How do we get there? Evaluating transformer neural networks as cognitive models for English past tense inflection

Anonymous ACL submission

Abstract

Neural network models have achieved good performance on morphological inflection tasks, including English past tense inflection. However whether they can represent human cognitive mechanisms is still under debate. In this work, we examined transformer models with different training size to show that: 1) neural models correlate with both human behaviors and cognitive theories' predictions on nonce verbs; and the model with small-size training data that matches parents' input distribution has the highest correlation; 2) neural models make different types of errors on regular and irregular verbs, exhibiting a clear distinction between regulars and irregulars. Therefore, we conclude that neural networks have the potential to be good cognitive models for English past tense.

1 Introduction

English past tense has been the subject of debate in human language processing for decades. The past tense has attracted so much attention because both adults and children exhibit a clear distinction between the regulars and irregulars. The regular form follows a formal rule: adding '-ed [*/d/,/t/,/ɪd/*]' to the verb stem as in 'help/helped'. This regular rule has also known to be productive with novel words (e.g. 'wug-wugged' Berko (1958)). The irregulars are categorized by phonological analogy, e.g. 'sing/sang', 'sink/sank', 'drink/drank/, 'begin/began' or learned by rote memory, e.g. 'go/went', 'do/did'. In human language processing, the debate of English past tense has been focused on the nature of the regular-irregular distinction, whether it is a *discrete* distinction that is governed by rules (e.g. Pinker and Prince, 1988), or a *gradient* distinction that is generated by phonological analogy (e.g. Bybee and Moder, 1983). The rule-based theory is also known as 'dual-route' theory, because it proposes human processes the regular items by applying the past tense rule, which involves procedural memory; and the irregular items are retrieved from memory involving declarative memory. The analogy theory claims that a single analogical process can handle both regulars and irregulars, also known as 'single-route' theory. Both theories have been supported by abundance studies with behavior, modeling and neuro-imaging data (e.g. Ullman et al., 1997; Tyler et al., 2005; Stockall and Marantz, 2006; Plunkett and Juola, 1999; Albright and Hayes, 2003; Ambridge, 2010). The debate is on-going and it's still unclear which theory better explains human past tense processing.

Rumelhart and McClelland (1986) (hence RM) proposed that past tense inflection can be learned by neural model. They constructed a connectionist model that learns to associate phonological features of the stem with phonological features of the past-tense forms. Since the early fixed-size feed-forward network can't handle sequences with varied lengths, they constructed wickelfeatures based on wickephones (Wickelgren, 1969) as input. Each wickelfeature is a phonological feature set of a trigram in the root verb, e.g. /ɛlp/ is represented as [<+vowel, +continuous, +unvoiced> + <+low, +liquid, +stop>]. The model successfully learned the regular and irregular forms. RM also claimed that the model mimics children's acquisition pattern (later being harshly criticized in Pinker and Prince (1988)). Modern neural networks with encoder-decoder can handle sequence with different lengths and achieved good performance in morphology inflection tasks across different languages (e.g. Cotterell et al., 2016). Despite neural model's high accuracy in past tense inflections, whether it can serve as a cognitive model and represent human behaviors is still unclear (Kirov and Cotterell, 2018; Corkery et al., 2019; Calderone et al., 2021). In addition, many psychologists and linguists are dismissive of neural networks as a cognitive model, because of the 'black box' nature of neural models. The neural networks might learn the past tense with a totally different mechanism, which is unrelated

to human behaviors and the existing cognitive theories. If RM's early connectionist model can be seen as an extension of the analogy theory (since it used phonological features as input), the modern networks that have raw phonemes as input do seem to be drifted away from major cognitive theories. It is hard to know what exactly neural networks are learning, rules or analogy, or something else.

In order to evaluate if neural networks can be good candidates for cognitive models, we need to show that neural networks not only model human behavior, but also are connected to the existing cognitive theories. In particular, we ask the following questions: 1) Do the neural networks model human adults and/or children's behavior? 2) Do the neural networks fit into the existing cognitive theories? If yes, rule-based theory or analogy theory? In this work, we begin by showing that transformer models with different training sizes all significantly correlate with human adult's data, but only the small-size model correlates with children's data. The models correlate more with the analogy theory on regular verb production; and the irregulars correlates more with rule-based theory. In addition, we also found that models make different types of errors for regulars and irregulars, showing that the transformer models also exhibit distinction between regulars and irregulars. We conclude that the neural networks have the potential to be cognitive models.

