UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Falling Through the Gaps: Neural Architectures as Models of Morphological Rule Learning

Permalink

https://escholarship.org/uc/item/1dp545wm

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

Author

Beser, Deniz

Publication Date

2021

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Falling Through the Gaps: Neural Architectures as Models of Morphological Rule Learning

Deniz Beser (beser@isi.edu)
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292

Abstract

Recent advances in neural architectures have revived the problem of morphological rule learning. We evaluate the Transformer as a model of morphological rule learning and compare it with Recurrent Neural Networks (RNN) on English, German, and Russian. We bring to the fore a hitherto overlooked problem, the morphological gaps, where the expected inflection of a word is missing. For example, 63 Russian verbs lack a first-person-singular present form such that one cannot comfortably say **oščušču* ('I feel'). Even English has gaps, such as the past participle of *stride*: the function of morphological inflection can be partial. Both neural architectures produce inflections that ought to be missing. Analyses reveal that Transformers recapitulate the statistical distribution of inflections in the training data, similar to RNNs. Models' success on English and German is driven by the fact that rules in these languages can be identified with the majority forms, which is not universal.[1]

Keywords: Morphology, Neural Networks, Cognitive Modeling

Introduction

A well-recognized aspect of language acquisition is the learning of productive morphological rules, such as the English past tense (e.g. *walk + /-d/; walked*) (Berko, 1958; Albright & Hayes, 2003; Lignos & Yang, 2018). Neural networks were first proposed as models of learning such morphological rules decades ago (Rumelhart & McClelland, 1985), and were faced with a seminal critique that highlighted their numerous linguistic inadequacies (Pinker & Prince, 1988). In light of the recent advancements in the field of natural language processing (NLP), the encoder-decoder recurrent neural network (RNN) architecture, a sequence-to-sequence model commonly used for machine translation (Sutskever, Vinyals, & Le, 2014), has been reconsidered as a model for learning productive morphological rules (Kirov & Cotterell, 2018). Corkery, Matusevych, and Goldwater (2019) compared the performance of the RNN architecture to human behavior on English past tense, and concluded that the fit to human data is weak. Similarly, McCurdy, Goldwater, and Lopez (2020) used the German plural to demonstrate that the RNN's behavior does not match human subjects and the model is susceptible to frequency distribution in the data, an argument first raised by Marcus, Brinkmann, Clahsen, Wiese, and Pinker (1995).

Evidently, the encoder-decoder RNN is insufficent as a cognitive model of morphological rule learning. However, RNNs are no longer regarded as the state-of-the-art system for language modeling in NLP. Vaswani et al. (2017) introduced the Transformer architecture, a neural architecture that relies on attention mechanism and does not use recurrence. The Transformer architecture has since outperformed RNNs across a wide range of tasks (Vaswani et al., 2017; Liu et al., 2019; Brown et al., 2020), including morphological inflection generation (Wu, Cotterell, & Hulden, 2021). The outstanding performance of the Transformer calls for an evaluation as a cognitive model for learning productive morphological rules.

In this paper, we first compare the performances of the Transformer architecture and the encoder-decoder RNNs across morphological phenomena in natural language that suggest humans seek certain structures in the languages they hear in order to learn productive morphological rules. Namely, we use the productivity observed in English past tense (Berko, 1958; Albright & Hayes, 2003) and the non-majority productivity in German plurals (Elsen, 2002; Bartke, Rösler, Streb, & Wiese, 2005; Zaretsky & Lange, 2015) which have previously been used for assessing RNNs as models of morphological learning (Kirov & Cotterell, 2018; Corkery et al., 2019; McCurdy, Lopez, & Goldwater, 2020). We then evaluate both architectures on a set of Russian verbs, an example of gaps in morphological productivity where native speakers of the language fail to generalize a productive rule and produce an inflected form (Sims, 2006; Pertsova, 2016; Gorman & Yang, 2017). Our analyses on the English and German languages show that the Transformer architecture is susceptible to the same analogical and distributional errors as RNNs. Likewise, both neural architectures produce inflected forms for the gapped Russian verbs while native speakers cannot. We argue that the models' predictions on the gapped words evince a fundamental challenge for both Transformers and RNNs as models of language learning and underline the need for additional structural priors for neural models to capture how humans expect and utilize structure in language data to generalize a rule productively.

