Evaluating Evaluation: Comparing Adversarial Approaches to Evaluating Neural Models of Morphological Inflection

Sarah Payne, QP2 Proposal
Stony Brook University
sarah.payne@stonybrook.edu

Abstract

Morphological inflection is a fundamental task in subword NLP, popularized by the recent SIGMORPHON shared tasks. For several years now, state-of-the-art neural models have reported extremely high average accuracy across languages on these tasks. This apparent saturation has led to the development of a range of adversarial evaluation practices, based on the common insight that traditional train-test splits don't control for whether the model has seen either the *lemma* or *feature set* separately in its training data. These evaluation practices, however, differ drastically in their results: while Goldman et al. (2022) reports that models fail to generalize to unseen *lemmas*, Kodner et al. (2022) find that the models have little trouble generalizing to unseen lemmas, but take a large performance hit when generalizing to unseen *feature sets*. In this Qualifying Paper, I will (**TODO:**)

1 Introduction

(**TODO:** Write the Introduction & maybe a related work section)

Generalization, the ability to extend patterns from known to unknown items, is a critical part of mor- phological competence. Morphological sparsity (**TODO:** talk about sparsity and acquisition)

So far, the best-performing models have been neural sequence-to-sequence models (Kann and Schütze, 2016; Canby et al., 2020)

Many subfields of NLP and machine learning in general suggested hard splits as means to improve the probing of models' ability to solve the under- lying task, and to make sure models do not simply employ loopholes in the data. The addition of unanswerable questions to question answering benchmarks (Rajpurkar et al., 2018), or the addition of expert-annotated minimal pairs (Gardner et al., 2020). Narayan et al. (2017) suggested using the WEBSPLIT data, 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. (2022) split their HEXAGONS data such that any one visual pat- tern 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. 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.

Motivation for the generalization task:

Pimentel et al. (2021) looked at lemma overlap as well but didn't control for featureset overlap

2 Defining the Task

2.1 Morphological Inflection as an NLP Task

In standard morphological inflection tasks, models are exposed to triples of `(lemma, feature set, inflected form)` during training. During evaluation, the model is given a `(lemma, feature set)` pair as input and the goal is to correctly predict the corresponding inflected form. For example, were the model to be given as input (walk, V;PAST), then we would expect it to output walked.

2.2 The Role of Overlap

In most versions of the SIGMORPHON shared task (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Goldman et al., 2023), train-test splits are created by randomly sampling from the available `(lemma, feature set, inflected form)` triples. While this approach entails that no triple occurring in the train set will occur in the evaluation set, as both Goldman et al. (2022) and Kodner et al. (2022) note, it ignores the fact that lemmas or feature sets that appear during train may reappear during test, since lemmas and feature sets can be combined independently.

To illustrate this point, consider the toy example below, taken from Kodner et al. (2023c). Though none of the triples appearing in the train set (1) re-appear in the evaluation set (2), the lemmas and feature sets seen in train *do* reappear individually in the evaluation set. For example, in e0, both the lemma and the feature set are attested separately in the training data: the lemma is attested in t0 and the feature set is attested in t1. By contrast, in e3, neither the lemma nor the feature set are attested in the training data. Though neither e0 nor e3 are attested *as entire triples* in the training data, one might expect that it would be easier for a model trained on (1) to generate the correct result for e0 than for e3.

(1) Example training set:

t0:	see	seeing	V;V.PTCP;PRS
t1:	sit	sat	V;PST

(2) Example evaluation set:

e0:	see	V;PST
e1:	sit	V;NFIN
e2:	eat	V;PST
e3:	run	V;PRS;3;SG

Indeed, this is the insight behind both Goldman et al. (2022) and Kodner et al. (2022): test pairs with novel lemmas or novel feature sets require a system to generalize along different morphological dimensions, and evaluation measures should control for this overlap so as to better measure models' ability to generalize along these dimensions. To formalize these dimensions, Kodner et al. (2022) define four types of overlap, which we repeat here:

- **both Overlap**: both the lemma and feature set of an evaluation pair are attested in train, though not together in the same triple (e0 in example 2)
- **lemmaOnly Overlap**: only the lemma is attested in training, and the feature set is novel (e1 in example 2)
- **featsOnly Overlap**: only the feature set is attested in training, and the lemma is novel (e2 in example 2)
- **neither Overlap**: neither the lemma nor the feature set is attested in train (e3 in example 2)

We additionally define **featsAttested** to be any evaluation triple for which the feature set is attested in train (i.e., featsAttested = both ∪ featsOnly) and **featsNovel** to be any evaluation triple for which the features are unattested in train (i.e. featsNovel = lemmaOnly ∪ neither). **lemmaAttested** and **lemmaNovel** are defined analogously.

