Frequency matters: Modeling irregular morphological patterns in Spanish with Transformers

Anonymous ACL submission

Abstract

Over the past decade, various studies have addressed how speakers solve the so-called 'The Paradigm Cell Filling Problem' (PCFP) (Ackerman et al., 2009) across different languages. The PCFP addresses a fundamental question in morphological processing: how do speakers accurately generate inflected forms of words when presented with incomplete paradigms? This problem is particularly salient when modeling complex inflectional systems. We focus on Spanish verbal paradigms, where certain verbs follow an irregular L-shaped pattern, where the first-person singular present indicative stem matches the stem used throughout the present subjunctive mood. We formulate the problem as a morphological reinflection task. Specifically, we investigate the role of input frequency in the acquisition of regular versus irregular L-shaped patterns in transformer models. By systematically manipulating the input distributions and analyzing model behavior, we reveal four key findings: 1) Models perform better on L-shaped verbs compared to regular verbs, especially in uneven frequency conditions; 2) Robust primacy effects are observed, but no consistent recency effects; 3) Memorization becomes more prominent as the proportion of L-shaped verbs increases; 4) There is a tendency to regularize L-shaped verbs when their consonant alternation pairs are rare or absent in the training data.

1 Introduction

A common generation task in morphology is morphological inflection, where a target form has to be generated from its corresponding lemma and feature tag, e.g., `(lemma:decir, target tag:<V;IND;PRS;1;SG>)` $\mapsto$ `digo`. A central challenge in understanding how speakers handle morphological inflection is the Paradigm Cell Filling Problem (PCFP) (Ackerman et al., 2009), which asks how speakers can reliably produce inflected forms of words when they are presented with incomplete paradigms.

To address the PCFP, encoder-decoder based neural networks have been used to simulate the learning and generation of inflected forms (Silfverberg and Hulden, 2018a; Wiemerslage et al., 2022). Our study extends this line of research by applying the PCFP to a morphomic pattern using encoder-decoder transformers. The morphomic pattern, as introduced by Spencer and Aronoff (1994), is a morphological pattern that exists independently of semantics or syntax. It is purely based on the form and structure of words. A key characteristic of morphomic patterns is their predictability within the verbal paradigm. The verb forms that are part of the pattern share morphological features, despite a lack of apparent semantic or syntactic motivation (Blevins, 2016; Maiden, 2018).

Maiden (2011, 2018, 2021) identified morphomic patterns across Romance languages. We focus on Spanish for data availability reasons (Herce and Allassonnière-Tang, 2024). To illustrate an example, the Spanish verb forms "digo" (1st person singular, indicative) and "digan" (3rd person plural, subjunctive) of the verb *decir* 'to say' (see Table 1) share the stem "dig-". This shared morphological feature is part of a morphomic pattern. However, there is no obvious semantic or syntactic property that links "digo" and "digan" while excluding "dicen" (3rd person plural, indicative) of the same verb, which uses a different stem "dic-".

Spanish exhibits several morphomic patterns, namely L-, N-, P- and F-shaped patterns (Maiden, 2018; Herce and Allassonnière-Tang, 2024). Of these, only the L-shaped pattern has been explored for human learnability (Nevins et al., 2015; Cappellaro et al., 2024). We choose L-shaped pattern for our study because it allows us to assess the cognitive plausibility of our neural network models. The L-shaped pattern is characterized by the use of a distinct stem form in the first person singu-

lar present indicative and all cells of the present subjunctive mood. For example, the irregular verb *decir* exhibits the L-shaped morphome pattern, as shown in Table 1.

'to say'	***Indicative***		***Subjunctive***	
	Orthographic	IPA	Orthographic	IPA
1SG	digo	d'igo	diga	d'iga
2SG	dices	d'ises	digas	d'igas
3SG	dice	d'ise	diga	d'iga
1PL	decimos	des'imos	digamos	dig'amos
2PL	decís	des'is	digáis	dig'ajs
3PL	dicen	d'isen	digan	d'igan

Table 1: A Spanish example of the Romance L-pattern, verb *decir* 'to say'. L-shaped pattern cells are shaded.

Spanish L-shaped verbs exhibit an interesting distribution in the lexicon: they are found in relatively few word types but demonstrate high token frequency, which is the frequency of occurrence of individual word forms (Maiden, 2011). The role of type frequency, which refers to the number of different words that follow a particular morphological pattern in morphological productivity has been well established in the linguistic literature (Bybee, 1995; Pierrehumbert, 2001; Bybee, 2003; Albright and Hayes, 2003; Baer-Henney and van de Vijver, 2012; del Prado Martın et al., 2004). These studies show that patterns with higher type frequency are more likely to be extended to novel forms, suggesting a strong correlation between type frequency and productivity. However, previous studies on L-shaped verbs challenge this established relationship. Nevins et al. (2015) and Cappellaro et al. (2024) find conflicting evidence on productivity of morphomes in human studies. This highlights the need for a computational approach that can systematically explore the factors influencing morphome productivity. Computational modeling allows us to manipulate type frequency in ways that would be challenging in human studies, enabling a more controlled investigation of its effects on morphomic pattern learnability and productivity.

In our study, we investigate the impact of type frequency on the learnability of morphomic patterns, specifically the L-shaped pattern, by implementing a morphological reinflection task framed as a PCFP using transformer models.

The morphological reinflection task aligns with the morphological framework of abstraction based on data directly available to speakers (i.e., inflection forms) (Blevins, 2006; Boyé and Schalchli, 2019), providing a realistic setting. We implement a multi-source setup of the morphological reinflection task (Kann et al., 2017), which uses multiple source form-tag pairs instead of one form-tag pair. The task here is to generate an inflected form from two source form-tag pairs and the target feature tag to predict the target inflected form, e.g., `(source form:digas, source tag:<V;SBJV;PRS;2;SG>, target tag:<V;IND;PRS;1;SG>)` $\mapsto$ `digo`. The choice of two-source setup is motivated by two key considerations: 1) identifying L-shaped verbs requires knowledge of at least two paradigm cells (one cell within the L-shaped pattern, one cell outside; see Table 1), and 2) previous research (Silfverberg and Hulden, 2018b; Liu and Hulden, 2020) shows that two-source form-tag pairs are sufficient for achieving high accuracy in paradigm completion, with no gains from additional sources.

