Learning Bidirectional Morphological Inflection Like Humans

Akiyo Fukatsu, Yuto Harada, Yohei Oseki
The University of Tokyo
{akiyofukatsu, harada-yuto, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

For nearly the past forty years, there has been discussion regarding whether symbolic representations are involved in morphological inflection, a debate commonly known as the Past Tense Debate. The previous literature has extensively explored whether neural models, which do not use symbolic representations can process morphological inflection like humans. However, current research interest has shifted towards whether neural models can acquire morphological inflection like humans. In this paper, we trained neural models, the recurrent neural network (RNN) with attention and the transformer, and a symbolic model, the Minimal Generalization Learner (MGL), under a human-like learning environment. Evaluating the models from the perspective of language acquisition, we found that while the transformer and the MGL exhibited some human-like characteristics, the RNN with attention did not demonstrate human-like behavior across all the evaluation metrics considered in this study. Furthermore, none of the models accurately inflected verbs in the same manner as humans in terms of morphological inflection direction. These results suggest that these models fall short as cognitive models of morphological inflection.

Keywords: Language acquisition, morphological inflection, computational modeling

1. Introduction

In the Past Tense Debate, it has been discussed whether humans use symbolic representations to process morphological inflection. To discuss this matter, neural and symbolic models were proposed as cognitive models for morphological inflection, and the focal point has been whether the neural models hold the psychological reality.

This debate started with the proposal of neural networks as cognitive models by Rumelhart and McClelland (1986) and the following rebuttal by Pinker and Prince (1988). Rumelhart and McClelland (1986) proposed a feed-forward neural network that does not involve symbolic representations, and argued that the morphological inflection can be processed via a single mechanism—neural networks. According to Rumelhart and McClelland (1986), their neural model replicated several phenomena characteristic of language acquisition such as the U-shaped learning curve. The U-shaped learning curve refers to the developmental path where children first produce target-like forms, then go through a period with erroneous outputs, and eventually start producing target-like forms again. When the accuracy dropped, human-like errors such as overregularization was also observed. Regarding such statements on language acquisition, Pinker and Prince (1988) pointed out that Rumelhart and McClelland (1986) manipulated the ratio of regular and irregular verbs in the inputs during training, which caused the U-shaped learning curve. Additionally, Rumelhart and McClelland (1986)'s model was suggested to learn inflectional patterns that were not found in humans.

After three decades, this debate was revived by Kirov and Cotterell (2018) along with the development of neural networks. Kirov and Cotterell (2018) applied the recurrent neural network (RNN) with attention (Bahdanau et al., 2014; Cho et al., 2014) to the English past tense, arguing that the RNN with attention that uses deep neural networks had overcome the problems suggested by Pinker and Prince (1988). Their work provoked consequent studies, which demonstrates that the accuracy of the RNN with attention is not consistent across trials (Corkery et al., 2019), and that the RNN with attention cannot learn inflectional patterns that are low-frequent but highly productive (McCurdy et al., 2020).

However, the previous literature that investigated the validity of deep learning models as cognitive models did not take language acquisition into consideration. Notably, the size of learning data is considerably larger than the inputs towards children, and the training data contains items that children will not hear during language acquisition.

To overcome this shortcoming, recent studies aim to conduct experiments under the more human-like learning environment. The objective of the shared task of SIGMORPHON 2022 (Kodner and Khalifa, 2022) was to learn morphological inflection with small data, and the target items are sampled, weighted with frequencies from the Child-Directed Speech (CDS) in CHILDES (MacWhinney, 2000).

Another problem in the previous literature is that the base form for morphological inflection is fixed a priori. For example, the related studies that focus on the English past tense (Kirov and Cotterell, 2018; Corkery et al., 2019) assume that English verbs are inflected from present forms (unsuffixed forms) to past forms (suffixed forms). In contrast, languages like Japanese does not have unsuffixed verb forms, and thus it is not obvious whether verbs are inflected from present forms to past forms or from past forms to present forms. For children, it is also part of the process of language acquisition to find out which form should be the base form for inflection so that other forms can be inflected efficiently (Albright, 2002). Therefore, the direction of morphological inflection should be considered in modeling of morphological acquisition.

In this paper, we test neural models, the RNN with attention (Bahdanau et al., 2014; Cho et al., 2014) and the transformer (Vaswani et al., 2017), and a symbolic model, the Minimal Generalization Learner (MGL: Albright and Hayes, 2002, 2003) bidirectionally (i.e., present→past and past→present) on CDS extracted from CHILDES (MacWhinney, 2000).

The main contributions of this paper are as follows. First, we trained models under experimental settings more realistic to the human learning environment and evaluated them from the perspective of language acquisition. The results suggest that the transformer and the MGL have some human-like characteristics, whereas the RNN with attention did not demonstrate human-like behavior across all the evaluation metrics considered in this study.

Second, we introduced the direction of morphological inflection (e.g., present→past and past→present) as a new evaluation metric. We found that all models cannot learn the direction of morphological inflection like humans, concluding that all models still fall short as cognitive models of morphological inflection.

2. Methods

2.1. Task

In a morphological inflection task, models receive tuples of two surface forms (e.g., present forms and past forms) as inputs during training. Learning the relationship between these tuples, the models then generate one form given the other in the test. We do not employ morphological attributes in this study, considering that children are not explicitly given morphological attributes possessed by each surface form.

2.2. Models

Morphological inflection tasks can be treated as string transformation tasks, and hence the neural models deploy the encoder–decoder architecture developed in translation studies. We trained two neural models, the RNN with attention (Bahdanau et al., 2014; Cho et al., 2014) and the small transformer (Vaswani et al., 2017; Wu et al., 2020), and one symbolic model, the MGL (Albright and Hayes, 2002, 2003).