3 Previous Work

While most previous work on morphological inflection has made use of random train-test splits which do not control for overlap, two lines of work have examined different aspects of overlap. Goldman et al. (2022, 864) focused on lemma overlap, arguing that 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" since the model may be able to memorize lemma-specific irregularities.

Specifically, Goldman et al. propose an evaluation strategy in which train-test splits are formed by splitting by *lemma* rather than by triple. As such, for any lemma $\mathcal{L}$ in the data, all triples of the form `(`$\mathcal{L}$`, feature set, inflected form)` are placed into the same set (either train or test); the *entire paradigm* for that lemma will occur in one set. In terms of the overlap types defined above, *every* triple in test will thus be `lemmaNovel`: either `featsOnly` or `neither`.

Goldman et al. re-split the data from the 2020 SIGMORPHON shared task using their proposed method and compare model performance on the original splits to performance on the lemma-based splits. They report an average drop in accuracy of about 30 percentage points from the original SIGMORPHON splits to their lemma-based splits, with the effect being the most significant for low-resourced languages. For example, while Germanic languages had an average drop of 23%, while Niger-Congo languages had an average drop of 39%. The worst drop in performance was approximately 95%. Goldman et al. conclude that models struggle to generalize to unseen lemmas.

In contrast to Goldman et al. (2022) , Kodner et al. (2022) report no loss in performance on unseen lemmas, but report a serious drop in performance on unseen feature sets in their analysis of the 6 systems submitted to SIGMORPHON 2022. Surprisingly, Kodner et al. report that all of the submitted systems actually performed *better* on `neither` overlap items than on `lemmaOnly` overlap items.

Kodner et al. evaluate test items with both unseen lemmas (`lemmaNovel`) and unseen feature sets (`featsNovel`) and compare these to items (**TODO:**) All systems perform better on items with attested fea- ture sets, but the gap in performance varies greatly from UBC's 32 points in the small training condition to OSU's 79 points in the large training condition.

The algorithm began by randomly partitioning a language's feature sets into OVERLAPPABLE and NON-OVERLAPPABLE sets and uniformly sampling the large training set from only those triples that contain feature sets in OVERLAPPABLE. If there were not enough triples with with feature sets in OVERLAPPABLE for a given language, then the OVERLAPPABLE partition was increased incremen- tally until enough training triples could be sampled. If there was insufficient data to create the large training set, then the small training set was sam- pled this way instead. If there was enough data, then the small training set was down-sampled uni- formly from the large training set. The test set was sampled from the remaining items, with half drawn from triples with feature sets in OVERLAPPABLE and half from triples with feature sets in NON-OVERLAPPABLE features. The development set was drawn from the remainder in the same fashion.

In preparation for this year's iteration, we found that the proportion of test items with seen feature sets varied greatly across languages in the 2018 task and may have been a major driver of performance. Indeed, ceiling effect for feature sets but not for lemma overlap

Across the six submitted systems and two baselines, the prediction of inflections with unseen features proved challenging. This was true even for languages for which the forms were in principle predictable, which suggests that further work is needed in designing systems tat capture the various types of generalization needed for the worlds languages. t provides a clear result: the gap between performance on test items attested and novel features does not generally improve even for these languages where it should, if the unfairness of the task were driving decreased performance on fusional languages. This shows that generalization to novel feature sets, that is, to previously unattested inflectional categories, remains a legitimate concern for nearly all the systems.

4 Misc

When they evaluate the top 3 systems on SIGMORPHON's 2020 shared task

All systems see a drop in performance, with average around 30 points and the lowest being 14 points for DeepSpin-02, which fares better for low-resource languages. (**TODO:** do this calculation).