Background

Majority Productivity: The English Past Tense

What is the past tense of the verb *google*? The English past tense is a prevalent example of morphological productivity in linguistic literature; a vast majority of English verbs follow a typical inflectional pattern (e.g. *walk-walked, kiss-kissed*) and a smaller subset of the verbs are exceptions (e.g.

[1] Our code and data are available at github.com/denizbeser/gaps

run-ran, think-thought) (Rumelhart & McClelland, 1985; Xu & Pinker, 1995; Albright & Hayes, 2003; Lignos & Yang, 2018). Children can learn these morphological rules from a young age despite sparsity of the data and even though they might not have heard the past tense of *'look'* before, they can effortlessly produce the past tense *'looked'* (Berko, 1958). They do not memorize all the words and their past tenses, but extract a productive rule and apply it to novel words. Importantly, this productive rule-learning is categorical in nature as children consistently apply to novel words and generally do not make analogical errors of 'overirregularization'; they categorically produce the past tense pair *bling-blinged*, despite the existence of an irregular pair *sing-sang*) (Berko, 1958; MacWhinney, 1978; Xu & Pinker, 1995). This rule-learning paradigm is further evindenced by patterns in language acquisition, including overgeneralization errors (Marcus et al., 1992) as children tend to apply learned rules to irregular verbs to produce regular inflections (e.g. *think-thinked*) and then correct the behavior over time. Previous work on the RNN's effectiveness on modeling the English past tense found that the model performs worse than rule-based systems and does not correlate strongly with human behavior (Corkery et al., 2019).

Since most of the English verbs follow this productive rule, a simple majority-rule could explain the productivity of the English past tense; however, there are telltale paradigms observed in language acquisition across different languages which show that the learning behavior is much more complicated, where the majority rule is not the productive one (Zaretsky & Lange, 2015; Gorman & Yang, 2017).

Non-Majority Productivity: The German Plural

While the English past tense can be expressed by a simple majority rule, sometimes no marker may form a majority, and even the least frequent marker may be applied productively (Elsen, 2002; Bartke et al., 2005; Yang, 2016). The German plural is a fitting example for non-majority productivity in rule learning as native speakers use multiple plural markers productively (Zaretsky & Lange, 2015; McCurdy, Goldwater, & Lopez, 2020). German has three grammatical genders and primarily five different types of suffixes that mark plural nouns (e.g. *die Frau - die Frauen*) partially categorized by gender. The most frequent of these markers is the /-(e)n/ with 48% type and 45% token ratio, and the least frequent of these is /-s/ with 4% type and 2% token ratio. The other three common suffix types for German plurals are /-e/, /-er/, and /-Ø/ denoting no change (Sonnenstuhl & Huth, 2002; Yang, 2016). See figure 1 for an overview of the distribution of plural markers in German.

German speakers certainly do not memorize the plural form of every noun; despite the lack of an apparent majority, they learn to produce plural forms using productive rules. However, there has been mixed arguments on the nature of the German plural: while the minority plural suffix /-s/ was first found to be applied to novel words (Clahsen, Rothweiler, Woest, & Marcus, 1992) and argued to be the default productive marker (Marcus et al., 1995), recent research suggests that the more frequent suffixes /-e/ and /-(e)n/ are also used productively on novel words (Zaretsky & Lange, 2015) and the inflections are influenced by both gender and phonological properties (McCurdy, Lopez, & Goldwater, 2020). Our analyses in this paper follow the perspective of the most recent studies on this subject that suggest there are multiple productive rules applied based on gender and phonological features (Zaretsky & Lange, 2015; McCurdy, Goldwater, & Lopez, 2020). Previous work on evaluating RNNs performance on German plurals have found that the model is biased toward the most frequent suffix (McCurdy, Goldwater, & Lopez, 2020) and is oversensitive to gender (McCurdy, Lopez, & Goldwater, 2020).

