Re-Evaluating the Evaluation of Neural Morphological Inflection Models

Jordan Kodner, Salam Khalifa,* Sarah Payne*
{jordan.kodner, salam.khalifa, sarah.payne}@stonybrook.edu
Department of Linguistics & Institute for Advanced Computational Science, Stony Brook University
Stony Brook, NY 11733 USA

Zoey Liu
liu.ying@ufl.edu
Department of Linguistics, University of Florida
Gainesville, FL 32611 USA

Abstract

Computational models of morphology acquisition have played a central role in debates over the nature of morphological representations. The apparent success of recent artificial neural network architectures for morphological inflection in natural language processing has renewed this debate. However, the actual suitability of these advanced neural models as models of human morphology acquisition remains uncertain. We argue that much of this confusion stems from inconsistent methods of training and evaluation. In this work, we demonstrate that more careful data set creation and an evaluation combining quantitative analysis and comparison with human development will put the evaluation of neural models on firmer ground.

Keywords: Linguistics, NLP, morphology, language acquisition, neural networks

Introduction

Computational models of morphological inflection burst onto the scene as part of the Past Tense Debate in the late 1980s (Rumelhart & McClelland, 1986; Pinker & Prince, 1988), where they were seen as providing insights into the cognitive computations underlying morphological learning and representation; see Pinker and Ullman (2002); McClelland and Patterson (2002); Seidenberg and Plaut (2014) for surveys. One feature of the debate was a push on one side for connectionist models, a family of artificial neural networks (ANNs). With the advent of more powerful "deep" ANNs in machine learning and natural language processing (NLP) in recent years, there has been renewed interest in ANNs as potential cognitive models (Kirov & Cotterell, 2018; Corkery, Matusevych, & Goldwater, 2019; McCurdy, Goldwater, & Lopez, 2020; Belth, Payne, Beser, Kodner, & Yang, 2021; Beser, 2021; Dankers, Langedijk, McCurdy, Williams, & Hupkes, 2021; Wiemerslage, Dudy, & Kann, 2022).

At the same time, morphological inflection has established itself as a (relatively) standardized task in NLP in the guise of the SIGMORPHON shared tasks (Cotterell et al., 2016, 2017, 2018; A. McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022). While submissions to these competitions aim primarily to maximize accuracy, the 2021 and 2022 competitions included explicitly cognitive sub-tasks (Kodner & Khalifa, 2022).

Results on the cognitive plausibility of ANNs are contradictory. We suspect that part of this can be explained by inconsistent success criteria. In terms of accuracy, performance on shared tasks is often near-ceiling. Models may instead be evaluated by how well their ratings correspond with human acceptability judgments. While these correlations are good, but sometimes problematic on English past tense (Corkery et al., 2019), they are not particularly human-like on German noun pluralization (McCurdy et al., 2020), another classic of the Past Tense Debate (Clahsen & Rothweiler, 1993).

Another alternative to pure accuracy or correlation with judgments is to evaluate the 'human-likeness' of productions: learning trajectories and errors. Taking this approach, even state-of-the-art models submitted to the shared task come up lacking. No evidence for *u*-shaped learning of English past tense (Marcus et al., 1992; Prasada & Pinker, 1993) was observed in Kodner and Khalifa (2022), and what Kirov and Cotterell (2018) report as *u*-shaped is really oscillating learning, something never reported in the developmental literature.

Research Goals

We aim to evaluate ANNs in terms of their overall accuracy and match to human learning trajectories. This approach takes inspiration from the developmental literature and early papers in the Past Tense Debate where errors were taken to provide particular insight into the representations underlying learning. We investigate the acquisition of English past tense (Marcus et al., 1992), German noun pluralization (Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995), and Arabic noun pluralization (Ravid & Farah, 1999), since much is known about acquisition patterns for these phenomena.

Our main methodological contribution is the creation of developmentally-plausible training data. If a model is to be evaluated from a cognitive perspective, it should learn from input that shares key properties with the input to language acquisition. Thus, training data is sampled from child-directed speech from CHILDES (MacWhinney, 2000) to the extent possible, following Belth et al. (2021) and Kodner and Khalifa (2022). Most SIGMORPHON shared tasks have sampled data from the UniMorph database of inflectional patterns, which is extracted primarily from Wiktionary (A. D. McCarthy et al., 2020; Batsuren et al., 2022) or from the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1996), such as Kirov and Cotterell (2018) and Wiemerslage et al. (2022). Neither data source corresponds well to early learner input.

