Morphological Inflection: A Reality Check

Jordan Kodner[1] **Sarah Payne**[1]* **Salam Khalifa**[1]* and **Zoey Liu**[2]
[1]Stony Brook University, Dept. of Linguistics & Institute for Advanced Computational Science
[2]University of Florida, Dept. of Linguistics
first.last@stonybrook.edu and liu.ying@ufl.edu

Abstract

Morphological inflection is a popular task in sub-word NLP with both practical and cognitive applications. For years now, state-of-the-art systems have reported high, but also highly variable, performance across data sets and languages. We investigate the causes of this high performance and high variability; we find several aspects of data set creation and evaluation which systematically inflate performance and obfuscate differences between languages. To improve generalizability and reliability of results, we propose new data sampling and evaluation strategies that better reflect likely use-cases. Using these new strategies, we make new observations on the generalization abilities of current inflection systems.

1 Introduction

Morphological inflection is a task with wide-reaching applications in NLP, linguistics, and cognitive science. As the reverse of lemmatization, it is a critical part of natural language generation, particularly for languages with elaborate morphological systems (Bender, 2009; Oflazer and Saraçlar, 2018). Since morphological inflection is a particular type of well-defined regular string-to-string mapping problem (Roark and Sproat, 2007; Chandlee, 2017), it is also useful for testing the properties of different neural network architectures. Within cognitive science and linguistics, computational models of inflection have a long history in arbitrating between competing theories of morphological representation and acquisition (surveyed in Pinker and Ullman, 2002; Seidenberg and Plaut, 2014), and inflection is often a focus of computational typology (Bjerva and Augenstein, 2018; Elsner et al., 2019).

However, despite the task's popularity, standard evaluation practices have significant weaknesses. We discuss three aspects of these practices which hamper investigators' ability to derive informative conclusions. **(1)** Uniform sampling, which creates unnatural train-test splits, **(2)** Evaluation of single data splits, which yields unstable model rankings, and **(3)** uncontrolled overlaps between train and test data components, which obscure diagnostic information about systems' ability to perform morphological generalizations.

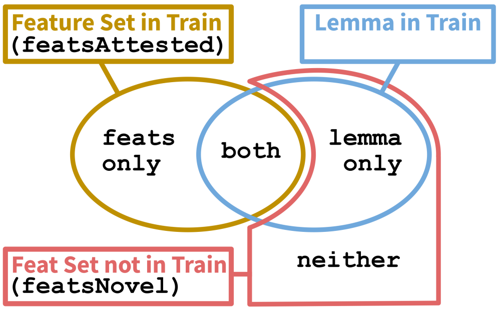

Figure 1: The four logically possible train-eval overlap types if evaluation data consists of (`lemma`, `feature set`) pairs: `both`, `featsOnly`, `lemmaOnly`, `neither`, as well as `featsAttested`= `both` ∪ `featsOnly` and `featsNovel`= `lemmaOnly` ∪ `neither`.

1.1 Practice 1: Uniform Sampling

Training and evaluation sets have been (with some exceptions) sampled uniformly by type from a corpus such as those available in the UniMorph Database (Kirov et al., 2018; McCarthy et al., 2020; Batsuren et al., 2022). While practical to implement for corpora that lack frequency information, uniform sampling is also unrealistic because morphological forms exhibit a highly skewed Zipfian distribution in any large text (Lignos and Yang, 2018). Thus, uniform sampling creates an unnatural bias towards low-frequency types. Since high frequency is correlated with irregularity across many but not all languages (Bybee, 1991; Fratini et al., 2014; Wu et al., 2019), this creates a bias towards more regular and reliable training items.

We provide two alternatives for producing realistic or challenging data sets: **(1)** a frequency-weighted sampling strategy to achieve a more real-

*Denotes equal contribution

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers*, pages 6082–6101
July 9-14, 2023 ©2023 Association for Computational Linguistics

istic distribution of out-of-vocabulary (OOV) lemmas and inflectional categories and better match practical use-cases or input during child language acquisition, and **(2)** a sampling strategy that explicitly balances OOV lemmas and inflectional categories in order to directly evaluate models' generalization ability along these dimensions.

1.2 Practice 2: Single Data Splits

The current practice in inflection evaluation, employed, for example, in the SIGMORPHON, CoNLL-SIGMORPHON and SIGMORPHON-UniMorph shared tasks 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), examines different models with one particular data set that is considered representative of the language or the inflection task at hand. This data set, and therefore all evaluation, usually consists of one pre-defined train-(dev-)test split.

However, this method is problematic because it implicitly assumes that the results from a single split are informative and generalizable. In reality, this assumption is untenable, particularly when facing severe data limitation (Liu and Prud'hommeaux, 2022), as is the case for the majority of languages in the world (cf. Blasi et al., 2022): In UniMorph 4, for example, data set size varies significantly across languages, with the smallest, Manx (Celtic, IE), containing only one lemma with 14 inflected forms, and the largest, Czech (Slavic, IE) containing approximately 800,000 lemmas with 50.3 million forms. If the performance on a single split is not necessarily representative, then the original model ranking derived from the one particular data split might also not generalize well.

The concerns outlined above were demonstrated in Liu and Prud'hommeaux (2022), which investigated model generalizability in low-resource morphological segmentation. Using data from 11 languages, they provided evidence that: **(1)** there are major differences in the numerical performance and rankings of each evaluated model type when using different splits from the same data set, and **(2)** even within a single split, large performance variability can arise for each model type when it is trained using different random seeds. These findings illustrate that common methods of model evaluation can lead to largely coincidental conclusions. We extend this approach to morphological inflection by applying multiple data splits, and evaluating variability between splits.

1.3 Practice 3: Uncontrolled Overlaps

The typical morphological inflection task paradigm presents (`lemma`, `inflected form`, `feature set`) triples during training and asks a system to predict inflected forms from (`lemma`, `feature set`) pairs during evaluation. Note that since the lemma and feature set can be combined independently, it is possible for either lemmas or feature sets that appeared during training to reappear during test without any individual triple violating train-on-test. Test pairs with OOV lemmas or feature sets require a system to generalize along different morphological dimensions. Performance is likely related to the relative rates of OOV lemmas and feature sets in the evaluation split, yet existing sampling strategies generally leave these variables uncontrolled.

We observe that uncontrolled OOV rates vary dramatically between different sampled data splits, and that uncontrolled sampling biases test sets towards "easier" items with in-vocabulary lemmas and feature sets. To remedy this, we argue that performance should be reported independently for items with each lemma/feature set overlap type regardless of sampling strategy. Furthermore, if a project's research goal is to evaluate the generalization ability of a model, lemma/feature set overlap-aware sampling should be used to ensure that a sufficient number of test items of each overlap type are present.

2 Defining Overlap

Morphological inflection requires generalization over two primary dimensions: to new lemmas ("*If I have witnessed the 2pl imperfective subjunctive with other verbs, how do I apply that to new verb X?*") and to new inflectional categories ("*If I have seen X inflected in several other categories, how do I create the 2pl imperfect subjunctive of X?*"). Because of the sparsity of morphological inflections in language use (Chan, 2008), both types of generalization are necessary during language acquisition as well as deployment of computational models.

As with many linguistic phenomena, the attestation of inflected forms follows an extremely sparse and skewed long-tailed distribution, as do attested lemmas ranked by the proportions of their potential paradigms that are actually attested (*paradigm saturation*; PS), and inflectional categories ranked by the number of lemmas with which they occur (Chan, 2008). For example, the median PS for Spanish verbs in millions of tokens of child-directed speech is equivalent to *two* of its three dozen possible forms, and the 2nd person plural imperfect subjunctive only occurs with two lemmas (cf. Lignos and Yang, 2018; Kodner, 2022).

Given the importance of both types of generalization, it is necessary to evaluate both to assess the abilities of a morphological learning model. In the evaluation made popular by the SIGMORPHON shared tasks, models are asked to predict inflected forms given (`lemma`, `feature set`) pairs, where feature sets can be seen as corresponding to inflectional categories or paradigm cells. Generalization across lemmas is required when an evaluation pair contains a lemma that was out-of-vocabulary (OOV) in training, and generalization across categories is required when an evaluation pair contains a feature set that was OOV. In all, there are four logically possible licit types of evaluation pairs distinguished by their lemma and feature overlap with training. These are expressed visually in Figure 1 along with two types which are unions of the other types:

`both` Overlap: Both the lemma and feature set of an evaluation pair are attested in the training set (but not together in the same triple).

`lemmaOnly` Overlap: An eval pair's lemma is attested in training, but its feature set is novel.

`featsOnly` Overlap: An eval pair's feature set is attested in training, but its lemma is novel.

