A Comprehensive Comparison of Neural Networks as Cognitive Models of Inflection

Adam Wiemerslage and **Shiran Dudy** and **Katharina Kann**
University of Colorado Boulder
`first.last@colorado.edu`

Abstract

Neural networks have long been at the center of a debate around the cognitive mechanism by which humans process inflectional morphology. This debate has gravitated into NLP by way of the question: Are neural networks a feasible account for human behavior in morphological inflection? We address that question by measuring the correlation between human judgments and neural network probabilities for unknown word inflections. We test a larger range of architectures than previously studied on two important tasks for the cognitive processing debate: English past tense, and German number inflection. We find evidence that the Transformer may be a better account of human behavior than LSTMs on these datasets, and that LSTM features known to increase inflection accuracy do not always result in more human-like behavior.

1 Introduction: The Past Tense Debate

Morphological inflection has historically been a proving ground for studying models of language acquisition. Rumelhart and McClelland (1985) famously presented a neural network that they claimed could learn English past tense inflection. However, Pinker and Prince (1988) proposed a dual-route theory for inflection, wherein regular verbs are inflected based on rules and irregular verbs are looked up in the lexicon. They highlighted several shortcomings of Rumelhart and Mc-Clelland (1985) that they claimed any neural network would suffer from.

This opened a line of work wherein cognitive theories of inflection are analyzed by implementing them as computational models and comparing their behavior to that of humans. A famous study in the area of morphology is the *wug test* (Berko, 1958), where human participants are prompted with a novel-to-them nonce word and asked to produce its plural form. Similarly, morphological inflection models are generally evaluated on words they have
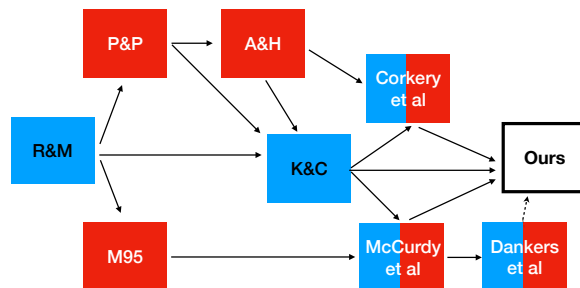

Figure 1: Summary of the past tense debate as it pertains to this work, color coded by evidence for (blue) or against (red) neural networks as a cognitively plausible account for human behavior.

not seen during training. However, since they are evaluated on actual words, it is impossible to meaningfully ask a native speaker, who knows the words' inflected forms, how likely different reasonable inflections for the words in a model's test data are. Thus, in order to compare the behavior of humans and models on words unknown to both, prior work has created sets of made-up nonce words (Marcus et al., 1995; Albright and Hayes, 2003).

English Past Tense English verbs inflect to express the past and present tense distinction. Most verbs inflect for past tense by applying the /-d/, /-ɪd/, or /-t/ suffix: allophones of the regular inflection class. Some verbs, however, express the past tense with a highly infrequent or completely unique inflection, forming the irregular inflection class. This distinction between regular and irregular inflection has motivated theories like the dual-route theory described above.

Prasada and Pinker (1993) performed a wug test for English past tense inflection in order to compare the model from Rumelhart and McClelland (1985) to humans with special attention to how models behave with respect to regular vs. irregular forms, finding that it could not account for human generalizations. Albright and Hayes (2003, A&H) gathered production probabilities – i.e., the normal-

ized frequencies of the inflected forms produced by participants – and ratings – i.e., the average rating assigned to a given past tense form on a well-formedness scale. They then implemented two computational models: a rule-based and an analogy-based model and computed the correlation between the probabilities of past tense forms for nonce verbs under each model and according to humans. They found that the rule-based model more accurately accounts for nonce word inflection.

After several years of progress for neural networks, including state-of-the-art results on morphological inflection (Kann and Schütze, 2016; Cotterell et al., 2016), this debate was revisited by Kirov and Cotterell (2018, K&C), who examined modern neural networks. They trained a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) with attention (Bahdanau et al., 2015) on English past tense inflection and in experiments quantifying model accuracy on a held out set of real English verbs, they showed that it addresses many of the shortcomings pointed out by Pinker and Prince (1988). They concluded that the LSTM is, in fact, capable of modeling English past tense inflection. They also applied the model to the wug experiment from A&H and found a positive correlation with human production probabilities that was slightly higher than the rule-based model from A&H.

Corkery et al. (2019, C&al.) reproduced this experiment and additionally compared to the average human rating that each past tense form received in A&H's dataset. They found that the neural network from K&C produced probabilities that were sensitive to random initialization – showing high variance in the resulting correlations with humans – and typically did not correlate better than the rule-based model from A&H. They then designed an experiment where inflected forms were sampled from several different randomly initialized models, so that the frequencies of each form could be aggregated in a similar fashion to the adult production probabilities – but the results still favored A&H. They hypothesized that the model's overconfidence in the most likely inflection (i.e. the regular inflection class) leads to uncharacteristically low variance on predictions for unknown words.