The reinflection task is particularly challenging due to the variability of the starting point (the source), which can be any other inflected form of the same lemma. This variability makes the task more cognitively plausible, reflecting the data sparsity encountered by human speakers, who never encounter all of the inflected forms of their language (Blevins et al., 2017).

The main aims of the study are: 1) To model the learning of the L-shaped morphome in Spanish using transformers for morphological reinflection; 2) To analyze the models' performance across varying input frequency distributions of regular and L-shaped verbs; 3) To conduct post-hoc analyses investigating: a) the impact of paradigm cell combinations on L-shaped learning, b) models' ability to memorize and generalize morphomic patterns, and c) sensitivity to the input frequency of consonant alternations.

We publish the dataset and code used in our study at `https://anonymous.4open.science/r/modeling_spanish_acl-7567/`.

2 Related Work

Silfverberg and Hulden (2018a) introduced the encoder-decoder approach to PCFP by formulating the problem as a morphological reinflection task. The following paragraphs will provide an overview of the methodologies used to address this problem over time.

Initial non-neural models focused on learning string edit rules from data using sequence-to-sequence models (Albright and Hayes, 2003; Durrett and DeNero, 2013) and string transductions

(Nicolai et al., 2015). Subsequently, Ahlberg et al. (2015) proposed a finite state construction and used a classifier to select the correct inflection. Similarly, Alegria and Etxeberria (2016) used a single Weighted Finite-State Transducer (WFST) (Novak et al., 2012) to model the mapping between lemmas and inflected forms. Taji et al. (2016) developed a morphological analyzer that learns from the training data and applies the learned patterns to reinflect test data. Nicolai et al. (2016) implemented a discriminative transducer based on Jiampojamarn et al. (2008) that searches for a series of character transformation rules to perform the inflection. Liu and Mao (2016) and Cotterell et al. (2017) applied affixing rules to generate inflected forms, which involved appending or altering affixes (prefixes, suffixes, infixes, etc.) according to morphological rules. More recently, Sherbakov and Vylomova (2022) used a non-neural system to predict inflected forms based on string patterns observed in training samples. Kwak et al. (2023) introduced an improved affixing system that incorporated additional linguistic information to better capture the complexities in morphological generation. These methods, while interpretable, often struggled with irregular forms and low-resource scenarios.

The introduction of neural-based sequence-to-sequence models marked a significant milestone in modeling morphological reinflection by learning complex morphological patterns without explicit rules (Kann and Schütze, 2016; Malouf, 2016; Faruqui et al., 2016). Subsequent studies built upon this approach, including the hard monotonic attention for strict alignment between input and output sequences (Wu and Cotterell, 2019), multi-source setups with bi-directional LSTMs (Kann and Schütze, 2016), character-level LSTMs (Silfverberg and Hulden, 2018a), and encoder-decoder based transformers (Wu et al., 2021). Recent developments include phonologically-aware embeddings (Guriel et al., 2023) to capture both orthographic and phonetic information of words. Other approaches include treating morphological reinflection as a classification problem (Shcherbakov and Vylomova, 2023), and the application of imitation learning (Makarov and Clematide, 2018). Lastly, Large Language Models have also been explored for morphological reinflection, including analyzing ChatGPT's capability in morphological generation across multiple languages (Weissweiler et al., 2023).

Our study implements a model that closely follows the formulation of the encoder-decoder transformer for character-level transduction proposed by Wu et al. (2021), due to its high performance on inflectional tasks across various languages (Cotterell et al., 2017, 2018; Vylomova et al., 2020).

3 Methodology

3.1 Model Architecture

We implement the model using `fairseq` (Ott et al., 2019), a PyTorch-based sequence modeling toolkit. The model consists of four layers with four attention heads, an embedding size of 256, and a hidden layer size of 1,024. We use the Adam Optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.001, 0.1 label smoothing, and a 1.0 gradient clip threshold. The model is trained for a maximum of 10,000 optimizer updates, with checkpoints saved every ten epochs. Beam search is used at decoding time with a beam width of five.

Hyperparameter tuning Hyperparameter tuning, particularly the batch size, plays a crucial role in seq2seq tasks (Wu et al., 2021; Popel and Bojar, 2018). We used varying batch sizes, from 32 to 3600, and observed that the impact of batch size on the accuracy of predicting L-shaped verbs is not uniform across frequency conditions (see Figure 7 in Appendix). We adopt a batch size of 400 following established practices in morphological tasks (Vylomova et al., 2020; Pimentel et al., 2021; Kodner and Khalifa, 2022).

3.2 Dataset construction

We use the Spanish verbal morphology dataset from the Universal Morphology (UniMorph) project[1]. The entries in the dataset are coded in the Unimorph scheme (Sylak-Glassman, 2016). For example, the label `V;IND;PRS;1;SG` , corresponding to a first person singular present tense verb form, such as `digo`, is decomposed into a set of morphosyntactic features: [POS=VERB, mood=INDICATIVE, tense=PRESENT, person=1, number=SINGULAR]. The data representation in UniMorph follows the structure: (`lemma, form, feature`). We transcribe the lemma and inflected form in the International Phonetic Alphabet (IPA) to capture phonological representations, resulting in an entry such as (`desir, digo, V;IND;PRS;1;SG`).

For the reinflection task, we convert these entries to two source form-tag pairs: the target feature

[1] https://unimorph.github.io/

tag and the target inflected form. For example, if the first source form is digo (1st person singular, indicative), the second source form is diga (1st person singular, subjunctive) and the target form is digas (2nd person singular, subjunctive), then the above entry is converted to a so-called *triple* entry such as (digo, V;IND;PRS;1;SG, diga, V;SBJV;PRS;1;SG, V;SBJV;PRS;2;SG, digas). The dataset contains 5,460 distinct lemmas, of which 300 are L-shaped lemmas, and 4,860 are NL-shaped lemmas, which results in 382,956 triples.