RNN with Attention The RNN with attention (Bahdanau et al., 2014; Cho et al., 2014) was first applied to morphological inflection tasks by Kann and Shültze (2016), and subsequently to the English past tense by Kirov and Cotterell (2018). The model uses bidirectional Long Short-Term Memory (LSTM) for the encoder, and unidirectional LSTM with attention for the decoder. The model is consisted of 4 encoder–decoder layers. Following the previous literature (Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020), we used the implementation by OpenNMT (Klein et al., 2018).

Small Transformer The transformer model developed by Vaswani et al. (2017) was adopted for morphological inflection tasks by Wu et al. (2020). This transformer model serves as the current baseline model replacing the RNN with attention. Since the original transformer model is designed to process a large amount of text, it is overpowered for morphological inflection tasks. To adjust the model size, Wu et al. (2020) proposed a small transformer consisting of 4 encoder–decoder layers and 4 self-attention heads. We use Wu et al. (2020)'s implementation.

Minimal Generalization Learner (MGL) The MGL has been used as the baseline model for symbolic models, as proposed by Albright and Hayes (2002, 2003). The model takes tuples of two surface forms as inputs, and extracts rules to inflect input forms to output forms. Then, the model assigns a reliability score to each rule by calculating $\frac{Hits}{Scope}$. Here, $Scope$ represents the number of items to which the rule's conditions can apply, and $Hits$ represents the number of items for which the inflected forms are correctly derived by that rule. Following Mikheev (1997)[1], this reliability score is penalized for low-scoped data because rules with higher scope (e.g., $\frac{Hits=800}{Scope=1000}$) should be more reliable than those with lower scope (e.g., $\frac{Hits=8}{Scope=10}$).

[1] The reliability score is converted into the confidence score (π) by the following equation.

$$\pi = \hat{p}^* - z_{(1-a)/2} \times \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n}}$$

where $\hat{p}^*$ is the smoothed probability $\frac{hits+0.5}{scope+1.0}$ to avoid zero in the numerator or denominator. α is a parameter called the confidence level, and higher α means greater penalty. We set α as 0.75. $\sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n}}$ is an estimated variance.

Train	Validation	Test	Direction
80% of K-IPA(L)	10% of K-IPA(L)	10% of K-IPA(L)	Present→Past Past→Present
		Wug	Present→Past Past→Present
80% of K-IPA(S)	10% of K-IPA(S)	10% of K-IPA(S)	Present→Past Past→Present
		Wug	Present→Past Past→Present
80% of CHILDES	10% of CHILDES	10% of CHILDES	Present→Past Past→Present
		Wug	Present→Past Past→Present

Table 1: Variables of experimental conditions.

2.3. Data

We prepared three training datasets, two test datasets, two directions of morphological inflection, leading to 12 conditions per model (Table 1). We first constructed three datasets and split them into 8:1:1 for training, validation, and test.

2.3.1. Training Data

We prepared three training datasets: K-IPA(L), K-IPA(S), and CHILDES. K-IPA(L) was developed to confirm that the models can acquire morphological inflection from the amount and type of inputs that adults receive. K-IPA(S) was created by randomly sampling a subset of K-IPA(L) to the same size as CHILDES. Given that CHILDES differs from K-IPA(L) not only in data size but also in the type of verbs, this dataset was prepared to distinguish the effect of training data size from that of training data type. CHILDES was constructed from CDS to approximate the inputs that children are likely to hear during language acquisition. The characteristics of each data are as follows. Each training data was split into 8:1:1 for training, validation, and test.

K-IPA(L) K-IPA(L) was constructed from two corpora: the Kyodai Text Corpus (Kurohashi and Nagao, 2003) and IPAdic (Asahara and Matsumoto, 2003). The Kyodai Text Corpus archives 40k sentences of news articles and editorials (20k each), chosen to cover the verbs that adult native speakers use on a daily basis. However, this corpus contained only 1366 verbs, not reaching the same vocabulary size as previous literature (e.g., Kirov and Cotterell, 2018; Corkery et al., 2019). To supplement the data, we also extracted verbs from IPAdic (Asahara and Matsumoto, 2003), a dictionary used in morphological analyzers, resulting in 5300 verbs. We then combined verbs obtained from these two corpora and eliminated the duplicates. As a result, we obtained 5502 pairs of present forms and past forms.

K-IPA(S) The dataset CHILDES differs from K-IPA(L) not only in size but also in type because the vocabulary that mothers use to address their children differs from the one that they use with adults. To distinguish the effect of training data size from that of training data type, we randomly sampled, 874 pairs of present forms and past forms, matching the size of CHILDES.

CHILDES One concern in creating a dataset from the Kyodai Text Corpus and IPAdic is that these corpora contain verbs that children are least likely to hear during language acquisition. Additionally, The data size is larger than the number of verbs used by parents addressing their children.

Given that the vocabulary that children hear may differ from that obtained from these corpora, we created another dataset based on CDS. To collect verbs, we used natural speech data of six Japanese-speaking children from the Miyata (Miyata, 1992) and MiiPro (Kokuritsu Kokugo Kenkyujo, 1981-1983) corpora in CHILDES (MacWhinney, 2000), which consists of 230k sentences. From these corpora, we obtained 874 pairs of present forms and past forms.

2.3.2. Test Data

For testing, we prepared two conditions: actual verbs and nonce verbs, namely "wug" verbs. The wug verbs are non-existent but follow morpho-phonological constraints in the target language. We adopt them to investigate whether the models can extend their knowledge to unknown words.