German Noun Plural McCurdy et al. (2020a, M&al.) applied an LSTM to the task of German noun plural inflection to investigate a hypothesis from Marcus et al. (1995, M95), who attributed the outputs of neural models to their susceptibility to the most frequent pattern observed during training, stressing that, as a result, neural approaches fail to learn patterns of infrequent groups.

German nouns inflect for the plural and singular distinction. There are five suffixes, none of which is considered a regular majority: /-(e)n/, /-e/, /-er/, /-s/, and /-∅/. M95 had built a dataset of monosyllabic German noun wugs and investigated human behavior when inflecting the plural form, distinguishing between phonologically familiar environments (rhymes), and unfamiliar ones (non-rhymes). The German plural system, they argued, was an important test for neural networks since it presents multiple productive inflection rules, all of which are minority inflection classes by frequency. This is in contrast to the dichotomy of the regular and irregular English past tense. M&al. collected their own human production probabilities and ratings for these wugs, and then compared those to LSTM productions. Humans were prompted with each wug with the neuter determiner to control for the fact that neural inflection models of German noun plurals are sensitive to grammatical gender (Goebel and Indefrey, 2000), and because humans do not have a majority preference for monosyllabic, neuter nouns (Clahsen et al., 1992).

The /-s/ inflection class, which is highly infrequent appears in a wide range of phonological contexts, which has lead some research to suggest it is the *default* class for German noun plurals, and thus the regular inflection, despite its infrequent use. M&al. found that it was preferred by humans in Non-Rhyme context more than Rhymes, but the LSTM model showed the opposite preference, undermining the hypothesis that LSTMs model human generalization behavior. /-s/ was additionally predicted less accurately on a held-out test set of real noun inflections when compared to other inflection classes.

They found that the most frequent inflection class in the training for the monosyllabic neuter contexts, /-e/, was over-generalized by the LSTM when compared to human productions. The most frequent class overall, /-(e)n/ (but infrequent in the neuter context), was applied by humans quite frequently to nonce nouns, but rarely by the LSTM. They additionally found that /-er/, which is as infrequent as /-s/, could be accurately predicted in the test set, and the null inflection /-∅/, which is generally frequent, but extremely rare in the monosyllabic, neuter setting was never predicted for the

wugs. We refer to McCurdy et al. (2020a) for more details on the inflection classes and their frequencies, and additional discussion around their relevance to inflection behavior.

Ultimately, M&al. reported no correlation with human production probabilities for any inflection class. They concluded that modern neural networks still simply generalize the *most frequent* patterns to unfamiliar inputs.

Dankers et al. (2021) performed in-depth behavioral and structural analyses of German noun plural inflection by a *unidirectional* LSTM *without* attention. They argued that these modeling decisions made a more plausible model of human cognition. In a behavioral test they found that, like humans but unlike M&al., their model did predict /-s/ more for non-rhymes than for rhymes, but the result was not statistically significant. They also found that /-s/ was applied with a high frequency and attributed this to sensitivity to word length. For a visual of all studies discussed in this section, see Figure 1.

Our Contribution   Most work on modern neural networks discussed here analyzes the same bidirectional LSTM with attention and draws a mixture of conclusions based on differing experimental setups. Dankers et al. (2021) changed the LSTM-based architecture, and found somewhat different results for German number inflection, though they did not investigate correlations with human ratings nor production probabilities in the same way as previous work. The limited variation of architectures in previous studies as well as inconsistent methods of comparison with human behavior prevent us from drawing definite conclusions about the adequacy of neural networks as models of human inflection.

Here, we present results on a wider range of LSTMs and a Transformer (Vaswani et al., 2017) model for both English past tense and German number inflection. We ask which architecture is the best account for human inflection behavior and, following M&al., investigate the actual model productions (and probabilities) for the German plural classes in order to qualitatively compare to human behavior. We additionally ask how architectural decisions for the LSTM encoder-decoder affect this correlation. Finally, we investigate the relationship between inflection *accuracy* on the test set and *correlation* with human wug ratings.

We find that the Transformer consistently correlates best with human ratings, producing probabilities that result in Spearman's ρ in the range of

0.47-0.71 for several inflection classes, which is frequently higher than LSTMs. However, when looking closely at the Transformer productions, it displays behavior that deviates from humans similarly to the LSTM in M&al., though to a lesser extent. While attention greatly increases LSTM *accuracy* on inflection, we also find that it does not always lead to better correlations with human wug ratings, and that the directionality of the encoder has more complicated implications. Finally, we find that there is no clear relationship between model *accuracy* and *correlation* with human ratings across all experiments, demonstrating that neural networks can solve the inflection task in its current setup without learning human-like distributions. While the Transformer experiment in this work demonstrates stronger correlations with human behavior, and *some* more human-like behaviors than before, our findings continue to cast doubt on the cognitive plausibility of neural networks for inflection.

