annotate a small dataset of true OOV words (neologisms) to test the models in real-world conditions. The overview of the data splits is in table 2. An example from the test-neologisms dataset can be found in table 3.

3.1. MorfFlex Morphological Dictionary

To build the train-dev-test split, we use the existing Czech morphological dictionary MorfFlex (Hajič et al., 2020), annotated with the morphological tagset of Hajič (2004). With more than 125M lemma-tag-form entries, it is relatively large compared to standard datasets in other languages.

[2] https://github.com/tomsouri/cz-inflect

[3] http://hdl.handle.net/11234/1-5471

[4] https://sklonuj.cz

Set	lemmas	forms	Source
train	360k	5.04M	MorfFlex
dev	44k	616k	MorfFlex
test-MorfFlex	44k	616k	MorfFlex
test-neologisms	101	1.4k	Čeština 2.0

Table 2: The four data splits, with lemma (paradigm table) counts and form counts.

lemma	tag	form
elektrořidič	S1	elektrořidič
elektrořidič	S2	elektrořidiče
elektrořidič	S3	elektrořidiči/elektrořidičovi
elektrořidič	S4	elektrořidiče
elektrořidič	S5	elektrořidiči
elektrořidič	S6	elektrořidiči/elektrořidičovi
elektrořidič	S7	elektrořidičem
elektrořidič	P1	elektrořidiči/elektrořidičové
elektrořidič	P2	elektrořidičů
elektrořidič	P3	elektrořidičům
elektrořidič	P4	elektrořidiče
elektrořidič	P5	elektrořidiči/elektrořidičové
elektrořidič	P6	elektrořidičích
elektrořidič	P7	elektrořidiči

Table 3: Example from the test-neologisms dataset. All 14 paradigm cells of lemma "*elektrořidič*" (driver of an electric car). The morphological tag is simplified to singular/plural (S/P) and 7 Czech cases (nominative, genitive, dative, accusative, vocative, locative and instrumental, numbered from 1 to 7). Some paradigm cells has two possible correct forms (those are separated by '/').

We start by filtering the data by selecting the noun paradigm table entries (460k out of 1M entries). Of these, we removed all nonbasic-variant forms (such as nonstandard variants), all negation forms and malformed or deficient paradigm tables. The removed portion of the noun entries form 2% of the noun entries and in the end, we acquired 449k noun paradigm tables. We then completed the incomplete paradigm tables (such as singular forms in tables of pluralia tantum or forms corresponding to non-flexible lemmas).

We experimentally verified that omitting the negated variants of lemmas in training data does not have a negative impact on the performance of the models: we compared the performance of the inflection model (trained on data with no negations) on the standard development set and on a negated variant of it, and observed comparable results on both datasets.

We finish by randomly splitting the data into three lemma-disjoint parts: train, dev and test set with lemma counts in the ratio 8:1:1 (we denote the test set by test-MorfFlex further in the text).

3.2. Neologisms

For evaluation in real-world OOV conditions, we build a new test set of true out-of-vocabulary words: neologisms. We considered several other options of what to use as the real OOV words, such as misspelled words, words with removed diacritics, proper nouns, but finally chose neologisms because they cannot be included in a dictionary by their very nature.

We draw new words from a dictionary of Czech neologisms Čeština 2.0 (Kavka and Škrabal et al., 2018).[5] Each entry contains the word or word phrase together with the explanation and usually also an example of usage in sentence or conversation.[6] We randomly chose 101 lemmas corresponding to nouns (not word phrases) that are not present in MorfFlex (Hajič et al., 2020).

The inflected forms were first automatically generated by the rule-based guesser sklonuj.cz and then carefully post-edited by one annotator. The annotator was one of the authors, a senior undergraduate student and a Czech native speaker. The annotator was instructed to first post-edit the inflections and then revisit the annotations from a global perspective to ensure overall consistency of the inflections. In case of doubts, the annotator was encouraged to consult a standard reference of the Czech language, the Internet Language Reference Book,[7] managed by Czech Language Institute of Czech Academy of Sciences. In case of multiple equally-correct forms in one paradigm cells, all of them were included.