`neither` Overlap: An evaluation pair is entirely unattested in training. Both its lemma and features are novel.

`featsAttested`: An eval pair's feature set is attested in training (`both` ∪ `featsOnly`)

`featsNovel`: An eval pair's feature set is novel (`lemmaOnly` ∪ `neither`)

For a concrete illustration, consider the training and evaluation sets provided in (1)-(2). Each evaluation pair exhibits a different kind of overlap.

(1) **Example Training Set**

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

(2) **Example Evaluation Set**

```
e0: see  V;PST       <-- both
e1: sit  V;NFIN      <-- lemmaOnly
e2: eat  V;PST       <-- featsOnly
e3: run  V;PRS;3;SG  <-- neither

featsAttested = {e0, e2}
featsNovel    = {e1, e3}
```

Computational work in morphological inflection has generally ignored these dimensions of evaluation. In the shared task, the four overlap types were uncontrolled before 2021, which contains one partial evaluation on `featsOnly` ∪ `neither` items. But, recognition of the value of these overlap types has grown recently. Goldman et al. (2022) showed that four models consistently struggle to generalize across lemmas, concluding that test sets should avoid lemma overlap altogether. However, this proposal removes the option to contrast performance on seen and unseen lemmas. Furthermore, they did not control for or evaluate feature overlap, so `both` vs. `lemmaOnly` and `featsOnly` vs. `neither` also cannot be distinguished. (3) summarizes their partition scheme, which distinguishes two overlap types. We call these `lemmaAttested` (= `both` ∪ `lemmaOnly`) and `lemmaNovel` (= `featsOnly` ∪ `neither`).

(3) **Goldman et al. (2022) Partition Types**

```
e0: sit  V;PST       <-- lemmaAttested
e1: see  V;NFIN      <-- lemmaAttested
e2: eat  V;PST       <-- lemmaNovel
e3: run  V;PRS;3;SG  <-- lemmaNovel
```

The 2022 SIGMORPHON-UniMorph shared task was the first to report results on all four overlap types (`both`, `featsOnly`, `lemmaOnly`, `neither`). Every system submitted to the shared task achieved much better performance on in-vocabulary feature sets (`both` and `featsOnly`) than OOV feature sets (`lemmaOnly` or `neither`). This discrepancy even held for languages for which a model should be able to generalize: highly regular agglutinative morphology for which this type of generalization is often transparent. On the other hand, lemma attestation produced a much smaller discrepancy. Following these observations, we focus our investigation on the four logical overlap types with extra emphasis on the `featsAttested` vs. `featsNovel` dichotomy. We address agglutinative languages specifically in Section 5.3

3 Data Sources and Preparation

We follow prior literature in providing training and evaluation data in UniMorph's format. Data sets were sampled from UniMorph 4 (Batsuren et al., 2022) and 3 (McCarthy et al., 2020)[1] aug-

[1] In some cases, UniMorph 4 was found to lack high-frequency items present in UniMorph 3. For example, English verbs *happen* and *run* are present in 3 and absent in 4. For languages where we determined this to be an issue, we sampled from deduplicated UniMorph 3+4 with tags normalized to 4.

mented with frequencies from running text corpora. When possible, frequencies were drawn from child-directed speech (CDS) corpora from the CHILDES database (MacWhinney, 2000), since one possible downstream application of the morphological inflection task is contribution to the computational cognitive science of language acquisition. CHILDES lemma and morphological annotations were converted into UniMorph format and intersected with UniMorph to create frequency lists.[2]

3.1 Languages

Languages were prioritized for typological diversity and accessibility of text corpora. Quantitative summaries of our frequency+UniMorph data sets are provided in Appendix B.

Arabic (Semitic, AA): Modern Standard Arabic frequencies were drawn from the diacritized and morphologically annotated Penn Arabic Treebank (PATB; Maamouri et al., 2004) and intersected with UniMorph 4 `ara` ∪ `ara_new`. Diacritized text is a requirement because orthographic forms drawn from undiacritized text are massively morphologically ambiguous. The text in the CHILDES Arabic corpora is undiacritized and thus unusable.

German (Germanic, IE): German was drawn from the Leo Corpus (Behrens, 2006), the only morphologically annotated German corpus in CHILDES, and intersected with UniMorph 3+4. Only nouns and verbs were extracted because annotation for adjectives is inconsistent.

English (Germanic, IE): English was included because it is heavily studied despite its relatively sparse morphology. Data was extracted from all morphologically annotated CHILDES English-NA corpora and intersected with UniMorph 3+4.[3] Only nouns and verbs were extracted due to inconsistent adjective annotation in both data sources.

Spanish (Romance, IE): Spanish exhibits a variety of fusional and agglutinative patterns. Data was extracted from all morphologically annotated Spanish CHILDES corpora intersected with Spanish UniMorph 3+4. Non-Spanish vocabulary was removed by intersecting with UniMorph. Only nouns and verbs were extracted.

Swahili (Bantu, Niger-Congo): Swahili morphology is highly regular and agglutinative with very large paradigms. Frequencies were drawn from Swahili Wikipedia dump 20221201 accessed through Huggingface (Wikimedia, 2022) and intersected with UniMorph 4 `swc` ∪ `swc.sm`. In cases where mapping inflected forms to UniMorph creates ambiguity due to syncretism, frequency was divided evenly across each triple sharing the inflected form. This ensured that the frequencies of inflected forms remain consistent with Wikipedia. Intersecting with UniMorph removed the large amount of non-Swahili vocabulary in the Wikipedia text.

Turkish (Turkic): Turkish is also highly regular and agglutinative with very large paradigms. Frequencies were drawn from Turkish Wikipedia dump 20221201 accessed through Huggingface, intersected with UniMorph 4, and processed identically to Swahili.

3.2 Data Splits

We employed three distinct sampling strategies to generate small (400 items) and large (1600) training, small (100) and large (400) fine-tuning, development (500), and test (1000) sets for each language.[4] Small training and fine-tuning are subsets of large training and fine-tuning. Each splitting strategy was applied five times with unique random seeds to produce distinct data sets.

UNIFORM: Raw UniMorph 3+4 corpora were partitioned uniformly at random. This approach is most similar to that employed by SIGMORPHON shared tasks, except for 2017 and 2022.

WEIGHTED: Identical to UNIFORM except splits were partitioned at random weighted by frequency. Small training+fine-tuning were sampled first, then additional items were sampled to create large training+fine-tuning. Training and fine-tuning sets were then split uniformly at random. Dev+test was next sampled by weight and then were separated uniformly. This frequency-weighted sampling is reminiscent of the 2017 shared task: it strongly biases the small training set towards high-frequency items and dev+test towards low-frequency items. Since most UniMorph forms do not occur in our corpora due to morphological sparsity, most triples had zero weight and were never sampled.

OVERLAPAWARE: Similar to the 2022 SIGMORPHON shared task. It enforces a maximum proportion of `featsAttested` pairs in the test set relative to train+fine-tuning: as close to 50% as pos-

[2] All data and code is available at `https://github.com/jkodner05/ACL2023_RealityCheck`.

[3] A full list of utilized English and Spanish CHILDES corpora is provided in Appendix A.

[4] Swahili large train and large fine-tune contain 800 and 200 items respectively due to the limited size of UniMorph.

sible without exceeding it. This ensures that there is ample representation of each overlap type in test. It is adversarial, since `featsNovel` pairs are expected to be more challenging than `featsAttested` pairs. This process also tends to increase the proportion of `lemmaOnly` items in the test set. Only items with non-zero frequency were sampled.

UNIFORM produces a heavy bias towards lower frequency words. For all languages and splits, the median frequency of sampled items is actually zero: that is, the majority of sampled items were not attested in our corpora. This is a consequence of the extreme sparsity of morphological forms discussed in Section 2. As a consequence, overlap between splits from different seeds is orders of magnitude lower for UNIFORM than the other strategies. WEIGHTED achieves the expected high-frequency bias in training sets relative to test sets.

Table 1 provides average means and standard deviations for the proportion of `featsAttested` and `featsNovel` in test sets relative to small and large train. OVERLAPAWARE consistently achieves a roughly 50-50 split with low variability across languages and seeds. The other strategies bias test sets heavily towards `featsAttested` with high variance across languages and seeds.[5]

Test vs S Train	μ %`featsAttested`	σ
UNIFORM	80.33%	19.50%
WEIGHTED	90.44	11.13
OVERLAPAWARE	48.81	0.98
Test vs L Train	μ %`featsAttested`	σ
UNIFORM	96.17%	5.55%
WEIGHTED	95.36	7.28
OVERLAPAWARE	49.92	0.17

Table 1: Language-by-language average mean percentage and standard deviation for proportion of `featsAttested` attested in test relative to small and large training. %`featsNovel`= 100 - %`featsAttested`.