2   Neural Morphological Inflection

2.1   Task Description

The experiments in this paper are centered around a natural language processing (NLP) task called morphological inflection, which consists of generating an inflected form for a given lemma and set of morphological features indicating the target form. It is typically cast as a character-level sequence-to-sequence task, where the characters of the lemma and the morphological features constitute the input, while the characters of the target inflected form are the output (Kann and Schütze, 2016):

$$\text{PST} \quad \text{c r y} \rightarrow \text{c r i e d}$$

Formally, let $\mathcal{S}$ be the paradigm slots expressed in a language and l a lemma in the language. The set of all inflected forms – or *paradigm* – π of l is then defined as:

$$\pi(l) = \left\{ \left(f_k[l], t_k \right) \right\}_{k \in \mathcal{S}} \tag{1}$$

$f_k[l]$ denotes the inflection of l which expresses tag t_k, and l and $f_k[l]$ represent strings consisting of letters from the language's alphabet Σ.

The task of morphological inflection can then formally be described as predicting the form $f_i[l]$ from the paradigm of l corresponding to tag t_i.

2.2   Models

Rumelhart and McClelland   The original model of Rumelhart and McClelland (1985)

preceded many of the features introduced by modern neural networks. For example, they use a feed-forward neural network to encode input sequences. This creates the requirement of coercing variable-length inputs into the fixed-size network. To solve this, they encode input words as fixed length vectors representing the phonological distinctive feature sets for each trigram in that word. The neural network is then trained to map the features of an input form to a feature vector of a hypothesized output form. The loss is computed between the input feature sets and the the feature set for an inflected output form encoded in the same way. At test time, they manually select candidate output forms for each input lemma in order to overcome the intractable decoding problem. The output form, then, is the candidate whose feature vector most closely resembles the model output. Beyond decoding problems, the order of input characters is not encoded, and unique words are represented with potentially identical phonological features.

LSTM The LSTM architecture (Hochreiter and Schmidhuber, 1997) overcomes several of the issues in Rumelhart and McClelland (1985), by way of a recurrent encoding and decoding mechanism, and reliance on character embeddings.

We experiment with several variations of the LSTM encoder-decoder (Sutskever et al., 2014; Cho et al., 2014) to test their behavior compared to humans. First, we vary *directionality* of the encoder under the assumption that bidirectional encoding leads to higher accuracy, but a unidirectional encoder may better resemble human processing. We additionally vary whether or not *attention* is used. Attention is typically a crucial feature to attaining high inflection accuracy. We expect that the same may also be true for assigning a cognitively plausible probability to a nonce inflection, by supplying the model with a mechanism to focus on only the relevant parts of the inflection.

This yields 4 LSTM-based variations. We refer these models as BiLSTMAttn (BA; from K&C, C&al., and M&al.), UniLSTMAttn (UA), BiLSTMNoAttn (BN), and UniLSTMNoAttn (UN; from Dankers et al. (2021)).

Transformer Finally, we present results for a Transformer sequence-to-sequence model (Vaswani et al., 2017), following the implementation proposed for morphological inflection by

Wu et al. (2021). Unlike LSTM-based models, the transformer employs a *self-attention* mechanism such that each character representation can be computed in parallel as a function of all other characters. The position of each character is encoded with a special positional embedding. This means that the relation between each character in a word can be represented directly, rather than through a chain of functions via the LSTM recurrence. It is considered to be state-of-the-art for morphological inflection in terms of accuracy, which makes it an important comparison for this study. Some work has called into question the cognitive plausibility of transformer self-attention in psycholinguistic experiments of word-level language models (Merkx and Frank, 2020) – claiming that the direct access it provides to past input is cognitively implausible. It is not clear though that these arguments apply to *character-level* models for inflection, wherein words do not necessarily need to be processed one character at a time.

Hyperparameters We implement all LSTMs with pytorch (Paszke et al., 2019) and borrow hyperparameters from previous work on morphological inflection. For the LSTMs, we use the hyperparameters from K&C, which were based on the tuning done by Kann and Schütze (2016). For the Transformer, we follow the hyperparameters from the best model in Wu et al. (2021), but set label-smoothing to 0. In preliminary experiments, we found no significant impact of label smoothing on accuracy nor correlation with human behavior across inflection classes.

For all architectures, we follow C&al. and train 10 randomly initialized models. At test time, we decode with beam search with a width of 12. We train for up to 50 epochs because the architectures with fewer parameters tend to converge more slowly.

MGL A&H implement the Minimal Generalization Learner (MGL), which learns explicit rules (e.g. insertion of /-ɪd/ if a verb ends in a /t/ or /d/) at varying levels of granularity. Each rule is associated with a confidence score for a given phonological environment based on its statistics in the train set. At test time, the rule with the highest confidence is applied to produce an inflection, and the confidences can be used to score various regular or irregular inflected forms. We compare to this model for English data, following previous work.