Estimates of child vocabulary knowledge for many languages (Fenson et al., 1994; Bornstein et al., 2004; Szagun,

*Denotes equal contribution

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Steinbrink, Franik, & Stumper, 2006), combined with observations of morphological development e.g., (Brown, 1973; Aksu-Koç, 1985; Ravid & Farah, 1999; Elsen, 2002; Deen, 2005) show that children acquire most of their morphological competence on the basis of just hundreds of types regardless of a language's morphological complexity. Thus we limit our training data to at most 1000 lemmas.

Additionally, we sample training and test data weighted by frequency, since this more closely approximates the human learning task and real-world application. Vocabulary acquisition is correlated with item frequency in the input (Goodman, Dale, & Li, 2008), so children will need to use their productive morphological knowledge to inflect low-frequency forms that they have not seen in their input.

Except for 2017 and 2022, the SIGMORPHON shared tasks sampled uniformly rather than weighted by frequency, and most recent work in cognitive modeling follows (e.g., Kirov and Cotterell (2018) discard CELEX lemma frequency as well). We compare weighted sampling with uniform sampling to determine how it influences ANN learners. Some prior work, including and following Kirov and Cotterell (2018), train their models on full morphological paradigms. However, since this does not at all reflect childhood linguistic experience, where the vast majority of potential inflected forms are never attested (Chan, 2008; Lignos & Yang, 2018), we do not consider a full-paradigm sampling strategy.

Finally, we make several data samples with unique random seeds in order to investigate the stability of model performance on different data sets. On a technical level, sampling has been shown to significantly affect performance in other morphology learning tasks (Liu & Prud'hommeaux, 2022), so a single sample cannot be taken as representative. By taking multiple samples, we aim to test the stability of learning systems. After all, every child receives their own unique input sample and yet develops similarly.

Experiments

Data sources and preparation

Three phenomena were investigated, all of which have been previously studied from developmental and computational perspectives: English past tense inflection, German noun pluralization, and Arabic noun pluralization. Original data for each language is taken from the 2022 SIGMORPHON developmental subtask and is all orthographical (diacritized for Arabic). English and German data were extracted from CHILDES (MacWhinney, 2000) child-directed speech (CDS) and intersected with UniMorph to remove errorful annotations; we then converted them to a standard format. Extracting forms from CDS provides frequency estimates for typical morphological input during acquisition and removes rare and unusual items from UniMorph. Frequencies for Arabic were extracted from the Penn Arabic Treebank (Maamouri, Bies, Buckwalter, & Mekki, 2004) because CHILDES Arabic corpora are not suitably annotated.

Following the format widely adopted for morphological inflection tasks in recent years, training items are presented as `(lemma, inflected form, morpho-syntactic feature set)` triples, and test items are `(lemma, feature set)` pairs, where the learner is asked to provide the appropriate inflected form. This is effectively the computational adaptation of the classic Wug test paradigm following Berko (1958). (1)-(2) provide example English and German training and test items.

(1)

English Training Items			English Test Items		
run	ran	V;PST	see	?	V;PST
walk	walked	V;PST	look	?	V;PST

(2)

German Training Items			German Test Items		
Lampe	Lampen	N;FEM;PL	Paar	?	N;NEUT;PL
Tanz	Tänze	N;MASC;PL	Amsel	?	N;FEM;PL

Data splits

We employed both uniform and frequency-weighted sampling strategies to generate training, fine-tuning, development, and test splits. Both sampling strategies were applied five times with unique random seeds to produce distinct data sets. We also compare these against the data splits adopted for SIGMORPHON 2022. The training+fine-tuning sets were sub-sampled into nested sets of size 100, 200, and so on in order to approximate a learning curve as vocabulary size increases. These sets were uniformly split into 80% training and 20% fine-tuning. The maximum English and Arabic sets contained 1000 items, while the maximum German sets were limited to 600 due to limited data. All dev sets had 500 items. Remaining items were assigned to test (Arabic: 496, English: 554, German: 600). Dev and test were kept consistent regardless of training size. No lemmas overlapped between the sets. Three sampling strategies were adopted:

UNIFORM: As in most prior work, data was partitioned uniformly at random without replacement, so that there were no differences in frequency between the sets. An advantage of this approach is that it can be performed on data sets with no frequency information, but it does not reflect the situation in language acquisition.