Even high-resourced languages, however, lose about 10 percentage points on average.

Goldman et al. argue that their results clearly show that generalizing inflection to unseen lemmas is far from being solved.

Used all 90 languages in the SIGMORPHON 2020 shared task.

The models used include:

- **Base LSTM:** character-based seq2seq model with a 1-layer bi-directional LSTM Encoder and a 1-layer unidirectional LSTM Decoder
- **chr-trm:** the character-level transformer baseline of Wu et al. (2021)
- **DeepSpin:** the system is composed of 2 bi-directional LSTM encoders with bi-linear gated Attention, 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. (Peters and Martins, 2020)

- **CULing**: another transformer, but with restructuring so that the model learns to inflect from any given cell in the inflection table rather than solely from the lemma. (Liu and Hulden, 2020)

5 Kodner et al

5.1 Kodner Khalifa Payne Liu

Arabic, German, English, Spanish, Swahili, Turkish Uniform, Weighted and OverlapAware

6 Introduction

The SIGMORPHON shared task (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner and Khalifa, 2022; Goldman et al., 2023)

UniMorph (McCarthy et al., 2020; Batsuren et al., 2022)

Goldman et al. (2022)

Kodner et al. (2023c)

Kodner and Khalifa (2022)

Kodner et al. (2023b)

Kodner et al. (2023a)

Morphological inflection is a fundamental task in sub-word NLP, popularized by the recent SIGMORPHON shared tasks; it has both practical and cognitive applications.

Morphological inflection is a popular task in subword NLP with both practical and cogni- tive applications.

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.

For years now, state-of-the- art systems have reported high, but also highly variable, performance across data sets and lan- guages. We investigate the causes of this high performance and high variability; we find sev- eral aspects of data set creation and evaluation which systematically inflate performance and obfuscate differences between languages. To improve generalizability and reliability of re- sults, we propose new data sampling and eval- uation strategies that better reflect likely use- cases. Using these new strategies, we make new observations on the generalization abilities of current inflection systems.

With average ac- curacy above 0.9 over the scores of all lan- guages, 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 in- flection models by employing harder train-test splits that will challenge the generalization ca- pacity of the models. In particular, as op- posed to the naïve split-by-form, we propose a split-by-lemma method to challenge the per- formance on existing benchmarks. Our exper- iments 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 general- izing inflection to unseen lemmas is far from being solved, presenting a simple yet effective means to promote more sophisticated models.

These instructions are for authors submitting papers to *ACL conferences using LaTeX. They are not self-contained. All authors must follow the general instructions for *ACL proceedings,[1] and this document contains additional instructions for the LaTeX style files.

The templates include the LaTeX source of this document (`acl_latex.tex`), the LaTeX style file used to format it (`acl.sty`), an ACL bibliography style (`acl_natbib.bst`), an example bibliography (`custom.bib`), and the bibliography for the ACL Anthology (`anthology.bib`).

[1] http://acl-org.github.io/ACLPUB/formatting.html

7 Sources of Difference

7.1 Training Data Size

7.2 SIGMORPHON Year

While Goldman et al.'s data came from the 2020 SIGMORPHON shared task, Kodner et al.'s came from the 2022 SIGMORPHON shared task. While SIGMORPHON 2020 contained 90 languages and SIGMORPHON 2022 contained 33, only *seven* languages were shared between these years. What's more, there is a notable difference in training size between the 2020 and 2022 versions of the shared task: while Kodner et al. capped their large training sets at 7,000 triples (smaller training sizes were used when fewer triples were available), Goldman et al. made use of the entirety of the available data, resulting in much larger training sizes on average. Indeed, while the mean training size for Goldman et al. was (**TODO:** sigmorphon, not goldman) Mean train size in 2022: 4452.939 (stdev: 3162.846837508738) Mean train size in 2020: 17488.933 (stdev: 26035.04482628495)

Due to this large difference in evaluation languages and training sizes, then, the results of Goldman et al. and Kodner et al. cannot be directly compared: (**TODO:**)

One notable difference between the SIGMORPHON shared tasks in 2020 and 2022 is with regards to training data size.