	Dev Acc	Test Acc		Prod. Prob.		Rating	
		reg	irreg	reg	irreg	reg	irreg
A&H MGL	-	**99.7**	**38.0**	.33	.30	.50	.49
K&C*	-	98.9	28.6	**.48**	.45	-	-
C&al. Agg.**	-	-	-	.45	.19	.43	.31
BiLSTMAttn	93.33	97.48 (.65)	9.05 (5.24)	.28	.36	.16	.46
BiLSTMNoAttn	76.37	82.72 (2.06)	7.62 (3.33)	.14	.44	.23	.35
UniLSTMAttn	92.45	96.53 (.68)	20.00 (4.38)	.35	.41	.40	.32
UniLSTMNoAttn	73.49	77.72 (1.64)	10.48 (10.24)	.22	.43	.28	.34
Transformer	**94.88**	99.21 (.53)	10.95 (11.46)	.38	**.47**	**.58**	**.58**

Table 1: English results for both regular (reg) and irregular (irreg) inflections for all architectures and metrics. Along with accuracy, we report Spearman's ρ between average model rating and our two human metrics. Standard deviations are given in parentheses.
*Trained and tested a different random split, **Trained and tested on all training data

3 Experiments

3.1 Languages and Data

We use the same data as previous work on English past tense, and German number inflection.

English We experiment with the English past tense data from A&H, following both K&C and C&al. For training, we split the CELEX (Baayen et al., 1996) subset produced by A&H, consisting of 4253 verbs (218 irregular), into an 80/10/10 random train/dev/test split following K&C.[1] We ensure that 10% of the irregular verbs are in each of the development and test set.

The English nonce words from A&H, used for computing the correlation of model rating with human ratings and production probabilities, comprise 58 made-up verb stems, each of which has 1 regular and 1 irregular past tense inflection. 16 verbs have an additional irregular form (58 regulars and 74 irregulars total). All English data is in the phonetic transcription provided by A&H.

German We also experiment with the German dataset from McCurdy et al. (2020a), who released train/dev/test splits consisting of 11,243 pairs of singular and plural nouns in the nominative case taken from UniMorph (McCarthy et al., 2020). They added gender, the only inflection feature provided, by joining UniMorph with a Wiktionary scrape.

The German wugs come from M95, who built a set of 24 monosyllabic nonce nouns: 12 of which are rhymes – resembling real words in their phonology, and 12 of which are non-rhymes – representing atypical phonology. Human ratings and production probabilities, however, are taken from M&al., who

administered an online survey to 150 native German speakers. Each participant was prompted with the nouns from M95 with the neuter determiner, and then asked to generate the plural form. Similar to A&H, after producing a plural for each noun, participants were asked to rate the acceptability of each potential plural form on a 1-5 scale. In their analysis, M&al. compare human and model behavior on 5 productive inflection classes, shown for our experiments in Table 3.

3.2 Evaluation Metrics

We evaluate models with respect to four metrics.

Accuracy This refers to raw accuracy on a set of real inflections that the model has not seen during training. Crucially, only the top prediction of a given model is considered, and the model's probability distribution over all predictions does not affect the score.

F1 We report F1 instead of accuracy for the German plural experiments following M&al. Here we classify each inflected form with its suffix (e.g. /-s/), and classify inflections that do not conform to the 5 inflection classes from M&al. as "other."

Production Probability Correlation Like previous work (Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020a), we compare model output probabilities with production probabilities from humans. The production probability of a form is calculated by counting all forms produced for a given lemma, and then normalizing them to obtain a probability distribution of the human productions. In keeping with most previous work and because we do not expect a linear relationship with the model ratings, we report Spearman's ρ. This is calculated *within* each inflection class,

---

[1]K&C use a subset of A&H, removing 164 regulars and 50 irregulars. We include them in our dataset.

meaning that, e.g., for English we report a regular and an irregular ρ. For example, the regular ρ for the set of lemmas {rife, drize, flidge} would be computed from the vector containing probabilities of the forms {rifed, drized, flidged} under the model, against the corresponding vector with human probabilities.

Rating Correlation   Finally, we compare model ratings to the average human rating of each form, again reporting ρ within inflection class. Here, rather than normalizing over production frequencies, humans were prompted with an inflection for a given lemma and asked to rate it on a scale that differed slightly between datasets. For each lemma, we thus get an average probability for a regular form, as well as for an irregular form.

3.3   Neural Network Wug Test

In order to compare our models to humans, we compute analogous values to the human ratings and production probabilities. We investigate two strategies: normalizing the inflected form counts output by our models, and computing the average probability of each form under our models.

Model Production Probability   Previous work (Corkery et al., 2019; McCurdy et al., 2020a) decoded outputs from multiple models and aggregated the resulting forms: given a lemma and a set of n models trained with different random seeds, an inflected form is sampled from each model, resulting in forms $f_1, ..., f_n$, where forms need not be unique. The frequency of each form is then normalized to obtain a probability distribution. For example, given the nonce lemma *rife*, the probability of the past tense form *rifed* is computed as

$$\frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1, & \text{if } f_i = \text{rifed} \\ 0, & \text{otherwise} \end{cases}$$

C&al. propose a version of this in their aggregate model, in which they sample 100 forms from each model , and normalize the resulting form frequencies. M&al., who instead train 25 randomly initialized models, perform the same aggregation over the *top* prediction of each model. We take the approach of M&al. (though we train only 10 models) to investigate model productions qualitatively. This metric is intuitively similar to quantifying human production probabilities if we consider one model to be one human participant.