Suffix	Singular	Plural	Gloss	Type	Token
-ø	der Daumen	die Daumen	"thumbs"	4320	87088
	die Mutter	die Mütter	"mothers"	(17%)	(29%)
-e	der Hund	die Hunde	"dogs"	6836	62239
	die Kuh	die Kühe	"cows"	27%	21%
-er	der Wald	die Wälder	"forests"	1067	10158
	das Huhn	die Hühner	"hens"	(4%)	(3%)
-(e)n	die Strasse	die Strassen	"streets"	12365	134492
	die Frau	die Frauen	"women"	48%	45%
-s	das Auto	die Autos	"cars"	1061	5468
	der Park	die Parks	"parks"	(4%)	(2%)

Figure 1: Types of German plural suffixes and their distribution with examples, based on the CELEX corpus (Sonnenstuhl & Huth, 2002; Yang, 2016).

Morphological Gaps: Defective Verbs in Russian

Besides the majority productivity of English past tense and non-majority rules observed in German plurals, there are numerous cases across languages where there is a gap in productivity, where the speakers fail to comfortably produce an inflected form, despite an apparent majority pattern (Sims, 2006; Pertsova, 2016; Gorman & Yang, 2017). For instance, consider the past participle of the English verb *stride*. Native speakers of English struggle to pick a past participle for this verb (*strode, stridden, strided, ...*) as there is no apparent productive rule for English past participle (Gorman & Yang, 2017).

In Russian, there is a defective gap of this kind in 63 second conjugation verbs (Sims, 2006; Pertsova, 2016). These defective verbs have stems ending in dental consonants (*t, d, s, z*), and in the first person singular (1SG) present form, they are expected to alternate with a palatal or alveo-palatal fricative (Pertsova, 2016). Yet, native speakers of Russian struggle to come up with an absolute 1SG word form for this set of second conjugation verbs (Pertsova, 2016). Some examples of these gapped verbs include *'to vacuum' (*pylesosit')*, *'to feel' (*oščutit')*, and *'to misbehave' (*škodit')*. We evaluate how this paradigm may pose a challenge for the Transformer and RNNs due to neural networks' distributional tendencies

	RNN				Transformer				Prediction	Prediction
	Train	**Dev**	**Test**	**/-d/ %**	**Train**	**Dev**	**Test**	**/-d/ %**	**Correlation**	**Similarity**
All Verbs	96.8	98.2	80.2	**97.6**	99.3	98.6	94.0	**99.6**	0.877	94.6
Irregulars	-	-	18.5	**96.3**	-	-	7.4	**92.6**	0.734	88.9

Table 1: Breakdown of model performance on the two English *test* sets, one that consists of regulars and irregulars, and one that consists of irregulars. The model is trained once on all verbs, a set of regular and irregular verbs, and then tested on both all verbs and a subset of irregular verbs. The /-d/ % column reports the percentage of predictions that use the /-d/ productive rule. The last two columns compare the Spearman correlation ($p<0.01$) and similarity of models' inflection predictions on the test set as explained in the Evaluation Overview section.

(Corkery et al., 2019; McCurdy, Lopez, & Goldwater, 2020)

Evaluations

Overview To compare the performances of the Transformer and RNN architectures on morphological patterns in each language, we train the separate instances of both architectures on pairs of bare and inflected word forms, where each word is represented as a sequence of space-separated characters (Kirov & Cotterell, 2018; Wu et al., 2021) In other words, the models learn to translate one sequence of characters to another. During testing, we perform 'wug tests' on the models by presenting novel words to the models and evaluating the inflected word forms the models predict (Berko, 1958). In our evaluations, we assess prediction accuracy on relevant test sets for the language and the distribution of inflections predicted by the models. We also measure the Pearson correlation coefficient and *prediction similarity* (measured as accuracy across predictions) between the inflection predictions of the RNN and Transformer architectures.