The wug test (Berko, 1958) was first developed to investigate children's ability to inflect nonce words based on actual verbs. During the acquisition of verbs, children will encounter and have to process unknown words. In psycholinguistic studies, it is well known that such verb-acquiring children can apply inflectional patterns to unknown words, inflecting them correctly. This experimental paradigm is well-established in psycholinguistic research and is also used in cognitive modeling to compare models' production with humans' production for unknown words (Albright and Hayes, 2003; Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020, among others).

Actual Verbs Of the dataset, 10% was used to test the models' accuracy. The test items were pseudo-randomly sampled from the datasets to ensure that low-frequency suffixes would appear in the test at least once.

Wug Verbs The wug dataset, constructed by Oseki et al. (2019), consists of 64 items, including 32 present forms and 32 past forms. Among these verbs, 16 items were vowel-final, and 16 were consonant-final verbs. In Japanese, vowel-final verbs end with either /i/ or /e/, resulting in 8 nonce items for each ending. The present form has 8 consonant-final endings (/tsu/, /u/, /mu/, /bu/, /nu/, /ku/, /su/ and /gu/), and thus there are 2 nonce items for each. The past form has 4 consonant-final endings (/ita/[2], /ida/, /nda/, /tta/), leading to 4 nonce items for each.

2.4. Experimental Settings

2.4.1. Notation

The data were originally notated in Japanese letters or Latin alphabets. However, feeding models with data in these notations is not realistic in the context of human learning environment, as children acquire language primarily through auditory inputs. To simulate the language acquisition environment more realistically, we fed the models with verbs represented in the International Phonetic Alphabet (IPA). Verbs were converted into IPA using Phonemizer (espeak).[3][4]

2.4.2. Hyperparameters

The RNN with attention has 100 hidden units per layer and utilizes Adadelta for optimization. The transformer has 1024 hidden units for the feed-forward layer and utilizes Adam for optimization. Both models have the embedding size of 256. The batch size was set to 32, which was validated in Ma and Gao (2022) with data size similar to our experiment. The maximum training epochs were set to 30 for K-IPA(L) and 100 for K-IPA(S) and CHILDES. Following the previous literature (Kirov and Cotterell, 2018; Wu et al., 2020), the early stopping was applied to the transformer, whereas the RNN with attention was trained until the maximum epochs to ensure complete training.

One concern in training neural models is that the result of a single initialization may not be reliable because the accuracy of neural models can be unstable across initializations (Corkery et al., 2019). Alternatively, Corkery et al. (2019) consider a single initialization as a single speaker and aggregate the results of multiple initializations. Following Corkery et al. (2019), we also aggregated the same number of initializations as the participants of the human data ($N = 39$). For the MGL, we only report the result of a single trial because the outputs of symbolic models are constant across trials.

2.4.3. Evaluation

We evaluate the models on their accuracy and correlation with humans. We averaged the accuracy scores across the conditions except the condition of interest:

(i) **Training Data Size**: A model should acquire morphological inflection from small data, given that children are data-efficient. Thus, a model is considered human-like if it does not show a decrease in accuracy when trained on K-IPA(S), compared to K-IPA(L).

(ii) **Training Data Type**: A model should acquire morphological inflection from CDS better than dictionary and adult-directed data. Thus, a model is considered human-like when it produced a higher accuracy when trained on CDS than K-IPA(S).

(iii) **Test Data Type**: A model should be able to extend their knowledge to wug verbs. Thus, a model is considered human-like if it does not show a decrease in accuracy when tested on wug verbs, compared to actual verbs.

(iv) **Correlation**: A model should show a correlation with human data. Thus, a model is considered human-like when their output distribution in the wug test is correlated with that of humans.

(v) **Direction**: A model should show a higher accuracy in the same direction of morphological inflection as humans. Thus, a model is considered human-like when it shows higher accuracy in the past→present direction than the present→past direction.

For correlation, we computed the scores using the production probability of humans and models. The production probability refers to the ratio of the participants who produce a particular output to the total number of participants. To align model production with human production, we calculated the production probability of the models by dividing the number of initializations that produced the forms observed in the human data with the total number of initializations ($N = 39$). The outputs in symbolic models like the MGL are constant across trials, and thus we used the confidence scores produced by the model. Any form not observed in human data was given a probability of 0.

[2] /ita/ in consonant-final verbs is always preceded by a vowel as in *kaita* ”write-PAST”. It should be notated that there are a few vowel verbs ending with /ita/ that follows a vowel such as *kuiru* “regret-PRES”–*kuita* “regret-PAST” although in most of vowel verbs, /ita/ follows a consonant.

[3] https://phonemizer?ref=morioh.com&utmsource=morioh.com

[4] We also conducted experiments with data converted into Latin alphabets. The results of items in the Latin alphabets are shown in Appendix A (Table 11).

	Model		
	RNN	Transformer	MGL
Large	95.42	95.65	92.98
	(±2.58)	(±3.21)	(±4.17)
Small	63.27	91.4	90.00
	(±16.82)	(±2.37)	(±5.39)
$\Delta(\uparrow)$	−32.15	−4.25	−**2.98**

Table 2: The effect of the training data size. The mean accuracy was calculated based on the models trained on the larger dataset, K-IPA(L), and the smaller dataset, K-IPA(S). The bold figure represents the best score, and the figures in parentheses represent the standard deviation.

	Model		
	RNN	Transformer	MGL
K-IPA(S)	63.27	91.4	90.00
	(±16.81)	(±2.37)	(±5.39)
CHILDES	67.95	90.85	90.24
	(±17.60)	(±5.82)	(±5.58)
$\Delta(\uparrow)$	**4.7**	−0.55	0.24

Table 3: The effect of the training data type. The mean accuracy was calculated based on K-IPA(S) and CHILDES. The bold figure represents the best score, and the figures in parentheses represent the standard deviation.