To investigate the role of input frequency, we implement three experimental conditions, each characterized by a different ratio of L-shaped to regular (henceforth, *NL-shaped*) verbs in the training set under a) a naturalistic frequency distribution with 10% L-shaped verbs and 90% NL-shaped verbs (henceforth, *10%L-90%NL condition*) to reflect a realistic frequency distribution of the Spanish language[2], and two counterfactual conditions with an increase in the frequency of L-shaped verbs, and a decrease in the frequency of the NL-shaped verbs: b) 50% L-shaped verbs and 50% NL-shaped verbs (henceforth, *50%L-50%NL condition*) and c) 90% L-shaped verbs and 10% NL-shaped verbs (henceforth, *90%L-10%NL condition*). The relative frequency of these counterfactual conditions is created to allow a direct comparison of the learnability of L-shaped verbs relative to NL-shaped verbs.

Data representation The input sequence for our model is structured as follows:

d i g a # <V;SBJV;PRS;1;SG> # d i g a s # <V;SBJV;PRS;2;SG> # <V;IND;PRS;1;SG>

The expected target output is the space-separated characters forming the target word, d i g o. We refer to this input-output sequence as a *combination*.

Data sampling To isolate the effect of relative type frequency, we used an identical set of lemmas across all three conditions. As mentioned above, only 300 L-shaped lemmas are found in the UniMorph dataset, therefore, the maximum number of L-shaped lemmas in the 90%L-10%NL condition is capped to 300, representing 90%L. Therefore, the training set contains 333 lemmas, and we need to sample 33 NL-shaped verbs (amongst the 4,860 NL-shaped verbs) to represent 10% of NL. Similarly, we sampled for the other two conditions (50%L-50%NL and 10%L-90%NL).

We implement a rigorous data splitting strategy to mitigate the risk of artificially inflated model performance due to lemma overlap between training and testing data (Goldman et al., 2022; Kodner et al., 2023). At the *lemma level*, we ensure no lemma overlap between the training, development, and test sets. We apply a 70-10-20 split ratio to the lemmas, with 70% (training), 10% (development), and 20% (testing). At the *combination level*, given that each lemma produces approximately 600 combinations and there are 333 lemmas in each condition (resulting in 199,800 combinations), we partitioned the data into four bins to manage computational complexity and maintain cognitive plausibility. Each bin preserves the condition-specific distribution of L and NL-shaped lemmas (e.g., 10%L-90%NL). Finally, at the *run level*, we implement three randomized runs for each combination bin to account for potential order effects during training. Specifically, we generate three distinct training sets for each of the four combination bins created at the combination level.

The training data comprises full inflection tables, with which the model inflects unseen lemmas. For the development and test data, every two-slot combination of given slots is used as input to predict the target form corresponding to the target Morphosyntactic description (MSD) tag. Across all frequency conditions, our sampling procedure yields a training set of 39,435 combinations, a development set of 4,455 combinations, and a test set of 44,220 combinations. To enhance the robustness of our evaluation, we maintain a constant test set for each combination bin. In total, we generate 12 such datasets for each frequency condition. See Figure 1 for an illustration of our data sampling procedure.

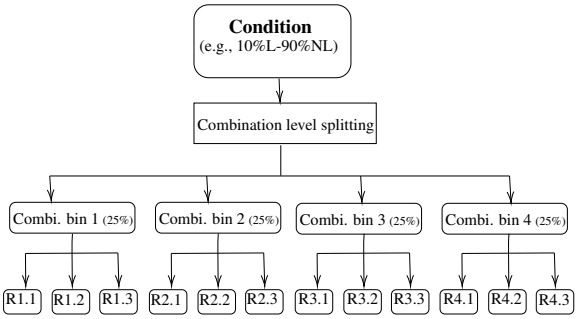

Figure 1: Flowchart showing the process for creating the dataset for each condition at the *combination level*, and *run level*.

[2]This is similar to the relative frequencies of L and NL-shaped verbs in the dataset which is 6% L and 94% NL.

4 Analysis and Results

We begin by presenting the sequence accuracies of the models across the three frequency conditions. Subsequently, we conduct a series of post-hoc analyses to understand the factors influencing learning of morphomic patterns under varying frequency conditions. In Section 4.1, we examine potential position biases in models. Section 4.2 investigates the models' sensitivity to memorization versus generalization effects (Hupkes et al., 2023). Lastly, in Section 4.3, we examine the models' behavior for specific phonological alternations.

We analyze the models' performance across varying input frequency distributions of regular and L-shaped verbs, addressing our research aim 2. We use two evaluation metrics: sequence accuracy and stem-only accuracy, as the errors can only occur in the suffixes and the stem of the predicted form (Kodner and Khalifa, 2022).

We first evaluate sequence accuracies (refer to Figure 2) across frequency conditions. In the 10%L-90%NL condition, NL-shaped verbs have a mean accuracy of 61.8%, which is 25.05% higher than the 36.75% mean accuracy of L-shaped verbs. Conversely, in the 90%L-10%NL condition, L-shaped verbs have a mean accuracy of 88.75%, outperforming NL-shaped verbs by 64.58%, which have a mean accuracy of 24.17%. In the 50%L-50%NL condition, L-shaped verbs have a mean accuracy of 72.31%, while NL-shaped verbs have a mean accuracy of 55.19%, with L-shaped verbs performing 17.12% better.

Subsequently, we analyze stem accuracies to isolate the models' performance on stem alternations across the frequency conditions (Appendix A.1). The results show a clear trend: the more frequent verb type in each condition consistently has better stem accuracy. In the 10%L-90%NL condition, NL-shaped verbs achieve a mean stem accuracy of 68.59%, which is 11.89% higher than the 56.7% mean stem accuracy of L-shaped verbs. In the 50%L-50%NL condition, L-shaped verbs have a mean stem accuracy of 77.24%, slightly better than NL-shaped verbs at 70.25%, with a difference of 6.99%. In the 90%L-10%NL condition, L-shaped verbs have a higher mean stem accuracy of 89.51%, outperforming NL-shaped verbs which have a mean stem accuracy of 31.09%, showing a difference of 58.2%.

These results show a clear performance difference between the models based on the distribution of verb types in the training data. In the 10%L-90%NL condition, which most closely approximates the natural distribution in Spanish, the models perform better on NL-shaped verbs than on L-shaped verbs. However, we find a learning advantage for L-shaped verbs in other conditions, suggesting the models might be learning specific characteristics of L-shaped verbs. To further understand the factors influencing the acquisition of morphomic patterns in the models', we conduct a series of post-hoc analyses.

