(Un)solving Morphological Inflection:
Lemma Overlap Artificially Inflates Models' Performance

Omer Goldman, David Guriel, Reut Tsarfaty

Bar-Ilan University

{omer.goldman,davidgu1312}@gmail.com,reut.tsarfaty@biu.ac.il

Abstract

In the domain of Morphology, Inflection is a fundamental and important task that gained a lot of traction in recent years, mostly via SIG-MORPHON's shared-tasks. With average accuracy above 0.9 over the scores of all languages, the task is considered mostly solved using relatively generic neural seq2seq models, even with little data provided. In this work, we propose to re-evaluate morphological inflection models by employing harder train-test splits that will challenge the generalization capacity of the models. In particular, as opposed to the naïve split-by-form, we propose a split-by-lemma method to challenge the performance on existing benchmarks. Our experiments with the three top-ranked systems on the SIGMORPHON's 2020 shared-task show that the lemma-split presents an average drop of 30 percentage points in macro-average for the 90 languages included. The effect is most significant for low-resourced languages with a drop as high as 95 points, but even high-resourced languages lose about 10 points on average. Our results clearly show that generalizing inflection to unseen lemmas is far from being solved, presenting a simple yet effective means to promote more sophisticated models.

1   Introduction

In recent years, morphological (re)inflection tasks in NLP have gained a lot of attention, most notably with the introduction of SIGMORPHON's shared tasks (Cotterell et al., 2016, 2017, 2018; Vylomova et al., 2020) in tandem with the expansion of Uni-Morph (McCarthy et al., 2020), a multi-lingual dataset of inflection tables. The shared-tasks sample data from UniMorph includes lists of triplets in the form of *(lemma, features, form)* for many languages, and the shared-task organizers maintain standard splits for a fair system comparison.

The best-performing systems to-date in all inflection shared-tasks are neural sequence-to-sequence models used in many NLP tasks. An LSTM-based model won 2016's task (Kann and Schütze, 2016), and a transformer came on top in 2020 (Canby et al., 2020). In 2020's task the best model achieved exact-match accuracy that transcended 0.9 macro-averaged over up to 90 languages from various language families and types. This trend of high results recurred in works done on data collected independently as well (e.g. Malouf, 2017, Silfverberg and Hulden, 2018, inter alia).

Interestingly, the averaged results of 2020's shared-task include languages for which very little data was provided, sometimes as little as a couple of hundreds of examples. This has led to a view considering morphological inflection a relatively simple task that is essentially already *solved*, as reflected in the saturation of the results over the year and the declining submissions to the shared tasks.[1] This also led the community to gravitate towards works attempting to solve the same (re)inflection tasks with little or no supervision (McCarthy et al., 2019; Jin et al., 2020; Goldman and Tsarfaty, 2021).

However, before moving on we should ask ourselves whether morphological inflection is indeed *solved* or may the good performance be attributed to some artifacts in the data. This was shown to be true for many NLP tasks in which slight modifications of the data can result in a more challenging dataset, e.g., the addition of unanswerable questions to question answering benchmarks (Rajpurkar et al., 2018), or the addition of expert-annotated minimal pairs to a variety of tasks (Gardner et al., 2020). A common modification is re-splitting the data such that the test set is more challenging and closer to the intended use of the models in the wild (Søgaard et al., 2021). As the performance on morphological inflection models seems to have saturated on high scores, a similar rethinking of the data used is warranted.

---

[1]The shared task of 2021 had seen only two submissions (Pimentel et al., 2021).

In this work we propose to construct more difficult datasets for morphological (re)inflection by splitting them such that the test set will include no forms of lemmas appearing in the train set. This splitting method will allow assessing the models in a challenging scenario closer to their desired function in practice, where training data usually includes full inflection tables and learning to inflect the uncovered lemmas is the target.

We show, by re-splitting the data from task 0 of SIGMORPHON's 2020 shared-task, that the proposed split reveals a greater difficulty of morphological inflection. Retesting 3 of the 4 top-ranked systems of the shared-task on the new splits leads to a decrease of 30 points averaged over the systems for all 90 languages included in the shared-task. We further show that the effect is more prominent for low-resourced languages, where the drop can be as large as 95 points, though high-resourced languages may suffer from up to a 10 points drop as well. We conclude that in order to properly assess the performance of (re)inflection models and to drive the field forward, the data and related splits should be carefully examined and improved to provide a more challenging evaluation, more reflective of their real-world use.