2 Background

2.1 Nonce verb experiment

With adult participants. One of the most replicated nonce verb experiments is Albright and Hayes (2003) (hence AH). They constructed an analogy model and a rule-based model which predicts an **acceptance score** for regular form and irregular form of the verb. To test the model, they created a set of 58 unique nonce verbs that are similar to the existing regular and irregular verbs in English. Each nonce verb has two possible past tense forms, the regular one which adds '-ed' [/d/, /t/, /ɪ/], and the irregular one that involves vowel change or other transformations. The analogy model's score is calculated based on the phonological similarity[1] of each nonce verb to the existing verbs in the CELEX (Baayen et al., 1995) database of English verbs (4253 verbs, 218 of which are irregulars). For example, for the nonce verb 'fleep /flip/', the score for regular past tense form 'fleep**ed** /flip**t**/' is calculated based on phonologically similarities to the regular verbs such as 'bleep, peep'; the score for irregular form 'flept /fl**ɛpt**/' is calculated based on the similarities to the irregular verbs such as 'sleep', 'weep'. The rule-based model's score is calculated based on the proportion of existing verbs that can be explained by certain linguistic rules. For example, for the nonce verb 'gleed /glid/', the regular form 'gleed**ed** /glid**ɪd**/' is formed based on the regular rule: '+ /ɪd/' if verb matches [X /d/,/t/__], e.g. 'wan**t**, nee**d**'. This rule could explain 87.2% past tense forms of the verbs ending in /d/ or /t/; thus the score for '/glidɪd/' is 0.872. The irregular form 'gled /gl**ɛd**/' is generated based on an irregular rule: '/i/>/ɛ/ if verb matches [X /r/,/l/ __/d/], e.g. 'bleed', 'read'. The irregular rule explains 79.3% past tens forms of verbs that matches [X /r/,/l/ __/d/], thus the score for '/glɛd/' is 0.793. In addition, 2 experiments with human adult participants on nonce verbs were conducted to evaluate the rule-based model and the analogy model. In Experiment 1, the participants produced the past tense form of each nonce verbs. In Experiment 2, participants rated each past tense form as well as produced them. In general, the human participants predominately produced the regular form for most of the nonce verbs. AH compared the analogy model's score and rule-based model's score with human participants' **production abilities** and **rating** on each nonce verb's regular and irregular past tense form. They concluded that the analogy model is better than rule-based model in predicting human nonce verb behavior.

With children participants. The nonce verb experiment has also been replicated on children. Ambridge (2010) selected 40 nonce verbs from AH and used the same analogy model and rule-based model to predict children's rating. He recruited children from 6-7 and 9-10 years old to rate the regular and irregular past tense forms of each verb. The analogy model's score has better correlation with children's **ratings** than rule-based model's score. Older children also showed more acceptance of irregular forms than younger children. Blything et al. (2018) used the same 40 nonce verbs and recruited children from 4 age groups (3-4 y/o, 5-6 y/o, 6-7 y/o and 9-10 y/o) for a production task. The older children produced more regular forms than the younger children. The analogy model also

[1]The phonological similarity is measured based on the natural class theory by Broe (1993).

performs better than the rule-based model across age groups in predicting **production probabilities**.

With neural models. Kirov and Cotterell (2018) (hence KC) revisited the past tense debate with modern sequence-to-sequence encoder-decoder model. They used a subset of verbs in the CELEX dataset, which contains 4039 verbs, 168 of which are irregular. They trained a biLSTM encoder-decoder model with 100 epochs. Their model reached near-perfect accuracy in both regulars and irregulars in the training. For the test set, the model achieved an accuracy of 98.9 for regulars in test and 28.6 for the irregulars. They also showed that the encoder-decoder model effectively models human behavior in nonce verbs. The correlation of model's nonce verb output is significantly correlated with human **production probabilities** (Spearman's $\rho = 0.48$ for regulars and $\rho = 0.45$ for irregulars).