- Sigmorphon 2022 vs. 2020 training sizes and overlap in languages
- The issues with the Goldman et al.sampling strategy

When re-splitting, we kept the same proportions of the form-split data, we split the inflection tables: 70%, 10%, 20% for the train, dev, and test sets. In terms of examples the proportions may vary as not all tables are of equal size. In prac- tice, the averaged train set size in examples terms was only 3.5% smaller in the lemma-split data, on average. (**TODO:** look at how much of a difference there was here) we split the inflection tables: 70%, 10%, 20% for the train, dev, and test sets. In terms of examples the proportions may vary as not all tables are of equal size. In prac- tice, the averaged train set size in examples terms was only 3.5% smaller in the lemma-split data, on average.

7.3 SIGMORPHON Year

7.4 Relationships Between Lemma & Feature Set Overlap

One possible explanation for the differing results between Goldman et al. (2022) and Kodner et al. (2022) is that lemma overlap and feature set overlap are correlated, and thus that the apparent effect of lemma overlap found by Goldman et al. is *actually* a result of feature set overlap. However, under Goldman et al.'s lemma-based splitting, feature set overlap between train and test *should* always be at ceiling: since entire paradigms are placed in either train or test, all feature sets in the test set should already be attested in train. For example, if we split the two verbal lemmas `walk` and `run` into train (3) and test (4), respectively, then each verb's entire paradigm is placed in its corresponding set. It is straightforward to see that the feature set overlap between (3) and (4) is 100%.

(3) An example lemma-based train set in the style of Goldman et al.:

walk	V;PST	walked
walk	V;NFIN	walk
walk	V;PRS;3;SG	walks

(4) An example lemma-based evaluation set in the style of Goldman et al.:

run	V;PST
run	V;NFIN
run	V;PRS;3;SG

There are two possible caveats to this point. Firstly, 100% feature set overlap will only be possible if *every* part of speech that appears in test also appears in train. However, this assumption is entirely tenable given a moderate data sample, and indeed all parts of speech appearing in the Goldman et al. test data also appear in the training data. Even with 100% part-of-speech overlap, however, there is the possibility of gaps: if `stride` is the only verbal lemma in our English training data, for example, then the past participle will not be attested in train. Even in cases of gaps, however, it is reasonable to expect feature overlap to remain near-ceiling given moderate data given the Goldman et al. lemma-based splitting strategy.

Indeed, when we examine the feature set overlap in the data used by Goldman et al. (Figure 4), we find that the overlap for most languages is 100%, and there is little relationship between feature set overlap and training size. However, not all languages achieve 100% feature set overlap: the most

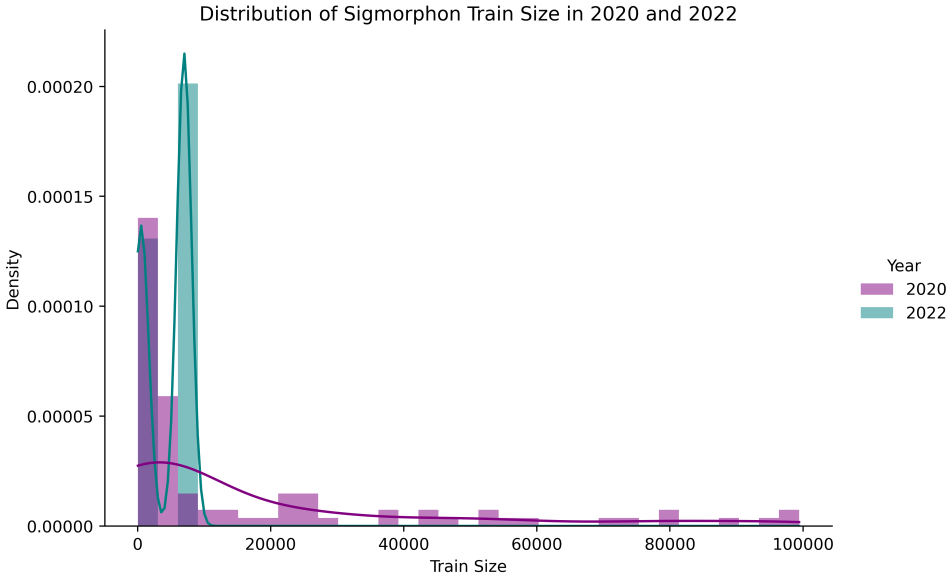

Figure 1: (**TODO:** write the caption)