By this process, we obtained the test-neologisms dataset. As it is disjoint from the training set and is drawn from a completely different source, it is expected to represent a greater challenge for the inflection systems.

4. Methods

4.1. Evaluation Metrics

Form accuracy (FA, see eq. 1) is computed over all forms (except those marked as non-existent in the gold data). A generated form is considered to be correct if it is equal to the gold form or if it is equal to one of the correct forms (in the case of the test-neologisms dataset which allows multiple gold forms in one paradigm cell).

[5] https://cestina20.cz/, in Czech only

[6] For manual annotation, a subset of all neologisms, namely all words beginning with 'e' and 'j', was selected.

[7] https://prirucka.ujc.cas.cz/en

$$\text{FA} = \frac{\#(\text{CORRECTLY PREDICTED FORMS})}{\#(\text{ALL EXISTENT GOLD FORMS})} \quad (1)$$

Full-paradigm accuracy (FPA, see eq. 2) is computed over all lemmas. A paradigm table generated for a lemma is considered to be correct if it contains correct form in every cell (except for the forms marked as non-existent in the gold data).

$$\text{FPA} = \frac{\#(\text{CORR. PREDICTED PARADIGM TABLES})}{\#(\text{ALL LEMMAS})} \quad (2)$$

4.2. Baseline Systems

We make use of several systems as baselines for performance comparison.

COPY The copy baseline ignores the training data and treats every lemma as inflexible during prediction: returns list of copies of the lemma as the predicted forms.

SKLONUJ.CZ Sklonuj.cz represents the only ready-to-use guesser for Czech. It is based on handcrafted rules and therefore has low recall. It does not use the training dataset.

SIG NONNEUR The first standard baseline we use from SIGMORPHON shared tasks (Pimentel et al., 2021) is the non-neural one. It extracts transformation rules from the training examples and during prediction, it uses a majority classifier to apply the most frequent suitable rule.

SIG TRM Furthermore, we evaluate the neural baseline from SIGMORPHON shared task (Pimentel et al., 2021), based on a vanilla Transformer with original hyperparameters from Wu et al. (2021). The default training is with batch size 400 for 20k steps with the best performing checkpoint on the dev data, evaluated at the end of each epoch.

SIG TRM-TUNE We observed that the default training setting is not ideal for our task, because on the training part of the Czech OOV Inflection Dataset, the optimizer finishes in less than 2 epochs. Consequently, we conducted experiments with increased batch size and the number of training steps and finally obtained best results with 150k train steps and batch size 800.

4.3. Retrograde Approach

The first approach finds a word in a database with the longest common suffix, and inflects according to it. We adapt the basic idea of the linguistic module in ASIMUT (Králíková and Panevová, 1990): deciding how the lemma inflects based on its ending segment. Unlike in ASIMUT, we do not extract the abstract paradigms manually but rather save all training words as possible paradigms and search amongst them for the most feasible during prediction. We call the approach Retrograde because it is based on retrograde lexicographical similarity of words and we denote the model RETRO.

The model relies on two properties of Czech: (i) when two lemmas share the same ending, they also inflect identically, and (ii) during inflection (by number and case), only the ending changes while the rest of the word remains the same. This mostly holds in Czech but not in all other languages (e.g., semitic languages). The retrograde model is therefore strongly language dependent and we do not expect it to work well in all languages.

When building the model, we start with a morphological dictionary that contains complete paradigm tables for all covered lemmas. We save all the lemmas together with their inflection tables in a retrograde trie such that we can efficiently search them based on the suffixes.

When inflecting a lemma X, we search in the database for lemma A, such that X and A are most similar (have the longest common ending segment), and inflect lemma X according to the paradigm of lemma A. In case of multiple lemmas A in the dictionary with the same longest common ending segment with lemma X, we inflect X according to all of them and combine the predictions performing majority vote for each paradigm cell. In case of a tie, we choose the form from the most frequent ones randomly.

The inflection of lemma X according to paradigm A is performed as follows (see table 4): remove the longest common suffix from lemma X and lemma A to obtain X-stem and A-stem. Then for each paradigm cell take the corresponding A-form and replace the A-stem by X-stem.