3. Results

Tables 2–6 display the averaged accuracy scores of the three models by each evaluation metric. We report the statistical significance of the observed differences for the neural models.[5]

3.1. Evaluation by Condition of Interest

Size of Training Data Table 2 illustrates the averaged accuracy and the delta scores, indicating that all models decreased in accuracy when trained on the smaller dataset compared to the larger one. Notably, the RNN with attention experienced a significant drop in accuracy when trained on the smaller datasets ($\beta = 3.589$, $p < .001$), whereas the transformer and the MGL exhibited relatively minor decreases in accuracy ($\beta = 1.317$, $p < .001$). Since an ideal cognitive model should be data-efficient, the transformer and the MGL are more human-like than the RNN with attention in terms of the effect of training data size.

	Model		
	RNN	Transformer	MGL
Actual Verb	78.41	90.18	94.31
	(±19.01)	(±4.40)	(±3.55)
Wug	32.95	95.09	87.84
	(±47.81)	(±2.66)	(±3.49)
$\Delta(\uparrow)$	−45.46	**4.91**	−6.48

Table 4: The effect of the test data type. The bold figure represents the best score, and the figures in parentheses represent the standard deviation.

Type of Training Data Table 3 displays the differences in accuracy when the models were trained on CHILDES as opposed to K-IPA in the same size. The results revealed that the RNN with attention produced higher accuracy when trained on CHILDES compared to K-IPA(S). The results revealed that the RNN with attention and the MGL showed a marginally higher accuracy when tested on CHILDES than K-IPA(S), but the statistical analysis suggests that the RNN with attention is less accurate when trained on CHILDES ($\beta = -0.1.901$, $p < .001$). In contrast, the transformer showed a marginal decrease in accuracy, though it was not statistically significant ($\beta = -1.112$, $p = .089$). K-IPA include verbs that children will not hear during language acquisition, and thus a cognitive model should learn morphological inflection better from CDS than dictionary and adult-directed data. From the perspective of training data type, the MGL is more human-like than the RNN with attention and the transformer.

Type of Test Data Table 4 presents the differences in model accuracy when tested on wug verbs in comparison to actual verbs. Notably, the RNN with attention and the MGL exhibited the decrease in accuracy when subjected to the wug test ($\beta = -2.014$, $p < .001$). The transformer's accuracy increased, but the effect of test type was not statistically significant ($\beta = 0.153$, $p = .893$). A cognitive model should be able to extend their knowledge to wug verbs. Since the transformer and the MGL did not show significant decrease in accuracy when tested on wug verbs than actual verbs, the transformer and the MGL are more human-like than the RNN with attention.

[5] To examine the statistical significance of the delta, we conducted regression analysis (see Appendix C). In this paper, we only report the results for the RNN with attention and the transformer due to the following reason. The accuracy scores for the MGL were obtained from a single trial, and the sample size was not sufficient to detect the significance of the effects. Moreover, the rule-based model like the MGL does not yield variations by trial, which does not fit the statistical analysis, which assumes variance in data.

	Model		
	RNN	Transformer	MGL
ρ	.31	**.51**	.36

Table 5: The correlation scores between the models and humans. The bold figure represents the best score, and the figures in parentheses represent the standard deviation.

	Model			
	Human	RNN	Transformer	MGL
Pres→Past	68.02 (±46.66)	85.44 (±11.16)	94.35 (±2.51)	88.48 (±3.98)
Past→Pres	76.60 (±42.35)	65.65 (±21.97)	90.91 (±5.25)	93.67 (±4.43)
Δ(↑)	8.58	−28.7	−3.44	−5.18

Table 6: The effect of direction. The figures in parentheses represent the standard deviation. Human accuracy was computed based on the data provided in Oseki et al. (2019).

Correlation with Humans We also investigated how similar the models' outputs are to humans' by computing the correlation scores between the production probabilities of the models and humans (Table 5). We found that the verb forms produced by the transformer are moderately correlated with those produced by humans. On the other hand, the RNN with attention and the MGL only showed a weak correlation with humans. These results suggest that the transformer is more human-like than the RNN with attention and the MGL in terms of similarity of output distribution.

Direction of Morphological Inflection Table 6 indicates averaged accuracy scores in the present→past direction and the past→present direction. All three models showed higher accuracy in the present→past direction, whereas humans showed higher accuracy in the past→present direction. Although the difference by the direction of morphological inflection was marginal in the transformer and the MGL ($\beta = -0.201$, $p = .894$), this result is not congruent with human production because humans distinctively showed higher accuracy in the past→present direction. The RNN showed a large decrease in accuracy in the past→present, compared to the present→past direction, and the effect of direction was marginally significant ($\beta = -0.824$, $p = .087$). Since an ideal model should produce higher accuracy in the same direction as humans, all models failed in terms of the choice of the base form for inflection.

	Model		
	RNN	Transformer	MGL
Max.	3.45	0.4	0
Min.	0	0	0
Mean	0.89	0.16	0
Std.	±1.20	±0.16	0

Table 7: Percentage of erroneous outputs that appeared in the training data.

Input	Answer	Prediction
to̞häida	to̞hägɰ$^{\beta}$	to̞ŋgäɾɰ$^{\beta}$
wäçitä	wäçiɾɰ$^{\beta}$	häiɾɰ$^{\beta}$
to̞çimɰ$^{\beta}$	to̞çindä	to̞ɾiättä

Table 8: The example erroneous outputs based on copying by the RNN with attention trained on CHILDES and tested on the wug verbs. The meanings of the predicted forms are to̞ŋgäɾɰ$^{\beta}$: *tongar*-PRES "to become sharp", häiɾɰ$^{\beta}$: *hair*-PRES "enter", to̞ɾiättä: *toriaw*-PAST "scramble".