Model Rating   Because the aggregate outputs method considers only the *most likely* prediction aggregated over the same architecture trained on the same dataset, we expect the prediction to typically be the same for each model. We instead report correlations with the *probability* of inflected forms under each model in Tables 1 and 3. K&C correlate this value with human production probabilities, and C&al. use this method in an experiment to compute individual model ratings.

More formally, given a lemma l and an inflected form f of length k, we compute

$$p(f \mid l) = p(f_1, ..., f_k \mid l) \tag{2}$$

$$= \prod_{1}^{k} p(f_i | f_{i-1}, l) \tag{3}$$

Where f_i is the ith character of f. We force the model to output each inflected form f to get its probability. In practice, we modify Equation 3 to compute a length-normalized probability because $p(f \mid l)$ becomes smaller as f increases in length. For f of length k, we have

$$p(f \mid l) = \sqrt[k]{\prod p(f_i | f_{i-1}, l)} \tag{4}$$

We expect computing ratings in this way to be similar to the aggregate model of C&al. described above. That is, the probability of a form f computed by aggregating n forms from a single model's probability distribution should approach $p(f \mid l)$, as $n \rightarrow \infty$. Finally, we compute the average probability of a form from all 10 randomly initialized models, and refer to it as the model rating.

4   Results

We present experimental results in Tables 1, 2, and 3 in terms of both inflection accuracy, and correlation with human behavior – our main focus. All correlations for neural models trained in this work are given with respect to **model rating**, and not the model production probability. We report results from training MGL on our data, and include the results reported K&C, C&al., and M&al. in the appropriate tables for reference.

4.1   English

For English, many of our models correlate better for irregulars than regulars, unlike previous work for which the strongest correlations occurred for regular verbs. As we do not have the same train

	Dev Acc.	Test F1					
		/-(e)n/	/-e/	/-∅/	/-er/	/-s/	other
M&al.	92.10	**95.00**	87.00	92.00	**84.00**	**60.00**	42.00
BiLSTMAttn	89.37	93.93 (0.6)	**88.08 (0.9)**	92.43 (0.6)	79.07 (5.1)	51.75 (4.6)	45.36 (4.0)
BiLSTMNoAttn	54.65	74.16 (1.9)	63.56 (2.4)	75.57 (2.1)	51.26 (3.7)	29.58 (7.7)	9.07 (0.6)
UniLSTMAttn	86.40	93.39 (0.6)	87.35 (1.0)	92.49 (1.1)	69.78 (5.3)	52.36 (4.5)	44.06 (5.8)
UniLSTMNoAttn	48.71	69.69 (2.2)	58.31 (2.4)	71.98 (1.7)	46.64 (5.2)	32.54 (7.7)	8.08 (0.4)
Transformer	91.04	92.93 (0.4)	87.81 (0.7)	**93.86 (0.3)**	65.44 (4.7)	57.89 (2.0)	**57.47 (4.5)**

Table 2: Average German F1s on all German plural inflections for all architectures. Standard deviation is given in parentheses. Dev accuracy for our experiments were decoded greedily.

	Prod. Prob						Rating					
	/-(e)n/	/-e/	/-∅/	/-er/	/-s/	avg.	/-(e)n/	/-e/	/-∅/	/-er/	/-s/	avg.
M&al.	.28	.13	-	.05	.33	.20	-	-	-	-	-	-
BiLSTMAttn	.11	.08	-.14	.24	.38	.20	.36	.44	.06	.36	.39	.32
BiLSTMNoAttn	**.44**	.08	-.12	.27	.39	**.30**	**.51**	.16	-.29	.30	.31	.20
UniLSTMAttn	.09	.16	-.13	**.36**	.39	.25	.22	.27	-.16	.46	.44	.25
UniLSTMNoAttn	.14	.15	**.08**	.17	.23	.15	.24	.16	-.17	.05	.20	.10
Transformer	.11	**.30**	-.13	.28	**.50**	.20	.48	**.59**	**.15**	**.50**	**.71**	**.49**

Table 3: German wugs Spearman's ρ for the average rating of each model with human production probabilities (left) and average human ratings (right). We report the macro average (avg.) over all inflection classes for both.

and test splits, it is difficult to draw conclusions from this result. We predominately focus on performance differences between the models trained in our experiments including MGL.

Accuracy The accuracies from this experiment generally reflect our expectations from prior work. The Transformer attains the highest test accuracy, LSTMs with attention always achieve higher accuracy than those without, and bidirectional LSTMs show modest improvements over their unidirectional counterparts. However, the unidirectional LSTMs outperform bidirectional counterparts for irregular accuracy (+2.86 and +10.95). Additionally, the Transformer has a low irregular accuracy, though with a very high standard deviation over all 10 runs, indicating at least one run was an outlier with much higher accuracy.