2 (Re)inflection and Memorization

Inflection and reinflection are two of the most dominant tasks in computational morphology. In the *inflection* task, the input is a lemma and a feature-bundle, and we aim to predict the respective inflected word-form. In *reinflection*, the input is an inflected word-form along with its features bundle, plus a feature-bundle without a form, and we aim to predict the respective inflected-form for the same lemma. The *training* input in SIGMORPHON's shared-tasks is a random split of the available *(lemma,form,features)* triplets such that no triplet occurring in the train-set occurs in the test-set.[2]

In such a setting, models can short-cut their way to better predictions in cases where forms from the same lemma appear in both the train and test data. This may allow models to memorize lemma-specific alternations that make morphological inflection a challenging task to begin with. Consider for example the notoriously unpredictable German plurality marking, where several allomorphs are associated with nouns with no clear rule governing the process. *Kind*, for example, is pluralized with the suffix *-er* resulting in *Kinder* tagged as NOM;PL. Assuming a model saw this example in the train set it is pretty easy to predict *Kindern* for the same lemma with DAT;PL features,[3] but without knowledge of the suffix used to pluralize *Kind* the predictions *Kinden* and *Kinds* are just as likely.

3 Related Work

Many subfields of NLP and machine learning in general suggested *hard splits* as means to improve the probing of models' ability to solve the underlying task, and to make sure models do not simply employ loopholes in the data.

In the realm of sentence simplification, Narayan et al. (2017) suggested the WEBSPLIT dataset, where models are required to split and rephrase complex sentences associated with a meaning representation over a knowledge-base. Aharoni and Goldberg (2018) found that some facts appeared in both train and test sets and provided a harder split denying models the ability to use memorized facts. Aharoni and Goldberg (2020) also suggested a general splitting method for machine translation such that the domains are as disjoint as possible.

In semantic parsing, Finegan-Dollak et al. (2018) suggested a better split for parsing natural language questions to SQL queries by making sure that queries of the same template do not occur in both train and test, while Lachmy et al. (2021) split their HEXAGONS data such that any one visual pattern used for the task cannot appear in both train and test. Furthermore, Loula et al. (2018) adversarially split semantic parsing for navigation data to assess their models' capability to use compositionality. In spoken language understanding Arora et al. (2021) designed a splitting method that will account for variation in both speaker identity and linguistic content.

In general, concerns regarding data splits and their undesired influence on model assessments led Gorman and Bedrick (2019) to advocate random splitting instead of standard ones. In reaction, Søgaard et al. (2021) pointed to the flaws of random splits and suggested adversarial splits to challenge models further. Here we call for paying attention to the splits employed in evaluating morphological models, and improve on them.

---

[2]This is true for all SIGMORPHON's inflection shared tasks, save the paradigm completion task of 2017.

[3]The addition of the dative marker *-n* is very regular.

Split	Accuracy		Edit Distance	
	Form	Lemma	Form	Lemma
DeepSpin-02	**0.90**	**0.76**	**0.23**	**0.58**
CULing	0.88	0.63	0.29	1.02
Base trm-single	**0.90**	0.53	**0.23**	1.32
Base LSTM	0.85	0.39	0.34	1.79
Average	0.88	0.58	0.27	1.18

Table 1: Exact-match accuracy and edit-distance for our baseline and 3 of the 4 top-ranked systems of SIGMORPHON's 2020 shared-task, all reported on the original split of the shared-task (form split) and on our harder lemma split. Best system per column is in **bold**.

	Accuracy	
Split	**Form**	**Lemma**
Afro-Asiatic	$0.93 \ (0.95)_\text{T}$	$0.51 \ (0.80)_\text{D}$
Austronesian	$0.78 \ (0.82)_\text{T}$	$0.45 \ (0.70)_\text{D}$
Germanic	$0.86 \ (0.88)_\text{D}$	$0.63 \ (0.74)_\text{D}$
Indo-Iranian	$0.93 \ (0.97)_\text{D}$	$0.55 \ (0.86)_\text{D}$
Niger-Congo	$0.95 \ (0.98)_\text{T}$	$0.56 \ (0.90)_\text{D}$
Oto-Manguean	$0.84 \ (0.86)_\text{T}$	$0.53 \ (0.60)_\text{D}$
Romance	$0.97 \ (0.99)_\text{T}$	$0.69 \ (0.86)_\text{D}$
Turkic	$0.95 \ (0.96)_\text{T}$	$0.64 \ (0.89)_\text{D}$
Uralic	$0.88 \ (0.90)_\text{C}$	$0.65 \ (0.72)_\text{D}$

Table 2: Aggregated results for the various language families. We provide the performance averaged across all systems, and in parenthesis the performance of the best system per family. The best system is identifiable in subscript: C - CULing, T - Base trm-single, D - DeepSpin-02. We include here only families with at least 3 languages in the data.

