Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network

Verna Dankers[*], **Anna Langedijk**[*], **Kate McCurdy**, **Adina Williams**, **Dieuwke Hupkes**
ILCC, University of Edinburgh ILLC, University of Amsterdam Facebook AI Research
{vernadankers, annalangedijk}@gmail.com, kate.mccurdy@ed.ac.uk,
{adinawilliams, dieuwkehupkes}@fb.com

Abstract

Inflectional morphology has since long been a useful testing ground for broader questions about generalisation in language and the viability of neural network models as cognitive models of language. Here, in line with that tradition, we explore how recurrent neural networks acquire the complex German plural system and reflect upon how their strategy compares to human generalisation and rule-based models of this system. We perform analyses including behavioural experiments, diagnostic classification, representation analysis and causal interventions, suggesting that the models rely on features that are also key predictors in rule-based models of German plurals. However, the models also display shortcut learning, which is crucial to overcome in search of more cognitively plausible generalisation behaviour.

1 Introduction

Language is a complex and mysterious system, which requires that speakers systematically generalise but also that they admit exceptions (Jackendoff and Audring, 2018; Bybee and Hopper, 2001; Pinker, 1998). A clear illustration of this is the domain of morphology, where suffixes and affixes can be productively used to express a particular grammatical property, but where there are also several words that follow irregular patterns for the same grammatical function. For example, while the past tense for most English verbs is formed by affixing *-ed* (*walk* → *walked*), the past tense of *break* is not *breaked* but *broke*. If an English speaker encounters an unknown form such as *treak*, they must decide whether to attach *-ed*, or instead go with the irregular form *troke*. Precisely because of such intricacies, the computational task of acquiring a morphological system capable of generalisation has a strong historical connection to cognitive science (e.g. Seidenberg and Plaut, 2014).

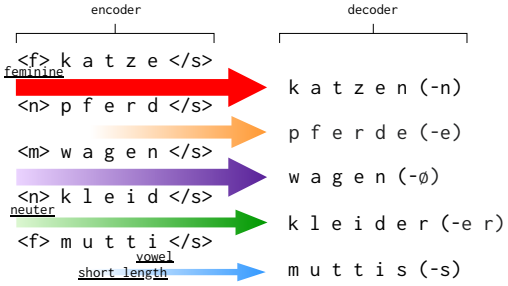


Figure 1: Illustration of how our models predict plural nouns. Line thickness indicates performance per plural class; colour gradients show how that performance increases while the encoder processes the word.

Making progress on how human minds process this interesting subdomain of language, however, is a challenging enterprise: probing the internal representations of human minds is difficult and, in some cases, potentially unethical. Several researchers have therefore opted to instead investigate neural models that show generalisation behaviour that is similar to humans in key aspects and use them to learn more about human language processing (for a prime example of this cycle, see Lakretz et al., 2019, 2021; Baroni, 2021). Using neural models in such a fashion starts with an accurate understanding of how they approach the phenomena of interest: without that, we are constrained in our attempts to use them to devise hypotheses about human language processing and compare them to existing theories (Hupkes, 2020; Baroni, 2021).

In that vein, we present a detailed examination of how *recurrent neural networks* (RNNs) process the complex German plural system which – contrary to the classical example of English past tense – features generalisation of multiple classes (Mar-

[*]Equal contribution.

cus et al., 1995; McCurdy et al., 2020; Belth et al., 2021). We ask what kind of representations such models learn, whether they support human-like generalisation, and we make a start with comparing their learnt solutions with existing models of German plural inflection. We train RNNs to predict the form of a German plural noun (or its *plural class*) from its singular form and grammatical gender and present an elaborate investigation of the resulting models. We perform *behavioural* analyses (§3) of the models' predictions as well as *structural* analyses of their internal representations, aimed at identifying the features that the models have learnt to associate with each of the classes (§4). For the latter, we use diagnostic classifiers (Hupkes et al., 2018) that we afterwards use to intervene in the model to establish causal connections between the internal encodings and the predictions the model generates from these encodings (§5).

We find that our networks show a mixture of cognitively plausible generalisation and reliance upon 'shortcuts' or heuristics (McCoy et al., 2019; Geirhos et al., 2020) (see Figure 1). On the negative side, the model's ability to cope with nouns in low-frequency plural classes is very brittle. Our behavioural analyses reveal that the models overrely on length as a heuristic to predict the rare class *-s*. However, we also find that our models correctly learn that key predictors of plural class include grammatical gender and the last few letters of a word, which are the same features considered by the recent decision-tree-based cognitive model of Belth et al. (2021). Our diagnostic classifiers additionally show how these predictors are encoded in the model's recurrent hidden representations. This interesting overlap between neural and rule-based models raises questions as to what neural models might teach us about the cognitive implementation of rule-based domains.[1]

2 Methods

German plural inflection comprises multiple plural classes with different frequencies (Clahsen et al., 1992; Marcus et al., 1995; Clahsen, 1999; McCurdy et al., 2020; Zaretsky and Lange, 2016). Most plural nouns are inflected with one of five suffixes (Clahsen et al., 1992): *-(e)n*, *-e*, *-∅*, *-er* and *-s* (Table 1 shows some examples). The suffixes *-e*, *-er*, and *-∅* sometimes combine with umlaut.[2] The literature on German plural generalisation has measured success rates based on correct suffixation (as opposed to a combination of the suffix and umlaut), and for simplicity, we keep this focus in the current study (McCurdy et al., 2020; Belth et al., 2021).