We examined the dependence of the retrograde model on the size of the training data by experimenting with random subsets. As expected, the accuracy steadily improves when using more training data (fig. 1). Nevertheless, even with relatively small number of training lemmas (400 compared to the total 360k) the retrograde model outperforms the rule-based sklonuj.cz model.

4.4. LSTM-Based seq2seq Models

The second approach uses LSTM-based sequence-to-sequence (seq2seq) architectures originally pro-

HRAD			ÚŘAD	
hr-ad	hr-ady		úř-ad	úř-ady
hr-adu			úř-adu	
hr-adu	...	$\longrightarrow$	úř-adu	...
...			...	
hr-adem	hr-ady		úř-adem	úř-ady

Table 4: Retrograde model: example of inflection according to a paradigm. Lemma X for inflection: *úřad* (office), found database lemma A with the longest common suffix: *hrad* (castle), the longest common suffix: *ad*.

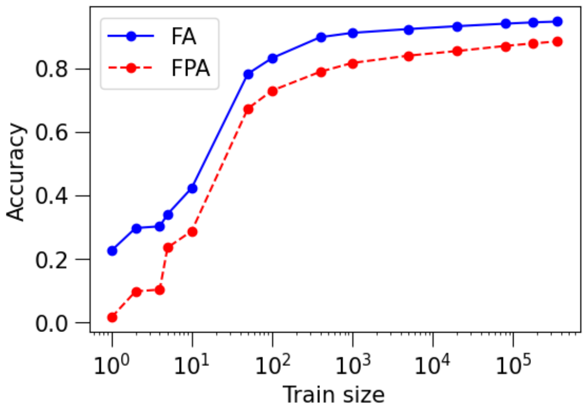

Figure 1: Retrograde model development evaluation with different sizes of training dataset. Train size = count of paradigm tables in the training dataset.

posed for the task of machine translation (Bahdanau et al., 2016). These architectures were broadly used in the SIGMORPHON shared tasks in recent years. We adapted the RNN-based encoder-decoder with soft attention as used by Kann and Schütze (2016). We used the implementation of the architectures as provided in the toolkit OpenNMT (Klein et al., 2017).[8]

[8] `OpenNMT-py v3.0.4`

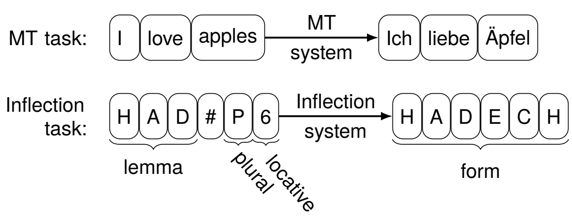

Figure 2: Input-output example for seq2seq models, and comparison of inflection task and the MT task. *had* (snake) is the lemma, "#" is a separator, `P6` is the morphological tag describing the target form (plural, locative case), *hadech* (snakes used as in *about snakes*) is the target inflected form.

4.4.1. Source-Target Data Representation

To be able to apply the MT architectures to our tasks, we formulate the inflection task as translation task using morphological tags, see fig. 2 for comparison of MT and inflection tasks, and for the example of input-output: the lemma plus tag as the source sequence, the inflected form as the target sequence.

Similarly to Kann and Schütze (2016), we use the individual characters of the source lemma followed by a separator and a 2-character morphological tag (describing the morphological categories of the target form) as input, and individual characters of the form as output. We investigated the usage of several different source-target representations, but obtained best results with this representation (although the differences in performance were marginal).

4.4.2. Hyperparameters

We perform hyperparameter tuning to adapt the architecture to the specifics of our task and the dataset. Batch size seems to be the most important: increasing it little by little from the original 20 to final 256 led to notable improvement, while adding the epochs by inflating the number of training steps with batch size fixed to 20 did not.

Kann and Schütze (2016) used 1 layer with 100 GRU units both in the encoder (bi-directional) and the decoder. We use LSTM units (Hochreiter and Schmidhuber, 1997) instead of GRU since it has been shown that LSTM performs better than GRU on larger datasets with shorter sequences (Yang et al., 2020).

Since our training dataset is much larger than the SIGMORPHON's 2016 dataset used by Kann and Schütze (2016), we examine extending the capacity of the network by increasing the number of hidden layers and their size and we experiment with the size of character and tag embeddings. Since the input and the output sequence share most of the vocabulary, we experiment with shared embeddings.