4 Experiments

In order to better assess the difficulty of morphological inflection, we compare the performances of 3 of the top-ranked system at task 0 (inflection) of SIGMORPHON's 2020 shared-task. We examined each system on both the the standard (form) split and the novel (lemma) split.

When re-splitting,[4] we kept the same proportions of the form-split data, i.e. we split the inflection tables 70%, 10% and 20% for the train, development and test set. In terms of examples the proportions may vary as not all tables are of equal size. In practice, the averaged train set size in examples terms was only 3.5% smaller in the lemma-split data, on average.[5]

4.1 The Languages

SIGMORHPON's 2020 shared-task includes datasets for 90 typologically and genealogically diverse languages from 14 language families. The languages are varied along almost any typological dimension, from fusional to agglutinative, small inflection tables to vast ones. They include mostly prefixing and mostly suffixing languages with representation of infixing and circumfixing as well. The languages vary also in use, including widely-used languages such as English and Hindi and moribund or extinct languages like Dakota and Middle High German.[6]

4.2 The Models

We tested the effects of lemma-splitting on our own LSTM-based model as well as 3 of the 4 top-ranked systems in the shared task.[7]

Base LSTM We implemented a character-based sequence-to-sequence model which consists of a 1-layer bi-directional LSTM Encoder and a 1-layer unidirectional LSTM Decoder with a global soft attention layer (Bahdanau et al., 2014). Our model was trained for 50 epochs with no model selection.[8]

Base trm-single The shared-task's organizers supplied various baselines, some based on a transformer architecture that was adapted for character-level tasks (Wu et al., 2021).[9] All baseline models include 4 encoder and 4 decoder layers, consisting of a multi-head self-attention layer and 2 feed-forward layers, equipped with a skip-connection. In every decoder layer a multi-head attention layer attends to the encoder's outputs. The network was trained for 4,000 warm-up steps and up to 20,000 more steps, each step over a batch of size 400. The model was examined with and without augmented data, trained separately on each language or each language family. One of the baseline setups, training a model per language without augmented data, made it to the top 4 systems and we include it here.

---

[4] The split was done randomly as is standard in SIGMORPHON tasks, although frequency-based sampling is also conceivable and is sometimes used, as in Cotterell et al. (2018).

[5] The newly-split data is available at https://github.com/OnlpLab/LemmaSplitting.

[6] The full list with the originally released data are at https://github.com/sigmorphon2020/task0-data.

[7] The best performing system, UIUC (Canby et al., 2020), did not have a publicly available implementation.

[8] The code is available at https://github.com/OnlpLab/LemmaSplitting.

[9] The code is available at https://github.com/shijie-wu/neural-transducer.

DeepSpin Peters and Martins (2020) submitted a recurrent neural network – dubbed DeepSpin-02.[10] The system is composed of 2 bi-directional LSTM encoders with bi-linear gated Attention (Luong et al., 2015), one for the lemma characters and one for the features characters, and a unidirectional LSTM Decoder for generating the outputs. The innovation in the architecture is the use of sparsemax (Martins and Astudillo, 2016) instead of softmax in the attention layer.[11]

CULing Liu and Hulden (2020)'s system is also based on the transformer architecture, with hyper-parameters very similar to *base trm-single*.[12] Their innovation is in restructuring the data such that the model learns to inflect from any given cell in the inflection table rather than solely from the lemma.

4.3 Results

Table 1 summarizes our main results. We clearly see a drop in the performance for all systems, with an average of 30 points. The table also shows that splitting the data according to lemmas allows discerning between systems that appear to perform quite similarly on the form-split data. The best system on the lemma-split data, DeepSpin-02, outperforms the second-ranked CULing system by about 13 points with both baseline systems performing significantly worse. The results in terms of averaged edit distance show the same trends.

DeepSpin-02 emerges victorious also in Table 2, where results are broken down by language family. The table shows that DeepSpin-02 is the best performer over all language families when data is lemma-split, in contrast to the mixed picture over the form-split data.

The average performance per language family seems to be controlled by training data availability. For example, Germanic languages show average drop of 23 points, while for Niger-Congo languages the drop is 39 points on average.