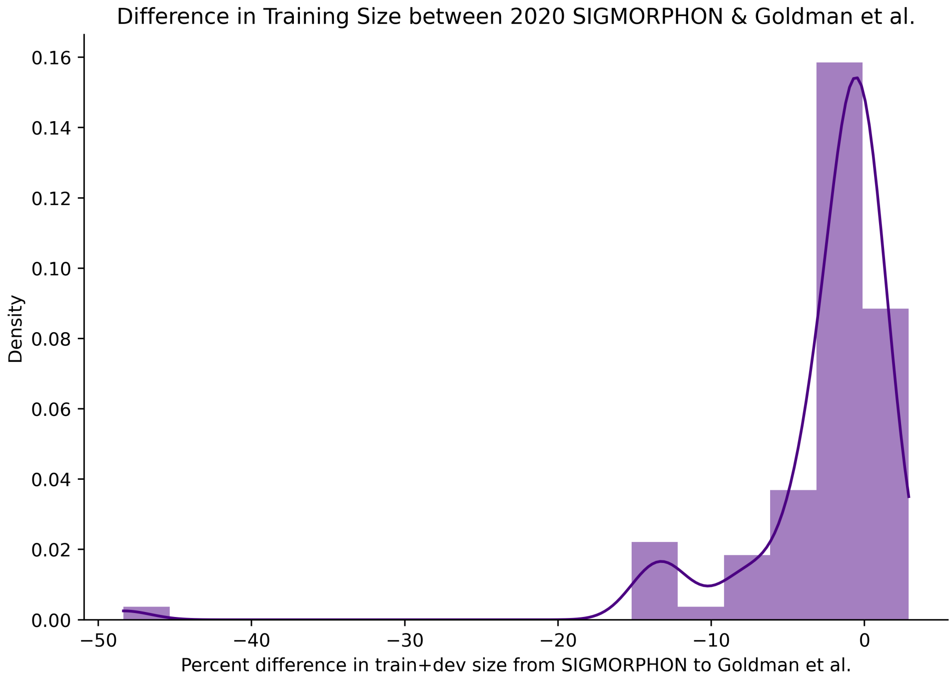

Figure 2: (**TODO:** write the caption)

dramatic outlier, Ludic, for example, has a feature overlap of just 0.303, meaning that less than a third of the feature sets appearing in the test set are attested in the train set. Preliminary examination of the Ludic data finds that this low overlap is due to just *three* of the 26 total lemmas appearing in the test data. One of these lemmas, `astuda` "go", appears with a whopping 131 feature sets.

We can ask whether it's reasonable to expect a model to generate all 131 of these feature sets by comparing the size of this test paradigm to paradigms of the same part-of-speech in train. In this case, `astuda` is a verb, and the average size of a verbal paradigm in the Ludic training data is just 1.562 (stdev: 0.864); even the largest verbal paradigm contains just 16 feature sets. A difference of this magnitude cannot be explained simply as the result of gaps or other naturalistic phenomena. Rather, it seems that cases where Goldman et al.'s sampling strategy does not yield 100% feature set

Figure 3: The proportion of feature sets appearing in test that have been seen in train, as a function of training data size. The relationship is not significant (Pearson's $r = 0.074$ ($p = p = 0.489$), Spearman's $\rho = -0.183$ ($p = 0.084$), Kendall's $\tau_B = -0.155$ ($p = 0.067$)).

overlap emerge as a result of data issues. Indeed, the corpus from which the Ludic data was drawn (Zaytseva et al., 2017) contains a large number of incomplete paradigms. Given these conditions, it is not surprising that the lemma-based splitting algorithm used by Goldman et al. will lead to low feature overlap. We investigate such issues with the data below.