3.2. Error Analysis: Copying

Neural models are known to copy items in training data and produce them as outputs (McCoy et al., 2021), which is a potential cause of errors. To investigate whether this copying behavior negatively affected the results, we counted the number of erroneous outputs that were included in the training data (Figure 7). Three erroneous outputs by the RNN with attention are shown as examples in Table 8.

The RNN with attention yielded the highest average rate of copying errors. The maximum error rate by condition was also highest for the RNN with attention. On the other hand, the transformer yielded the lower average and maximum rate for copying errors. It should be noted that the MGL, by its definition, does not copy training items because it extracts suffixes from stems during training.

4. Discussion

In this study, we examined the impact of the following conditions on model performance: Training data size (large and small), training data type (K-IPA and CHILDES), test data type (actual verbs and wug), correlation, and direction (present→past and past→present). The results for each condition are summarized in Table 9. In this section, we discuss the implications of these results.

	Model		
	RNN	Transformer	MGL
Training size	✗	✗	✓
Training type	✗	✗	✓
Test type	✗	✓	✗
Correlation	✗	✓	✗
Direction	✗	✗	✗

Table 9: Summary of results.

4.1. Effect of Training Data

4.1.1. Data Size

Previous research typically utilized datasets ranging from 4K to 10K examples (Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020). In our study, we created an equally large dataset. However, it's important to note that the CDS data contains significantly fewer verbs, approximately 870 lexemes in the case of Japanese. To investigate if the success of neural models depends on data size, we trained the models on smaller training data.

Our findings show that all models achieved higher accuracy with the larger dataset. The RNN with attention was most sensitive to data size changes, experiencing a significant accuracy decrease. In contrast, both the transformer and the MGL maintained accuracy well. Notably, the MGL exhibited minimal accuracy decline with smaller datasets due to its strong generalizability, facilitated by symbolic representations. Relative to the MGL, the transformer also maintained accuracy with smaller datasets. Children also possess generalization ability as they extract patterns from limited inputs. In this regard, the transformer and the MGL outperform the RNN with attention as cognitively-adequate models of morphological inflection.

4.1.2. Data Type

In this condition, we examined whether models can learn morphological inflection better from CDS than dictionary and adult-directed speech data.

Table 3 illustrates the comparison between K-IPA(S) and CHILDES datasets. the MGL achieved marginally higher scores when trained on CHILDES compared to K-IPA(S), while the transformer's accuracy showed a marginal decrease when trained on CHILDES.

These results indicate that the simplicity in CDS may have been beneficial to the explicit rule-induction model. Child-Directed Speech (CDS) typically contains shorter and simpler verbs compared to adult-directed speech, which often includes more complex verbs. The presence of longer and more complex verbs in adult speech datasets may result in difficulty finding the morpheme boundaries and inflectional patterns. Given the fact that all models learned better from the larger dataset than the smaller dataset (Section 4.1.1), the longer and more complex items in adult-directed speech, may necessitate a larger training dataset for effective learning.

The findings suggest that the MGL is more human-like than the RNN with attention and the transformer in terms of training data type. However, it should be notated that the difference of accuracy in the MGL is also marginal. Thus, further investigation may be necessary to conclude that the symbolic models learn better from CDS than adult-directed speech.

4.2. Effect of Test Data

A critical aspect of human linguistic knowledge lies in the ability to generalize observed patterns to unseen data. Children acquire language proficiency with limited exposure, and this capacity for generalization plays a pivotal role in successful language acquisition.

However, testing the generalization power of models using a subset of training data may not suffice, as some verbs share partial stems and can be solved based on similar items within the training data. To address this limitation, we conducted tests using the wug test, designed to assess whether models can generalize inflectional patterns to out-of-domain data. Additionally, the wug test controls for the number of verb types, mitigating the influence of high-frequency verb types on accuracy.

The results reveal that the RNN with attention experienced a significant decrease in accuracy in the wug test compared to the test with the subset data, while the MGL exhibited only a marginal decrease. Remarkably, the transformer achieved even higher accuracy, indicating its robust generalization capacity.

The transformer's robustness when faced with out-of-domain data may be attributed to its exceptional generalization capacity. Recent work by Ma and Gao (2022) tested transformers on English past tense, demonstrating that the model successfully generalizes regular patterns, even when regular verbs constitute less than 10% of the training data. In our study, the training data also encompassed a limited number of items for some verb types; for instance, there were only 11 t-final verbs in the CHILDES dataset. Despite the small dataset, the transformer effectively generalized inflectional patterns to out-of-domain data.

4.3. Correlation with Humans

Computing the correlation scores between the models and humans, we found that the transformer showed higher correlation scores with humans than the RNN with attention and the MGL. Corkery et al. (2019) report that MGL's correlation with human production data as $\rho = .35$ (regular) and $\rho = .36$ (irregular), and the correlation scores for the RNN with attention are $\rho = .45$ (regular) and $\rho = .19$ (irregular). The averaged correlation scores in this study seems to fall within the comparable range for the RNN with attention and the MGL.

The RNN with attention and the MGL both only showed a weak correlation with humans, but their inconsistency with human production seem to result from different reasons. The MGL produced forms observed in human data, but with considerably higher production probability than human production. In contrast, the RNN with attention are more apt to produce forms not observed in human data. This differences are likely to result from the symbolic model's robustness to data size; the MGL is so robust to data size that it overextends a single output form for a certain input form (See Appendix B). The transformer also have tendency to produce certain forms highly productively, but the model shows gradual distribution of outputs without producing unobserved forms. This tendency possibly resulted in higher correlation with humans than the other two models.