Implementation Details The Transformer models use eight attention heads, have two layers, model dimension of 312, and feed-forward network dimension of 512. The decoder uses beam search with beam size of 12. The Transformer models are trained for 20,000 steps.

The RNN implementation is kept identical to Kirov and Cotterell (2018). The RNNs are trained for 100 epochs, with a batch size of 20. Both encoder and decoder RNNs are bidirectional LSTMs (Schuster & Paliwal, 1997) with two layers and 100 hidden units, with a vector size of 300. The RNN decoder also uses beam search with beam size of 12. Both the RNN and the Transformer implementations are based on OpenNMT (Klein, Kim, Deng, Senellart, & Rush, 2017).

The English Past Tense

Method We train the models on pairs of English verbs with the infinitive and past-tense form sampled from English CELEX (Baayen, Piepenbrock, & Gulikers, 1995). The verbs are represented in phonetic form. The dataset uses type frequency for training; i.e each verb appears once in the training dataset. The dataset had 4,000 datapoints for training, 500 for development, and 500 for testing. During testing we use test accuracy, the prediction rate of the /-d/ majority productive rule, and the correlation and similarity between the models. In addition to evaluating the model performance on the complete test set which contains both regular and irregular English past tense verbs, we also measure the performance on the irregular subset of test verbs.

Results and Discussion The results of English evaluation are shown in Table 1. The Transformer architecture achieves a higher prediction accuracy and /-d/ prediction rate than the RNN when tested on all 500 test verbs. This trend is reversed when the models are tested on the irregular verbs; in this case, the Transformer has a lower test accuracy and /-d/ prediction rate than the RNN. There is a high resemblance between the predictions made by the models; when tested on all verbs, the correlation between the models is 0.877, and 94.6% of predictions are identical. When tested on irregular verbs, the correlation and the similarity between the models are both lower, 0.734 and 88.9%, respectively. Both models achieve high /-d/ rule prediction rates, meaning that they prioritize the correct productive rule. However, when this rule prediction rate is compared across test conditions, we see that both models have a lower rate when tested on irregular verbs. The drop in the /-d/ prediction rate when tested on irregulars, which is higher for the Transformer than for the RNN, suggests that both models, especially the Transformer, are susceptible to analogical errors that are made by adults but not by children (Berko, 1958; Albright & Hayes, 2003) and that neither model does not in fact learn a single productive rule.

The German Plural

Method 1: Model Comparison To compare the models on the acquisition of the German plural, we train both networks with singular-plural verb pairs marked with their gender. [2] The training, development, and test datasets for German had 4000, 500, and 500 datapoints, respectively. After training is complete, each model is wug-tested on the two test sets. In the first dataset, all test nouns are marked with neuter gender. In the second dataset, each noun is marked with its correct gender. We then evaluate the prediction rates of each plural marker on different subsets of the test datasets, separated by the original gender marker of the nouns in the dataset.

Method 2: Human Data Comparison In our second evaluation for German, we use data provided by McCurdy, Goldwater, and Lopez (2020) to compare the two models with

[2] We use the German CELEX dataset (Baayen et al., 1995)