Class	Example		Frequency	Length
	Singular	Plural		
-(e)n	Frau	Frauen	44.7%	11.7
-e	Hund	Hunde	26.3%	11.0
-∅	Wagen	Wagen	16.9%	11.3
-er	Wald	Wälder	3.5%	10.4
-s	Auto	Autos	5.4%	8.0
-?	Cello	Celli	3.2%	10.6

Table 1: The German plural system as represented in the Wiktionary dataset with examples, along with inflection class frequency and average (singular form) word length.

2.1 Dataset

In our experiments, we use the Wiktionary dataset,[3] which contains orthographic representations of pairs of singular and plural forms of German nouns in the nominative case.[4] Given a control token indicating grammatical gender and a singular form (a sequence of discrete characters followed by a stop token, e.g. `<f> f r a u </s>`), the model is trained to predict the plural form (`f r a u e n`). Nouns that did not have a gender listed were excluded. The training, validation, and test splits consist of 46k, 6.5k, and 6.6k instances, respectively. Masculine, feminine and neuter nouns appear 23k, 25k and 11k times, respectively. Table 1 indicates the frequency of each of the plural classes, with the average length of the inputs per class. Notice that the nouns that take the *-s* class are, on average, 8 characters long, while the overall average length is 11 characters.

To label plural forms predicted by the model, we consider whether the plural form has one of the five acceptable suffixes (*-(e)n*, *-e*, *-∅*, *-er* or *-s*) and whether the singular form appears in the plural form (after removing umlauts). Predictions that belong to one of these five classes are considered *well-formed*. Otherwise, the input belongs to the unknown inflection class *-?*.

[1] The data and implementation are available here.

[2] How exactly this works is a theoretically open domain (Alexiadou and Müller, 2008; Müller, 2015; Trommer, 2020).

[3] The dataset is available here.

[4] In German, the mapping from orthographic representations to phonological representations is comparatively straightforward (Neef, 2004, 2011); which is why we use this orthographic form for our investigations.

2.2 Model

We study a recurrent encoder-decoder model implemented with the OpenNMT library (Klein et al., 2017). Modern RNNs trained to perform sequence-to-sequence tasks, including morphological inflection, typically have a bidirectional encoder and an attention mechanism (Corkery et al., 2019; McCurdy et al., 2020). While effective in terms of task accuracy, that setup is further away from how humans process the task at hand (incrementally and in one go) than unidirectional models (see also Hupkes, 2020; Christiansen and Chater, 2016; Baroni, 2021). Furthermore, in models with attention, there is no bottleneck between the encoder and decoder that forces the model to create one localised representation of the input and its potential plural class. We, therefore, study unidirectional models without attention. The encoder and decoder consist of two-layer unidirectional LSTMs, a hidden dimensionality of 128, character embeddings of size 128 and a dropout of 0.1 between layers. The Adadelta (Zeiler, 2012) optimiser is used, with a batch size of 64. During evaluation, we apply beam search with a beam of size five. We train five models with different random initialisations. All results presented are averaged over those models.

Rule-based comparison Belth et al. (2021) propose a cognitive model of morphological learning which uses recursive application of the frequency threshold defined by the Tolerance Principle (Yang, 2016) to identify productive rules, resulting in a decision tree. The model checks at each node whether to keep traversing the tree, apply a learnt rule, or match the input form to a stored exception.For German plural inflection, their model relies upon grammatical gender and the last few characters of the input word as features in the decision tree. We train their model on our dataset for comparison.[5]

3 Behavioural results

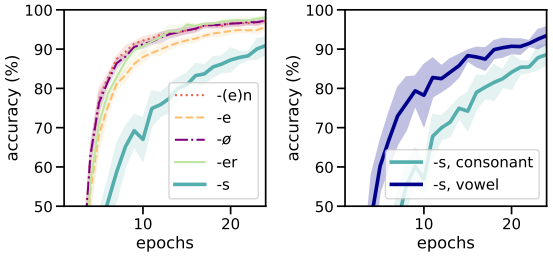
In Table 2, we summarise the models' performance after 25 epochs. With 92.3%, our suffix accuracies are competitive, despite the unidirectionality and removal of attention. The RNNs outperform the rule-based model of Belth et al. (2021) on unseen data.[6] Figure 2a shows the RNNs' training curve per plural class, with all classes undergoing rapid

[5]We make examples of these decision trees available here. A part of the model is visualised in Appendix E.

[6]The Belth et al. model does not handle stem changes such as umlaut, which negatively impacts its full noun accuracy.