4.4. Effect of Model Architecture

The investigated models differ in their architectures, and certain outputs appears to be resulted from these architectural differences. We found that the RNN with attention had a higher error rate in producing items identical to those in the training data (Table 7). The maximum error rate of the RNN with attention is 3.45%, whereas that of the transformer and the MGL is 0.4% and 0%, respectively.

For further investigation, we examined the outputs generated by the RNN with attention (Table 8). The erroneous outputs indicates that the RNN with attention copied items from the training data, even when those items had different inflectional patterns due to the different final sounds of the stems. If we define analogy as the strategy to process unseen items based on familiar words in memory, this output pattern suggests that the RNN with attention relies on the analogy-based processing to inflect target forms.

This strategy seems to work well under the conditions where the RNN with attention was tested on the subset of the training data but not in the wug test. The difference between the subset data and the wug test is whether the test items share the same domain with the training data. The analogy-based processing relies on the similarity between seen items and unseen items, and thus the out-of-domain data like the wug verbs possibly nullified the advantage of analogy.

4.5. Effect of Direction

We trained models bidirectionally (i.e., present→past and past→present) and found that all models did not show higher accuracy in the present→past direction than the past→present direction. This is unsurprising, given that inflectional patterns in the present→past direction are generally more systematic than the past→present direction. In the present→past direction, all mappings can be distinguished by the final sound of the stem, except for r-final and vowel-final verbs.

However, this trend contradicts human production patterns, as humans tend to produce inflected forms more accurately in the past→present direction than the present→past direction. Several studies in Japanese (Vance, 1987, 1991; Klafehn, 2003, 2013) report that native speakers of Japanese struggle to inflect verbs in the present→past direction. This odd human behavior have led the previous literature to conclude that humans do not use rules in morphological inflection. As an alternative account, these studies suggest that humans rely on analogical processing.

The findings in the present study and the previous studies suggest that human knowledge of inflection may not solely derive from inputs. Thus, higher human accuracy in the past→present direction possibly resulted from other factors such as children's inductive bias. In fact, Tatsumi et al. (2018) report that Japanese-acquiring children were more biased towards past forms; that is, 1) they were more accurate in producing past forms than present forms; and 2) they tended to produce past forms in the context of present tense. Taken together, children are likely to have some bias towards past forms, but this bias is not acquired by pure inductive learning from inputs. Hence, further investigation is needed to discover the mechanism underlying the process to determine the base form for inflection and the direction in which they learn morphological inflection.

In summary, the current models of morphological inflection show human-like performance in the evaluation metrics that have been used previously used (such as data size), but they all failed in light of direction of morphological inflection. This result suggests that both neural and symbolic models still fall short as cognitive models of morphological inflection.

5. Conclusion

In this paper, we provided the human-like learning environment and investigated the morphological acquisition in neural and symbolic models. We found that the transformer and the MGL have some characteristics that are human-like, whereas the RNN with attention failed in all evaluation metrics. More importantly, by introducing the direction of morphological inflection as a new evaluation metric, we demonstrated that none of them reproduced the process in which children find the base form for morphological inflection.

Therefore, we conclude that all models still face challenges to be cognitive models of morphological inflection. These results also suggest that the morphological acquisition in humans is not achieved by pure inductive learning, but governed by some other factor like an inductive bias. For more human-like morphological acquisition, future studies should explore hybrid models combining inductive learning with some bias attested in language acquisition.

6. Acknowledgments

This work was supported by JST PRESTO Grant Number JPMJPR21C2, Japan. We sincerely thank the reviewers for their valuable feedback and the editors for their guidance and support.

7. Ethical considerations

The present study uses three corpus, of all which are provided for the research purpose. Among these corpus is the CHILDES Corpus (MacWhinney, 2000), which contains minors. In this corpus, data is prepossessed in the way that that eliminates any confidential information such as the use of pseudo names. We also used only verbs that appeared in texts and transcriptions of these corpora, and we confirmed that no personal information was included.

8. Limitations

There are two limitations in creating the experimental setting that are realistic for language acquisition. In the morphological inflection task, tuples of two surface forms are given at a time to models, which is not expected to occur frequently in child language acquisition. These surface forms are likely to appear in utterances in different situations, and children have to identify which surface forms share the lexeme. Thus, it is required in future studies to develop an experimental paradigm where two words are given separately so that the process to relate words is part of the task.

In the same vain, we fed the models with surface forms only, but in reality children are given semantic and pragmatic information as target verbs are embedded in sentences. Thus, future studies are expected to investigate the morphological inflection along with semantic information.

9. Bibliographical References

Adam Albright. 2002. *The Identification of Bases in Morphological Paradigms*. Ph.D. thesis, University of California, Los Angeles.

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with Minimal Generalization. In *Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGMORPHON)*, pages 58–69, USA.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90(2):119–61.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

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.

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, Florence, Italy.

Katharina Kann and Hinrich Shültze. 2016. Single-model Encoder-Decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 555–560.

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*, pages 651–665.

Terry Klafehn. 2003. *Properties of Japanese Verbal Inflection*. Ph.D. thesis, University of Hawaii.

Terry Klafehn. 2013. Myth of the wug test: Japanese speakers can't pass it and English speaking children can't pass it either. In *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, pages 170–184.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

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

Kokuritsu Kokugo Kenkyujo. 1981-1983. *Yoji no kotoba shiryo [A Record of Child-Mother Speech] (I-VI)*. Shuei Shuppan, Tokyo.

Xiaomeng Ma and Lingyu Gao. 2022. How do we get there? evaluating transformer neural networks as cognitive models for English past tense inflection. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1101–1114, Online only. Association for Computational Linguistics.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. 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.

Andrei Mikheev. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23:405–423.