Corkery et al. (2019) (hence CMS) also conducted the a similar nonce verb experiments on encoder-decoder models and did not find such strong correlations. They adopted the model architecture in KC and trained the model on all 4253 verbs as in AH and 4039 verbs in KC. They used the beam probabilities of each regular and irregular form to calculate the correlation with human data. They showed that with different random initializations, the model's output correlates with human **production probability** differently, ranging from $\rho = 0.1$ - 0.6 for regulars and $\rho = 0.2$ - 0.4 for irregulars. They wondered if these models should be treated as individual participants instead of an averaged representation. Therefore, they further trained 50 individual models with same training data and hyperparameters and sampled 100 past tense forms from each model to have an aggregated model result. The aggregated model shows better correlations with human **rating** data, but still not as good as the analogy model. CMS also suspected that 100 training epochs might lead to model overfitting, and training for less time might have better correlations with human data. Reducing training epochs to 10 achieved the best correlation with human data, but resulted in bad accuracy on real verbs.

2.2 Acquisition pattern of past tense

English speaking children's past tense error has been one of the most widely studied phenomenon in linguistics and psychology. The past tense acquisition has been characterized by **overregularization error** and U-shape learning curve. (e.g. Plunkett and Marchman, 1991; Marcus et al., 1992; Xu and Pinker, 1995; Maratsos, 2000; Maslen et al., 2004). Overregularization errors are the incorrect past forms of irregular verbs when children add '-ed [/d/,/t/,/ɪd/]' to the stem. The most common type of overregularization errors is 'Stem+ed', e.g. '*drawed', '*falled', '*maked'. Children also attach '-ed' to the irregular form ('Past+ed'), such as '*boughted', '*felled', '*tored'. In addition, previous studies also found other rare errors such as incorrect vowel change, e.g. 'bring-*brang' on irregulars. The accuracy of past tense verbs exhibits a U-shape developmental pattern: when children first produce past tense verbs, they produce the correct regular and irregular verbs; then they start to make overregularization errors, causing the accuracy to drop; finally they go back to produce the correct regular and irregular forms. Under the macro U-shape pattern for all irregular verbs, there are also micro U-shape pattern for individual irregulars where the irregulars oscillate between correct and incorrect forms.

RM claimed that they successfully modeled the macro U-shape learning curve: the irregulars were initially produced correctly, followed by overregularization errors and went back to the correct form. However, Pinker and Prince (1988) pointed out that they achieved this by manipulating the input distribution by training the first several epochs only on irregular verbs. KC kept the input distribution constant and did not captured the macro U-shape. However, they modeled the oscillations for different irregular verbs, e.g. stem: 'mislead, epoch 8: 'misleaded', epoch 21: 'misled', epoch 24: '*misleaded', epcho 41: 'misled'. In addition, the model made 'Stem+ed' erros on irregulars, but not other types of overregularization. The model also made some errors on regulars, and most of them involve vowel change, e.g. 'try: /traɪd/-/traʊd/'.

2.3 Evaluating model

Human Behaviors. In this work, we first correlate the model's output on nonce verbs with both **production probability** and **rating** data for adults and children. We also test model's output on real English verbs. If the model mimics adult's behavior, we expect the model's output on real verbs to show some distinction between regular and irregular verbs. If the model mimics the children's

behavior, we expect to the model to output overregularization errors.

Cognitive theories. We correlate the model's output on nonce verbs with the **acceptance score** predicted by rule-based model and analogy model reported in AH. The summary of evaluating methods is shown in Table 1.

Verbs	Nonce	Real
Adults	1, 2	Distinction in regulars vs irregulars
Children	1, 2	Overregularization
Rule-based	3	
Analogy	3	

1 = Production Probability, 2 = Rating
3 = Acceptance Score

Table 1: Evaluating methods for human behavior and cognitive model prediction

3 Methods

3.1 Architecture and hyperparameters

We use transformer model for our training. The transformer model is a self-attention-based encoder-decoder model that is able to process sequential data in a parallel manner, which is different from the LSTM models. The transformer model has achieved great success in complex tasks like machine translation and language generation. Since the datasets for our character-level morphological inflection task are significantly smaller than traditional transformer tasks, we employed a smaller transformer with 2 layers in the encoder (1 attention layer, 1 feed-forward layer) and 3 layers in the decoder (2 attention layers, 1 feed-forward layer). Layer normalization is applied to the output of encoder and decoder. Positional embedding layers are used to capture the positional information. We use 6 self-attention heads, embedding size is 256 and hidden size of feed-forward layer is 1024. The transformer model has $\sim$ 5.83M parameters. Training was done using Adadelta optimization (Zeiler, 2012) with batch size of 32. We train 100 epochs for each model.