Model	Measure	Train	Validation	Test
Bidirec. & att.	noun	97.4±.3	92.9±.2	93.0±.2
	noun[-1]	97.8±.3	93.9±.2	94.0±.1
Unidirectional	noun	95.8±.5	87.8±.6	87.9±.5
	noun[-1]	97.4±.2	92.2±.2	92.3±.3
Belth et al.	noun	99.9±0	78.8±0	78.2±0
	noun[-1]	99.9±0	89.2±0	89.0±0

Table 2: Accuracy (*noun*) and final letter accuracy (*noun[-1]*), with standard deviations over seeds.



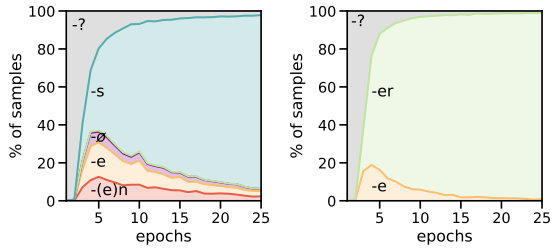
(a) All classes (b) *-s*

Figure 2: Training accuracy across epochs for the five plural classes, and for the *-s* nouns with stems ending in vowels and consonants, separately.

increases in performance during the first 5 epochs and less rapid but still substantial increases between epochs 5 and 10. Particularly notable is the curve for samples from the *-s* class, which is learnt more slowly than the other classes. Figure 2b details the training curve for the *-s* class by separating inputs ending in a consonant, and those ending in a vowel, which suggests that mostly the inputs from the former class are learnt later during training.

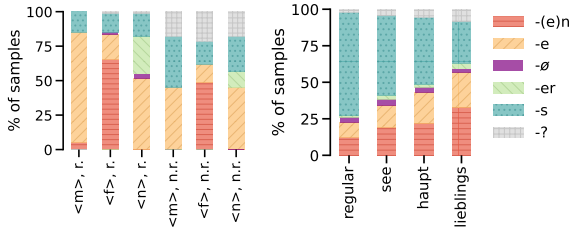
Overgeneralisation Throughout training, samples can be assigned suffixes different from their target suffix. This rarely happens for majority classes (*-(e)n*, *-e*, *-ø*) (with the exception of *-e*, which, during one epoch, is predicted as *-(e)n* in 5% of the cases), but is more frequent for the rarer, minority classes *-er* and *-s*. The models tend to generalise the suffixes of majority classes to the minority classes, a phenomenon traditionally referred to as *overgeneralisation* (Feldman, 2005). Maximum overgeneralisation typically occurs early on during training, as shown in Figure 3 (c.f. Korrel et al., 2019; Hupkes et al., 2020; Dankers et al., 2021).

Wug testing Next, we apply the trained models to 24 nonce word stimuli from Marcus et al. (1995). Of these stimuli, 12 are *rhymes* – i.e. phonologically familiar words rhyming with an existing word – and 12 *non-rhymes* – i.e. phonotactically atypical words. We feed them to the network following



(a) *-s* (b) *-er*

Figure 3: Cumulative distribution of the predicted suffix for training samples with *-s* and *-er* as plural class.



(a) Nonce predictions (b) *-s* compound predictions

Figure 4: Distribution of plural classes, (a) for nonce words with three gender tags ('r.' marks rhymes, and 'n.r.' non-rhymes), or (b) for Wiktionary validation data for *-s*, in the 'regular' format and as compounds with the words indicated prepended to the singular noun.

each of the gender tokens; the distribution over plural classes for the predictions of converged models is shown in Figure 4a. Similar to previous work (McCurdy et al., 2020, which only considered the neuter tag) non-rhymes fall more often into the unknown class (/?). Different from McCurdy et al. (2020) though, *-s* predictions are more frequent than *-er*, and are also more frequent for non-rhymes than for rhymes.[7] Figure 4a also illustrates that gender impacts the models' predictions: the feminine tag seems related to *-(e)n* predictions, and the neuter tag to *-er* predictions.

Enforcing gender In wug testing, changing the gender changes the model's predictions. Is this the case for Wiktionary data as well? To find out, we compare the model's predictions for samples from the validation set with its predictions for the same samples force-fed with new gender control tokens. Figure 5 visualises the corresponding results. Introducing the feminine control token has the most prominent effect: the vast majority of model predictions changes to *-(e)n*. Providing the masculine or

[7]We trained 95 further model instances with the same setup, to approach the sample size of speakers tested by McCurdy et al. (2020), and found that the pattern of increased *-s* production on non-rhymes is not statistically reliable.

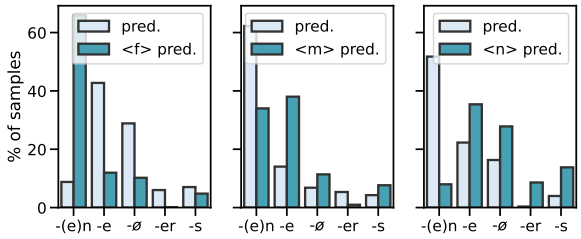


Figure 5: Predicted plural classes change when new gender control tokens (<f>, <m>, <n>) are enforced for nouns from the Wiktionary validation data that normally would not have that grammatical gender.

neuter one reduces the amount of *-(e)n* predictions, increasing *-e*, *-ø* and *-s* predictions. The plural class *-er* only appears associated with the neuter grammatical gender. Taken together, the impact of gender on both wug data and Wiktionary data suggests the model has learnt to strongly rely on the gender markers.