Susanne Miyata. 1992. Wh-questions of the third kind: The strange use of wa-questions in Japanese children. *Bulletin of Aichi Shukutoku Junior College*, 31:151–155.

Yohei Oseki, Yasutada Sudo, Hiromu Sakai, and Alec Marantz. 2019. Inverting and modeling morphological inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 170–177, Florence, Italy.

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.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of English verbs. In David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*, volume 2, chapter On Learning the Past Tenses of English Verbs. MIT Press, Cambridge, MA.

Tomoko Tatsumi, Ben Ambridge, and Julian Pine. 2018. Testing an input-based account of children's errors with inflectional morphology: an elicited production study of Japanese. *Journal of Child Language*, 45:1144–1173.

Timothy J Vance. 1987. *An introduction to Japanese phonology*. State University of New York Press, Albany, NY.

Timothy J Vance. 1991. A new experimental study of Japanese verb morphology. *Journal of Japanese Linguistics*, 13:145–166.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction.

10. Language Resource References

Asahara, Masayuki and Matsumoto, Yuji. 2003. *ipadic version 2.7.0 User's Manual*. [link].

Kurohashi, Sadao and Nagao, Makoto. 2003. *Building a Japanese parsed corpus while improving the parsing system*. Kluwer Academic Publishers.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates.

11. Appendices

A. Accuracy scores in all conditions

We computed the accuracy score for each condition. Table 10 presents accuracy scores of the models tested on actual verbs. Table 11 presents accuracy scores of the models tested on wug verbs.

B. Accuracy by Input Type

The frequency of each ending form differs, which may affect the accuracy of the models. Thus, we averaged the accuracy rate by each input form, and aligned the results with that of humans (Figures 1 and 2).

C. Statistical Analysis

For the analysis of statistical significance, we modeled the model's accuracy (binary as "correct" or "incorrect") by fitting generalized linear mixed-effect models. We used the software R (R Core Team, 2023) and the R package `glmer` (Bates et al., 2015). The fixed and random effects in a concatenated form were specified as follows in R:

$$\begin{aligned}
Y \sim & TrainingSize+ \\
& TrainingType+ \\
& Direction+ \\
& TestType+ \\
& Direction \times Test+ \\
& TrainingSize \times Direction+ \\
& TrainingSize \times TestType+ \\
& TrainingType \times Direction+ \\
& TrainingType \times TestType+ \\
& TrainingSize \times Direction \times TestType+ \\
& TrainingType \times Direction \times TestType+ \\
& (1|Input)
\end{aligned} \tag{1}$$

Table 12 shows the results for the RNN with attention, and Table 13 presents the results for the transformer.

		Model					
		RNN_attn		Transformer		MGL	
Train	Test	Pres→Past	Past→Pres	Pres→Past	Past→Pres	Pres→Past	Past→Pres
CHILDES	Actual	89.77	59.74	92.36	82.26	97.73	88.51
CHILDES	Wug	73.24	49.04	94.95	93.83	90.32	84.38
K-IPA(S)	Actual	81.93	50.13	92.45	87.89	96.59	91.95
K-IPA(S)	Wug	73	48	93.11	92.15	87.1	84.38
K-IPA(L)	Actual	95.8	93.07	94.2	91.9	96.92	94.18
K-IPA(L)	Wug	98.88	93.91	99.04	97.44	93.33	87.5

Table 10: The accuracy of models trained on data in IPA

		Model					
		RNN_attn		Transformer		MGL	
Train	Test	Pres→Past	Past→Pres	Pres→Past	Past→Pres	Pres→Past	Past→Pres
CHILDES	Actual	0.03	0.025	91.69	83.64	97.73	89.66
CHILDES	Wug	0.14	0.023	97.52	93.38	93.55	84.38
K-IPA(S)	Actual	0	0.01	90.69	85.71	96.59	91.95
K-IPA(S)	Wug	0.02	0	94.71	91.35	90.32	84.38
K-IPA(L)	Actual	96	93.6	93.76	87.89	96.92	95.64
K-IPA(L)	Wug	98	98	98.08	96.72	93.33	87.5