WEIGHTED: Sub-sampling was weighted by frequency without replacement. The first 100 training+fine-tuning set was sampled from the entire data set, followed by 100 more items to create the 200 training+fine-tuning set, and so on. After training was completely sampled, dev+test were sampled together from the remainder and then split uniformly at random. Thus, the smallest training sets are skewed towards the highest-frequency items and dev+test are skewed towards the lower frequency items. We take this to reflect the situation during language acquisition: high-frequency items are more likely to be present in a typical child's input and thus can be memorized, while low-frequency items are less likely to be presented to any given child and will often need to be inferred from morphological knowledge constructed over on average higher-frequency forms.

SIGM22: The SIGMORPHON 2022 split was included for comparison. It also relied on weighted sampling, but only one split was produced. All test sets contained 600 items. Dev sets varied (Arabic: 343, English: 454, German: 500).

Models

We evaluated three neural models and one non-neural model, chosen for their performance in recent shared tasks.

CHR-TRM (Wu, Cotterell, & O'Donnell, 2019) is a character-level transformer previously holding the state-of-the-art in SIGMORPHON task and serving as a baseline in 2021 and 2022. We used the provided default hyperparameters for small training conditions.

CLUZH-GR (Wehrli, Clematide, & Makarov, 2022) is a character-level transducer which outperformed CHR-TRM in 2022. While the SIGMORPHON submission was optimized for each language individually, we used consistent hyperparameters to facilitate comparison. **CLUZH-B4** replaces the greedy decoding of the GR with beam decoding, size=4.

NONNEUR (Cotterell et al., 2017) has been used as a baseline since 2017. It extracts lemma-form mappings from training and trains a majority classifier with the associated feature sets. We trained NONNEUR on the combined training and fine-tuning sets so that each model was exposed to the same data in some way during training.

Quantitative Analysis

In this section, we present quantitative analyses of model performance. Throughout this section, we report results on the test set; we carried out the same series of examinations for the dev set and there was no observable qualitative difference in the findings. All reported accuracies are exact match.

Effect of training size

We start with analyzing the effect of different training sizes on the overall accuracy. After deriving the evaluation results from the two sampling strategies (UNIFORM and WEIGHTED) for all languages, we combined them together and performed linear regression (with the programming language R). The model predicts the overall accuracy score as a function of the training size, controlling for the language, the model type, and the sampling strategy, along with interactions between each of the aforementioned four fixed effects:

ACCURACY∼LANG * MODEL * TRAIN SIZE * SAMPLING STRAT

Based on the regression model, there is a weak yet significant effect of training size overall (β=0.02, $p < 0.001$); this suggests that while more training data yield higher accuracy scores, the effect is not as pronounced as one might expect. The role of training size appears to be consistent regardless of the specific sampling strategy, evidenced by the lack of significant interaction effects between the two factors. The interaction between the training size and model type is most significant for CHR-TRM (β=0.03, $p < 0.01$). Across languages, this tendency between CHR-TRM and training size was more pronounced for German (β=0.07, $p < 0.001$). This is confirmed visually in Figures 1-3. Performance increases as training size increases for all models, but a much sharper increase is observed for CHR-TRM than the others. The same patterns were observed on the SIGM22 data.

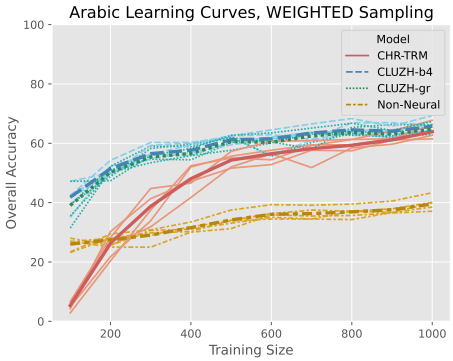

Figure 1: Learning curves for Arabic nouns. Thin lines = individual seeds and thick lines = averages across seeds.

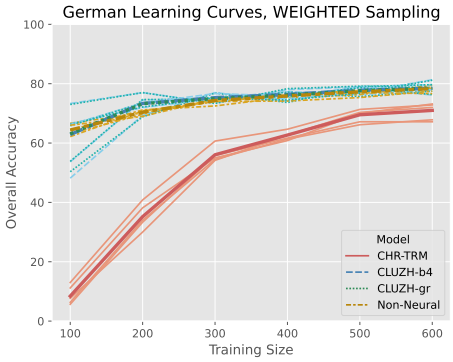

Figure 2: Learning curves for German nouns. Thin lines = individual seeds and thick lines = averages across seeds.