7.5 (Lack of) Data Quality

Though Ludic is by far the most glaring example, the pattern of large paradigms in test causing low feature overlap extends to other languages for which overlap is less than 100%. To quantify this, we measure the percent increase in paradigm size for **problematic lemmas** — those those appearing with at least one unattested feature set in test — compared to the average paradigm size of words with the same part of speech in the training data. In other words, we calculate:

$$\textsc{Diff} = 100 * \frac{\texttt{mean}(\textsc{ProbLem}) - \texttt{mean}(\textsc{Train})}{\texttt{mean}(\textsc{Train})} \quad (1)$$

Where PROBLEM indicates the paradigm size of the problematic lemmas and TRAIN indicates the paradigm size of the same POS in train.

Indeed, we find that the mean percent difference in paradigm size between the problematic lemmas and the corresponding train lemmas is a whopping 437.683% (stdev: 871.027%), with the large standard deviation resulting from a number of large positive outliers. To give a point of comparison, we also measure the percent increase from the average training paradigm size to the *maximum* test paradigm size across all POS for all languages for which there is 100% overlap. Here, the mean percent increase is only 11.511% (stdev: 34.054%); indeed, an unpaired T-test finds a significant difference between these two measures of increase ($t = 5.375$, $p < 10^{-6}$) The difference in these distributions is visualized in Figure 4.

7.6 Model Types

7.7 Splitting Strategy

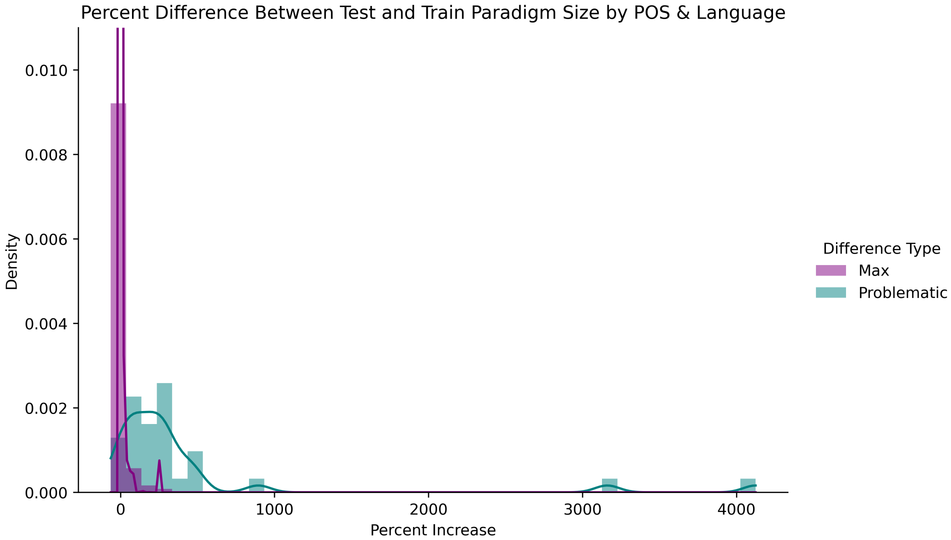

Figure 4: The proportion of feature sets appearing in test that have been seen in train, as a function of training data size. The difference in distributions is significant (unpaired $t = 5.375,\ p < 10^{-6}$).

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. In *Interspeech*.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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, 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.

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.

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.

Jordan Kodner and Salam Khalifa. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 157–175, Seattle, Washington. 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.

Jordan Kodner, Salam Khalifa, Sarah RB Payne, and Zoey Liu. 2023a. Re-evaluating the evaluation of neural morphological inflection models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Jordan Kodner, Salam Khalifa, and Sarah Ruth Brogden Payne. 2023b. Exploring linguistic probes for morphological generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8933–8941, Singapore. Association for Computational Linguistics.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023c. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw Me a Flower: Processing and Grounding Abstraction in Natural Language.

Transactions of the Association for Computational Linguistics, 10:1341–1356.

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.

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.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA. PMLR.

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 Twelfth 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. 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.

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.

Nina Zaytseva, Andrew Krizhanovsky, Natalia Krizhanovsky, Natalia Pellinen, and Aleksndra Rodionova. 2017. Open corpus of veps and karelian languages (vepkar): preliminary data collection and dictionaries. In *Corpus Linguistics*, volume 2017, pages 172–177.