We obtained the best result with LSTM-based seq2seq trained for 13 epochs with Adam (Kingma and Ba, 2015) with default values of βs, with learning rate 0.001 and warm-up 4k steps, batch 256, with 2 layers of size 200, shared embedding of dimension 128, bi-directional encoder and with Luong attention (Luong et al., 2015); full configuration files are in the attachment.

We denote this model LSTM further in the text.

4.5. Transformers

We performed several experiments with current state-of-the-art Transformer-based seq2seq architecture implemented by OpenNMT. We used the same source-target data representation as for LSTM-based seq2seq models, described in section 4.4.1 (see fig. 2).

Although Wu et al. (2021) claimed that a small-capacity Transformer needs to be used in the inflection task, we achieved surprisingly good results with a high-capacity setting recommended for MT.[9] Only minor changes in the hyperparameters (hidden layer size, embedding dimension, dropouts, number of training steps and batch size) led to a model surpassing our extensively tuned LSTM model in both the form accuracy and the full-paradigm accuracy.

The Transformer has the following parameters: 6 layers of size 256, trainable embeddings dimension 256 (for single-character tokens representing words and morphological tags), 8 attention heads, feed-forward of size 2048, trained for 40k steps with batch size 1024 and accumulation count 4 (effective batch size 4096) with Adam with "noam" decay, starting at learning rate 2, with $\beta_2 = 0.998$. For regularization it uses layer normalization and dropout 0.2, attention dropout 0.2 and label smoothing 0.1. Full configuration files are in the attachment.

We denote this model TRM further in the text.

4.6. Model Ensembling

In addition to experiments with individual models, we investigated combining all the baselines and our models into ensembles: for every target form, the combined models vote, and in case of a tie a random form from the most frequent predictions is chosen.

We explored all possible combinations of models, and achieved best performance on the development set with the combination of two baselines and our 3 models: SIG NONNEUR, SIG TRM-TUNE, RETRO, LSTM and TRM. We denote it by ENSEMBLE further in the text.

5. Results and Systems Comparison

We evaluated all our systems, the baselines and the ENSEMBLE on the test-MorfFlex dataset and the test-neologisms dataset, and compared the performance on both form accuracy (FA) and full-paradigm accuracy (FPA). We measured statistical significance of the the differences on both metrics using the non-parametric approximation permutation test algorithm (Fay and Follmann, 2002; Gandy, 2009) with significance level 0.05 and with 10k resamplings. The results are presented in table 5.

[9] adapted from https://github.com/ymoslem/OpenNMT-Tutorial/blob/main/2-NMT-Training.ipynb

test	MorfFlex		neologisms	
model	FA	FPA	FA	FPA
COPY	22.59	1.48	13.13	0
SKLONUJ.CZ	88.88	74.43	86.22	55
SIG NONNEUR	94.78	88.15	**89.49**	**71**
SIG TRM	95.47	87.29	87.53	63
SIG TRM-TUNE	96.17	90.15	86.51	55
RETRO	94.85	88.64	89.34	**71**
LSTM	96.16	89.80	86.95	58
TRM	**96.18**	**90.44**	87.24	61
ENSEMBLE	96.35	90.70	90.43	64

Table 5: Evaluation of systems (FA and FPA in percent) on test-MorfFlex (left) and test-neologisms (right). Upper section: SIG baseline systems by other authors (Cotterell et al., 2017; Pimentel et al., 2021), trained on Czech OOV Inflection Dataset. Midsection: our systems. Bottom section: ENSEMBLE combines SIG NONNEUR, SIG TRM-TUNE, RETRO, LSTM and TRM.

5.1. Test-MorfFlex

On test-MorfFlex, the best performing model is TRM, which achieves 90.44% in the full-paradigm accuracy and is statistically significantly better than all other models. In the form accuracy, it achieves 96.18%, but LSTM and SIG TRM-TUNE perform only slightly worse, and the differences between them are not statistically significant. All these three models are significantly better than the rest of the models. They are followed by SIG TRM baseline and then by RETRO model and SIG NONNEUR baseline. The RETRO model statistically significantly outperforms the SIG NONNEUR baseline in both metrics. All models (except for the COPY baseline) are significantly better than SKLONUJ.CZ.