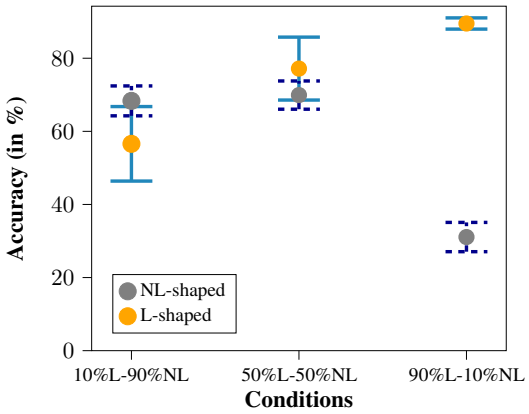

Figure 2: Mean and confidence intervals of overall sequence accuracies for L-shaped and NL-shaped verbs across frequency conditions (10%L-90%NL, 50%L-50%NL and 90%L-10%NL conditions).

4.1 Cell combinations

We examine the influence of paradigm cell combinations on the acquisition of L-shaped verbs, in line with our research aim 3a. Our analysis reveal robust primacy effects but recency effects are inconsistent.

Transformers are prone to *position bias*, disproportionately focusing on specific token positions due to architectural constraints (Dufter et al., 2022). We examine position bias in transformer through two psycholinguistically grounded metrics: primacy and recency effects. The primacy effect is a cognitive bias in which humans tend to remember and be influenced by the first pieces of information they are exposed to more than information presented later on (Asch, 1946). The recency effect refers to the tendency for humans to more easily recall items at the end of a list compared to items in the middle of the list (Marshall and Werder, 1972).

Each cell combination consists of three parts: the source 1 cell , the source 2 cell, and the target cell. We classify these cells based on their position relative to the L-shaped morphomic pattern: cells

within the L-shaped morphomic pattern are labeled as *In*, while those outside are labeled as *Out*. For example, a combination with source 1 as digo (In), source 2 as dices (Out), and target form as diga (In) is categorized as In-Out-In.

	10%L-90%NL			50%L-50%NL			90%L-10%NL		
Cell combi.	L (%)	NL (%)	L/NL	L (%)	NL (%)	L/NL	L (%)	NL (%)	L/NL
In-In-In	45.61	60.64	0.7	82.56	53.75	1.63	90.02	24.14	6.85
In-Out-Out	4.44	34.84	0.08	57.18	34.58	1.78	91.53	40.66	2.86
In-In-Out	27.17	62.57	0.4	63.94	54.19	1.15	87.66	22.85	5.07
In-Out-In	40.26	45.4	0.9	80.47	45.83	1.92	91.98	48.72	6.07
Out-In-In	40.62	64.77	0.57	76.17	57.19	1.41	89.54	16.93	15.23
Out-In-Out	37.68	66.68	0.54	70.49	59.58	1.18	87.43	16.53	6.58
Out-Out-In	30.39	59.23	0.45	59.8	52.11	1.08	87.06	22.42	7.44
Out-Out-Out	25.56	58.87	0.38	69.77	56.19	1.23	86.7	18.4	6.19

Table 2: Cell combination accuracies for 10L-90NL (left), 50L-50NL (middle), and 90L-10NL (right). The mean accuracies in percentage are calculated for separately by verb types (L denotes L-shaped verbs and NL denotes NL-shaped verbs) and by cell combinations (e.g., In-In-In). L/NL denotes the ratio of the mean accuracies of the L-shaped vs NL-shaped verbs. For a visualization of this table, see Appendix A.2.

To investigate potential primacy and recency effects, we compare pairs of cell combinations where the target cell aligns with either source 1 (primacy) or source 2 (recency) (Table 2). For example, **In**-Out-**In** allows for a primacy effect, while Out-**In**-**In** allows for a recency effect. We evaluate these effects by identifying minimally different pairs of cell combinations, such as comparing In-Out-In (primacy) to Out-Out-In to assess the presence of a primacy effect. The results show a consistent primacy effect for L-shaped verbs in the 10%L-90%NL and 50%L-50%NL conditions, although this effect is weaker in the 90%L-10%NL condition. For instance, in the 10%L-90%NL condition, the accuracy of the In-Out-In cell combination (40.26%) exceeds that of Out-Out-In (30.39%), indicating an apparent primacy effect. However, we do not detect a consistent recency effect across conditions.

We also find that 'In' targets are predicted more accurately than 'Out' targets for L-shaped verbs across all frequency conditions, suggesting a potential bias towards these 'In' cells.

4.2 Memorization and Generalization

We examine the models' ability to memorize and generalize morphomic patterns, focusing specifically on stem-final consonant triples under varying frequency conditions, addressing our research aim 3b. Our analysis shows that in most frequency conditions, the models show higher accuracy for memorized stem-final consonants compared to generalized ones, with the exception of the balanced 50%L-50%NL condition. We also find that memorization improves as the frequency of L-shaped verbs increases.

In order to generate accurate predictions, the transformer model must balance between memorization and generalization (Arpit et al., 2017; Zhang et al., 2021). This balance is particularly crucial when modeling L-shaped morphomes as it involves irregular stem alternations that do not follow straightforward phonological or semantic rules. Thereby, the model must rely on memorization to reproduce the specific alternations of seen verbs. At the same time, it must be able to extend to novel alternations. We assess how varying the distribution of L-shaped verbs in the training data affects the models' ability to memorize and generalize stem-final consonant triples. We focus on triples formed by the stem-final consonants of the first source form, second source form, and target form. For example, given the forms (`tɾadˈusen` (source 1), `tɾaduskˈamos` (source 2), `tɾadˈuskan` (target)), the stem-final consonant triple consists of `s`, `sk`, and `sk`. We quantify memorization as the models' ability in reproducing seen stem-final consonant triples, and generalization as their ability to correctly predict unseen triples. We treat memorization and generalization as two distinct knowledge states.