4 Experimental Setup

One non-neural and three neural systems were evaluated. These were chosen based on their availability and performance in recent shared tasks:

CHR-TRM (Wu et al., 2021) is a character-level transformer that was used as a baseline in 2021 and 2022. We used the hyper-parameters suggested by the original authors for small training conditions.

CLUZH-GR and **CLUZH-B4** (Wehrli et al., 2022) is a character-level transducer which substantially outperformed CHR-TRM in the 2022 shared task. The results submitted for the shared task are from an elaborate ensemble model optimized for each language. For this work, we evaluate two published variants with consistent hyper-parameters across languages, CLUZH-GR with greedy decoding and CLUZH-B4 with beam decoding, beam size = 4.

NONNEUR (Cotterell et al., 2017) has been used as a baseline in SIGMORPHON shared tasks since 2017. It heuristically extracts transformations between lemmas and inflected forms and applies a majority classifier conditioned on the associated feature sets. NONNEUR was trained on combined training and fine-tuning sets so that each architecture was exposed to the same amount of data.

5 Results

This section presents our analyses of the results. All evaluations report exact match accuracy. *Overall accuracy* refers to average accuracy on an entire evaluation set. *Average overall accuracy* refers to the mean of overall accuracy over all five seeds. See Appendix C for full breakdowns by language and architecture.

5.1 Effect of Training Size

We begin by comparing average overall accuracy for each training size. All reported analyses focus on test, but there were no observable qualitative differences in behavior between dev and test. We summarize the results in Table 2, broken down by overlap partition and sampling strategy. The large training size consistently leads to higher accuracies than small training. Across languages, the average accuracy score difference between the two training sizes is 9.52%. Taking Arabic as an illustrative example, the score difference between the two training sizes ranges from 1.74% to 19.32% depending on model type and splitting strategy, with an average of 12.05%.

Test vs S Train	`featsAttested`	`featsNovel`
UNIFORM	70.47%	33.57%
WEIGHTED	79.25	22.77
OVERLAPAWARE	79.60	31.13
Test vs L Train	`featsAttested`	`featsNovel`
UNIFORM	80.00%	55.57%
WEIGHTED	85.94	23.74
OVERLAPAWARE	86.22	35.51

Table 2: Overall accuracy across languages by overlap type in test.

[5] See Appendix B for breakdowns by language, training size, and overlap partitions.

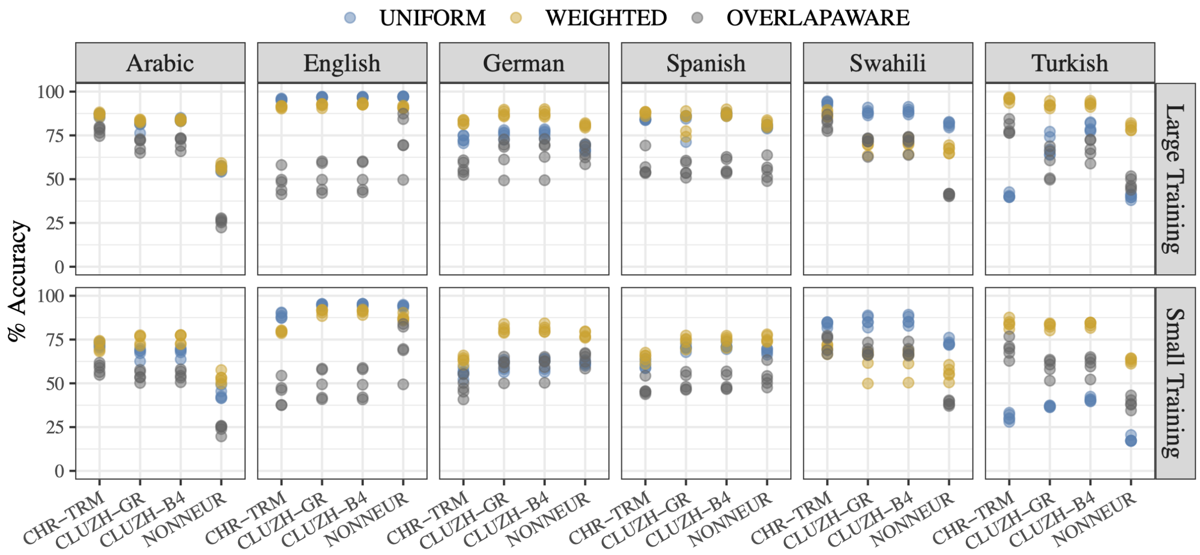

Figure 2: Overall accuracy for each language/seed by training size, sampling strategy, and model type.

5.2 Effect of Sampling Strategy

We next turn to measuring the effect of sampling strategy on overall accuracy. Figure 2 provides a visualization of accuracy by sampling strategy across seeds broken down by training size, language, model type. Using Arabic as an illustration, for large training, WEIGHTED sampling leads to the highest average overall accuracy across model types (77.76%), while OVERLAPAWARE sampling yields the lowest (61.06%); comparing the results from the three sampling strategies given each of the four model types, WEIGHTED consistently results in the highest accuracy for all model types except for CLUZH-B4, where UNIFORM sampling (83.84%) leads to a performance slightly better than that of WEIGHTED (83.82%). We make similar observations for small training: WEIGHTED and OVERLAPAWARE result in the highest and the lowest average overall accuracy, respectively, across model types for Arabic (68.82% vs. 47.81%). WEIGHTED sampling leads to a higher accuracy compared to the other two strategies for every model type other than CHR-TRM, where the result from UNIFORM sampling (71.90%) is again slightly higher than that of WEIGHTED (71.60%).

When considering other languages, we also find some variation. WEIGHTED sampling also yields the highest average accuracy scores across model types for Arabic, German, Spanish, and Turkish for both training sizes, except for Spanish under the large training condition with CLUZH-GR, where UNIFORM leads. In contrast, UNIFORM consistently results in the highest average accuracy on English and Swahili for both training sizes.

Across languages, the average accuracy from WEIGHTED is the highest for both large (83.75%) and small (74.22%) training sizes, followed by UNIFORM (large: 79.20%, small: 66.16%). OVERLAPAWARE always yields the lowest accuracy. These observations align with our expectations about the adversarial nature of OVERLAPAWARE, where challenging `featsNovel` (Table 2) constitutes a much larger proportion test set (Table 1).

5.3 Effect of Overlap

We now provide an analysis of accuracy scores by overlap partition. Figure 3 provides a visualization of accuracy by partition across seeds broken down by training size, language, model type. Using Arabic again as an illustration, the average accuracy across model types and sampling strategies for large training is much higher for `featsAttested` (77.70%) than for `featsNovel` (41.92%), somewhat higher accuracy is achieved for both (79.53%) than for `featsOnly` (77.28%), and higher accuracy is achieved for `lemmaOnly` (49.12%) than for `neither` (41.92%). This ranking is consistent across model types, sampling strategies, and training sizes. Scores from these two overlap partitions are also higher than those from `lemmaOnly` and `neither`.

These patterns hold across languages. Specifically, we observe two general tendencies. First, the accuracy averaged across model types and sampling strategies is always substantially higher for `featsAttested` than it is for `featsNovel`; the average accuracy difference between the two is

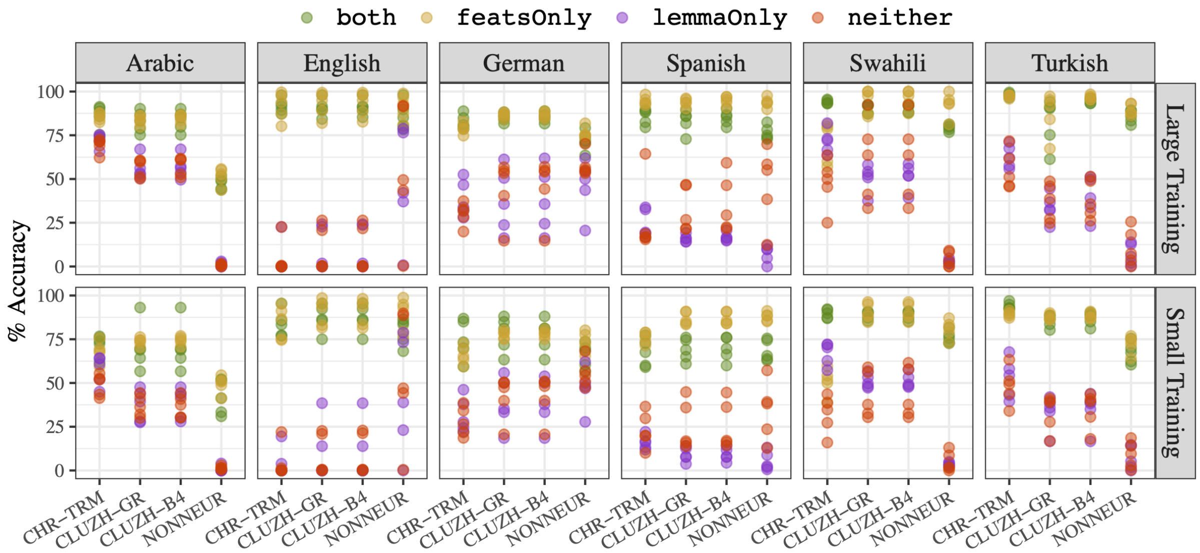

Figure 3: Accuracy on OVERLAPAWARE splits for each partition/seed by training size, language, and model type. `featsAttested` = `both` (green) and `featsOnly` (gold). `featsNovel` = `lemmaOnly` (violet) and `neither` (red).