The success of the Transformer-based models suggests that the Transformer architecture is indeed suitable for the inflection task, even in the OOV conditions, at least when the training data is plentiful.

The ENSEMBLE outperforms all the models in both metrics, showing that the errors made by the models are somehow complementary. Moreover, if we knew how to choose the best among the forms predicted by all the models and baselines, we could achieve 99.3% in the form accuracy and 97.3% in the full-paradigm accuracy.

5.2. Test-Neologisms

The results are quire different on the test-neologisms dataset. There is a large drop in the performance of all models as compared to the performance on test-MorfFlex, most pronounced in the performance of the neural models, especially SIG TRM-TUNE with almost 10% drop in the form accuracy and 35% in the full-paradigm accuracy.

The RETRO model and SIG NONNEUR baseline perform comparably and are statistically significantly better in the form accuracy than all other models.[10] The differences between the neural models and SKLONUJ.CZ are not statistically significant.

The overall drop in performance is understandable: the models were trained on data that come from the same distribution as test-MorfFlex, but from a completely different distribution than test-neologisms.

The dominance of the RETRO model and the SIG NONNEUR baseline could be (at least partially) caused by the fact that test-neologisms contains high percentage (37%) of compounds, blends or words derived by prefixing, whose ending segment is an existing word present in MorfFlex (and thus possibly present in the training data). Those words are especially convenient for the RETRO model since the simple algorithm is able to ignore the prefix and inflect the word correctly.

The ENSEMBLE outperforms all models in the form accuracy, but not in the full-paradigm accuracy. The upper bound accuracy, when choosing the predictions from all the models and baselines, is 96.5% in the form accuracy and 82.2% in the full-paradigm accuracy, which shows that there is still room for improvement when using ensembles of current models.

5.3. SIGMORPHON 2022 Evaluation

In order to evaluate the robustness of our seq2seq systems and to compare them to established approaches on a well-known dataset, we evaluated our LSTM and TRM models on the SIGMORPHON 2022 data (Kodner et al., 2022). Specifically, we evaluate the performance on all 16 development languages[11] that included large training dataset and test data for the feature overlap (OOV) evaluation condition.

The datasets differ from our setting in several aspects: (i) the training dataset is smaller (~2k lemma-tag-form entries) compared to our dataset (~5M entries), even in the large data condition, (ii) there are datasets for 16 different languages and none of them is Czech, and (iii) the data consist not only of nouns but also contain other part-of-speech.

[10]In the full-paradigm accuracy, almost no difference was stastistically significant due to low number of paradigm tables in test-neologisms dataset.

[11]ang = Old English, ara = Modern Standard Arabic, asm = Assamese, got = Gothic, hun = Hungarian, kat = Georgian, khk = Khalkha Mongolian, kor = Korean, krl = Karelian, lud = Ludic, non = Old Norse, pol = Polish, poma = Pomak, slk = Slovak, tur = Turkish, vep = Veps (Kodner et al., 2022)

To be able to run our models on the SIGMORPHON data, we convert the data to our format by tokenizing the lemma and the word form to individual characters, add the special separator token to the end of the source sequence and then add the morphological features one by one. Once the model produces the output in our format, we convert it back to the SIGMORPHON format and evaluate it using the official evaluation script.[12]

We trained the LSTM model for 260k steps with batch size 256 (approx. 9.5k epochs), and the TRM for 40k steps with effective batch size 4096 (approx. 23.4k epochs) and we chose the checkpoint with best performance on dev.

We present the results in the feature overlap (OOV) condition in table 6. We compare the performance of our systems with the neural and non-neural baseline and with all 5 submitted systems evaluated in the feature overlap (OOV) condition.

The LSTM model achieves the best score in 4 out of 16 languages and the TRM model in other 5 languages. Averaged over all languages, our systems take the second and third place (TRM with 86.1%, LSTM with 85.3%, respectively).

We suspect that the Transformer approach lag behind LSTM in some languages might result from an interplay between the corpus size and the morphological complexity of the language. Some of the SIGMORPHON corpora are relatively small for ML training. We hypothesize that Transformers might benefit from plentiful training data, but the influence of morphological complexity of the language remains to be accounted for.