Using mixed-effects logistic regression, we examine how frequency conditions and knowledge states influence prediction accuracy. Logistic mixed-effects models are implemented using the `glmer` function from the `lme4` package (Bates et al., 2015) in `R`. Our model predicts accuracy (`prediction_status`: correct vs. incorrect) with two fixed effects - frequency conditions and knowledge state conditions - and two random intercepts: `triples` and `model` (among 12 models). We implement the following model structure:

```
glmer(prediction_status ~
knowledge_state * frequency_condition +
(1|triples) + (1|model),  data=df,
family="binomial")
```

To interpret the results of our fitted model, we use the `emmeans` package (Lenth, 2024) to calculate estimated marginal means (EMMs). This way, we can estimate the predicted probabilities of correct predictions for different combinations of knowledge states and frequency conditions. In the condition where the frequency distribution of verb types is equal (50%L-50%NL condition), we observe

a slight advantage for generalization over memorization, with predicted probabilities for generalized stem-final consonant pairs being 0.113 higher than for memorized stem-final consonant pairs. However, in the 10%L-90%NL condition, memorized stem-final consonant pairs show a 0.036 higher predicted probability. In contrast, in the 90%L-10%NL condition, this advantage increases to 0.304 (see Appendix A.3 for detailed results).

In terms of memorization, we find a positive correlation between the frequency of L-shaped verbs in the training data and the probability of correct predictions. The highest probability of correct predictions occurs in the 90%L-10%NL condition (0.754), followed by 50%L-50%NL condition (0.627) and least in the 10%L-90%NL condition (0.252). This suggests that increased exposure to L-shaped verbs enhances the model's ability to memorize stem-final consonant triples.

For generalization, the highest probability of correct predictions occurs in the balanced 50%L-50%NL condition (0.74). Interestingly, the probabilities are lower in the skewed conditions, with the 10%L-90%NL condition having a probability of 0.216 and the 90%L-10%NL condition having a probability of 0.450.

4.3 Consonant pair analysis

We investigate the models' performance on specific stem-final consonant pairs of L-shaped verbs, focusing on the alternating pairs comprising the stem-final consonant of the forms in the 'Out' cells and that of the 'In' cells within the paradigm (as discussed in section 4.1), addressing our aim 3c. Our analysis shows that the models are sensitive to the input frequency of consonant alternations, indicating that they have not fully acquired the abstract morphological patterns.

For example, for the lemma desir, the consonant pair is [s]-[g], where [s] is the stem-final consonant of the out cells and [g] is found in forms sharing the L-shaped pattern. The most frequent pairs in the dataset are [s]-[sk], with 141 occurrences, followed by [n]-[ng] and [ç]-[x], with 53 and 25 occurrences, respectively. This skewed distribution naturally results in varying ratios for each experimental run due to our data sampling process (as shown in Section 3).

We examine the models' sensitivity to consonant pair frequencies across the frequency conditions and assess how varying proportions of L-shaped in the input affect the learning of consonant alternations. Across all three runs of the 10%L-90%NL condition, [s]-[sk] appears frequently in both test (3-4 times) and training sets (8-13 times). Some pairs like [lç]-[lx] and [s]-[g] appear in test sets but are rare or absent in training, which might pose difficulty for models in applying patterns to novel combinations. In the 50%L-50%NL condition, [s]-[sk] remains the most frequent pair, appearing 14-22 times in test sets and 15-23 times in training sets. Other pairs like [n]-[ng] and [s]-[g] also appear but less frequently. In the 90%L-10%NL condition, [s]-[sk] appears even more frequently (22-32 times in test sets and 97-104 times in training), while other pairs remain less common. Details are given in Appendix A.4. A confusion matrix in Appendix A.5 summarizes the top 5 most erroneous consonant pairs for L-shaped verbs.

The main frequency effect can still be found consistently across these consonant pairs. In the 10%L-90%NL condition, [s]-[sk] achieves 68.6% accuracy, while [s]-[g] and [n]-[ng] reach 26.2% and 77.4%, respectively. Accuracy increases in the 50%L-50%NL condition to 89.8%, 60.6%, and 83.6%, respectively. In the 90%L-10%NL condition, accuracies further improve to 93.3%, 92.6%, and 91.4%. Looking beyond accuracies, we find that the models still make systematic errors, often defaulting to more frequent lemma consonants rather than altered ones (e.g., predicting [s]-[s] instead of [s]-[sk]).

The models perform worse for less frequent or unseen alternations (such as [s]-[g]) compared to more frequent alternations ([s]-[sk]). This indicates that the models have not fully acquired the abstract morphological patterns. In these cases, models tend to regularize L-shaped verbs in datasets, as erroneous predictions often result in non-alternating pairs. These results suggests that the models are relying heavily on frequency-based pattern matching rather than acquiring true morphological competence.

5 Conclusion

In this paper, we examine the learning capabilities of transformer models with respect to morphomic pattern, specifically the L-shaped pattern in Spanish. We conduct a series of post-hoc analyses to understand the factors influencing the learning of morphomic patterns under varying frequency conditions.

Transformer models have shown remarkable

ability to learn irregular patterns in related language tasks, such as English past tense inflection, German noun plurals and Arabic noun plurals (Kodner and Khalifa, 2022; Kakolu Ramarao et al., 2022). However, morphomic patterns are complex linguistic patterns that are hard to acquire (Nevins et al., 2015). The models' ability to capture morphomic patterns was not guaranteed given these patterns' unique characteristics: they operate independently of semantics, syntax, and phonology. In our study, the models' performance on L-shaped verbs, especially in conditions with uneven frequency, indicates that transformers have developed some level of competence in recognizing and applying morphomic patterns from the training data.

We first look at the learning strategies of the transformer with respect to input ordering. We observe a clear primacy effect in models' processing of L-shaped verbs, suggesting that the models are more influenced by the first source cell in making predictions. We also observe that models have higher accuracy in predicting target forms that are part of L-shaped pattern compared to those outside the L-shaped pattern within a morphological paradigm. This bias suggests that the models are capturing some aspects of the overall paradigm structure, particularly the distribution of irregular forms within the L-shaped pattern.

In our investigation of models' strategies for balancing memorization and generalization, we show that the type frequency of L-shaped verbs impacts the models' ability to both memorize and generalize stem-final consonant alternations. For memorization, we observe a positive correlation between the frequency of L-shaped verbs in the training data and the model's ability to correctly reproduce seen stem-final consonants. This suggests that increased exposure to L-shaped verbs enhances the model's ability to retain and apply stem-final consonant alternations. This finding aligns with studies which show morphological patterns with higher type frequency to be more productive (Bybee, 1995; Pierrehumbert, 2001; Baer-Henney and van de Vijver, 2012, *inter alia*). However, when it comes to generalization, we observe a non-linear relationship between the frequency of L-shaped verbs in the training data and the model's ability to produce unseen stem-final consonants. We find that generalization performance peaks in the balanced condition (50%L-50%NL), but decreases in skewed conditions (90%L-10%NL and 10%L-90%NL conditions). While increased exposure to L-shaped verbs enhances memorization, it may not necessarily lead to better generalization.