Table 11: The accuracy of models trained on data in Latin alphabets

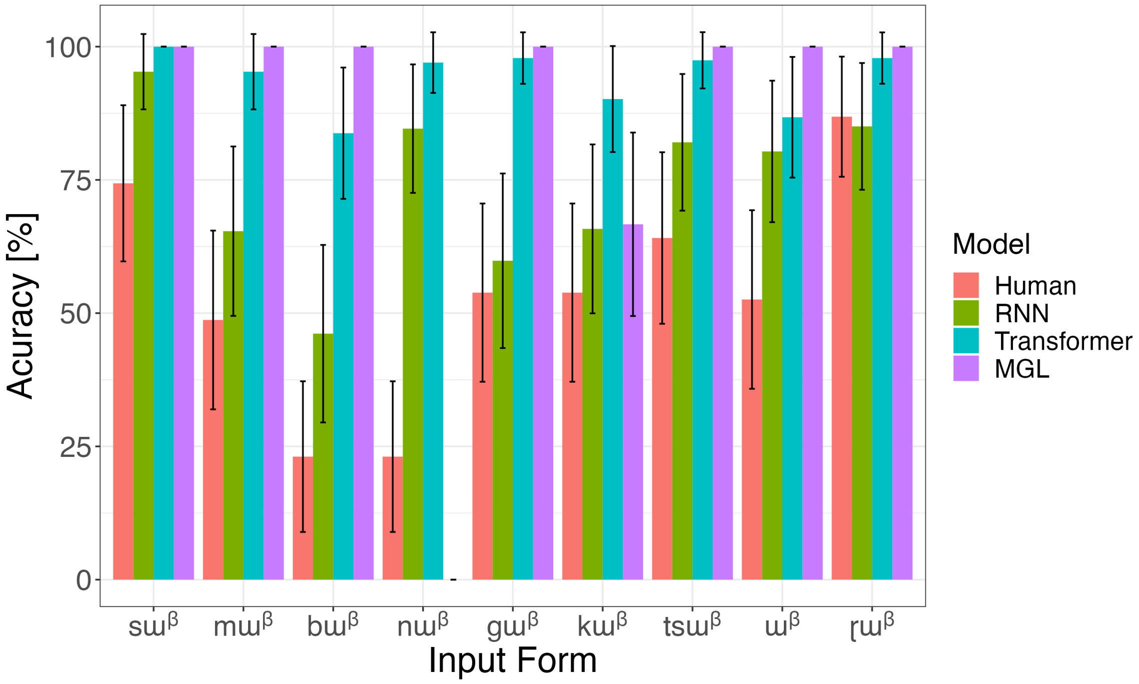

Figure 1: Accuracy by each input form in the present→past direction.

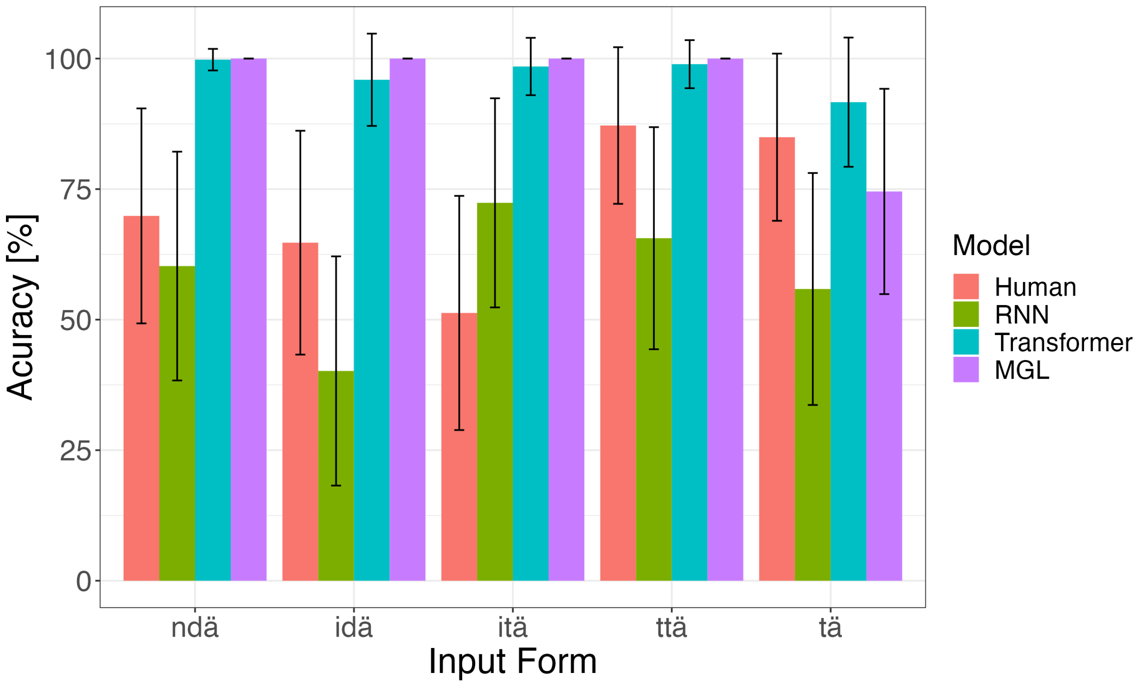

Figure 2: Accuracy by each input form in the past→present direction.

	Estimate	Std. Error	z value	Pr(>|z|)
((Intercept)	3.158	0.297	10.645	.001 ***
Training size	3.589	0.221	16.206	.001 ***
Training type	-1.901	0.360	-5.280	.001 ***
Direction	-0.824	0.482	-1.711	.087 .
Test	-3.540	0.796	-4.447	.001 ***
Direction × Test	3.051	1.161	2.628	.009 **
Training size × Direction	2.035	0.487	4.175	.001 ***
Training size × Test	1.502	0.357	4.205	.001 ***
Training type × Direction	4.925	0.688	7.154	.001 ***
Training type × Test	1.968	0.374	5.259	.001***
Training size × Direction × Test	-2.762	0.637	-4.337	.001 ***
Training size × Direction × Test	-4.970	0.706	-7.038	.001 ***

Signif. codes: 0 "***" .001 "**" .05 "." .1 " " 1

Table 12: Results from logistic regression models predicting the accuracy of prediction by the RNN with attention from training size, training type, direction of morphological inflection, and test type, in IPA.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.666	0.363	23.901	.001 ***
Training size	1.317	0.273	4.815	.001 ***
Training type	-1.112	0.740	-1.504	.133
Direction	0.030	0.426	0.070	.944
Test	0.153	1.132	0.135	.893
Direction ×	-0.201	1.505	-0.133	.894
Training size × Direction	-0.093	0.379	-0.245	.806
Training size × Test	0.174	0.356	0.488	.625
Training type × Direction	4.006	0.955	4.194	.001 ***
Training type × Test	1.476	0.765	1.930	0.054
Training size × Direction × Test	0.796	0.540	1.473	0.141
Training size × Direction × Test	-3.988	0.992	-4.020	.001 ***

Signif. codes: 0 “***” .001 “**” .05 “.” .1 “ ” 1

Table 13: Results from logistic regression models predicting the accuracy of prediction by the transformer from training size, training type, direction of morphological inflection, and test type, in IPA.