Enforcing length The high frequency of *-s* predictions for nonce words is remarkable, given the relative rarity of the *-s* plural. We observe, however, that the nonce words overall are rather short (4.6 characters), and that the nouns from the *-s* class are the shortest in Wiktionary (see Table 1). To investigate whether the model has learnt a causal connection between input length and emitting *-s*, we sample an equal number of nouns from the Wiktionary validation set of each gender that are balanced for whether their singular form ends in a vowel or a consonant. We then lengthen them by prepending nouns of three lengths ("See", "Haupt" or "Lieblings") to form compounds, which simulates a length increase without altering the target's plural class (which is generally determined by the second noun in a compound in German). Our results confirm that the models emit *-s* less often for longer inputs (see Figure 4b), suggesting that they rely on length as a shortcut for predicting *-s*.[8]

4 Diagnostic classification

We now look into how and where the plural classes are encoded by the model. To do so, we use *diagnostic*, or *probing classifiers* (DCs, Adi et al., 2017; Belinkov et al., 2017; Hupkes et al., 2018; Conneau et al., 2018), commonly used to estimate the extent to which hidden representations of a

[8]Since copying the stem of longer words might be difficult, we focus on suffix accuracy only in Figure 4b. Appendix A discusses a related experiment for nonce words.

Class	Gender	n final letters		Both	
		$n=1$	$n=2$	$n=1$	$n=2$
-(e)n	90.3	76.1	87.2	93.4	95.6
-e	59.8	55.8	74.9	73.7	87.3
-ø	0.0	67.4	88.2	79.0	92.5
-er	0.0	0.0	41.1	55.6	78.6
-s	0.0	39.0	49.0	41.2	55.8
Macro F_1	30.0	47.7	68.1	68.6	82.0

Table 3: Performance (F_1) of the plural class of the models' outputs for the validation set, for several majority baselines, conditioned on 1) gender tag, 2) final letter(s) of the singular form, or 3) both.

neural model reflect a specific linguistic property. DCs are simple classifiers that are trained to predict that property from the representation. The DC's performance on new data is assumed indicative of whether the linguistic property was, in fact, encoded.[9] Our experiments with DCs target: i) how well different hidden representations encode plural classes; ii) how the DC performance evolves over time while processing an input; iii) what is special about the hidden representations' neurons that are most salient to the DC.

4.1 Setup

At the end of a training epoch, we extract the model's representations for a subset of the training data and the validation data. The training data subset contains an equal number of samples for each plural class. We record the hidden state $\vec{h}_t^{l,e}$, memory cell state $\vec{c}_t^{l,e}$ and the activations for the input, forget and output gates $\vec{i}_t^{l,e}$, $\vec{f}_t^{l,e}$ and $\vec{o}_t^{l,e}$, from the encoder layers $l \in \{1, 2\}$, with $1 \leq t \leq m$; m being the length of the input. The hidden state and memory cell state from the ultimate time step of the encoder form the initialisation of the decoding LSTM. In the absence of an attention mechanism, these representations form the information bottleneck between the encoder and the decoder.

We train DCs to predict the plural class a model will assign to an input from intermediate time steps. If the to be predicted suffix can accurately be inferred by the DCs, and this generalises to unseen examples, that strengthens the hypothesis that the suffixes are distinctly encoded in the hidden representations of the model. The targets used to train the DCs are the plural class of a model's prediction, rather than the true target class. We only train and evaluate DCs on well-formed model predictions, which means the amount of samples available for training and evaluation changes across epochs. Training lasts for 50 epochs, with a batch size of 16, a learning rate of .00025 and Adam as optimiser. We train five DCs per model and evaluate the DCs using the Wiktionary validation data, with F_1-scores per plural class and the macro average across classes.

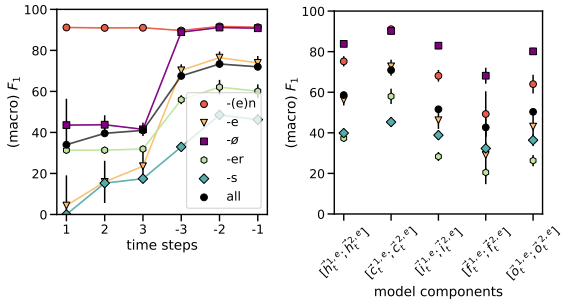
We compare with rule-based baselines, where the plural class is estimated from the grammatical gender or the final characters of singular nouns. The F_1-scores are provided in Table 3. Less well-informed baselines predicting one class only, or predicting at random according to the frequencies of the different classes, obtain macro-averaged F_1-scores of 12.5 and 19.6, respectively.

4.2 DC results

We first consider the difference between different model components (i.e. hidden states, gates) and processing steps. For every input token, we consider the first and the last three time steps. At time step 1, the model processes the gender tag, followed by the first and second character of the noun in time steps 2 and 3. Time steps -3 and -2 correspond to the last two characters of the noun; time step -1 is the *end-of-sequence* (EOS) token. We train separate DCs for every time step, using representations from the 25$^{\text{th}}$ (i.e., final) epoch. Figure 6 reports the F_1-scores of the DCs, per plural class; the macro-averaged F_1-score is shown in black.