While the models demonstrate some ability to learn and apply the L-shaped pattern, they exhibit a clear preference for regular patterns when encountering unfamiliar consonant alternations. This suggests that the models simply rely on frequency-based pattern and have not fully acquired abstract morphological rules.

Furthermore, the models' superior performance on regular verbs in the 10%L-90%NL condition (a close approximation to natural language distribution) validates the results from Nevins et al. (2015)'s study. In their study with Spanish speakers on a wug-test-like inflection task, 71.9% of the participants showed a preference to NL-shaped responses over L-shaped responses. This similarity in behavior between transformers and human participants motivates a deeper comparison. In the future, we aim to compare our model's performance with the experimental results from human speakers, such as those from Nevins et al. (2015) and Cappellaro et al. (2024). Through this comparison, we could assess how effectively our computational approach captures the linguistic and cognitive phenomena observed in human morphological processing in the context of morphomic patterns.

Limitations

We acknowledge several limitations in our current study that could be addressed in future research. First, we did not explore alternative computational approaches. For example, Recurrent Neural Networks (RNNs), as used in studies like modeling German plurals (Dankers et al., 2021), or the other approaches outlined in Section 2.

Second, we did not perform probing analyses to investigate the internal representations of the model but instead relied solely on post-hoc analyses.

Third, to rigorously evaluate the model's generalization capabilities, it would be beneficial to use test data that is entirely unattested in the training set. This means that both the lemmas (word stems) and the feature tags (e.g., tense, number, person) should be novel to the model. This approach, similar to Kodner et al. (2023), would provide a more robust assessment of the model's ability to generalize morphomic patterns to unseen data.

Finally, we did not account for the morphological complexity of the verbs. Some verbs have prefixes, and some do not, therefore, some lemmas share the same stems. On the one hand, this ren-

ders the lemma-level train-development-test splitting procedure less effective and might artificially inflate the accuracies of our models. On the other hand, this is arguably ecologically more valid as human learners do get exposed to both morphologically complex and simple verbs.

Ethics Statement

All the models we use are small, which significantly reduces the computational resources required for training and inference. All data used in the study are from open datasets. To promote transparency and reproducibility, all data and code used in this study are publicly available.

The involved university does not require IRB approval for this kind of study, which uses publicly available data without involving human participants. We do not see any other concrete risks concerning dual use of our research results. Of course, in the long run, any research results on AI methods based on large language models could potentially be used in contexts of harmful and unsafe applications of AI. But this danger is rather low in our concrete case.

References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1029, Denver, Colorado. Association for Computational Linguistics.

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

Iñaki Alegria and Izaskun Etxeberria. 2016. EHU at the SIGMORPHON 2016 shared task. a simple proposal: Grapheme-to-phoneme for inflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 27–30, Berlin, Germany. Association for Computational Linguistics.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Solomon E Asch. 1946. Forming impressions of personality. *The journal of abnormal and social psychology*, 41(3):258.

Dinah Baer-Henney and Ruben van de Vijver. 2012. On the role of substance, locality, and amount of exposure in the acquisition of morphophonemic alternations. *Laboratory Phonology*, 3(2):221–249.

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.

James P. Blevins. 2006. Word-based morphology. *Journal of Linguistics*, 42(3):531–573.

James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

James P. Blevins, Petar Milin, and Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In *Perspectives on Morphological Organization*, pages 139–158. BRILL.

Gilles Boyé and Gauvain Schalchli. 2019. Realistic data and paradigms: the paradigm cell finding problem. *Morphology*, 29(2):199–248.

Joan Bybee. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5):425–455.

Joan Bybee. 2003. *Phonology and language use*, volume 94. Cambridge University Press.

Chiara Cappellaro, Nina Dumrukcic, Isabella Fritz, Francesca Franzon, and Martin Maiden. 2024. The cognitive reality of morphomes. evidence from italian. *Morphology*, 34(1):33–71.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108, Online. Association for Computational Linguistics.

Fermın Moscoso del Prado Martın, Aleksandar Kostić, and R Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94(1):1–18.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2022. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 864–870. Association for Computational Linguistics.

David Guriel, Omer Goldman, and Reut Tsarfaty. 2023. Morphological inflection with phonological features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 613–622, Toronto, Canada. Association for Computational Linguistics.

Borja Herce and Marc Allassonnière-Tang. 2024. The meaning of morphomes: distributional semantics of spanish stem alternations. *Linguistics Vanguard*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08: HLT*, pages 905–913, Columbus, Ohio. Association for Computational Linguistics.

Akhilesh Kakolu Ramarao, Yulia Zinova, Kevin Tang, and Ruben van de Vijver. 2022. HeiMorph at SIGMORPHON 2022 shared task on morphological acquisition trajectories. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 236–239, Seattle, Washington. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 514–524. Association for Computational Linguistics.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

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.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Alice Kwak, Michael Hammond, and Cheyenne Wing. 2023. Morphological reinflection with weighted finite-state transducers. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 132–137, Toronto, Canada. Association for Computational Linguistics.

Russell V. Lenth. 2024. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.0.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, SIGMORPHON 2020, Online, July 10, 2020*, pages 153–161. Association for Computational Linguistics.

Ling Liu and Lingshuang Jack Mao. 2016. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 36–40, Berlin, Germany. Association for Computational Linguistics.

Martin Maiden. 2011. Allomorphy, autonomous morphology and phonological conditioning in the history of the Daco-Romance present and subjunctive. *Transactions of the Philological Society*, 109(1):59–91.

Martin Maiden. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford University Press.

Martin Maiden. 2021. The morphome. *Annual Review of Linguistics*, 7:89–108.

Peter Makarov and Simon Clematide. 2018. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882, Brussels, Belgium. Association for Computational Linguistics.

Robert Malouf. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistics Papers*, 6:122–129.

Philip H Marshall and Pamela R Werder. 1972. The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior*, 11(5):649–653.