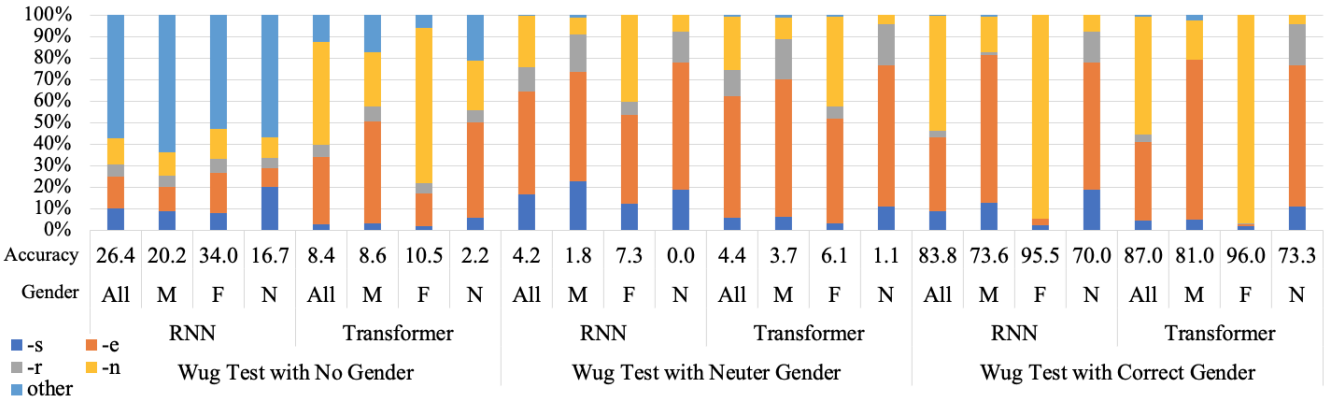


Figure 2: The cumulative distribution of predicted inflections for German nouns across model types and test datasets. Prediction accuracy on test sets separated by gender are marked on the x axis.

the human behavior. Similar to McCurdy, Goldwater, and Lopez (2020), we train 25 separate instances of the RNN and Transformers. For each model instance, we use a different 4000/500/500 train/dev/test random split from 5000 German CELEX verbs pairs in orthographic form.[3] We test the models on the CELEX test set, and use the experiment test set provided by McCurdy, Goldwater, and Lopez (2020) for human behavior comparison.

Results and Discussion The test performance for both models with the distribution of the predictions for each test set are shown in Figure 2. The models achieve a high test accuracy and make similar predictions when tested on an unseen test set of 500 novel nouns marked with the correct gender. The correlation of predictions of two architectures is 0.67 ($p<0.01$) when tested with correct gender. The accuracy of both models drop significantly when tested in no-gender and neuter gender test sets, which aligns with RNNs oversensitivity to gender (McCurdy, Lopez, & Goldwater, 2020). Both models produce a higher number of inflections that are different from the plural markers when the test set is not marked with gender information. While this positively indicates that the models learn some associations between a noun's gender and its inflected form, it also shows that the models' weakness in making inference with partial information and the predicted inflections may depend too much on the grammatical gender.

When evaluated with the second method based on McCurdy, Goldwater, and Lopez (2020), neither model is found to be correlated strongly to human behavior across any condition, as shown in Table 2. When the performance is measured across suffixes (Table 2, left), both models perform best ($F1 >= 0.90$) with the /-(e)n/ suffix, which has the highest type frequency in the German corpus, but perform relatively poorly on other suffixes. Despite the lack of a strong and statistically significant correlation between the two model's predictions, the similar trend in the precision, recall, and F1 scores of the models across suffixes suggest that the Transformer performs similarly to the RNN and also struggles to generalize productive rules in minority classes.

	RNN			Transformer			Correlations		
	Pr.	Rec.	F1	Pr.	Rec.	F1	R - T	R - H	T- H
/-(e)n/	.91	.90	.90	.94	.91	.93	-.25	-.07	.23
/-e/	.79	.74	.74	.74	.90	.81	.41	.00	.05
/-Ø/	1.00	.61	.74	1.00	.42	.57	.00	.00	.00
/-s/	.40	.29	.30	.71	.21	.31	.32	.44	.01
/-r/	.51	.48	.45	.70	.39	.48	**.55**	.22	.44
other	.00	.00	.00	.00	.00	.00	.00	-.31	.00

Table 2: Model performance (precision, recall, F1) on each suffix in CELEX test set, averaged across 25 splits (left). Spearman correlation between models and human data (RNN-Transformer, RNN-Human, Transformer-Human) when tested on the experiment data set (right). Only one of the correlations were statistically significant ($p<0.01$), marked in bold. The Transformer did not predict any /-(e)n/ when tested on this test set.