Effect of sampling strategy

We now turn to investigate the effect of each sampling strategy (UNIFORM vs WEIGHTED) on the evaluation results. Considering all languages together, the average overall accuracy over all training sizes derived from UNIFORM is slightly higher on average (67.17%)[1] than that from WEIGHTED (65.24%). The score difference between the two sampling strategies is the largest for English (2.77%), and the lowest for Arabic (0.91%). SIGM22 also employed a weighted sampling strategy, which also leads to a lower average overall accuracy score (65.29%) than UNIFORM. These results strengthen our finding above that UNIFORM overall leads to inflated performance for evaluations of morphological inflection models.

The score discrepancy between UNIFORM and WEIGHTED is clearest at smaller training sizes, where the largest differences are found for English (UNIFORM 66.32% vs. WEIGHTED 59.45% at 100 training). When comparing individual model types, CHR-TRM showed the largest discrepancy at smaller training sizes (UNIFORM 14.83% vs WEIGHTED 7.42% at 100 training; UNIFORM 42.69% vs WEIGHTED 30.28% at 300 training).

[1] Accuracies are reported as percent correct for readability.

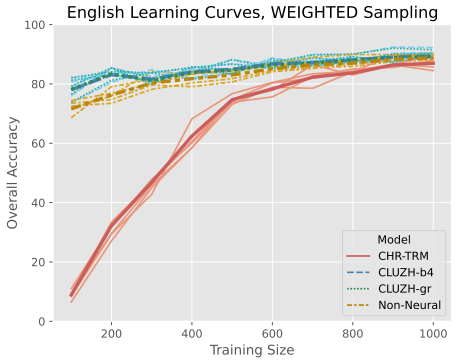

Figure 3: Learning curves for English verbs. Thin lines = individual seeds and thick lines = averages across seeds.

Variation across random seeds

Our analysis thus far attends to accuracy scores averaged across random seeds. This raises the question: how much variability is there in model performance across random seeds? Given the five random seeds of each unique combination of different languages, sampling strategies, and model types, we calculated two metrics, which we refer to hereafter as *score range* and *random seed variability*. The former measures the difference between the lowest and the highest accuracy, while the latter computes the standard deviation of the accuracy scores across the 5 random seeds.

We first analyzed the score variation across random seeds for each language. Across sampling strategy, training sizes and model types, we found the average score ranges for Arabic (6.26%) and German (5.57%) to be higher than that for English (5.14%); the results of average random seed variability for the three languages follow the same trend (Arabic: 2.49%; German: 2.25%; English: 2.06%).

We then analyzed the score range for each sampling strategy. Overall, UNIFORM led to slightly a higher score range (5.75%) compared to WEIGHTED (5.59%); on the other hand, the average random seed variability for the two sampling strategies is comparable (UNIFORM: 2.28%; WEIGHTED: 2.26%). However, there appears to be more variable results when looking at the accuracy scores for individual languages when combined with different sampling strategies. For Arabic, UNIFORM and WEIGHTED yielded very close average score range values (6.33% vs 6.19%), as well as mean random seed variability values (2.51% vs. 2.47%). By contrast, for German, WEIGHTED led to more variable score values; whereas UNIFORM resulted in a higher mean score range and average random seed variability for English.

Lastly, we studied the relationship between training sizes and score variation, specifically, whether larger training sizes would lead to less model performance variation and thereby more reliable evaluation results. To address that, we again utilized linear regression analysis. We fit two models here. For both models, we included the training size as a fixed effect, controlling for the effects of language, model types, and sampling strategies (with interaction terms between all aforementioned fixed effects); one model predicts the score range value, while the other predicts the random seed variability:

SCORE RANGE/SEED VARIABILITY $\sim$
LANG * MODEL * TRAIN SIZE * SAMPLING STRAT

Based on results from the regression models, it seems that the training size has pronounced negative effects on score range (β=-0.007, $p < 0.05$) as well as random seed variability (β=-0.003, $p < 0.05$). That said, these coefficient values are quite small; this suggests that although as training size gets larger, there is a tendency for less variable model performance on average, this tendency is relatively weak.

Linguistic Analysis

For a model to be cognitively plausible, it should not only achieve high performance but also learn in a similar way to children. This section evaluates the neural models' outputs in terms of how well they reproduce observed patterns in child language acquisition.

Arabic Noun Pluralization

Arabic nouns form plurals in two ways: by suffixation (*sound plurals*) or by stem mutation (*broken plurals*). There are two sound plural suffixes in the nominative, feminine (FEM) *-āt*, and masculine (MASC) *-ūn*. The relationship between gender and sound plural ending is generally reliable, but some (generally non-human) MASC nouns take the FEM sound plural.

Broken plurals can be divided into many subclasses by which templatic pattern defines the output of their stem mutations. In Modern Standard Arabic (MSA), there are approximately 30 broken plural patterns (J. J. McCarthy & Prince, 1990), though the exact count depends on the level of abstraction assumed for the templatic pattern.