49.75% for the large training, and 48.02% for small training. This is reflected in a full breakdown by overlap type: higher accuracy is consistently achieved for `both` and `featsOnly`, than for `neither` and `lemmaOnly`. This large asymmetry corresponds to our expectations regarding the effect of feature overlap on performance.

We provide three sub-analyses to further investigate this asymmetry and compare it with the lemma-based division advocated for by (Goldman et al., 2022). First, we compute the average accuracy difference between `lemmaAttested` (`both` ∪ `lemmaOnly`) and `lemmaNovel` (`featsOnly` ∪ `neither`). The score difference between `lemmaAttested` and `lemmaNovel` is less than 2% averaged across languages for both training sizes, which is an order of magnitude smaller than the difference between `featsAttested` and `featsNovel`. This trend is consistent with the results of the 2022 SIGMORPHON shared task, which also found a much greater impact of feature set attestation than lemma attestation.

Second, we measure the correlation between the proportion of `featsAttested` items (number `featsAttested` items divided by the size of the dev or test set), and overall accuracy (average accuracy on an entire dev or test set), as well as between the proportion of `lemmaAttested` and overall accuracy. We used Spearman's ρ, which assesses if there is any monotonic (not necessarily linear) relationship between the two variables.[6] If ρ between an overlap type and overall accuracy is high, it would suggest that the distribution of overlaps is an important driver of performance. `lemmaAttested` shows little correlation (small: 0.01, large: -0.10). However, we find substantial positive correlations for `featsAttested` (small: 0.69, large: 0.68).

Third, we compute the correlation between the accuracy score of individual partitions and the overall accuracy score on UNIFORM and WEIGHTED vs. on OVERLAPAWARE. This demonstrates to what extent evaluation results based on each overlap partition resemble those captured by the overall accuracy and how it differs when overlaps are controlled during sampling. If the correlation is small, it suggests that the performance on a particular overlap partition is largely independent of the others and should be evaluated independently.

When overlaps are not explicitly controlled, correlations are particularly strong for `featsAttested` because this partition makes up a large majority of the test set (Table 3). These partitions are also the ones that tend to show the highest performance, which is then reflected in the overall accuracy. However, for OVERLAPAWARE, correlations are higher between overall accuracy and the challenging partitions: `featsNovel`, `lemmaOnly`, and `neither`. They are also higher not only for `featsNovel`, but also `lemmaAttested`, and `lemmaNovel` even though these overlaps were not explicitly controlled. This demonstrates that OVERLAPAWARE sampling better balances individual partitions in its overall accuracy scores and can be expected to produce

[6] ρ falls in the range [-1,1], where -1 is a perfect negative correlation and 1 is a perfect positive correlation.

a more challenging evaluation. However, all partitions should be evaluated regardless of sampling strategy.

Overlap Partition	Uncontrolled ρ	Controlled ρ
`featsAttested`	0.97	0.45
`featsNovel`	0.16	0.93
`lemmaAttested`	0.84	0.88
`lemmaNovel`	0.78	0.82
`both`	0.89	0.49
`featsOnly`	0.73	0.21
`lemmaOnly`	0.24	0.89
`neither`	-0.04	0.85

Table 3: Correlation between average accuracy for each overlap partition and average overall accuracy across the six languages. Uncontrolled = WEIGHTED and UNIFORM. Controlled = OVERLAPAWARE.

Up to this point, we have considered all languages in the analysis. However, whether or not it is reasonable to expect a system to achieve high accuracy on `featsNovel` items varies typologically. For languages with highly regular and agglutinative morphologies, such as Swahili and Turkish, each feature in a feature set roughly corresponds to a single affix in a certain order with a limited number of allomorphs. For these languages, this dimension of generalization should often be straightforward. For languages with mixed systems, like Spanish and Arabic, and languages with fusional systems like English, the individual members of a feature set often do not have direct bearing on the inflected form. For these languages, generalization to a novel feature set is sometimes impossible when it cannot be inferred from its component features. The same problem applies to lemmas with erratic stem changes or suppletion.

Thus, if a model type can generalize to novel feature sets, one would expect that the accuracy gap between `featsAttested` and `featsNovel` would be lower for Swahili and Turkish than for the other languages. However, the gaps for these are actually larger than for German or Arabic. One would also expect the correlation between the proportion of `featsAttested` in the data and overall accuracy to be lower for Swahili and Turkish, however this is not borne out either. These findings, provided in Table 4, reveal that current leading inflection models do not necessarily generalize well to novel feature sets even in precisely the cases where they should be able to.

5.4 Model Ranking

In this section, we analyze how performance varies across the four model types. We first compare model performance based on the average overall accuracy. Averaged across the six languages, CLUZH-B4 ranks among the highest, while NONNEUR consistently achieves the lowest performance.

Train Size	Language Strategy	Avg. Score Difference	`featsAttested` $\sim$Accuracy ρ
Small	Arabic	33.00%	0.57
	Swahili	40.04	0.63
	German	40.35	0.23
	Turkish	41.96	0.83
	Spanish	52.60	0.75
	English	74.10	0.66
Large	Arabic	35.79%	0.44
	German	36.19	0.73
	Swahili	39.26	0.64
	Turkish	52.14	0.59
	Spanish	61.01	0.64
	English	80.17	0.82

Table 4: Avg. score difference between `featsAttested` and `featsNovel` and correlation between proportion `featsAttested` and overall accuracy by language/training size, ranked by score difference.

large: CLUZH-B4 (78.32%) > CHR-TRM (78.07%) > CLUZH-GR (76.17%) > NONNEUR (65.82%)
small: CLUZH-B4 (68.58%) > CLUZH-GR (67.97%) > CHR-TRM (64.76%) > NONNEUR (58.97%)

Model rankings for individual languages are much more variable, especially for large training. There is not a single model ranking that holds for every language. While CLUZH-B4 yields the best performance for three languages (German, Spanish, and Turkish), CHR-TRM outperforms other model types for Arabic and Swahili, and NONNEUR leads to the highest accuracy for English. There is less variation in model rankings for small training; the same model ranking was observed for German, English, and Spanish (NONNEUR > CLUZH-B4 > CLUZH-GR > CHR-TRM). Notably, for each individual language, the model rankings were always inconsistent between the two training sizes.

Several trends emerge in model rankings by overlap partition. First, the model rankings based on the overall accuracy do not hold for the overlap partitions except for Arabic and Swahili large training. Second, within each overlap partition, model rankings are more stable across languages for small train than large. Third, on average, CLUZH-B4 outperforms the other model types on partitions with feature overlap whereas CHR-TRM leads on partitions without feature overlap. These tendencies resonate with our proposal in Section 2: future models of morphological inflection should be evaluated based on alternative metrics in addition to overall accuracy. They also reveal difference generalization strengths across models.

When comparing performance by sampling strategy, we found lower variability for each language. For example, with UNIFORM large training, two model rankings turn out to be the most frequent, each observed in two languages. Among the models, CLUZH-B4 and CHR-TRM achieve the best performance. For small training, one model ranking holds for three out of the six languages (CLUZH-B4 > CLUZH-GR > CHR-TRM > NONNEUR). Considering both training sizes, there are no noticeable differences in terms of the most frequent model ranking across the three sampling strategies. For UNIFORM and WEIGHTED, the neural systems are always ranked among the highest for both training sizes; yet for OVERLAPAWARE with small training, NONNEUR achieves the highest performance for German, English, and Spanish.