Correlation The trend in accuracy for attentional LSTMs is not strictly true for correlation. LSTMs *without* attention typically correlate with humans slightly better than their attentional counterparts for irregulars. Additionally, unidirectional models result in higher *regular correlations*, which is in contrast to the higher irregular accuracy. Irregular correlations are fairly similar across LSTMs with the exception of the BiLSTMAttn correlation with human ratings, which is much higher than the other LSTM correlations. We also reproduce previous results showing that A&H's rule-based model,

MGL, is better correlated than any LSTM model. The transformer, however, is correlated most highly with humans among all experiments that we ran.

4.2 German

We refer to F1 in Table 2, and correlation with humans in Table 3. Notably all models typically correlate better with human ratings than with production probabilities, though those two metrics have a positive linear relationship (r=0.75). Intuitively, the task of assigning a probability to a form is more like the human rating task than decoding the single most likely form. We present a graph of model production probabilities and model ratings for the German wugs in Figure 2.

F1 F1 scores follow a similar trend to English. In contrast to the very small performance gap in Dankers et al. (2021), LSTMs with attention clearly perform better in terms of F1 than without – though our training dataset from M&al. is much smaller than the one they used, which might amplify the gap. Directionality has much less effect on F1 than attention for German, with the unidirectional LSTMs actually outperforming the bidirectional ones for the infrequent /-s/ class in our experiments. The Transformer attains high (though not necessarily the highest) F1 scores for every class.

Correlation The Transformer clearly correlates most highly with human ratings, attaining a moderate correlation (0.48-0.59) for /-e/, /-(e)n/, and /-er/,
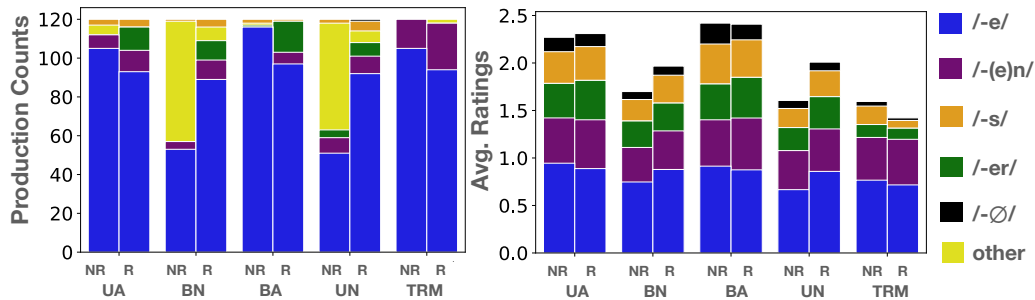
Figure 2: German plural productions (left) and average probabilities (right) for each architecture in Rhyme (R) and Non-Rhyme (NR) contexts for all lemmas and all random initializations. Shorthands are used for architectures – UA refers to `UniLSTMAttn`, whereas BN refers to `BiLSTMAttn`, for example.

and a high correlation (0.71) for /-s/. All architectures correlate poorly with /-∅/, despite very high F1. Looking more closely at /-∅/, it consistently receives very low ratings (as is the case for human ratings), and it was never produced as a model's best output as can be seen in Figure 2. However, there are only 3 /-∅/ inflections in the training data that fit the same phonological context as the wugs. Across all contexts though, /-∅/ is a very common inflection in the training data, which explains its high accuracy on the test set.

There is no clear trend between LSTMs in terms of correlation with human production probability, with rather low ρ overall. However in the case of human ratings, LSTMs with attention always correlate better than those without, with the exception of the most frequent class overall in the training data, /-(e)n/. `BiLSTMNoAttn` is most strongly correlated for /-(e)n/, in contrast to its lower F1 – demonstrating that removing attention leads to a lower F1, but also to a more human-like probability of /-(e)n/.

Regarding directionality, unidirectional LSTMs always outperform their bidirectional counterparts for the infrequent /-s/ class in our experiments. `UniLSTMAttn` correlates better with humans than any LSTM for the infrequent classes /-er/ and /-s/. However, `BiLSTMAttn` has the highest correlation for the frequent /-e/ and /-(e)n/.

5 Analysis

We mainly analyze the correlation between (average) model ratings and human *ratings*. *We find that the Transformer correlates best with human ratings with few exceptions*, indeed it attains a statistically significant positive correlation for all inflection classes in both languages, with the exception of /-∅/ in German. It is also highly accurate, as in previous work (Wu et al., 2021).

Regarding LSTM architectural decisions, unsurprisingly, attention and bidirectionality typically increase accuracy in both languages. The positive effect of attention is similar for correlations with some exceptions. Attention almost always leads to better correlations in German, with the interesting exception of /-(e)n/. Given that humans rate /-(e)n/ most highly on average, the higher correlation could be because without attention, LSTMs are very sensitive to the high /-en/ frequency in the training set. The attentional LSTMs might learn the monosyllabic, neuter context that applies to the wugs, for which there are very few /-(e)n/ training examples. Despite slightly higher accuracy for bidirectional LSTMs, unidirectional LSTMs tend to attain higher correlations with both human metrics for English, especially for the more frequent *regular* inflections.