We annotated the predictions of the best performing models of seed 0 for training sizes 200, 400, 600, 800, and 1,000. The annotation made a distinction between the two sound plurals (FEM and MASC) to detect sound-to-sound errors `Snd`→`Snd`, while broken plurals are annotated as `Br` regardless of the plural pattern. This level of granularity was adopted from Dawdy-Hesterberg and Pierrehumbert (2014).

Ravid and Farah (1999) identify two kinds of *u*-shaped learning in Palestinian Arabic-learning children: **(1)** they begin by accurately distinguishing MASC and FEM sound plurals followed by over-application of the FEM to MASC forms before returning to high accuracy. **(2)** they also go through a period of over-applying FEM sound plurals to what should be broken plurals. There are very few instances of FEM-to-MASC sound and broken-to-broken errors.

Figure 4 provides a breakdown of error types at each annotated training size for CLUZH-B4, the best overall performing model type, in the style of Dawdy-Hesterberg and Pierrehumbert (2014). We make some key observations. First, learning is monotonic. Neither type of *u*-shaped learning is observed. Second, `Br`→`Snd` errors are indeed relatively common. Third, however, most errors are `Br`→`Br` (e.g., *nabiyy-*nab* instead of *nabiyy-ʔanbiyāʔ*) or `Snd`→`Br` (e.g., *maʃrūb-*maʃārīb* instead

of *maʃrūb-maʃrūbāt*), which are very rare developmentally, while Snd→Snd errors are quite rare in CLUZH-B4's productions even though they dominate developmentally. Across all annotated predictions, 32 Snd→Snd are MASC→FEM, while 20 are the reverse. FEM→MASC errors are proportionately far more common than what is attested developmentally.

An error analysis of the CLUZH-GR and CHR-TRM uncovered qualitatively similar patterns. Overall, the ANNs make errors that are not similar to children's errors.

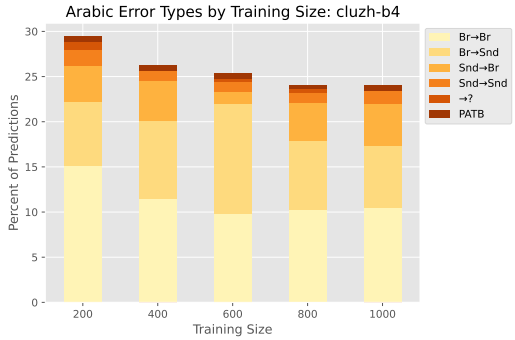

Figure 4: Breakdown of CLUZH-B4's Arabic errors across training sizes. Color indicates error type. PATB are due to annotation errors in the original data. ? are nonsense outputs.

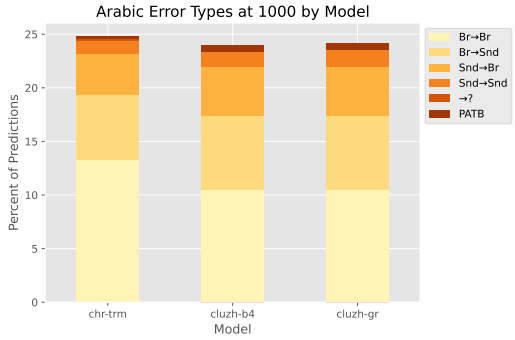

Figure 5: Breakdown of each neural model's Arabic errors at maximum training. Color indicates error type.

English Past Tense

For English, we analyzed predictions made by all models on the full training set (Figure 7) and predictions of the CLUZH-B4 model trained on each nested set in increments of 100 (Figure 6); all analyses were done on seed 0. Model predictions were classified into three types: -(e)d (regular, e.g., *smear-smeared*), irreg (irregular or analogized to irregular, e.g., *fly-flew, tweet-*twet*), or ? (unnatural predictions, e.g., *correspond-*correspood*). Because the goal of this task was not to capture the idiosyncrasies of English orthography, near-correct regular predictions such as *obey-obeid* or *trim-trimed* were annotated as -(e)d.