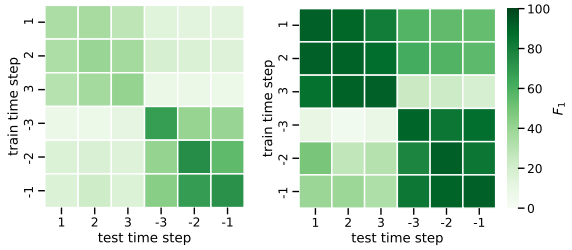
In Figure 6a, we visualise results for DCs trained on the concatenated hidden and memory cell states, for multiple time steps in the encoder. Figure 6b summarises the performance in the last time step for the remaining model components (full figures are in Appendix B). The graphs show that DCs trained on hidden and memory cell states consistently outperform DCs trained on gates, and that the memory cell state alone captures nearly the same amount of information as the concatenated hidden and memory cell states. For all components, performance increases when the model has processed more characters. A remarkable exception to this is the *-(e)n* class, for which performance immedi-

[9] This argument has been problematised (e.g. Voita and Titov, 2020; Pimentel et al., 2020b; Sinha et al., 2021): the DC may *learn* the task instead of *extracting* information. To assess the extent of this problem for our case, we ran linguistically meaningless control experiments (Hewitt and Liang, 2019) and show that our DCs can reach up to 41% macro-averaged F_1-scores there. The setup and results are listed in Appendix B. Furthermore, our experiments with *interventions* (§5) causally link our DC experiments to the model's behaviour.



(a) $[\vec{h}_t^{1,e}; \vec{h}_t^{2,e}; \vec{c}_t^{1,e}; \vec{c}_t^{2,e}]$ (b) Remaining components

Figure 6: Performance for DCs trained and evaluated with data per time step, separately. Negative time steps are relative to the position of the EOS token (position -1). In (a), performance is shown for the concatenation of the hidden and memory cell states. (b) shows the final time step for the remaining model components.



(a) All classes (b) *-(e)n*

Figure 7: DC F_1-score when training on representations from one time step, and testing on representations from another, (a) averaged over all plural classes, and (b) shown for *-(e)n* only.

ately peaks at the first time step. Considering the behavioural analyses indicating a large impact of the feminine gender tag on this class, the DC may have learnt that this is a strong predictor for that category. The gender tag is fed during the first time step, and may be encoded still towards the end.

Focusing on the DC performance for the concatenations of the hidden and memory cell states, these F_1-scores are either similar to the highest baseline performance in Table 3 (for *-(e)n*, *-ø*, *-e*) or even sub-par compared to those scores (for *-s* and *-er*). Taken together with the impact of gender observed for *-(e)n*, and the fact that the scores increase as the word is being processed, this suggests that the last few letters and the grammatical gender are essential features in predicting the plural class.

Generalisability across time steps Following Giulianelli et al. (2018), we now test how well DCs generalise across time steps to get an indication of when consistent representations of the plural classes are formed. Considering again epoch 25, we test our DCs trained on the concatenation of the hidden and memory cell states from one time step and evaluate on another, for time steps 1, 2, 3, -3, -2 and -1. We show the results in Figure 7, where the diagonal corresponds to results in Figure 6a, and the off-diagonal entries represent generalisation across time steps. Classifiers trained on early time steps do not generalise to representations close to the end. This is unsurprising, considering that features learnt by early DCs cannot be based on the noun, since it has not been processed yet. Given the average input length of 11, time steps 3 and -3 will typically be far apart, which is why time step 3 need not generalise to time step -3 for the majority of the inputs. Yet, generalisation is not good even among early or late time steps, with the exception of the *-(e)n* class (see Figure 7b), for which there are blocks visible in the upper left and bottom right corners, suggesting that the DC relies on the same feature in multiple time steps. The absence of blocks in the lower left and upper right corner implies that that feature is differently encoded at the beginning than at end of processing – e.g. because the hidden representations store more information later on. For the remaining classes, even the last two time steps do not generalise perfectly to one another, which either means that the plural class is not decided until the end or that the decision is encoded in multiple ways, with the DCs in different time steps picking up on different features.

Performance over epochs Figure 8 shows the F_1-scores for DCs trained on the hidden and memory cell states for different training epochs. Because the number of well-formed predictions changes over the course of training, the size of the dataset available for training DCs increases over time. Nonetheless, the DCs' performance on evaluation data remains stable, or even slightly decreases over time. A potential cause could be that inputs for which the model learns to emit a class after the initial epochs are atypical nouns for which the model memorises a suffix to emit, but whose features do not generalise towards new inputs.

4.3 Dissecting the representation space

To better understand the features that the DCs rely on, we train sparse DCs by applying L_0 regularisation to the DCs' parameters, that reduces the number of non-zero weights in the classifier. These DCs are trained on representations from epoch five, when performance of the DC peaked in Figure 8.