Andrew Nevins, Cilene Rodrigues, and Kevin Tang. 2015. The rise and fall of the L-shaped morphome: diachronic and experimental studies. *Probus: International Journal of Latin and Romance Linguistics*, 27(1):101–155.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, Denver, Colorado. Association for Computational Linguistics.

Garrett Nicolai, Bradley Hauer, Adam St Arnaud, and Grzegorz Kondrak. 2016. Morphological reinflection via discriminative string transduction. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 31–35, Berlin, Germany. Association for Computational Linguistics.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, Donostia–San Sebastián. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Janet Pierrehumbert. 2001. Stochastic phonology. *Glot international*, 5(6):195–207.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*.

Andreas Shcherbakov and Ekaterina Vylomova. 2023. Does topological ordering of morphological segments reduce morphological modeling complexity? a preliminary study on 13 languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 120–125, Dubrovnik, Croatia. Association for Computational Linguistics.

Andreas Sherbakov and Ekaterina Vylomova. 2022. Morphology is not just a naive bayes–unimelb submission to sigmorphon 2022 st on morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 240–246.

Miikka Silfverberg and Mans Hulden. 2018a. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Miikka Silfverberg and Mans Hulden. 2018b. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2883–2889. Association for Computational Linguistics.

Andrew Spencer and Mark Aronoff. 1994. Morphology by itself: Stems and inflectional classes. *Language*, 70:811.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, Center for Language and Speech Processing, Johns Hopkins.

Dima Taji, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. The Columbia University - New York University Abu Dhabi SIGMORPHON 2016 morphological reinflection shared task submission. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 71–75, Berlin, Germany. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Adam Wiemerslage, Shiran Dudy, and Katharina Kann. 2022. A comprehensive comparison of neural networks as cognitive models of inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1933–1945, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

A Appendices

A.1 Accuracies

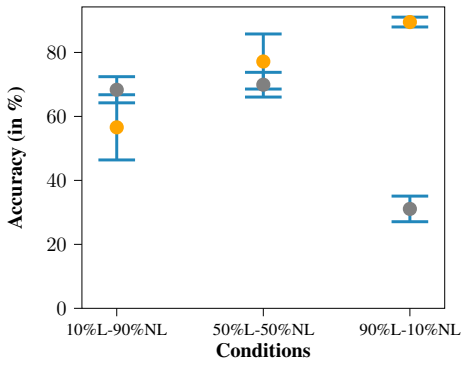

Figure 3: Mean and confidence intervals of stem sequence accuracies for L-shaped and NL-shaped verbs across frequency conditions. Gray: NL-shaped, Orange: L-shaped.

A.2 Cell combinations

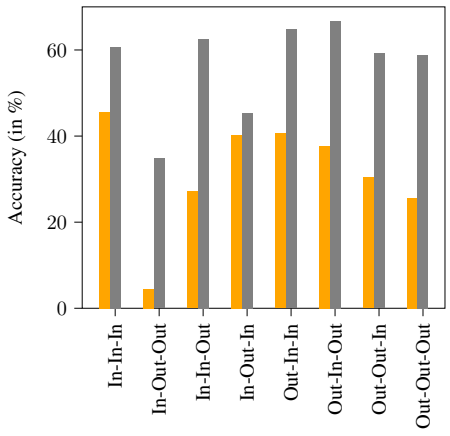

Figure 4: Cell combination accuracies for 10L-90NL condition. Gray: NL-shaped, Orange: L-shaped.

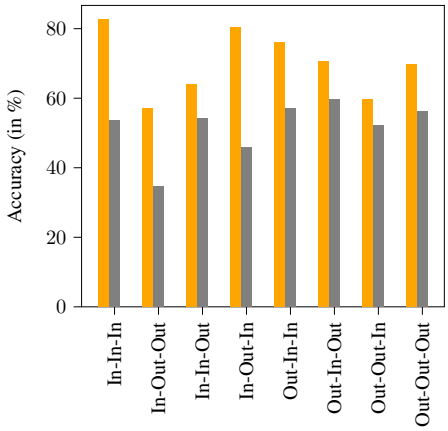

Figure 5: Cell combination accuracies for 50L-50NL condition. Gray: NL-shaped, Orange: L-shaped.

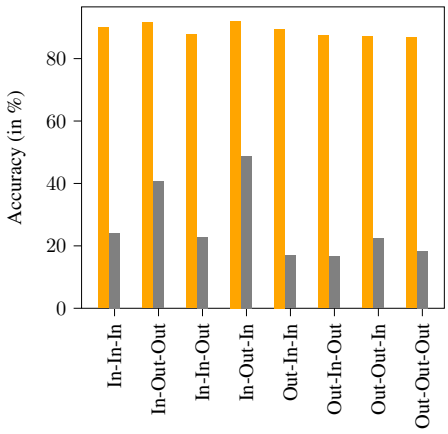

Figure 6: Cell combination accuracies for 90L-10NL condition. Gray: NL-shaped, Orange: L-shaped.

A.3 Memorization and Generalization task

Knowledge state	Condition	Prob.	SE	asymp.LCL	asymp.UCL
Generalization	10%L-90%NL	0.216	0.0150	0.188	0.247
Memorization	10%L-90%NL	0.252	0.0183	0.218	0.290
Generalization	50%L-50%NL	0.740	0.0260	0.686	0.788
Memorization	50%L-50%NL	0.627	0.0351	0.556	0.693
Generalization	90%L-10%NL	0.450	0.0353	0.382	0.520
Memorization	90%L-10%NL	0.754	0.0301	0.690	0.808

Table 3: Estimated marginal means for knowledge state and frequency condition combinations

Memorization					
Contrast	odds.ratio	SE	null	z.ratio	p.value
10%L-90%NL / 50%L-50%NL	0.2009	0.02336	Inf	-13.806	<.0001
10%L-90%NL / 90%L-10%NL	0.1100	0.01394	Inf	-17.420	<.0001
50%L-50%NL / 90%L-10%NL	0.5476	0.08795	Inf	-3.749	0.0005
Generalization					
Contrast	odds.ratio	SE	null	p.value	
10%L-90%NL / 50%L-50%NL	0.0969	0.00949	Inf	-23.833	<.0001
10%L-90%NL / 90%L-10%NL	0.3369	0.03802	Inf	-9.642	<.0001
50%L-50%NL / 90%L-10%NL	3.4750	0.47015	Inf	9.207	<.0001

Table 4: Pairwise comparisons of frequency condition levels for each knowledge state

A.4 Consonant-pair analysis

Stem-final consonants	No. of occurrences
[s]-[sk]	141
[n]-[ng]	53
[ç]-[x]	25
[s]-[g]	15
[]-[jg]	14
[rç]-[rx]	10
[nç]-[nx]	10
[l]-[lg]	4
[lç]-[lx]	4
[s]-[sg]	2
[b]-[p]	1

Table 5: Number of alternating stem final-consonant pairs of all the L-shaped verbs.