5.5 Variability across Random Seeds

Analysis so far relies on accuracy scores averaged across random seeds. The final component of our analysis investigates how much variation arises due to random data sampling. Given the five random seeds for each combination of language, sampling strategy, overlap partition, and model type, we calculated the *score range*, which is the difference between the lowest and the highest overall accuracy, as well as the standard deviation of the accuracy scores across the seeds, which we refer to as *random seed variability*.

We first considered the score range for overall accuracy for each language. For large training, the mean score range spans from 4.41% for Arabic, to 8.38% for English; the mean random seed variability follows the same trend (1.73% to 3.54%). For every language, the score range and random seed variability for the large training size are consistently larger than those derived from small training. In both cases, score ranges are non-negligible.

Train Size	**Sampling Strategy**	**Score Range**	**Random Seed Variability**
Small	UNIFORM	4.51%	1.84%
	WEIGHTED	6.33	2.57
	OVERLAPAWARE	12.13	5.01
Large	UNIFORM	3.99%	1.68%
	WEIGHTED	4.08	1.66
	OVERLAPAWARE	13.06	5.50

Table 5: Average score range and random seed variability across languages for each sampling strategy for both training sizes.

Next, for each language, we analyze the average score range for each sampling strategy and model type separately. Comparing results from the three sampling strategies in Table 5, OVERLAPAWARE sampling consistently yields the highest score range and random seed variability. This indicates that OVERLAPAWARE, despite exhibiting the least variability in overlap partition sizes, is also the most variable in terms of model performance. This likely suggests that it is not just feature set attestation in general, but also exactly which feature sets that happen to appear in train vs. test drive performance. Finally, when looking at results for each individual model type, CLUZH-GR demonstrates the most variable performance. Its average score range (9.47% for large training, 7.94% for small) and its average random seed variability (4.03% for large training, 3.31% for small) end up being the highest.

6 Conclusions

We investigated the roles that sampling strategy, random seeds, and overlap types play in evaluating and analyzing the results of morphological inflection tasks and conclude that common practices leave much to be desired. We argue for frequency-weighted splitting to achieve more realistic train-test distributions and feature/lemma overlap-aware sampling for directly investigating the generalization abilities of different models. The high score range observed for overlap-aware sampling relative to other strategies suggests that which feature sets happen to appear in train vs. test play a major role in the ability of a model to generalize, though future work would need to confirm this.

Regardless of sampling strategy, evaluation items of each overlap type should be used in addition to an overall analysis. The evaluation in this work reveals that all model types under investigation struggle to generalize to unseen feature sets, even for languages where that should be possible, a fact that has been overlooked in prior studies. Finally, results drawn from one data split are unlikely to be representative, so multiple splits should be made with different random seeds and compared, particularly for shared tasks and leader boards where final model rankings matter.

Limitations

Our suggested approaches have two primary practical limitations: First, WEIGHTED sampling is restricted to languages with available running text sources for extracting frequencies. A project on *extremely* low-resource languages (e.g., Liu et al., 2022) may be restricted to UNIFORM and OVERLAPAWARE sampling. Second, as the number of seeds increases, so do requirements for training time and/or computing power. A shared task, for example, might limit itself to only a few seeds in order to assure on-time submissions. Future work would benefit from a wider selection of model architectures, along with more sampling strategies, and of course a wider sample of typologically diverse languages.

Notably, this work reproduces the effect observed in the SIGMORPHON 2022 shared task (Kodner et al., 2022), which found a substantial performance hit for `featsNovel` relative to `featsAttested`, but not `lemmaNovel` relative to `lemmaAttested`. However, both this work and the shared task fail to replicate the effect observed in Goldman et al. (2022), which reports a 95% performance hit on `lemmaNovel` vs. `lemmaAttested`. This may have something to do with differences in splitting algorithms, unmeasured feature overlap in Goldman et al. (2022), or choice of model architectures.

Ethics Statement

To the best of our knowledge, all results published in this paper are accurate, and we have represented prior work fairly to the best of our abilities. All data sources are free and publicly available, except for the Penn Arabic Treebank (Maamouri et al., 2004), which is accessible through the LDC.[7] No sensitive data was used which could violate individuals' privacy or confidentiality. Authorship and acknowledgements fairly reflect contributions.

Acknowledgements

We thank Charles Yang, Jeffrey Heinz, Mitch Marcus, and the audience at Stony Brook University AT-LaC for their helpful discussion. Experiments were performed on the SeaWulf HPC cluster maintained by RCC and the Institute for Advanced Computational Science (IACS) at Stony Brook University and made possible by National Science Foundation (NSF) grant No. 1531492. The second author gratefully acknowledges funding through the IACS Graduate Research Fellowship and the NSF Graduate Research Fellowship Program under NSF Grant No. 2234683.

[7] `https://catalog.ldc.upenn.edu/LDC2005T20`

References

Javier Aguado-Orea and Julian M Pine. 2015. Comparing different models of the development of verb inflection in early child Spanish. *PloS One*, 10(3):e0119613.

Janet Bang and Aparna Nadig. 2015. Learning language in autism: Maternal linguistic input contributes to later vocabulary. *Autism Research*, 8(2):214–223.

Elizabeth Bates, Inge Bretherton, and Lynn Sebestyen Snyder. 1991. *From first words to grammar: Individual differences and dissociable mechanisms*, volume 20. Cambridge University Press.

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.

Heike Behrens. 2006. The input–output relationship in first language acquisition. *Language and cognitive processes*, 21(1-3):2–24.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

MM Berl, LM Balsamo, B Xu, EN Moore, SL Weinstein, JA Conry, PL Pearl, BC Sachs, CB Grandin, C Frattali, et al. 2005. Seizure focus affects regional language networks assessed by fMRI. *Neurology*, 65(10):1604–1611.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Lynn Bliss. 1988. The development of modals. *The journal of applied developmental psychology*, 9:253–261.

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3):380–420.

Lois Masket Bloom. 1970. *Language development: Form and function in emerging grammars*. Ph.D. thesis, Columbia University.

John Neil Bohannon III and Angela Lynn Marquis. 1977. Children's control of adult speech. *Child Development*, pages 1002–1008.

Susan R Braunwald. 1971. Mother-child communication: the function of maternal-language input. *Word*, 27(1-3):28–50.

Michael R Brent and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

Roger Brown. 1973. *A first language: The early stages.* Harvard University Press, Cambridge, MA.

Joan L Bybee. 1991. Natural morphology: The organization of paradigms and language acquisition. *Crosscurrents in second language acquisition and linguistic theories*, 2:67–92.

Giuseppe Capelli, Victoria Marrero, and María José Albala. 1994. Aplicación del sistema morfo a una muestra de lenguaje infantil. *Procesamiento del Lenguaje Natural*, 14.

Erwin Chan. 2008. *Structures and distributions in morphological learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Jane Chandlee. 2017. Computational locality in morphological maps. *Morphology*, 27(4):599–641.

Eve V Clark. 1978. Awareness of language: Some evidence from what children say and do. In *The child's conception of language*, pages 17–43. Springer.

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.

Barbara L Davis and Peter F MacNeilage. 1995. The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, 38(6):1199–1211.

Martha Jo-Ann Demetras. 1986. *Working Parents' Conversational Responses to their two-year-old sons*. The University of Arizona.

Marty Demetras. 1989. Changes in parents' conversational responses: A function of grammatical development. *ASHA, St. Louis, MO*.

Marty J Demetras, Kathryn Nolan Post, and Catherine E Snow. 1986. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of child language*, 13(2):275–292.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and speech*, 49(2):137–173.

David K Dickinson and Patton O Tabors. 2001. *Beginning literacy with language: Young children learning at home and school.* Paul H Brookes Publishing.

Micha Elsner, Andrea D Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L King, Luana Lamberti Nunes, et al. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):53–98.

Andrea Feldman. 1998. *Constructing grammar: fillers, formulas, and function*. Ph.D. thesis, University of Colorado at Boulder.

Viviana Fratini, Joana Acha, and Itziar Laka. 2014. Frequency and morphological irregularity are independent variables. Evidence from a corpus study of Spanish verbs. *Corpus Linguistics and Linguistic Theory*, 10(2):289–314.

Catherine Garvey and Robert Hogan. 1973. Social speech and social interaction: Egocentrism revisited. *Child Development*, pages 562–568.

Virginia C Gathercole. 1986. The acquisition of the present perfect: Explaining differences in the speech of Scottish and American children. *Journal of Child Language*, 13(3):537–560.

Susan A Gelman, John D Coley, Karl S Rosengren, Erin Hartman, Athina Pappas, and Frank C Keil. 1998. Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child development*, pages i–157.

Ronald Bradley Gillam and Nils A Pearson. 2004. *TNL: test of narrative language*. Pro-ed Austin, TX.

Jean Berko Gleason. 1980. The acquisition of social speech routines and politeness formulas. In *Language*, pages 21–27. Elsevier.