It is also interesting that we achieved high score particularly in Slavic languages (Polish (pol), Pomak (poma) and Slovak (slk)). We can see that although we focused specifically on Czech morphology when tuning our setup, the models perform particularly well when trained and evaluated also on other Slavic languages.

These results show robustness of our seq2seq systems: although they were tuned for good performance on inflection of Czech nouns, they are suitable for inflecting also other parts-of-speech and other languages.

6. Error Analysis

We perform error analysis of the model predictions on the dev set of the Czech OOV Inflection Dataset.

6.1. Proper vs. Common Nouns

Across all the models (except for COPY baseline), almost 70% or more of the incorrectly predicted forms are forms of proper nouns, while the total

[12]https://github.com/sigmorphon/2022InflectionST/tree/main/evaluation

Lang	Submitted systems					Baselines		Ours	
	CLUZH	Flexica	OSU	TüM	UBC	Neural	NonNeur	LSTM	TRM
ang	**76.6**	64.4	73.7	71.9	74.1	73.4	68.7	76.3	75.5
ara	81.7	65.5	78.7	78.5	65.5	81.9	50.8	79.2	**82.6**
asm	83.3	75.0	75.0	**91.7**	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
hun	93.5	62.9	93.1	92.8	91.5	**94.4**	73.1	92.8	**94.4**
kat	96.7	95.7	96.7	96.7	96.7	97.3	96.7	97.3	**97.8**
khk	94.1	47.1	94.1	94.1	88.2	94.1	88.2	**100.0**	94.1
kor	**71.1**	55.4	50.6	56.6	60.2	62.7	59.0	49.4	62.7
krl	87.5	69.8	85.9	57.8	85.4	57.8	20.8	**89.1**	85.9
lud	87.3	92.0	92.9	93.4	88.2	**94.3**	93.4	89.2	92.0
non	85.2	77.0	85.2	80.3	**90.2**	88.5	80.3	83.6	88.5
pol	**96.1**	85.9	94.9	74.0	95.7	74.4	86.3	**96.1**	95.6
poma	76.1	54.5	70.1	69.4	73.3	74.1	47.8	75.2	**76.3**
slk	93.5	90.0	92.2	70.4	**95.7**	71.1	92.4	95.2	**95.7**
tur	93.7	57.9	**95.2**	80.2	92.9	79.4	66.7	**95.2**	92.9
vep	**71.5**	58.8	70.0	57.5	68.8	59.2	60.4	70.7	68.8
average	**86.3**	68.3	83.9	78.6	83.8	80.0	72.2	85.3	86.1

Table 6: SIGMORPHON 2022 comparison – Feature Overlap (FA metric in percent): A test pair's feature set is attested in training, but its lemma is novel. Except for results of our systems, the table was adopted from Kodner et al. (2022, Tables 17, 18 – feature rows). The systems are: CLUZH (Wehrli et al., 2022), Flexica (Sherbakov and Vylomova, 2022), OSU (Elsner and Court, 2022), TüM (Merzhevich et al., 2022), UBC (Yang et al., 2022), and the SIGMORPHON 2022 baselines (Kodner et al., 2022).

subset of dev set	FA	FPA
proper nouns	90.83	73.29
common nouns	98.40	93.91
overall	96.00	90.12

Table 7: TRM model performance on proper vs. common nouns.

percentage of proper nouns in the dev set is only 31.68%.

We compare the performance when evaluated on the corresponding parts of the dev set separately. The performance of all models improves substantially when running on common nouns only, and gets worse on the proper nouns subset. We show the differences of performance of the TRM model in table 7. Most noticable is the poor performance on the FPA on proper nouns. This trend is similar in the rest of the models, with the only exception of SKLONUJ.CZ, which has extremely poor performance on proper nouns (72.30% FA, 29.87% FPA), but on common nouns it is much closer to the rest of the models (96.36% FA, 87.40% FPA). This is caused by the fact that it is not able to inflect a lot of proper nouns and simply returns nothing for them.