On the full training set, both CLUZH-GR and CLUZH-B4 make only over-regularization and over-irregularization errors, with over-regularizations dominating. By contrast, over-regularizations are in the minority for CHR-TRM, which

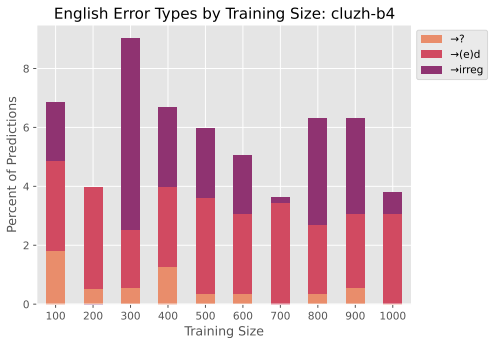

Figure 6: Breakdown of CLUZH-B4's English errors at each training size. Color indicates error type.

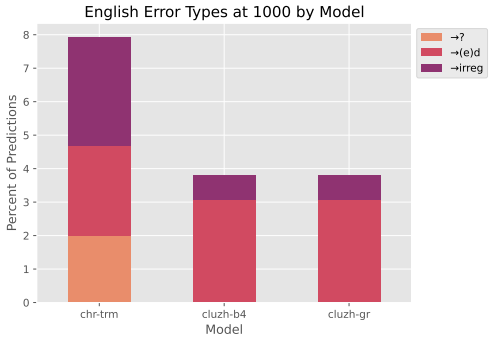

Figure 7: Breakdown of each neural model's English errors at maximum training. Color indicates error type.

produces a greater proportion of unnatural errors and over-irregularizations. Since children produce orders of magnitude more over-regularization than over-irregularization errors (Marcus et al., 1992; Xu & Pinker, 1995), no model demonstrates a human-like error distribution.

Further, the dominance of over-regularizations is not consistent across training sizes for CLUZH-B4 (Figure 6): both the overall error rate and the distribution of errors oscillate significantly, and the rates of over-irregularization and unnatural errors are generally much higher than expected. Additionally, children often exhibit *u*-shaped learning when they learn that *-(e)d* is productive and liberally overapply it to irregular verbs (Marcus et al., 1992; Prasada & Pinker, 1993). Though CLUZH-B4 exhibits a temporary spike in error rate at 300 words (Figure 6), this spike is caused by an increase in *over-irregularization* rather than over-regularization. Thus, the ANNs do not fit well with developmental findings: they fail to exhibit child-like error distributions, and the best performing model does not exhibit developmental regression.

German Noun Pluralization

German nouns are inflected with one of five (nominative and accusative) plural patterns: *-(e)n*, the most frequent, especially among FEM nouns, *-e*, the second most frequent, *-∅*, *-(e)r*, and *-s*, the least frequent. This is interesting because *-s* appears to be the default form of last resort despite its low frequency (Marcus et al., 1995). Thus, German noun plu-

rals provide a way to distinguish productivity from frequency, something that the English past tense cannot do.

To investigate whether the models generalized *-s*, plurals were annotated to indicate which of the five pluralization types is applied: `-e`, `-(e)s`, `-(e)r`, `-(e)n`, or `-0`. Since the goal here was not to perfectly capture German orthography, failure to capture consonant doubling (*-inen* instead of expected *-innen*) was not penalized. Model predictions that did not fit into any of these categories (usually due to unnatural stem-internal changes) were labeled *?*. Figure 8 breaks down error types by training size on seed 0 for CLUZH-B4 the overall best performing model. At 200 and above, the over-application of *-e* is consistently the plurality or majority error type. Developmentally, early application of *-e* is consistent with child development, where *-e* and *-∅* are acquired very early (Gawlitzek-Maiwald, 1994).

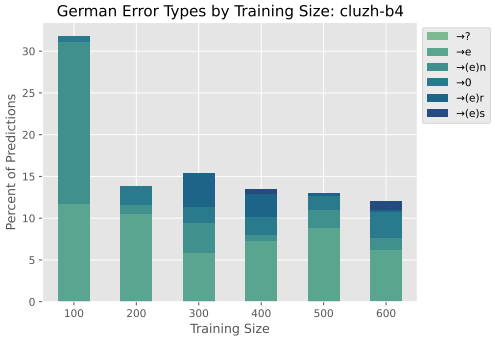

Figure 8: Breakdown of CLUZH-b4's German errors at each training size. Color indicates error type.

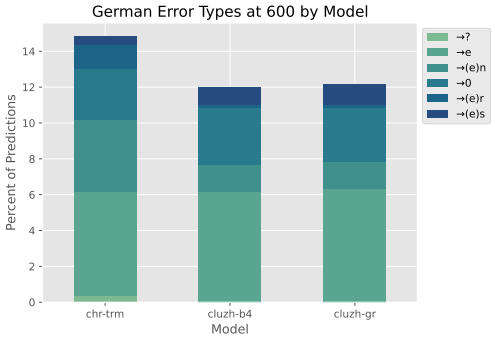

Figure 9: Breakdown of each neural model's German errors at maximum training. Color indicates error type.