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.

William S Hall and William C Tirre. 1979. The communicative environment of young children: Social class, ethnic, and situational differences. *Center for the Study of Reading Technical Report; no. 125*.

John Heilmann, Susan Ellis Weismer, Julia Evans, and Christine Hollar. 2005. Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Patholog*, 14:40–51.

Roy Patrick Higginson. 1985. *Fixing: Assimilation in language acquisition*. Ph.D. thesis, Washington State University.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jordan Kodner. 2022. Computational Models of Morphological Learning. Oxford University Press.

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.

Stan A Kuczaj II. 1977. The acquisition of regular and irregular past tense forms. *Journal of verbal learning and verbal behavior*, 16(5):589–600.

Constantine Lignos and Charles Yang. 2018. Morphology and language acquisition. *Cambridge handbook of morphology*, pages 765–791.

Josetxu Linaza, María Eugenia Sebastián, and Cristina del Barrio. 1981. Lenguaje, comunicación y comprensión: Conferencia a nual de la sección de psicología del desarrollo de la british psychological society. *Infancia y Aprendizaje*, 4(sup1):195–197.

Zoey Liu and Emily Prud'hommeaux. 2022. Data-driven model generalizability in crosslinguistic low-resource morphological segmentation. *Transactions of the Association for Computational Linguistics*, 10:393–413.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings* *of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Susana López Ornat. 1997. What lies in between a pre-grammatical and a grammatical representation? Evidence on nominal and verbal form-function mappings in Spanish from 1; 7 to 2; 1. *Contemporary perspectives on the acquisition of Spanish*, 1:3–20.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Brian MacWhinney. 1991. The CHILDES language project: Tools for analyzing talk. *Journal of Speech, Language and Hearing Research*, 40:62–74.

Brian MacWhinney. 2000. *The CHILDES Project: The Database*, volume 2. Psychology Press, Abingdon-on-Thames.

Brian MacWhinney and Catherine Snow. 1985. The child language data exchange system. *Journal of Child Language*, 12(2):271–295.

María del Carmen Aguirre Martínez and Sonia Mariscal Altares. 2005. *Cómo adquieren los niños la gramática de su lengua: perspectivas teóricas*. Editorial UNED.

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.

Lorraine McCune. 1995. A normative study of representational play in the transition to language. *Developmental psychology*, 31(2):198.

Rosa Montes. 1987. Secuencias de clarificación en conversaciones con niños (morphe 3-4). *Universidad Autónoma de Puebla*.

Colleen E Morisset, Kathryn E Barnard, and Cathryn L Booth. 1995. Toddlers' language development: Sex differences within social risk. *Developmental Psychology*, 31(5):851.

Katherine Nelson. 2006. *Narratives from the crib*. Harvard University Press.

Rochelle S Newman, Meredith L Rowe, and Nan Bernstein Ratner. 2016. Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of child language*, 43(5):1158–1173.

Johanna G Nicholas and Ann E Geers. 1997. Communication of oral deaf and normally hearing children at 36 months of age. *Journal of Speech, Language, and Hearing Research*, 40(6):1314–1327.

Anat Ninio, Catherine E Snow, Barbara A Pan, and Pamela R Rollins. 1994. Classifying communicative acts in children's interactions. *Journal of communication disorders*, 27(2):157–187.

Kemal Oflazer and Murat Saraçlar. 2018. *Turkish natural language processing*. Springer.

Ann M Peters. 1987. The role of imitation in the developing syntax of a blind child. *Text-Interdisciplinary Journal for the Study of Discourse*, 7(3):289–309.

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.

Steven Pinker and Michael T Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

V Remedi. 2014. *Creación de corpus de datos sobre estudio longitudinal de adquisición de lenguaje de* *una niña de la región central de Argentina*. Ph.D. thesis, National University of Córdoba.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*, volume 4. Oxford University Press.

Pamela Rosenthal Rollins. 2003. Caregivers' contingent comments to 9-month-old infants: Relationships with later language. *Applied Psycholinguistics*, 24(2):221–234.

Jacqueline Sachs and KE Nelson. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, 4:1–28.

R Keith Sawyer. 2013. *Pretend play as improvisation: Conversation in the preschool classroom*. Psychology Press.

Mark S. Seidenberg and D. Plaut. 2014. Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, 38 6:1190–228.

Melanie Soderstrom, Megan Blossom, Rina Foygel, and James L Morgan. 2008. Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, 35(4):869–902.

Richard A Sprott. 1992. Children's use of discourse markers in disputes: Form-function relations and discourse in child language. *Discourse Processes*, 15(4):423–439.

Patrick Suppes. 1974. The semantics of children's language. *American Psychologist*, 29(2):103.

Virginia Valian. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, 40(1-2):21–81.

Lori J. van Houton. 1986. The role of maternal input in the acquisition process: The communicative strategies of adolescent and older mothers with the language learning children. In *The Proceedings of the Boston University Conference on Language Development*.

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.

Amye Warren-Leubecker. 1982. *Sex differences in speech to children*. Ph.D. thesis, Georgia Institute of Technology.

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.

Richard M Weist, Aleksandra Pawlak, and Karen Hoffman. 2009. Finiteness systems and lexical aspect in child Polish and English. *Linguistics*, 47(6):1321–1350.

Wikimedia. 2022. Wikimedia Downloads.

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.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

Karina Hess Zimmermann. 2003. *El desarrollo linguístico en los años escolares: análisis de narraciones infantiles*. Ph.D. thesis, El Colegio de México.

A English and Spanish Data Sources

A.1 English

The following CHILDES corpora were used to create the English data set. Utterances from speaker *CHI were excluded: Bates (Bates et al., 1991), Bliss (Bliss, 1988), Bloom (Bloom, 1970; Bloom et al., 1974), Bohannon (Bohannon III and Marquis, 1977), Braunwald (Braunwald, 1971), Brent (Brent and Siskind, 2001), Brown (Brown, 1973), Clark (Clark, 1978), Davis (Davis and MacNeilage, 1995), Demetras (Demetras, 1986, 1989), EllisWeismer (Heilmann et al., 2005), Feldman (Feldman, 1998), Garvey (Garvey and Hogan, 1973), Gathercole (Gathercole, 1986), Gelman (Gelman et al., 1998), Gillam (Gillam and Pearson, 2004), Gleason (Gleason, 1980), Hall (Hall and Tirre, 1979), Higginson (Higginson, 1985), HSLLD (Dickinson and Tabors, 2001), Kuczaj (Kuczaj II, 1977), MacWhinney (MacWhinney, 1991), McCune (McCune, 1995), Morisset (Morisset et al., 1995), Nadig (Bang and Nadig, 2015), Nelson (Nelson, 2006), NewEngland (Ninio et al., 1994), NewmanRatner (Newman et al., 2016), Nichols-TD (Nicholas and Geers, 1997), Peters (Peters, 1987), POLER (Berl et al., 2005), Post (Demetras et al., 1986), Providence (Demuth et al., 2006), Rollins (Rollins, 2003), Sachs (Sachs and Nelson, 1983), Sawyer (Sawyer, 2013), Snow (MacWhinney and Snow, 1985), Soderstrom (Soderstrom et al., 2008), Sprott (Sprott, 1992), Suppes (Suppes, 1974), Tardif (MacWhinney, 2000), Valian (Valian, 1991), VanHouten (van Houton, 1986), VanKleeck (MacWhinney, 2000), Warren-Leubecker (Warren-Leubecker, 1982), Weist (Weist et al., 2009).

A.2 Spanish

The following CHILDES corpora were used to create the Spanish data set. Utterances from speaker `*CHI` were excluded: Aguirre (Martínez and Altares, 2005), ColMex (MacWhinney, 2000), Fernandez/Aguado (MacWhinney, 2000), GRERLI (MacWhinney, 2000), Hess (Zimmermann, 2003), Linaza (Linaza et al., 1981), Marrero (Capelli et al., 1994), Montes (Montes, 1987), AguadoOrea/Pine (Aguado-Orea and Pine, 2015), Ornat (López Ornat, 1997), Remedi (Remedi, 2014), SerraSole (MacWhinney, 2000).

B Splitting Strategy Data Summaries

This appendix contains Tables 6-9.