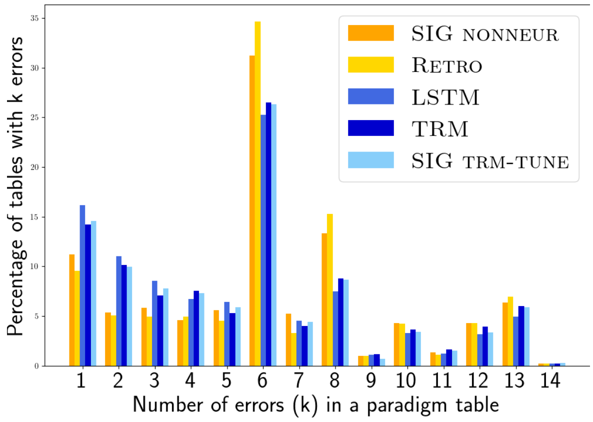

Figure 3: Histogram: the percentage of paradigm tables with given number of errors, counted for all incorrectly generated paradigm tables, for each model separately.

6.2. Distribution of Error Counts

We focused on the error counts amongst the incorrectly generated paradigm tables, and on the percentages of errors made by the models in the individual paradigm cells (`S1` up to `P7`, `S`=singular, `P`=plural, `1-7`=case as in table 1).

The cells `S6`, `P1` and `P7` are the most difficult to predict for all systems, while the easiest one is `S1` (typically equal to the lemma).

We count the number of errors in each incorrectly predicted paradigm table, and for every error count (1 up to 14) we plot the percentage of tables with that number of errors for each of the models (TRM, SIG TRM-TUNE, LSTM, RETRO and SIG NONNEUR) separately (fig. 3). Clearly, the most common number of errors amongst all models is 6 (more than 1/4 of all incorrect paradigm tables). More interestingly, the errors are in the same cells across all the models: for every model, more than 90% of tables with 6 errors have the errors in the cells `S2`, `S3`, `S4`, `S6`, `P1` and `P5`. This probably reflects a property of the language itself: individual paradigms differ in these cells more than in the other cells.

Moreover, we can see that the non-neural models (RETRO, SIG NONNEUR; shades of orange) behave similarly, the neural models (LSTM, TRM, SIG TRM-TUNE; shades of blue) also behave similarly, but these two model groups behave differently. The neural models have higher percentage of small number of errors (1 up to 4 errors), while the non-neural tend to make more errors (especially 6, 8, and 13 errors). We believe this is because the neural models generate each form for a lemma independently without using the concept of paradigms, thus easily making occasional errors in individual forms. On the other hand, the non-neural models implicitly or explicitly use the concept of paradigms, and thus are more likely to either choose the paradigm correctly and make no errors, or incorrectly and make many errors.

7. Conclusion

We examined the understudied topic of inflection in out-of-vocabulary (OOV) conditions.

To this end, we created a lemma-disjoint train-dev-test split of a large pre-existing Czech morphological dictionary MorfFlex, and we also manually annotated a new small Czech test set of neologisms. We release this data as the Czech OOV Inflection Dataset.[13]

We studied three approaches to inflect OOVs: retrograde approach, LSTMs and Transformers. We thoroughly tested these approaches on our dataset, as well as OOV test sets for 16 other languages from the SIGMORPHON 2022 shared task.

We find that on our dataset, Transformer reaches the best results on test-MorfFlex, whereas the retrograde approach beats both neural models on test-neologisms. On the SIGMORPHON data, our seq2seq models achieve state-of-the-art results for 9 out of 16 languages.

We release our inflection system as a Python library.[14]

[13] http://hdl.handle.net/11234/1-5471

[14] https://github.com/tomsouri/cz-inflect

Limitations

As the Czech OOV Inflection Dataset encompasses all noun entries from the large Czech morphological dictionary MorfFlex (Hajič et al., 2020), with the exception of 2% cleaned entries, we assume that we did not introduce any significant bias when constructing the dataset.

The manually annotated test-neologisms is a subset of a corpus of Czech neologisms Čeština 2.0 (Kavka and Škrabal et al., 2018): for annotation, all words starting with 'e' and 'j' were selected. This process cannot be generally viewed as random and entirely representative. Nevertheless, we assume that the first character of a lemma does not have a significant influence on the way the word inflects. This assumption is supported by the fact that Czech is mostly a suffixing language. Another possible bias of the test-neologisms might be stemming from the fact that the underlying corpus of Czech neologisms contains many compounds.