3.2 Models and Data

Modeling Adults. To counter the overfitting problem mentioned in CMS, we decide to reduce the training data instead of reducing the number of epochs. We randomly sampled 500, 1500 and 3000 verbs as training data from 4039 verbs used in KC. We believe these data should be sufficient to model the verbs that adults have been exposed to. We also adopt CMS's idea that each model should be treated as an individual participant. CMS changed the initializations of each model to generate different 'participants'. We change the training data for each model by randomly generating 30 samples with 500 verbs, 1500 verbs and 3000 verbs to created 30 'participants' for each training size. We aggregate 30 participant models' output for each training size to produce the **models' production probability**. In the training data, the average proportion of irregular is 4% for models with 500, 1500 and 3000 verbs.

Modeling Children. Children are exposed to less verbs than adults with higher proportion of irregulars. To better model the verbs that children are exposed to, we generate the training data based on real-life parents' input verbs. We selected 8 children's corpora in the CHILDES database (MacWhinney, 2000) that contain overregularization errors. We included each child's first recording file to the first file where they made overregularization errors, and aggregated the parents' the past tense verbs, which contains 246 unique past tense verbs (65 irregular verbs) [2]. The irregular proportion is 26%, which is higher than other training datasets. We randomly generated 30 samples with 246 verbs in CELEX dataset matching the numbers of regular and irregular verbs in the parents' input as our training set and aggregate these models output to produce production probability. The detailed proportion of regular and irregular verbs in each training set is shown in Table 2.

Data size	Regular % /-d/	/-t/	/-ɪd/	Irregular irr%
500	50 (2.2)	19 (2.2)	27 (0.7)	4 (0.7)
1500	51 (1.2)	18 (0.9)	27 (0.9)	4 (0.4)
3000	51 (0.5)	18 (0.4)	27 (0.4)	4 (0.2)
246	42	22	10	26

Table 2: The mean proportions of regulars and irregulars (standard deviation in brackets) averaged over 30 samples of training data with different size

Test Data. We evaluate the models on the nonce verbs and real English verbs. We use all 58 unique nonce verbs for comparing adult's behavior, matching AH, and 40 nonce verbs matching Ambridge

[2]The detailed summary of parent's data in shown in Table11 in Appendix.

(2010) and Blything et al. (2018) to compare children's behavior. We also randomly selected 150 regular verbs (50 for */d/*, */t/* and */ɪd/*) and 20 irregular verbs from the CELEX dataset as the testing data for real English verbs.

4 Experiments

4.1 Experiment 1: Evaluating on nonce verbs

Our first experiment aims to evaluate if the model's production probability correlates with adult's behavior, children's behavior and cognitive models' scores on the nonce verb set. First we report the train and validation accuracy as a sanity check in Table 3. The three large-size model achieved almost perfect accuracy, showing that the model successfully learned the past tense forms. The small-size model has relatively low accuracy, but the model's performance is still decent considering only 246 verbs were used in training. This result confirms that neural models have no difficulty learning past tense forms even with small training data.

Data size	Train %	Validation %
246	98.53 (0.08)	89.59 (1.39)
500	99.29 (0.05)	98.49 (0.72)
1500	99.52 (0.05)	98.67 (0.32)
3000	99.50 (0.05)	98.82 (0.31)

Table 3: Mean accuracy of training set and validation set (standard deviations in brackets) averaged over 30 samples for each data size. Train-val split is 90-10.

4.1.1 Correlation with adults' behavior

KC only calculated the Spearman's correlation (ρ) with the Experiment 1's production probability (Exp 1. Prob.) in AH. CMS calculated the correlation with Exp 1. Prob. and ratings using both Spearman's ρ and Pearson's r. We use both **Exp 1 production probability** and **total production probability** (Total Prob.), and **rating** to calculate the correlation with ρ and r. The results are listed in Table 4.