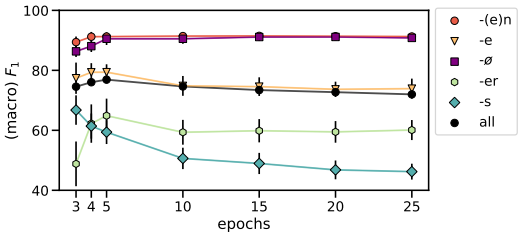


Figure 8: DCs trained on the concatenated hidden and memory cell states for seven training epochs.

The L_0-norm is not differentiable and cannot simply be added to the training objective as a regularisation term, but is incorporated through a mask that is multiplied with the weights of the linear layer, where a collection of non-negative stochastic gates determine which weights are set to zero. We refer the reader to Louizos et al. (2018) for a detailed explanation of this regularisation technique. The sparse DCs are trained for 50 epochs, with a learning rate of .001; the L_0 component in the loss is weighted by a hyperparameter $\lambda = .005$.

We use sparse DCs with, approximately, 95% of the hidden dimensions excluded, to visualise the non-zero weighted dimensions per output class, using t-SNE projections. We only include dimensions that are not pruned by five DCs trained with different random seeds.[10] We then visually inspect the representations by considering features such as the RNN's predicted plural class, the class the DC predicts, the grammatical gender, the singular noun length, and the last few letters of a word. We have four main observations: (1) The gender tags are grouped for *-(e)n* and *-er* – e.g. see Figure 9b for *-er*. This corresponds to the fact that in Figure 5, masculine and neuter proved predictive of *-(e)n* and *-er*, respectively. Furthermore, the feminine tag is grouped for *-e* and *-ø*, but the three tags are scattered for *-s*, as shown in Figure 9a. (2) For all classes, the length is an important organisational feauture in the representation space, but for all classes except *-s*, the DC still predicts that class for nearly all input lengths – e.g. compare Figures 9c and 9d. (3) The final letter of the singular noun is a prominent organisational feature too; there are clusters of 'e', 't' and 'r', in particular, that are among the top five most frequent final letters of nouns in Wiktionary (see, for example, Figure 9e). Vowels other than 'e' are typically scattered across the representation space, except for *-s*, for which

[10] We inspect the visualisations separately for five models, and include findings that hold for multiple models.

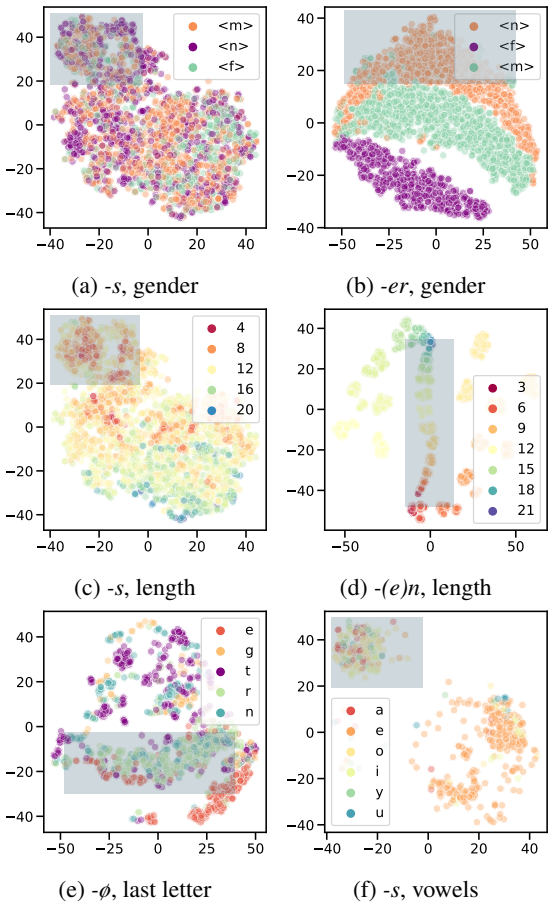
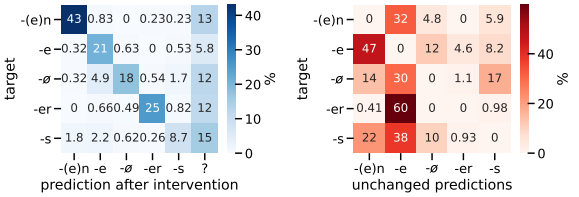


Figure 9: T-SNE visualisations of hidden and memory cell states. The dimensions t-SNE uses vary per figure and are those most relevant to the plural class in the caption. Colour schemes show (a, b) gender tags, (c, d) lengths of singular nouns, (e) the most frequent last letters of singular nouns, (f) vowels occurring as the last letter. Grey approximately marks the area in which the DC predictions match the plural class in the caption.

they cluster (Figure 9f). (4) Lastly, while all predicted classes cluster together when we select the dimensions from the sparse DC for that class, *-e* cannot easily be localised, potentially due to the fact that two other suffixes can involve adding an 'e' to the singular noun (i.e. *-(e)n* and *-er*).