The Russian Morphological Gap

Method In order to test the models' behavior on the gaps observed in Russian, we train and evaluate multiple separate instances of both models on 5 different datasets consisting of pairs of non-gapped Russian verbs in present infinitive and their present first person singular form with Cyrillic alphabet.[4] Each training dataset consists of all or a subset of the Russian verbs and their inflected forms, selected based on the criteria listed on Table 3. We use a 80% training, 10% and 10% testing split for each dataset.[5] During testing for each dataset, we wug-test the models on the corresponding 10%

[3]The RNN is trained for 50 epochs and the Transformer is trained for 10,000 steps.

[4]Training and testing word forms are obtained from the Russian Dictionary dataset: https://github.com/Badestrand/russian-dictionary

[5]Training steps for the Transformer are increased proportionally to the training dataset size.

Dataset	RNN	Transformer	Correlation	Similarity
All Verbs	88.5	82.1	0.855	93.0
1st and 2nd Conjugation	82.2	83.1	0.848	93.0
1st Conjugation	84.8	86.4	0.879	96.0
2nd Conjugation	83.9	84.1	0.888	90.6
Dental-stem 2nd Conj	83.8	84.6	0.844	93.7

Table 3: Model comparison on Russian verbs. The reported numbers are accuracy scores on the unseen test set within the dataset per model, and finally the Spearman correlation prediction similarities between the predictions made by the models. All correlation scores have *p<0.01*.

English	Russian Inf.	RNN	Transformer
to protest	бузить	бужу'	бу'жу
to yell	голосить	гололошу'	голоси'ю
to dream	грезить	грежу'	гре'жу
to itch	зудеть	зуде'ю	зуде'ю

Table 4: Examples from the predictions produced by the RNN and the Transformer models on the set of gapped verbs.

test set, as well as a special test set that consists of the gapped verbs.[6]

Since it is technically infeasible to provide the neural sequence-to-sequence models with an explicit option to reject inflecting a given word other than using the models' implicit ability to return the original form, in order to assess whether the models have any difficulty in producing inflected forms resembling the cognitive block humans experience, we compare the models' prediction confidence across the test sets by measuring the likelihood of the predicted inflected forms. If the models struggle to produce inflected forms for the gapped verbs, we would expect to see a statistically significant drop in the models' prediction confidence for those verbs.

Results and Discussion

The results of the Russian evaluation are shown in Tables 3 and 4. Regardless of the verb data used to train the models, both neural architectures achieve similar accuracy scores. The architectures make similar predictions, as indicated by statistically significant correlations and high predicted inflection similarities between the predictions made by the two models across all conditions. Examples of the gapped verbs and corresponding inflected forms produced by the neural models are presented in Table 4. We find no significant change in the models' prediction confidences across the regular and gap test sets. When trained on the 'all verbs' dataset, the change in prediction confidences across the test sets is less than 10% for both models, a decrease from 0.95 ($\sigma = 0.12$) to 0.86 ($\sigma = 0.18$) for the RNN model, and a decrease from 0.46 ($\sigma = 0.11$) to 0.39 ($\sigma = 0.13$) for the Transformer model, which suggests that neither model captures the gaps in productivity.

[6]We use the list of gapped words provided by Pertsova (2016).

Conclusion & General Discussion

We evaluated the Transformer architecture as a model for morphological rule learning and compared it to encoder-decoder RNNs. In our evaluations, we have used three telltale phenomena in morphology that show how children and native speakers expect certain structure in language that cannot be extracted from data per se. Specifically, we have evaluated models on the English past tense where the productive morphological rule is the majority form (Berko, 1958; Albright & Hayes, 2003; Lignos & Yang, 2018), the German plural where multiple non-majority morphological rules are applied productively (Marcus et al., 1995; Zaretsky & Lange, 2015; McCurdy, Goldwater, & Lopez, 2020), and finally the gaps observed in 63 Russian dental-stem second conjugation verbs for which native speakers cannot comfortably produce a first-person-singular present form (Sims, 2006; Pertsova, 2016).