Conversely in German, the bidirectional LSTMs correlate better for the more frequent /-(e)n/ and /-e/ classes, but `UniLstmAttn` correlates better for the rarer /-er/ and /-s/ classes. The dichotomy between just one highly productive class in English and several productive classes in German may explain the first observation: if unidirectional LSTMs *overfit* to the frequent class, then they might appear to correlate better in English, but not German. However, this would not explain the German class correlations for *infrequent* inflections, which could be explored in future work.

German Model Productions The model production counts in Rhyme versus Non-Rhyme contexts were important for the conclusion in M&al. that `BiLSTMAttn` is not a good model of human behavior. We thus investigate this in Figure 2.

Most of the criticisms from M&al. apply to the productions in our experiments as well. One new observation is that, without attention, LSTMs pre-

	reg	irreg	/-(e)n/	/-e/	/-∅/	/-er/	/-s/
r	0.44	-0.31	0.01	0.80	0.73	0.70	0.83

Table 4: Pearson r between model acc. (or F1), and correlation with human ratings within infl. class (n=5).

	BA	BN	UA	UN	Trm
r	-0.57	-0.33	-0.37	-0.39	-0.38

Table 5: Pearson r between model acc. (or F1), and correlation with human ratings within model (n=7).

dict many "other" forms for NR contexts, but not for R. This likely means that Non-Rhymes lead to decoding errors for these models due to the unfamiliar context. Additionally, despite several behaviors that differ from humans in the Transformer productions, its second most produced inflection class is /-(e)n/, like humans, and unlike any LSTM model. The right side of Figure 2 instead displays the average model rating of each inflection class, on which we base our correlations in Tables 1 and 3.

The average model rating of an inflection class represents the probability assigned to it averaged over all 10 randomly initialized models and all 24 lemmas. The /-e/ inflection accounts for a much smaller amount of the probability mass on average than its production probability. The preference for /-e/ in the NR context, which diverges from human ratings, is smaller by this metric for the Transformer and LSTMs with attention. Furthermore, /-(e)n/ has a more reasonable average probability for most models when compared to the human ratings in M&al., despite the preference for Rhymes, which diverges from human behavior. However, for /-s/ the Transformer shows a much higher average probability for Non-Rhymes than for Rhymes, which is more in line with human ratings.

Overall, this means model *ratings* of German noun plurals look more similar to human ratings than model *productions* do to human productions. The Transformer is a better account for human behavior than the LSTM, though it still diverges in some ways. Dankers et al. (2021) warned that the /-s/ behavior may be explainable by a simple heuristic though, so this behavior may not actually indicate cognitive plausibility.

Accuracy vs. Correlation The task of predicting the *most likely* inflection for an unknown word (measured by accuracy or F1) is not the same as rating *multiple inflections* (measured by Spearman's

ρ). We thus investigate the relationship between these two tasks by measuring Pearson's r between them to see if better inflection models in terms of accuracy are also more human-like. First, we consider the relationship for all models and inflection classes in both datasets and find no correlation (r = -0.17, n=35). However, some inflection classes or models may behave differently than others. We refer to Table 5 to investigate this relationship within each architecture. In Table 4, we check the correlation within each inflection class. There is not sufficient data to draw statistically significant conclusions in either case, but the correlations that we report can still characterize the relationship in our experiments. We find that all architectures show a negative correlation. This implies that models are more accurate for inflection classes on which they correlate poorly with humans, and vise versa. However, Table 4 shows that all German inflection classes have a positive correlation between the two metrics, with the exception of /-(e)n/. This is likely because /-e(n)/ is highly frequent in the training set, but is less suitable for the monosyllabic, neuter wugs. Neither English inflection class shows a strong relationship, though.

6 Conclusion

We ask which neural architecture most resembles human behavior in a wug test. We introduce results on a wider range of architectures than previous work and find that the Transformer, a state-of-the-art model for morphological inflection, frequently correlates best with human wug ratings. Despite this, a closer look at model ratings and productions on German plural inflection shows that neither model closely resembles human behavior. We also find that, while attention is crucial for LSTM inflection accuracy, it does not always lead to higher correlations with humans. Additionally, the often less accurate unidirectional model sometimes correlates better than its bidirectional counterpart, especially in the case of infrequent German plural classes. Finally, while for some inflection classes more accurate models correlate better with humans, there is no clear relationship between the two metrics overall. Future work might consider behavior when hyperparameters are tuned to maximize plausibility of the probability distribution rather than accuracy. Additionally, these results motivate a closer look at the effect of LSTM encoder directionality with respect to inflection class frequency.

Limitations

This work is limited by the scope of languages and inflection categories that our models are tested on. We present results for two specific inflection categories in two languages. Previously, McCurdy et al. (2020b) ran experiments on neural network behavior for the German plural wugs used here, which brought into question some of the conclusions found in prior work for English past tense inflection. We thus believe that expanding this work to new inflection phenomenon and new languages may introduce results where the findings here do not necessarily hold.

Acknowledgments

References

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The celex lexical database (cd-rom).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Harald Clahsen, Monika Rothweiler, Andreas Woest, and Gary F. Marcus. 1992. Regular and irregular inflection in the acquisition of german noun plurals. *Cognition*, 45(3):225–255.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. 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*, pages 3868–3877.