Rating: *Between Regular and Irregular*: All the models are significantly correlated with the adult's rating for both regulars and irregulars. The correlation with regulars are generally higher than the irregulars, but most of the differences are not significant. Only for model with 246 verbs and 1500 verbs, the Spearman's ρ is significantly higher for regulars than the irregulars. *Among models*: The model with 246 verbs has highest correlation with regulars and irregulars. Using the Fisher r-to-z transformation, we found that the model with 246 verbs has significantly higher correlation in regular ratings than model with 1500 and 3000 verbs. There is no significant differences detected in the irregular correlations. Increasing the training size of the model does not result in higher correlation. Instead, small-size model seems to correlate with adult ratings better. Our models correlate with the rating better than CMS and KC.

Production probability: *Between Regular and Irregular*: All models are significantly correlated with total production probability for regulars. For irregulars, only the model with 3000 verbs is not significantly correlated with total production probability. In general, the correlation for regulars are higher than irregulars, but there is no significant differences. *Among models*: Similar to the rating, the model with 246 verbs has higher correlation with total production probability. There is no significant differences among correlations detected. Only the model with 246 verbs correlates with Exp 1. production probability better than CMS and KC.

Summary: In general, our models show significant correlations with production probability and rating for both regulars and irregulars. The models have higher correlations with regulars than irregulars. Model with 246 verbs correlates with adult's production probability and rating better than other models. It is puzzling that models with more training verbs did not have better correlation. One possible explanation is that the irregular proportion in the model with 246 verbs (26%) is higher than other models, which better represents the verbs distribution that adults exposed to.

4.1.2 Correlation with Cognitive Models

Between Regular and Irregular: All models are significantly correlated with analogy score for regulars. Model with 246, 500 and 1500 verbs are correlated with rule-based score in Pearson's r for regulars. The correlations with rule-based score is not significantly different from the analogy score for regulars. For irregulars, only model with 1500 verbs is significantly correlated with analogy score; models with 246, 500 and 1500 verbs are significantly correlated with rule-based score. It seems that analogy score better correlates with regulars and rule-based score better correlates with irregulars. *Among models*: For regulars, the model with 246 verbs has the highest correlation with analogy score and rule-based score, and is significantly

		Regular					Irregular				
		Adult Behavior			Cognitive Model Acceptance Score		Adult behavior			Cognitive Model Acceptance Score	
Size		Exp1. Prob.	Total Prob.	Rating	Rule based	Analogy	Exp1. Prob.	Total Prob.	Rating	Rule based	Analogy
246	ρ	**0.53**	**0.67**	**0.71**†	0.26	**0.57**	**0.45**	**0.52**	**0.51**	**0.31**	0.17
	r	**0.67**	**0.76**	**0.77**	**0.48**	**0.58**	**0.61**	**0.75**	**0.66**	**0.34**	0
500	ρ	**0.36**	**0.47**	**0.49**	0.11	**0.34**	**0.27**	**0.38**	**0.34**	0.11	0.01
	r	**0.37**	**0.47**	**0.53**	**0.35**	**0.35**	0.20	**0.35**	**0.38**	0.25	0.02
1500	ρ	**0.37**	**0.50**	**0.59**†	0.24	**0.35**	0.21	**0.28**	**0.33**	**0.39**	**0.32**
	r	0.22	**0.41**	**0.46**	0.25	**0.27**	-0.02	0.21	**0.30**	**0.34**	0.1
3000	ρ	0.20	**0.31**	**0.41**	0.2	**0.26**	0.04	0.2	**0.28**	**0.34**	0.25
	r	**0.42**	**0.50**	**0.52**	**0.33**	**0.32**	-0.04	0.2	**0.29**	**0.33**	0.09
CMS	ρ	*0.45*		*0.43*			*0.19*		*0.31*		
	r	*0.30*		*0.40*			*0.17*		*0.40*		
KC	r	*0.48*					*0.45*				

Prob. = Production Probability

† indicates a significant difference between regular and irregular

Table 4: Correlations between model's production probability vs. adult data and cognitive models' score. Significant correlations highlighted in bold. CMS and KC didn't report significance level.