	Train		Test	
Arabic	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	0.46	0	0.47	0
Weighted	57.53	18	26.44	12
OverlapAware	6.72	2	6.46	2
English	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	9.71	0	1.24	0
Weighted	1840.51	362	122.55	67
OverlapAware	182.29	5	163.22	5
German	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	0.14	0	0.18	0
Weighted	111.99	20	9.56	5
OverlapAware	25.46	2	30.42	2
Spanish	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	0.12	0	0.13	0
Weighted	119.15	29	13.89	8
OverlapAware	25.50	2	21.97	2
Swahili	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	40.13	0	38.38	0
Weighted	518.95	88	8.11	4
OverlapAware	130.00	3	143.39	3
Turkish	$\mu\mu$	μM	$\mu\mu$	μM
Uniform	26.63	0	26.6	0
Weighted	4854.13	1252	588.76	348
OverlapAware	436.41	12	397.94	12

Table 6: Average training and test item mean corpus frequency ($\mu\mu$) and median frequency (μM).

Arabic	J_{LTrain}	J_{Test}
Uniform	0.10	0.05
Weighted	9.90	3.17
OverlapAware	1.56	1.07
English	J_{LTrain}	J_{Test}
Uniform	0.12	0.09
Weighted	32.12	8.86
OverlapAware	4.78	3.31
German	J_{LTrain}	J_{Test}
Uniform	0.13	0.06
Weighted	27.80	8.16
OverlapAware	7.69	4.98
Spanish	J_{LTrain}	J_{Test}
Uniform	0.08	0.06
Weighted	27.81	8.07
OverlapAware	6.89	4.65
Swahili	J_{LTrain}	J_{Test}
Uniform	3.06	3.74
Weighted	41.20	24.06
OverlapAware	11.97	15.95
Turkish	J_{LTrain}	J_{Test}
Uniform	0.10	0.11
Weighted	27.91	7.66
OverlapAware	3.37	2.21

Table 7: Average Jaccard similarity quantifying overlap between large training samples (J_{LTrain}) across random seeds and similarity between test samples (J_{Test}) across seeds. $J \in [0, 100]$ where 100 indicates that all UniMorph triples appear in all training sets

	Raw UniMorph			UniMorph×Freq		
	#L	#F	#T	#L	#F	#T
Arabic	12815	567	834113	11628	300	56035
English	399758	11	716093	8370	6	16528
German	39417	113	599141	4460	44	10501
Spanish	65689	175	1286348	3592	117	11337
Swahili	184	257	15149	180	225	3725
Turkish	3579	883	570420	1649	242	24332

Table 8: Type frequencies for lemmas (#L), feature sets (#F), and triples (#T) for each language data set. Raw UniMorph (3+)4 and intersected with frequency.

C Detailed Results

This appendix contains Tables 10-11.

Overall Test vs S Train	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	15.02 (*25.29*)	65.31 (*33.2*)	6.25 (*8.32*)	13.43 (*14.57*)	80.33 (*19.50*)	19.67 (*19.50*)
Weighted	25.69 (*15.61*)	64.75 (*25.01*)	6.97 (*10.67*)	2.59 (*2.42*)	90.44 (*11.13*)	9.56 (*11.13*)
OverlapAware	13.27 (*13.43*)	35.54 (*13.96*)	14.92 (*15.20*)	36.27 (*14.71*)	48.81 (*0.98*)	51.19 (*0.98*)

Overall Test vs L Train	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	30.58 (*32.47*)	65.59 (*35.62*)	2.83 (*4.56*)	1.00 (1.40)	96.17 (*5.55*)	3.83 (*5.55*)
Weighted	50.59 (*16.38*)	44.76 (*21.74*)	4.24 (*7.22*)	0.39 (*0.58*)	95.36 (*7.28*)	4.64 (*7.28*)
OverlapAware	23.94 (*14.76*)	25.97 (*14.84*)	25.17 (*14.14*)	24.91 (*14.05*)	49.92 (*0.17*)	50.08 (*0.17*)

Ara Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	3.12 (*0.26*)	66.38 (*4.22*)	1.32 (*0.35*)	29.18 (*4.02*)	69.50 (*4.14*)	30.50 (*4.14*)
Weighted	13.02 (*1.18*)	77.52 (*1.33*)	2.06 (*0.40*)	7.40 (*1.14*)	90.54 (*1.53*)	9.46 (*1.53*)
OverlapAware	3.06 (*0.62*)	44.62 (*0.92*)	3.30 (*0.72*)	49.02 (*1.14*)	47.68 (*0.57*)	52.32 (*0.57*)

Ara Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	15.82 (*1.03*)	80.82 (*2.10*)	0.78 (*0.26*)	2.58 (*1.08*)	96.64 (*1.30*)	3.36 (*1.30*)
Weighted	39.38 (*1.17*)	57.42 (*0.78*)	1.66 (*0.62*)	1.54 (*0.46*)	96.80 (*0.77*)	3.20 (*0.77*)
OverlapAware	10.40 (*1.31*)	39.50 (*1.24*)	10.82 (*0.84*)	39.28 (*0.86*)	49.90 (*0.11*)	50.10 (*0.11*)

Deu Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	1.16 (*0.52*)	97.42 (*1.09*)	0.00 (*0.00*)	1.42 (*0.84*)	98.58 (*0.84*)	1.42 (*0.84*)
Weighted	12.08 (*0.50*)	85.90 (*1.34*)	0.74 (*0.43*)	1.28 (*0.70*)	97.98 (*1.11*)	2.02 (*1.11*)
OverlapAware	4.70 (*1.40*)	45.20 (*1.50*)	4.90 (*1.13*)	45.20 (*1.19*)	49.90 (*0.15*)	50.10 (*0.15*)

Deu Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	4.38 (*0.34*)	95.42 (*0.43*)	0.00 (*0.00*)	0.20 (*0.13*)	99.80 (*0.13*)	0.20 (*0.13*)
Weighted	36.38 (*1.24*)	63.50 (*1.24*)	0.08 (*0.07*)	0.04 (*0.05*)	99.88 (*0.10*)	0.12 (*0.10*)
OverlapAware	14.74 (*3.32*)	35.26 (*3.32*)	14.96 (2.28)	35.04 (2.28)	50.00 (*0.00*)	50.00 (*0.00*)

Eng Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	0.10 (*0.11*)	99.68 (*0.26*)	0.00 (*0.00*)	0.22 (*0.29*)	99.78 (*0.29*)	0.22 (*0.29*)
Weighted	10.62 (*0.82*)	89.38 (*0.82*)	0.00 (*0.00*)	0.00 (*0.00*)	100.00 (*0.00*)	0.00 (*0.00*)
OverlapAware	1.94 (*0.61*)	48.06 (*0.61*)	3.02 (*0.39*)	46.98 (*0.39*)	50.00 (*0.00*)	50.00 (*0.00*)

Eng Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	0.38 (*0.07*)	99.62 (*0.07*)	0.00 (*0.00*)	0.00 (*0.00*)	100.00 (*0.00*)	0.00 (*0.00*)
Weighted	31.26 (*0.91*)	68.74 (*0.91*)	0.00 (*0.00*)	0.00 (*0.00*)	100.00 (*0.00*)	0.00 (*0.00*)
OverlapAware	7.16 (*2.48*)	42.84 (*2.48*)	12.04 (*0.63*)	37.96 (*0.63*)	50.00 (*0.00*)	50.00 (*0.00*)

Spa Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	3.88 (*0.56*)	84.60 (*2.48*)	0.46 (*0.22*)	11.06 (*1.99*)	88.48 (*2.14*)	11.52 (*2.14*)
Weighted	28.40 (*1.40*)	63.54 (*1.78*)	5.94 (*0.71*)	2.12 (*0.67*)	91.94 (*1.12*)	8.06 (*1.12*)
OverlapAware	15.02 (*3.78*)	34.00 (*3.71*)	15.54 (*2.03*)	35.44 (*1.94*)	49.02 (*0.17*)	50.98 (*0.17*)

Spa Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	16.72 (*0.61*)	83.28 (*0.61*)	0.00 (*0.00*)	0.00 (*0.00*)	100.00 (*0.00*)	0.00 (*0.00*)
Weighted	53.30 (*1.58*)	44.76 (*1.64*)	1.74 (*0.49*)	0.20 (*0.23*)	98.06 (*0.69*)	1.94 (*0.69*)
OverlapAware	28.08 (*4.52*)	21.90 (*4.53*)	28.02 (*4.10*)	22.00 (*4.10*)	49.98 (*0.04*)	50.02 (*0.04*)

Swc Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	70.98 (*2.51*)	11.12 (*2.25*)	16.02 (*1.50*)	1.88 (*0.50*)	82.10 (*1.76*)	17.90 (*1.76*)
Weighted	52.24 (*5.04*)	15.10 (*1.39*)	30.00 (*4.90*)	2.66 (*0.94*)	67.34 (*5.76*)	32.66 (*5.76*)
OverlapAware	40.68 (*1.10*)	7.04 (*1.07*)	46.52 (*1.33*)	5.76 (*1.41*)	47.72 (*0.30*)	52.28 (*0.30*)