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.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. 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*, pages 94–108.

Rainer Goebel and Peter Indefrey. 2000. A recurrent network with short-term memory capacity learning the german-s plural. *Models of language acquisition: Inductive and deductive approaches*, pages 177–200.

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

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Christo Kirov and Ryan Cotterell. 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.

Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

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

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020a. 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*, pages 1745–1756.

Kate McCurdy, Adam Lopez, and Sharon Goldwater. 2020b. Conditioning, but on which distribution? grammatical gender in German plural inflection. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 59–65, Online. Association for Computational Linguistics.

Danny Merkx and Stefan L Frank. 2020. Human sentence processing: Recurrence or attention? *arXiv preprint arXiv:2005.09471*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

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

David E Rumelhart and James L McClelland. 1985. On learning the past tenses of english verbs. Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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

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.

Model	Hyperparams.
BiLSTMAttn	0.93M
BiLSTMNoAttn	0.90M
UniLSTMAttn	0.56M
UniLSTMNoAttn	0.54M
Transformer	7.41M

Table 6: Number of parameters in each model.

A Individual Model Variance

In figure A.2, we show the variance, via boxplots, when correlating with human *ratings*. Models typically have higher correlations with ratings than with production probabilities, but the two are linearly related in our results. Similar to the findings of C&al., who compared to production probabilities, we find that individual BiLSTMAttn models vary quite a bit with respect to correlation with humans. For English, some models vary far less, for example BiLSTMAttn has a much lower variance with respect to both regulars and irregulars than BiLSTMAttn. Similarly, the Transformer often correlates the same across different random initializations, with the exception of a few outliers. Turning to the German boxplots in A.2b, we see similarly low variance for the transformers, and typically higher variance for most LSTMs. For architectures that vary more, i.e. LSTMs, we often see a higher correlation when the ratings are first averaged (as reported in Table 1 and 3), but the same is often not true for English.
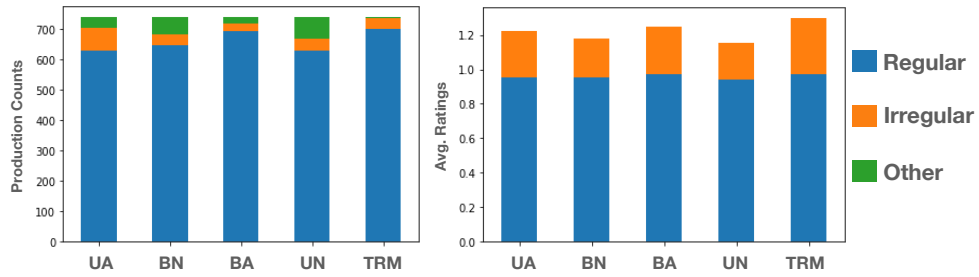
Figure A.1: English past tense productions (left) and average probability (right) for each architecture for all lemmas and all random initializations.
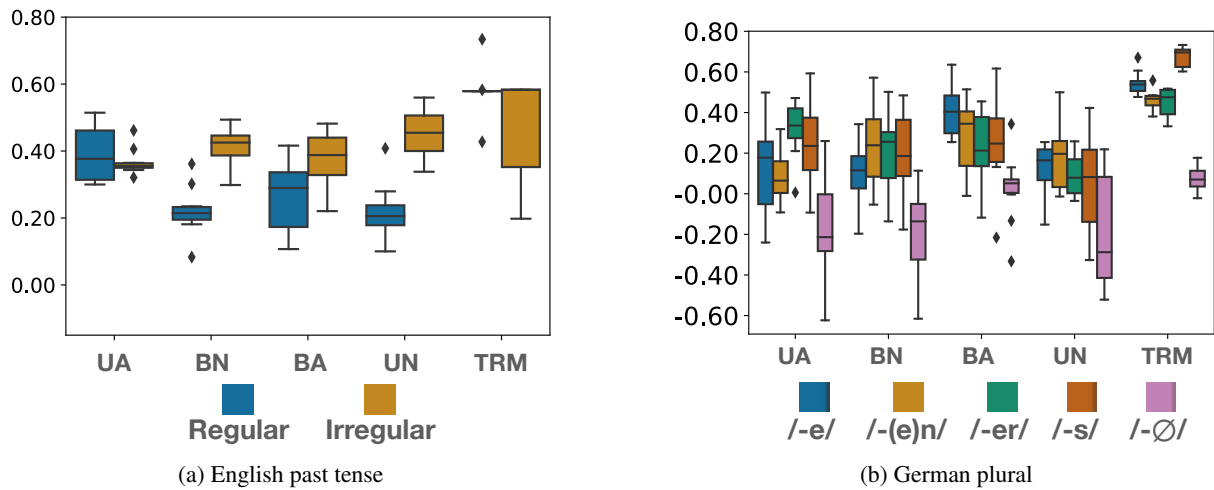

(a) English past tense

(b) German plural

Figure A.2: Boxplots of Spearman's correlation for individual models with respect to average human ratings