		Regular						Irregular					
Model		Rating		Production Probability				Rating		Production Probability			
Size		Age 6-7	Age 9-10	Age 3-4	Age 5-6	Age 6-7	Age 9-10	Age 6-7	Age 9-10	Age 3-4	Age 5-6	Age 6-7	Age 9-10
246	ρ	-0.03	**0.34**	0.12	-0.02	0.11	**0.36**	0.31	0.14	**0.53**	**0.6**	**0.56**	**0.44**
	r	0.01	0.32	0.11	0.03	0.12	0.29	**0.48**	0.1	**0.63**	**0.59**	**0.57**	**0.47**
500	ρ	-0.17	0.14	0.02	0.15	0.15	0.24	0.21	0.1	0.31	**0.36**	0.27	0.2
	r	-0.07	0.11	0.01	0.15	0.17	0.12	**0.35**	0.09	0.14	0.16	0.08	0
1500	ρ	-0.22	0.18	0	0.02	-0.02	0.27	0.23	0.28	0.04	0.05	0.2	0.19
	r	-0.05	0	-0.1	0.02	-0.06	0.15	0.27	0.1	-0.06	-0.06	-0.04	-0.08
3000	ρ	-0.12	0.09	0.06	0.08	0.1	0.19	0.15	0.27	-0.12	-0.11	0.09	0
	r	-0.1	-0.01	-0.11	0	0.03	0.09	0.26	0.1	-0.08	-0.08	-0.05	-0.09

Rating data are from Ambridge (2010). Production Probability data are from Blything et al. (2018).

Table 5: Correlations between model's production probability vs children's rating and production Probability

higher than model with 1500 verbs and 300 verbs. For irregulars, the correlations of rule-based score are not significantly different among models. **Summary:** The models better correlate with analogy score for regulars, and rule-based score for irregulars. This result seems to suggest that the neural network might have separate mechanisms: for regulars, it behaves more like analogy model that learns the phonological similarities of regulars; for irregulars, it behaves more like rule-based model that learns different levels of rules.

4.1.3 Correlation with children's behavior

Rating: Only three pairs of significant correlations were found in ratings, as shown in Table 5. Model with 246 verbs is significantly correlated with regular ratings for age 9-10. Model with 246 and 500 verbs are significantly correlated with irregular ratings for age 6-7. No other models are correlated with children's rating data. **Production Probability**: Model with 246 verbs is significantly correlated with irregulars for all age groups, and only correlated with regulars for age 9-10. There is also a significant correlation found between model with 500 verbs and age 5-6 for irregulars. No other

significant correlations were found.

Summary: Model with 246 verbs is highly correlated with children's irregular production probability across all age groups, but not with regulars. None of the other models correlate with children's data. We expect the model with 246 verbs to perform better than other models since the it matches parent's input distribution. However, it is baffling why it only correlates with irregulars but not regulars. One possible explanation could be found in the similar dichotomy in the correlation with rule-based model and analogy model. Since the model with 246 verbs also only correlates with rule-based model for irregulars, the mechanism to process irregulars for model and children might be more closer to what rule-based model describes, therefore resulting in high correlation.

4.2 Experiment 2: Evaluating on real verbs

In this experiment, we aim to conduct an error analysis on the models' real verb output to see if there's differentiation between regulars and irregulars and if the models make any overregularization errors.

First, we report the test accuracy on the real verb set, listed in Table 6. The large-size models (with 500, 1500 and 3000 verbs) reached near-perfect accuracy for the regular verbs and the small-size model's accuracy is poor. Also, all model's achieved some accuracy on irregular verbs.

	Regulars %			Irr
Size	/-d/	/-t/	/-d/	irr %
246	80 (5.4)	89 (4.2)	49 (8.8)	17 (4.6)
500	98 (1.7)	97 (1.7)	96 (3.3)	5 (3.2)
1500	99 (1.2)	98 (1.4)	99 (1.2)	13 (4.7)
3000	99 (1.2)	99 (1.3)	99 (2.2)	27 (3.6)

Table 6: Mean accuracy of test set with 170 verbs (standard deviations in brackets) averaged over 30 samples for each data size. There might be some overlapping in the training data and test data, since training data are generated randomly.

4.2.1 Distinction between regulars and irregulars

We analyzed all the errors made by each model with different data size and roughly divided them into 5 categories. **1. No change:** the model output is the same as the root, e.g. 'oversee': /oʊvərsi/ - */oʊvərsi/, 'teach': /tiʧ/ - */tiʧ/ **2. Plural /d/**: the model erroneously produced multiple /d/s at the end of the verb, e.g. 'withdraw': /wɪθdrɔ/ - */wɪθdrɔ*ddddddddd*/. **3. Allomorphy**: the model either output a wrong regular ending to a regular verb, e.g. 'bribe': /braɪb/ - */braɪb*t*/; or output a regular ending to an irregular verb, e.g. 'retell': /ritɛl/ - */ritɛl*d*/. **4. Consonant change**: the model erroneously changed the consonant in the root, e.g. 'secure': /sɪkj*ʊ*r/ - */sɪk*t*ʊ*r*d/, 'force-feed': /fɔrsʃi d/ - */fɔrs*t*id/'. **5. Vowel change**: the model erroneously changed the vowel in the root, e.g. 'rewrite': /rir*aɪ*t/ - */rir*oɪ*t/', 'giggle': /gɪgəl/ - */g*a*gəld/.