5 Interventions

Until now, we have only been able to relate DC features to models' behaviour by making adaptations in the inputs fed to the model. To strengthen these results, we now ask: can we also change the models' behaviour *without* changing the input, by changing the input's hidden representation? To do so, we use *interventions* (Giulianelli et al., 2018) that halt the model while it processes inputs and change its representations using the DC. We moni-



(a) Changed predictions (b) Unchanged predictions

Figure 10: The results of interventions, showing target class distribution for (a) interventions that changed the suffix of the prediction, and (b) those that did not.

tor the effect to establish a causal link between the DC's results and the model's predictions.

Setup Following Giulianelli et al. (2018), the DC findings are linked to models' behaviour by adapting the hidden representations through the signal provided by DCs, while monitoring the impact on the models' predictions. We perform interventions on the hidden and memory cell state from the final encoder time step, by running the RNN, halting it after the encoder processed the input and intervening on the decoder's initialisation before it predicts an output. Assuming $\vec{h}$ is the hidden representation, we use the DC as follows: $\vec{h} \leftarrow \vec{h} - \alpha \nabla_{\vec{h}} L_{DC}(\vec{h})$. We perform interventions with respect to the true plural classes of samples from the validation set for which the prediction is well-formed but not correct, with α empirically set to 2.0.[11]

Results Figure 10a summarises the impact of interventions during the fifth epoch. For well-formed model predictions that have been assigned the wrong plural class, we can change the model's prediction to the right class in up to 43% of the samples with target *-(e)n* confirming that the information detected by DCs is partially also used by the model. For *-e*, *-ø* and *-er*, a smaller, yet still substantial percentage can be corrected (18-25%). However, this comes at a cost; some previously well-formed predictions are no longer well-formed (*-?* in the figure). In most of these cases, though, it is not the plural class that was corrupted, but the rest of the noun – i.e. the intervention sometimes negatively impacts the decoder's ability to recover all of the noun's characters. That the DC has picked up on class-specific features can be deduced from the fact that the interventions either change the class to the correct one, or make the predictions less well-formed, but hardly ever cause the model to emit a different incorrect plural class.

In many cases, intervening does not lead to any changes in the models' predictions, as shown in Figure 10b. Predictions belonging to the plural class *-e* are typically immune to interventions, which may, again, be due to the fact that two other classes (i.e. *-(e)n* and *-er*) contain 'e' as part of their suffix.

[11] The success of interventions depends on α. Figure 14a in Appendix C illustrates how.

6 Related work

Our analysis draws upon current research investigating neural model representations. We apply these techniques to German plural generalisation, a challenging domain with an extensive cognitive and linguistic literature.

Morphological inflection in neural networks Recently, others have explored the potential linguistic and cognitive implications of morphological generalisation in neural networks. Malouf (2017) visualised the representation space learnt by RNNs to draw connections with more traditional linguistic categories. King et al. (2020) and Gorman et al. (2019) grouped sequence-to-sequence model errors into linguistically meaningful categories. Neural models have been used to estimate the information theoretic contribution of meaning to gender (Williams et al., 2019) and of meaning and form to gender and declension class (Williams et al., 2020). McCarthy et al. (2020) used grammatical gender classes to track phylogenetic relationships between related languages, while others used them to model morphological learnability (Elsner et al., 2019; Cotterell et al., 2019; Forster et al., 2021).

Probing has been used to investigate how linguistic information is encoded in neural model representations (Alain and Bengio, 2017; Hupkes et al., 2018; Goodwin et al., 2020; Ravichander et al., 2020), including morphological structure (Torroba Hennigen et al., 2020). Much recent debate has focused on appropriate methods for probing (Belinkov, 2021; Hewitt and Liang, 2019; Hall Maudslay et al., 2020; Pimentel et al., 2020a; Ravichander et al., 2021; White et al., 2021). Our work applies probing to German plural inflection and bolsters it using causal interventions.

German plurals and the past tense debate In a wider context, our work fits within the (in)famous *past tense debate*, one of the longest and most vigorous conflicts in cognitive science (e.g. Seidenberg and Plaut, 2014), which contrasted neural network models of English past tense inflection

(Rumelhart and McClelland, 1986) against theories of generalisation which emphasised a need for symbolic rules (Pinker and Prince, 1988).

German plurals have been an important phenomenon for this debate. Dual-route theorists argued that German speakers show rule-based generalisation for one plural class – the numerically rare *-s* class – and analogical generalisation of the other classes (Clahsen et al., 1992; Marcus et al., 1995; Clahsen, 1999). This account has been contested by schema theories of German plural generalisation (Köpcke, 1988; Bybee, 1995). Later experiments also cast doubt on the dual-route account of speaker preference for *-s* (Hahn and Nakisa, 2000; Zaretsky and Lange, 2016; McCurdy et al., 2020), and recent rule-based models of German plural inflection model all plural classes with a unified approach (Yang, 2016; Belth et al., 2021). The speaker preference for *-s* on unusual inputs, such as the non-rhyme words developed by Marcus et al. (1995), has been claimed as a key signature of human-like generalisation in contrast to neural network models (Clahsen, 1999). Similar to Goebel and Indefrey (2000), our neural model shows this behaviour, although one should be careful in interpreting this given the length shortcut observed for *-s*. The question of how rules might be represented neurally is still open to debate and investigation. Our work continues to weaken the original empirical objections to connectionist models (see also Kirov and Cotterell 2018).