Swc Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	91.82 (*0.65*)	4.34 (*0.77*)	3.66 (*0.53*)	0.18 (*0.16*)	96.16 (*0.63*)	3.84 (*0.63*)
Weighted	72.62 (*2.51*)	6.86 (*1.18*)	20.12 (*2.43*)	0.40 (*0.22*)	79.48 (*2.63*)	20.52 (*2.63*)
OverlapAware	47.64 (*1.07*)	2.04 (*1.08*)	48.70 (*0.98*)	1.62 (*0.89*)	49.68 (*0.29*)	50.32 (*0.29*)

Tur Test vs STrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	10.88 (*0.63*)	32.64 (*2.15*)	19.68 (*0.90*)	36.80 (*1.16*)	43.52 (*1.88*)	56.48 (*1.88*)
Weighted	37.80 (*1.51*)	57.06 (*1.13*)	3.06 (*0.78*)	2.08 (*0.41*)	94.86 (*1.03*)	5.14 (*1.03*)
OverlapAware	14.24 (*1.67*)	34.30 (*1.45*)	16.22 (*0.72*)	35.24 (*0.66*)	48.54 (*0.28*)	51.46 (*0.28*)

Tur Test vs LTrain	`both%` (σ)	`featsOnly`	`lemmaOnly`	`neither`	`featsAttested`	`featsNovel`
Uniform	54.36 (*0.81*)	30.06 (*0.75*)	12.52 (*1.21*)	3.06 (*0.72*)	84.42 (*1.35*)	15.58 (*1.35*)
Weighted	70.62 (*1.33*)	27.30 (*1.26*)	1.88 (*0.61*)	0.20 (*0.11*)	97.92 (*0.53*)	2.08 (*0.53*)
OverlapAware	35.64 (*1.06*)	14.30 (*1.04*)	36.50 (*1.52*)	13.56 (*1.47*)	49.94 (*0.08*)	50.06 (*0.08*)

Table 9: Language-by-language average mean percentage of each overlap type in test sets relative to small and large training. Standard deviations are (*italicized*). OverlapAware targets a `featsAttested` relative to large train as close to 50% as possible without exceeding it. %`featsAttested` = %`both` + %`featsOnly` and %`featsNovel` = %`lemmaOnly` + %`neither`.

nonneur Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	70.92	66.75	17.16	19.10	67.50	16.94	59.83
Weighted	67.86	77.93	8.15	13.07	74.98	9.91	68.79
OverlapAware	66.47	75.43	17.79	26.55	73.39	24.63	48.30
nonneur Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	73.59	66.00	21.85	25.75	71.66	31.72	70.33
Weighted	75.35	83.62	8.06	9.17	79.15	7.61	76.1o
OverlapAware	74.52	82.49	18.57	29.31	77.84	24.33	51.03

chr-trm Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	70.02	61.05	58.61	30.48	67.70	39.36	65.33
Weighted	79.18	69.36	43.60	26.20	75.08	36.15	72.27
OverlapAware	80.28	72.46	38.15	30.86	78.06	35.97	56.67
chr-trm Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	79.60	76.61	63.85	39.92	79.51	55.72	78.82
Weighted	89.42	85.42	59.62	37.81	89.48	52.64	88.56
OverlapAware	89.78	86.56	45.65	38.87	89.83	43.92	66.85

cluzh-b4 Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	77.09	71.75	57.13	33.22	73.87	39.72	70.29
Weighted	78.35	86.22	26.18	21.40	83.67	22.63	78.09
OverlapAware	79.97	84.86	30.43	32.00	83.66	32.16	57.38
cluzh-b4 Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	88.14	79.80	72.66	47.34	86.02	69.86	85.42
Weighted	86.14	90.39	20.63	20.93	88.22	17.71	85.83
OverlapAware	88.31	91.81	35.35	41.20	89.78	37.68	63.70

cluzh-gr Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	75.72	70.77	55.27	31.89	72.83	38.27	69.21
Weighted	77.79	85.91	25.75	21.22	83.28	22.38	77.72
OverlapAware	79.78	84.50	29.98	31.49	83.28	31.78	57.00
cluzh-gr Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	85.15	75.83	65.54	43.43	82.83	65.00	82.24
Weighted	84.65	89.17	20.17	17.13	86.89	17.01	84.52
OverlapAware	85.76	89.64	33.91	40.04	87.42	36.12	61.74

Table 10: Average percent accuracy across seeds and models on the test set by architecture.

Overall Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	73.44	67.58	47.05	28.67	70.47	33.57	66.16
Weighted	75.79	79.86	25.92	20.47	79.25	22.77	74.22
OverlapAware	76.62	79.31	29.09	30.22	79.60	31.13	54.84
Overall Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	81.62	74.56	55.97	39.11	80.00	55.57	79.20
Weighted	83.89	87.15	27.12	21.26	85.94	23.74	83.75
OverlapAware	84.59	87.63	33.37	37.36	86.22	35.51	60.83

Ara Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	72.52	67.86	54.84	50.58	68.06	50.80	62.80
Weighted	73.82	73.15	35.79	23.98	73.24	26.54	68.82
OverlapAware	63.77	66.33	33.42	30.97	66.14	31.11	47.81
Ara Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	83.60	76.52	62.57	44.31	77.67	48.62	76.76
Weighted	79.92	78.95	38.29	23.67	79.34	31.04	77.76
OverlapAware	75.07	76.36	46.49	45.99	76.09	46.09	61.06

Deu Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	63.61	60.00	–	28.27	60.06	28.27	59.65
Weighted	78.22	76.73	26.06	16.48	76.91	20.18	75.81
OverlapAware	73.90	73.88	38.98	41.80	74.12	41.60	57.84
Deu Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	75.37	73.07	–	73.33	73.16	73.33	73.14
Weighted	85.35	84.37	25.00	0.00	84.72	14.58	84.64
OverlapAware	81.22	82.00	40.02	44.25	81.84	43.24	62.54

Eng Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	97.22	93.34	–	0.00	93.35	0.00	93.14
Weighted	76.90	88.43	–	–	87.20	–	87.20
OverlapAware	84.30	88.53	17.10	19.14	88.45	18.99	53.72
Eng Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	95.66	96.49	–	–	96.48	–	96.48
Weighted	84.25	95.26	–	–	91.83	–	91.83
OverlapAware	89.96	92.11	17.81	19.80	91.95	19.32	55.63

Spa Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	75.09	71.24	46.87	39.58	71.35	39.67	67.67
Weighted	65.97	83.03	10.02	8.36	77.74	9.59	72.22
OverlapAware	68.60	84.40	9.94	27.14	79.90	21.92	50.35
Spa Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	84.09	83.39	–	–	83.50	–	83.50
Weighted	80.73	92.16	24.60	38.89	85.94	24.74	84.77
OverlapAware	82.57	94.20	16.06	35.42	87.92	24.83	56.37

Swc Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	89.68	69.89	63.61	31.14	87.02	60.08	82.22
Weighted	80.41	75.56	29.41	26.04	79.27	29.12	62.79
OverlapAware	85.83	78.31	43.16	31.05	84.79	41.75	62.28
Swc Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	90.74	58.56	59.70	6.25	89.26	57.27	88.01
Weighted	82.30	77.40	40.77	33.75	81.88	40.66	73.36
OverlapAware	88.53	88.42	44.11	43.24	88.56	44.01	66.14

Tur Test vs S Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	42.51	43.14	22.85	22.46	42.99	22.61	31.51
Weighted	79.46	82.24	28.32	27.51	81.15	28.41	78.46
OverlapAware	83.33	84.42	31.93	31.23	84.18	31.43	57.03
Tur Test vs L Train	both%	featsOnly	lemmaOnly	neither	featsAttested	featsNovel	overall
Uniform	60.24	59.34	45.65	32.55	59.94	43.08	57.33
Weighted	90.80	94.75	6.93	10.00	91.91	7.70	90.16
OverlapAware	90.21	92.67	35.72	35.44	90.94	35.59	63.23

Table 11: Language-by-language average percent accuracy across seeds and models on the test set. Dashes indicate overlap partitions with size zero.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☑ A2. Did you discuss any potential risks of your work?
Conclusions and Limitations.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Many corpora were used, all cited in the introduction and/or Appendix A. Code and data are available through the link provided in the paper

☑ B1. Did you cite the creators of artifacts you used?
Citations in-line and in Appendix A

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics Statement.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We processed wordlists

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Throughout the paper and in detail in Appendix. Our data sets have also been made available

C ☑ Did you run computational experiments?

Section 4-5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and Acknowledgements

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Throughout the paper

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Throughout the paper and Appendix B

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Our code is available

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.