Finally, the two most notable limitations of the Czech OOV Inflection Dataset is the restriction to nouns only, and the fact that it contains only the Czech language; we leave the other parts of speech and other languages for future work. We nevertheless assume that the presented results can be generalized to other languages, as evidenced by extensive evaluation of all methods also on 16 languages of the SIGMORPHON shared task data.

Of the presented methods, the retrograde approach (section 4.3) is expected to be the most limited in generalization across languages, as it exploits the shared similarity in suffix inflection between lemmas in the Czech language.

Ethical Considerations

All manual annotations and evaluations within the work described in this paper were done by one male member of the team. However, as the morphological inflection in Czech is relatively straightforward and follows grammatical rules, we do not expect differences in annotation results in a mixed team.

No personal information has been among the lemmas extracted from the morphological dictionary. Both our neural methods, LSTM and Transformer, were trained from scratch on the training data, and we did not utilize any pre-trained LLMs, which might have contained personal information or biases.

Also, as we are not using any pre-trained LLMs, our methods are relatively cheap and efficient.

The authors declare that they are not aware of any conflict of interest related to the work published herein.

Acknowledgements

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).

This work has also been supported by the Grant Agency of the Czech Republic under the EXPRO program as project "LUSyD" (project No. GX20-16819X).

We thank the Bernard Bolzano Endowment Fund for the contribution for covering the travel expenses to present this work.

We also thank the anonymous reviewers for their valuable comments.

8. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

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.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2013. Robust multilingual statistical morphological generation models. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 158–164, Sofia, Bulgaria. Association for Computational Linguistics.

Micha Elsner and Sara Court. 2022. OSU at SigMorphon 2022: Analogical inflection with rule features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 220–225, Seattle, Washington. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning.

Michael P. Fay and Dean A. Follmann. 2002. Designing monte carlo implementations of permutation or bootstrap hypothesis tests. *The American Statistician*, 56(1):63–70.

Axel Gandy. 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511.

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.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Linguistic Data Consortium, University of Pennsylvania.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70, Berlin, Germany. Association for Computational Linguistics.

Martin Kavka and Michal Škrabal et al. 2018. *Hacknutá čeština*. Jan Melvil Publishing.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

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, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 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.

Květoslava Králíková and Jarmila Panevová. 1990. ASIMUT - a method for automatic information retrieval from full texts. *Explizite Beschreibung der Sprache und automatische Textbearbeitung*, XVII.

Ling Liu and Mans Hulden. 2021. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. *CoRR*, abs/2104.06483.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation.

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.

Tatiana Merzhevich, Nkonye Gbadegoye, Leander Girrbach, Jingwen Li, and Ryan Soh-Eun Shim. 2022. 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*, pages 204–211, Seattle, Washington. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. 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*, pages 229–259, Online. Association for Computational Linguistics.

Andreas Sherbakov and Ekaterina Vylomova. 2022. 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*, pages 240–246, Seattle, Washington. Association for Computational Linguistics.

Eleonora Slavíčková. 1975. *Retrográdní morfematický slovní češtiny*, 1. edition. Academia.

Tomáš Sourada. 2023. *Automatic inflection in Czech language*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague.

Jana Straková, Milan Straka, and Jan Hajič. 2014. 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*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 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.

Silvan Wehrli, Simon Clematide, and Peter Makarov. 2022. 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*, pages 212–219, Seattle, Washington. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *CoRR*, abs/2005.10213.

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.

Changbing Yang, Ruixin (Ray) Yang, Garrett Nicolai, and Miikka Silfverberg. 2022. Generalizing morphological inflection systems to unseen lemmas. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 226–235, Seattle, Washington. Association for Computational Linguistics.

Shudong Yang, Xueying Yu, and Ying Zhou. 2020. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence IWECAI*, pages 98–101.

9. Language Resource References

Hajič, Jan and Hlaváčová, Jaroslava and Mikulová, Marie and Straka, Milan and Štěpánková, Barbora. 2020. *MorfFlex CZ 2.0*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University.