Over-application of *-(e)n* dominates at 100 training items then falls off. It is the default ending for FEM nouns, and CLUZH-B4 applies it to FEM near-categorically and often incorrectly applies it to MASC and neuter NEUT nouns. It also over-extends the *-s* plural for a time, particularly around 300-400 items in training, which is consistent with developmental expectations (Elsen, 2002). Figure 9 shows each neural model's error types at maximum training on seed 0. For each of the three models, over-application of *-e* is dominant followed by *-(e)n* and *-∅*. The error rate is unacceptably high, but learning trajectories are roughly what would be expected from a cognitively-plausible learning model.

Discussion

Our quantitative analysis shows that UNIFORM sampling tends to inflate performance relative to more realistic WEIGHTED sampling, thus we advocate for adopting weighted sampling when possible to evaluate artificial neural networks (ANNs). Choice of random seed yielded score ranges as high as 6 percentage points, which is substantial. Evaluation across several random seeds highlights the sensitivity of a model to sampling effects, where low sensitivity is likely more cognitively plausible, consistent with the observation that children show consistent development despite each having received unique input samples.

Our linguistic analysis focuses on how well models' learning trajectories and error types correspond to observed human learning. Results on German were promising. Productive inflection patterns emerged in roughly the expected order, and CLUZH-B4 demonstrates over-application of *-s*. Results for Arabic and English were less promising. Models did not show patterns of *u*-shaped learning which are robustly attested in child learners, nor did they reproduce attested strong asymmetries between different error types.

Future studies of this type may be repeated with phonological transcriptions rather than orthography, especially for English. It is unclear what effect orthography has on performance. In one sense, irregularities in English spelling add complexity. However, spelling also collapses some spoken distinctions such as the three-way phonologically conditioned allomorphy of *-(e)d* (/-t/, /-d/, /-əd/), thus decreasing chances for error. Taken together, this work demonstrates that our conclusions about ANNs as cognitive models of morphology are sensitive to the often implicit design decisions made during evaluation. We hope that this line of work will be extended to more models, languages, and morphological phenomena.

Acknowledgments

S.P. gratefully acknowledges funding through the Institute for Advanced Computational Science (IACS) Graduate Research Fellowship and the National Science Foundation (NSF) Graduate Research Fellowship Program under NSF Grant No. 2234683. Some experiments were performed on the SeaWulf HPC cluster maintained by RCC, and IACS at Stony Brook University and made possible by NSF grant No. 1531492. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IACS or the NSF.

References

Aksu-Koç, A. A. (1985). The Acquisition of Turkish. *The Cross-linguistic Studies of Language Acquisition. Vol. 1: The Data*, 839–876.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium.

Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., ... Vylomova, E. (2022, June). UniMorph 4.0: Universal Morphology. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 840–855). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.89

Belth, C., Payne, S., Beser, D., Kodner, J., & Yang, C. (2021). The Greedy and Recursive Search for Morphological Productivity. In *Proceedings of the 43th annual conference of the cognitive science society* (p. 2869-2875).

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2-3), 150–177. doi: 10.1080/00437956.1958.11659661

Beser, D. (2021). Falling Through the Gaps: Neural Architectures as Models of Morphological Rule Learning. In *Proceedings of the 43th annual conference of the cognitive science society* (p. 1042-1048).

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., ... Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, *75*(4), 1115–1139.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press. doi: 10.4159/harvard.9780674732469

Chan, E. (2008). *Structures and distributions in morphological learning*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Clahsen, H., & Rothweiler, M. (1993). Inflectional rules in children's grammars: Evidence from German participles. In *Yearbook of morphology 1992* (pp. 1–34). Springer.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019, July). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3868–3877). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1376 doi: 10.18653/v1/P19-1376

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., ... Hulden, M. (2018, October). The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection* (pp. 1–27). Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K18-3001 doi: 10.18653/v1/K18-3001

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., ... Hulden, M. (2017, August). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 shared task: Universal morphological reinflection* (pp. 1–30). Vancouver: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K17-2001 doi: 10.18653/v1/K17-2001

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016, August). The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology* (pp. 10–22). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W16-2002 doi: 10.18653/v1/W16-2002

Dankers, V., Langedijk, A., McCurdy, K., Williams, A., & Hupkes, D. (2021, November). Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network. In *Proceedings of the 25th conference on computational natural language learning* (pp. 94–108). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.conll-1.8 doi: 10.18653/v1/2021.conll-1.8

Dawdy-Hesterberg, L. G., & Pierrehumbert, J. B. (2014). Learnability and generalisation of Arabic broken plural nouns. *Language, cognition and neuroscience*, *29*(10), 1268–1282.