We tabulated each model's different types of error in contingency Table 7 and conducted chi-square analysis to test if there is association between error types and regularity. Since some cell numbers are lower than 5, we used Fisher's exact test instead of chi-square test. The p-value is significant for model with 246 verbs, 500 verbs and 1500 verbs, suggesting that these models make different errors for regulars and irregulars. There is no significant distinction in error types for regulars and irregulars in model with 3000 verbs, probably due to the low number of errors. The error type associations with regularity are different for model with 246, 500 and 1500 verbs, as shown in Table 8. All three models tend to make Plural /d/ and Allomorphy errors on irregulars. Model with 246 and 500 verbs tend to make No change and Vowel change errors on regulars. Model with 500 and 1500 verbs tend to make Consonant change errors on irregulars. The differences in the regular-irregular association might be explained the low number of errors on regulars in model with 500 and 1500 verbs.

Si-	246		500		1500		3000	
ze	R	I	R	I	R	I	R	I
1	591	44	60	42	6	57	7	43
2	4	83	3	275	1	78	0	19
3	31	62	7	88	2	107	4	32
4	134	48	11	85	8	116	7	48
5	466	115	60	37	31	52	14	48
p	**<.001**		**<.001**		**<.001**		0.14	

p=Fisher's test p value, R=regular, I=irregular,
1=No change, 2=Plural /d/, 3=Allomorphy
4=Consonant Change, 5=Vowel Change

Table 7: Contingency table of the frequency of errors of different type in models with different size. The Fisher's exact p-value is significant for three models, highlighted in bold.

The distinction between regular error type and irregular error type is very interesting. We won-

Size	246	500	1500
1.No change	Reg	Reg	Irr
2.Plural /d/	Irr	Irr	Irr
3.Allomorphy	Irr	Irr	Irr
4.Consonant Change	Reg	Irr	Irr
5.Vowel Change	Reg	Reg	Irr

Table 8: The different types of errors each model tend to make on regulars or irregulars

der how the model learned this distinction: is it learned based on the verb stem or the past tense forms? To further investigate this distinction, we trained 6 more models with only regular verbs with training size ranging from 500 - 3000 and tested it on the same real verb test set. Since there is no irregular verbs in the training data, we expect model to produce the regular past tense ('+ed') for the irregulars. The 6 models all have 100 accuracy on regulars and 0 accuracy in irregulars. However, we only found 2 '+ed' errors on the irregulars: 'deal': /dild/, 'retell':/ritɛld/. All the models produced Plural /d/ errors on the rest of the 18 irregular verbs. This result further confirms that the model learned the regular-irregular distinction, and suggests that the distinction is learned from verb stem.

4.2.2 Overregularization Errors on irregulars

We found all three types of overregularization errors in our model output, as listed in Table 9. In addition, the model also made many novel errors, such as incomplete suffix (e.g. rewrite - */riraɪtɪ/), double suffix (e.g. awake - */əweɪktd/) and truncation (e.g. stand - */stæn/). A more careful qualitative analysis on these errors should help us to understand more of the model's behavior.

Type	Examples
Stem+ed	deal - */dild/*, stick - */stɪkt/*
Past+ed	sink - */sæŋkt/*, awake - */əwəʊkt/*
Incorrect vowel change	swing-*/swæŋ/*, oversee-*/oversɛ/*

Table 9: Examples of overregularization errors made by models

5 Discussion

In this work, we showed that neural networks can be potential cognitive models by connecting transformer models with human behaviors and cognitive theories. We found that all neural models have significant correlations with adult behavior's in both regulars and irregulars. Small-size model correlates with children's irregular behavior, but not the regulars. The models correlate with rule-based model on regulars and with analogy model on irregulars. The dichotomy in correlations with cognitive theories and children's data suggested that the model's behavior and children's behavior on irregular verbs are more closer to what rule-based theory describes. The summary of correlation is listed in Table 10. We also found overregularization errors the models make that are similar to children's errors. Although the models make many non-human like errors, we show that these errors exhibit a clear distinction between regulars and irregulars. The model possibly learned the regular-irregular distinction from the verb stem instead of the past tense forms. The error data also confirms that models mimic human behavior.