In order to further examine the relation between the amount of training data and drop in performance we plotted in Figure 1 the drop per system and per language against the size of the avail-
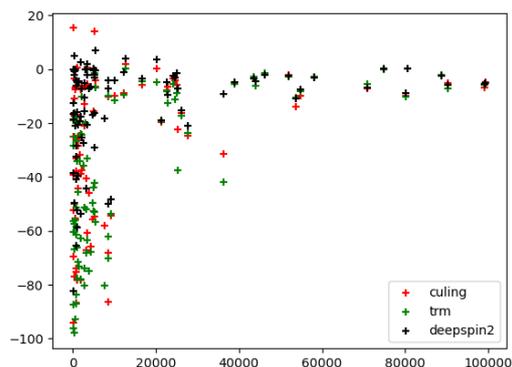

Figure 1: Performance drop for the various systems when moving from form to lemma split as a function of the size of the train data. The effect is clearly more significant for lower-resourced languages.
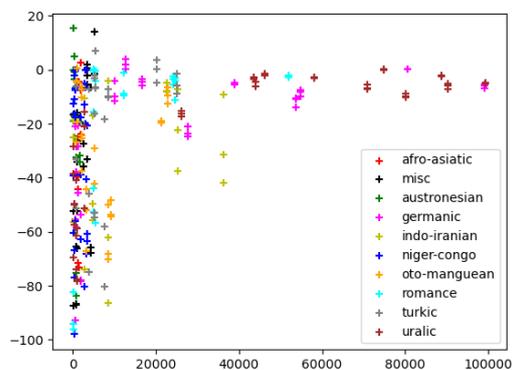

Figure 2: Performance drop for the various language families when moving from form to lemma split as a function of the size of the train data. We include here only families with at least 3 languages in the data, the rest are classified under *misc*.

able train data, color-coded to indicate systems. It shows that the major drops in performance that contributed the most to the overall gap between the splits are in those low-resourced language. Remarkably, for some systems and languages the drop can be as high as 95 points. On the other hand, on high-resourced languages with 40,000 training examples or more, all systems didn't lose much. The analysis also shows the advantage of DeepSpin-02 in the lower-resourced settings that made it the best performer overall.

When color-coding the same broken-down data for linguistic family membership rather than system, as we do in Figure 2, it becomes clear that there is no evidence for specific families being eas-

---

ier for inflection when little data is provided. The figure does show the remarkable discrepancy in annotation effort, as the high-resourced languages mostly belong to 2 families: Germanic and Uralic.

5 Discussion

We proposed a method for splitting morphological datasets such that there is no lemma overlap between the splits. On the re-split of SIGMORPHON's 2020 shared-task data, we showed that all top-ranked systems suffer significant drops in performance. The new split examines models' generalization abilities in conditions more similar to their desired usage in the wild and allows better discerning between the systems in order to point to more promising directions for future research — more so than the original form-split data on which all systems fared similarly. The new splitting method is likely to lead to more sophisticated modeling, for instance, in the spirit of the model proposed by Liu and Hulden (2021). The suggested move to a harder split is not unlike many other NLP tasks, in which challenging splits are suggested to drive the field forward. We thus call for morphological studies to carefully attend to the data used and expose the actual difficulties in modelling morphology, in future research and future shared tasks.

Acknowledgements

References

Roee Aharoni and Yoav Goldberg. 2018. Split and rephrase: Better evaluation and stronger baselines. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 719–724, Melbourne, Australia. Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Siddhant Arora, Alissa Ostapenko, Vijay Viswanathan, Siddharth Dalmia, Florian Metze, Shinji Watanabe,

and Alan W Black. 2021. Rethinking end-to-end evaluation of decomposable tasks: A case study on spoken language understanding.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Marc Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and Julia Hockenmaier. 2020. University of Illinois submission to the 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 137–145, Online. Association for Computational Linguistics.

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.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco,

Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2021. Minimal supervision for morphological inflection.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

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.

Royi Lachmy, Valentina Pyatkin, and Reut Tsarfaty. 2021. Draw me a flower: Grounding formal abstract structures stated in informal natural language.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 153–161, Online. Association for Computational Linguistics.

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.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

André F. T. Martins and Ramón F. Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1614–1623. JMLR.org.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

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.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Ben Peters and André F. T. Martins. 2019. IT–IST at the SIGMORPHON 2019 shared task: Sparse two-headed models for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 50–56, Florence, Italy. Association for Computational Linguistics.

Ben Peters and André F. T. Martins. 2020. One-size-fits-all multilingual models. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 63–69, Online. 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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. 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, 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.

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.