Deen, K. U. (2005). *The Acquisition of Swahili* (Vol. 40). John Benjamins Publishing.

Elsen, H. (2002). The acquisition of German plurals. In *Morphology 2000: Selected papers from the 9th morphology meeting, vienna, 25-27 february 2000* (Vol. 218, p. 117).

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Gawlitzek-Maiwald, I. (1994). How do children cope with variation in the input? The case of German plurals and compounding. In R. Tracy & E. Lattey (Eds.), *How tolerant is Universal Grammar? essays on language learnability and language variation* (p. 225-266). Tübingen: Niemeyer.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531. doi: 10.1017/s0305000907008641

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Kodner, J., & Khalifa, S. (2022, July). SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. In *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonol-*

ogy, and morphology (pp. 157–175). Seattle, Washington: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.sigmorphon-1.18

Kodner, J., Khalifa, S., Batsuren, K., Dolatian, H., Cotterell, R., Akkus, F., ... Vylomova, E. (2022, July). SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 176–203). Seattle, Washington: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.sigmorphon-1.19

Lignos, C., & Yang, C. (2018). Morphology and language acquisition. *Cambridge handbook of morphology*, 765–791.

Liu, Z., & Prud'hommeaux, E. (2022). Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation. *Transactions of the Association for Computational Linguistics*, *10*, 393–413.

Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004). The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *Nemlar conference on arabic language resources and tools* (Vol. 27, pp. 466–467).

MacWhinney, B. (2000). *The CHILDES Project: The Database* (Vol. 2). Abingdon-on-Thames: Psychology Press. doi: 10.1162/coli.2000.26.4.657

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, *29*(3), 189–256.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*.

McCarthy, A., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., ... others (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* (pp. 229–244).

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., ... Yarowsky, D. (2020, May). UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th language resources and evaluation conference* (pp. 3922–3931). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.483

McCarthy, J. J., & Prince, A. S. (1990). Foot and Word in Prosodic Morphology: The Arabic Broken Plural. *Natural Language & Linguistic Theory*, *8*, 209-283.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*(11), 465-472.

McCurdy, K., Goldwater, S., & Lopez, A. (2020, July). Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1745–1756). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.159 doi: 10.18653/v1/2020.acl-main.159

Pimentel, T., Ryskina, M., Mielke, S. J., Wu, S., Chodroff, E., Leonard, B., ... Vylomova, E. (2021, August). SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 229–259). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.sigmorphon-1.25 doi: 10.18653/v1/2021.sigmorphon-1.25

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193. doi: 10.1016/0010-0277(88)90032-7

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463. doi: 10.1016/s1364-6613(02)01990-3

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and cognitive processes*, *8*(1), 1–56.

Ravid, D., & Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language*, *19*(56), 187–206.

Rumelhart, D. E., & McClelland, J. L. (1986). *On learning the past tenses of English verbs*. Cambridge, MA: MIT Press.

Seidenberg, M. S., & Plaut, D. (2014). Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate. *Cognitive science*, *38 6*, 1190-228.

Szagun, G., Steinbrink, C., Franik, M., & Stumper, B. (2006). Development of vocabulary and grammar in young German-speaking children assessed with a German language development inventory. *First Language*, *26*(3), 259-280.

Vylomova, E., White, J., Salesky, E., Mielke, S. J., Wu, S., Ponti, E. M., ... Hulden, M. (2020, July). SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 1–39). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.sigmorphon-1.1 doi: 10.18653/v1/2020.sigmorphon-1.1

Wehrli, S., Clematide, S., & Makarov, P. (2022, July). CLUZH at SIGMORPHON 2022 Shared Tasks on Morpheme Segmentation and Inflection Generation. In *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 212–219). Seattle, Washington: Association for Computational Linguistics. Retrieved from

https://aclanthology.org/2022.sigmorphon-1.21
doi: 10.18653/v1/2022.sigmorphon-1.21

Wiemerslage, A., Dudy, S., & Kann, K. (2022, December). A Comprehensive Comparison of Neural Networks as Cognitive Models of Inflection. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1933–1945). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.emnlp-main.126

Wu, S., Cotterell, R., & O'Donnell, T. (2019, July). Morphological Irregularity Correlates with Frequency. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5117–5126). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1505
doi: 10.18653/v1/P19-1505

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*(3), 531-556.