Correlation	Regular	Irregular
Adults	✓	✓
Children	×	✓
Rule-based	×	✓
Analogy based	✓	×

Table 10: Summary of correlations of model vs adult, children and rule-based theory and analogy based theory

One important difference of our neural models and KC, CMS is that we manipulated the training data. We showed that model with small-size training data with high proportion of irregulars correlates better with human behavior and cognitive models' score. However, the small-size model that replicates parents' verb distribution generally have lower accuracy than human children. If we can improve the accuracy without flooding the model with more training data, we could better demonstrate that neural networks can be good cognitive models.

To further evaluate neural networks, there are many other potential aspects that can be explored, such as a more careful error analysis, inflections in other languages, or visualizing hidden layers to help us understand what the neural networks learned. We hope that our evaluation could motivate more future explorations of neural networks as cognitive models.

References

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

Ben Ambridge. 2010. Children's judgments of regular and irregular novel past-tense forms: New data on the english past-tense debate. *Developmental Psychology*, 46(6):1497.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

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

Lois Bloom. 1973. *One word at a time: The use of single word utterances before syntax*, volume 154. Walter de Gruyter.

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

Ryan P Blything, Ben Ambridge, and Elena VM Lieven. 2018. Children's acquisition of the english past-tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*, 42:621–639.

Michael B Broe. 1993. *Specification theory: the treatment of redundancy in generative phonology*. Ph.D. thesis, University of Edinburgh.

Joan L Bybee and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language*, pages 251–270.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 196–204.

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.

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

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.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Michael Maratsos. 2000. More overregularizations after all: new data and discussion on marcus, pinker, ullman, hollander, rosen & xu. *Journal of Child Language*, 27(1):183–212.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178.

Robert JC Maslen, Anna L Theakston, Elena VM Lieven, and Michael Tomasello. 2004. A dense corpus study of past tense and plural overregularization in english.

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.

Kim Plunkett and Patrick Juola. 1999. A connectionist model of english past tense and plural morphology. *Cognitive Science*, 23(4):463–490.

Kim Plunkett and Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

David. E. Rumelhart and James L. McClelland. 1986. *On Learning the Past Tenses of English Verbs*, page 216–271. Cambridge, MA, USA.

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

Linnaea Stockall and Alec Marantz. 2006. A single route, full decomposition model of morphological complexity: Meg evidence. *The mental lexicon*, 1(1):85–123.

Lorraine K Tyler, Emmanuel A Stamatakis, Brechtje Post, Billi Randall, and William Marslen-Wilson. 2005. Temporal and frontal systems in speech comprehension: An fmri study of past tense processing. *Neuropsychologia*, 43(13):1963–1974.

Michael T Ullman, Suzanne Corkin, Marie Coppola, Gregory Hickok, John H Growdon, Walter J Koroshetz, and Steven Pinker. 1997. A neural dissociation within language: Evidence that the mental

dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of cognitive neuroscience*, 9(2):266–276.

Wayne A Wickelgren. 1969. Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, 76(1):1.

Fei Xu and Steven Pinker. 1995. Weird past tense forms. *Journal of child language*, 22(3):531–556.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

A Appendix

Tokens		Parent's Regular			Parent's Irregular
Child	Files	*/-d/*	*/-t/*	*/-ɪd/*	*irr*
Adam[1]	18	18	18	3	36
Eve[1]	5	5	7	3	18
Sarah[1]	33	13	17	0	33
Peter[2]	14	1	3	0	8
Naomi[3]	20	9	9	4	27
Allison[4]	6	8	4	1	18
April[5]	2	5	5	1	17
Fraser[6]	90	83	44	17	62

1.Bloom (1973), 2.Bloom et al. (1974), 3.Sachs (1983), 4.Bloom (1973), 5. Higginson (1985), 6.Lieven et al. (2009)

Table 11: Summary of each parent's regular verb and irregular verb tokens