In our evaluations on the English past tense, we observed that the Transformer architecture, perhaps due to having more power in capturing distributional properties, is more susceptible to analogical errors when tested on irregular verbs. Our analyses on German plurals suggest that the Transformer, similar to the RNN, is sensitive to the grammatical gender of nouns, is biased towards the most majority suffixes, and does not correlate with human behavior. When evaluated on the gapped Russian verbs, both neural architectures produce inflected first-person-singular forms that are not produced by native speakers, suggesting that the models extract patterns from data and apply them to novel examples without employing the necessary linguistic priors.

As evinced by the models' predictions across the English past tense, the German plurals, and the Russian gaps, a fundamental challenge for neural architectures such as Transformers and RNNs as models of language learning is the need for additional structural priors that can enable learning categorical productive rules. Incorporating such priors into the neural architectures may not only make them more suitable for cognitive modeling, but also strengthen their generalization capabilities.

In sum, we did not find any evidence that would suggest Transformers are better cognitive models of morphological rule learning than RNNs. Both neural architectures recapitulate the statistical distribution of inflection alternatives in the training data, and hence are susceptible to similar analogical and distributional errors, suggesting that these models do

not seek the necessary linguistic structures sought by humans during morphology acquisition. We also suggest that morphological gaps can be used for measuring the effectiveness of neural architectures as cognitive models in the language domain.

Acknowledgements

I would like to thank Prof. Charles Yang for his guidance throughout this research and during my studies at the University of Pennsylvania. I would also like to thank Dr. Asli Rehber Beser for her support and reviews, Anton Relin for assisting with the Russian language, and the anonymous reviewers for valuable suggestions.

References

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

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). Celex2 ldc96l14. *Web Download. Philadelphia: Linguistic Data Consortium*.

Bartke, S., Rösler, F., Streb, J., & Wiese, R. (2005). An erp-study of german 'irregular'morphology. *Journal of Neurolinguistics*, *18*(1), 29–55.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901).

Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G. F. (1992). Regular and irregular inflection in the acquisition of german noun plurals. *Cognition*, *45*(3), 225–255.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. In *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, volume 1: Long papers* (pp. 3868–3877). Association for Computational Linguistics.

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

Gorman, K., & Yang, C. (2017). When nobody wins. *Competition in inflection and word formation. Dordrecht: Springer. Google Scholar*.

Kirov, C., & Cotterell, R. (2018). Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, *6*, 651-665.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017, July). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, system demonstrations* (pp. 67–72). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-4012

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the society for research in child development*, 1–123.

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, i–178.

McCurdy, K., Goldwater, S., & Lopez, A. (2020, July). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1745–1756).

McCurdy, K., Lopez, A., & Goldwater, S. (2020). Conditioning, but on which distribution? grammatical gender in german plural inflection. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 59–65).

Pertsova, K. (2016). Transderivational relations and paradigm gaps in russian verbs. *Glossa: a journal of general linguistics*, *1*(1).

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

Rumelhart, D. E., & McClelland, J. L. (1985). *On learning the past tenses of english verbs.* (Tech. Rep.). CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Sims, A. D. (2006). *Minding the gaps: inflectional defectiveness in a paradigmatic theory*. Unpublished doctoral dissertation, The Ohio State University.

Sonnenstuhl, I., & Huth, A. (2002). Processing and representation of german-n plurals: A dual mechanism approach. *Brain and Language*, *81*(1-3), 276–290.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 3104–3112). Curran Associates, Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wu, S., Cotterell, R., & Hulden, M. (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 (pp. 1901–1907). Association for Computational Linguistics.

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of child language*, *22*(3), 531–556.

Yang, C. (2016). *The price of linguistic productivity.* Cambridge, MA: MIT Press.

Zaretsky, E., & Lange, B. P. (2015). No matter how hard we try: Still no default plural marker in nonce nouns in modern high german. In *A blend of malt: Selected contributions from the methods and linguistic theories symposium* (pp. 153–178).