	Consonant-pairs	Freq. in Test	Freq. in Train
Run 1	[s]-[sk]	3	8
	[n]-[ng]	2	4
	[nç]-[nx]	1	1
	[lç]-[lx]	1	0
Run 2	[s]-[sk]	3	13
	[n]-[ng]	1	2
	[nç]-[nx]	1	1
	[s]-[g]	1	0
Run 3	[s]-[sk]	4	10
	[s]-[g]	2	1

Table 6: Frequency of stem-final consonants in train and test in 10%L-90%NL condition.

	Consonant-pairs	Freq. in Test	Freq. in Train
Run 1	[s]-[sk]	19	20
	[n]-[ng]	3	10
	[nç]-[nx]	2	0
	[s]-[g]	3	2
	[l]-[lg]	1	0
	[ ]-[jg]	3	0
	[lç]-[lx]	1	1
Run 2	[s]-[sk]	22	15
	[ç]-[x]	2	5
	[n]-[ng]	3	7
	[nç]-[nx]	1	1
	[lç]-[lx]	1	1
	[s]-[g]	2	3
	[s]-[sg]	1	0
Run 3	[s]-[sk]	14	23
	[s]-[g]	4	0
	[n]-[ng]	2	12
	[ ]-[jg]	3	4
	[l]-[lg]	1	1
	[rç]-[rx]	1	2
	[nç]-[nx]	2	1
	[ç]-[x]	2	4

Table 7: Frequency of stem-final consonants in train and test in 50%L-50%NL condition.

	Consonant-pairs	Freq. in Test	Freq. in Train
Run 1	[s]-[sk]	32	97
	[ç]-[x]	5	18
	[ ]-[jg]	5	5
	[nç]-[nx]	2	7
	[rç]-[rx]	3	6
	[n]-[ng]	6	40
	[lç]-[lx]	1	2
	[s]-[g]	2	11
Run 2	[s]-[sk]	22	104
	[ç]-[x]	2	5
	[rç]-[rx]	1	7
	[s]-[g]	4	10
	[l]-[lg]	2	2
	[ ]-[jg]	4	8
	[nç]-[nx]	3	7
	[s]-[sg]	1	1
Run 3	[s]-[sk]	25	103
	[nç]-[nx]	4	6
	[s]-[g]	2	11
	[ç]-[x]	7	18
	[lç]-[lx]	2	1
	[n]-[ng]	10	35
	[l]-[lg]	2	2
	[rç]-[rx]	1	6
	[ ]-[jg]	1	9

Table 8: Frequency of stem-final consonants in train and test in 90%L-10%NL condition.

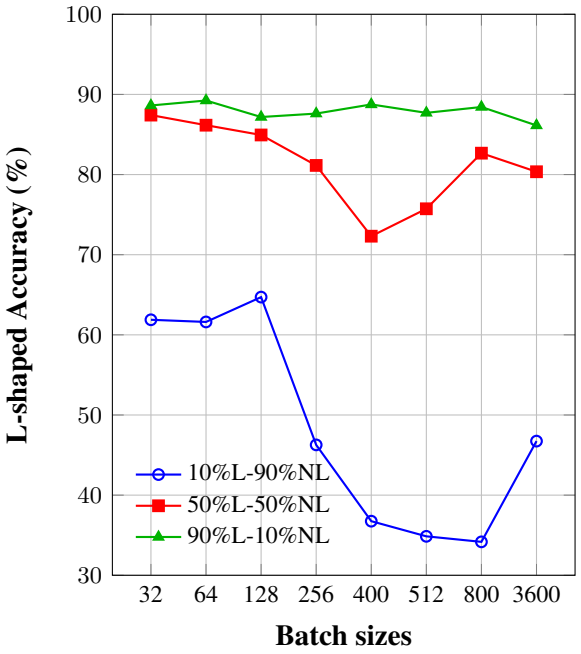

Figure 7: Effect of batch sizes

A.5 Confusion Matrix

10%L-90%NL condition										
	P [s]-[s]	P [s]-[sk]	P [s]-[g]	P [n]-[ng]	P [s]-[]	P [s]-[d]	P [s]-[sg]	P [n]-[n]	P [n]-[rb]	Accuracy
G [s]-[sk]	303	**857**	0	0	48	41	0	0	0	68.61%
G [s]-[g]	235	21	**99**	0	8	0	15	0	0	26.19%
G [n]-[ng]	0	0	0	**281**	0	0	0	56	26	77.41%
50%L-50%NL condition										
	P [s]-[sk]	P [s]-[g]	P [n]-[ng]	P [s]-[s]	P [s]-[d]	P [s]-[ʃ]	P [s]-[jg]	P [n]-[n]	P [n]-[mp]	Accuracy
G [s]-[sk]	**6170**	32	0	523	73	70	0	0	0	89.84%
G [s]-[g]	112	**613**	0	210	0	0	76	0	0	60.63%
G [n]-[ng]	0	0	**853**	0	0	0	0	139	29	83.55%
90%L-10%NL condition										
	P [s]-[sk]	P [n]-[ng]	P [ç]-[x]	P [s]-[s]	P [s]-[rs]	P [n]-[n]	P [n]-[sk]	P [n]-[n]	P [ç]-[ç]	Accuracy
G [s]-[sk]	**9473**	0	0	570	36	70	0	0	0	93.34%
G [n]-[ng]	0	**3394**	0	0	0	130	0	139	0	92.66%
G [ç]-[x]	0	0	**1954**	0	0	0	0	0	183	91.44%

Table 9: Confusion matrix for the top 5 most erroneous consonant-pairs (considering mean values) for L-shaped verbs across conditions. G = Gold, P = Prediction. Bold indicates the correct predictions.