7 Discussion & Conclusion

For more than 30 years, the field of inflectional morphology has been a testing ground for broader questions about generalisation in language, centred around the extent to which explicit rules are required. In this discussion, neural networks are traditionally considered as an alternative to the explicit representation of rules. However, recent studies have shown that such models show interesting generalisation patterns – sometimes comparable to behaviour observed in humans (Corkery et al., 2019; Kirov and Cotterell, 2018). This raises the question of what kind of solution is implemented by neural networks to process language in seemingly rule-governed domains, how these solutions relate to rule-based models, and what it teaches us about human processing of inflectional morphology. Our study takes a step in this direction by exploring how an RNN encodes generalisation behaviour.

We find that an RNN shows a mixture of human-like generalisation and reliance upon 'shortcuts'. The models correctly learn that key predictors of plural class include grammatical gender and the last few letters of a word, which are the same features used by the recent rule-based cognitive model of Belth et al. (2021). Our DCs show how these predictors are largely encoded in the hidden representations of the encoder. Variation in the classifiers' performance may reflect that some plural classes are encoded more consistently than others; for instance, feminine gender is highly predictive of the *-(e)n* class. Alternatively, the decoder may decide the plural class for some inputs.

On the other hand, the models' ability to cope with nouns in low-frequency plural classes is very brittle. The DCs perform worst for the minority classes, it proved hard to change the model's predictions to *-s* in the interventions, and behavioural analyses suggested the model overly relies on length as a shortcut to predict this class. By contrast, we see model bias for the frequent class *-e* in overgeneralisation behaviour (§3), in representation space dispersion (§4), and in resistance to interventions (§5). We speculate that the character-based RNN may conflate the *-e* class with the 'e' character that appears in the *-(e)n* and *-er* classes.

In summary, we contribute a detailed analysis of how an RNN processes the complex task of plural inflection in German. Interestingly, we find cognitively plausible generalisation behaviour through learnt representations which echo recent rule-based models. Future work could address the broader questions raised by these findings, such as what constitutes a rule given overlap in strategy between neural and rule-based models, and how a mechanistic understanding of how neural networks approach seemingly rule-governed domains might contribute to understanding how such generalisation is instantiated in the human brain.

Acknowledgements

We thank Elia Bruni and Ryan Cotterell for their feedback on this paper. KM is supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. VD is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Y. Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *ICLR 2017, Workshop Track Proceedings*.

Artemis Alexiadou and Gereon Müller. 2008. Class features as probes. In Asaf Bachrach and Andrew Nevins, editors, *Inflectional Identity*, volume 18 of *Oxford Studies in Theoretical Linguistics*, pages 101–155. Oxford University Press, Oxford.

Marco Baroni. 2021. On the gap between theoretical and computational linguistics. Keynote at EACL2021.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *CoRR*, abs/2102.12452.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 861–872.

Caleb Belth, Sarah Payne, Deniz Beser, Jordan Kodner, and Charles Yang. 2021. The Greedy and Recursive Search for Morphological Productivity. In *CogSci*.

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

Joan L Bybee and Paul J Hopper. 2001. *Frequency and the emergence of linguistic structure*, volume 45. John Benjamins Publishing.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39.

Harald Clahsen. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22(6):991–1013.

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

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 2126–2136.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics (TACL)*, 7:327–342.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2021. The paradox of the compositionality of natural language: a neural machine translation case study. *CoRR*, abs/2108.05885.

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L. King, Luana Lamberti Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):53.

Naomi Feldman. 2005. *Learning and overgeneralization patterns in a connectionist model of the German plural*. Master's thesis, University of Vienna.

Martina Forster, Clara Meister, and Ryan Cotterell. 2021. Searching for search errors in neural morphological inflection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Main Volume*, pages 1388–1394, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248.

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

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1958–1969.

Kyle Gorman, Arya D McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but ok: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151.

Ulrike Hahn and Ramin Charles Nakisa. 2000. German Inflection: Single Route or Dual Route? *Cognitive Psychology*, 41(4):313–360.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Dieuwke Hupkes. 2020. *Hierarchy and interpretability in neural models of language processing*. Ph.D. thesis, University of Amsterdam.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ray Jackendoff and Jenny Audring. 2018. Morphology and memory: toward an integrated theory. *Topics in cognitive science*.

David King, Andrea Sims, and Micha Elsner. 2020. Interpreting sequence-to-sequence models for Russian inflectional morphology. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 481–490.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: open-source toolkit for neural machine translation. In *ACL (System Demonstrations)*.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. 2019. Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–11.

Klaus-Michael Köpcke. 1988. Schemas in German plural formation. *Lingua*, 74(4):303–335.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, page 104699.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)*, pages 11–20.

Christos Louizos, Max Welling, and Diederik P Kingma. 2018. Learning sparse neural networks through L_0 regularization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

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

Arya D. McCarthy, Adina Williams, Shijia Liu, David Yarowsky, and Ryan Cotterell. 2020. Measuring the similarity of grammatical gender systems by comparing partitions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5664–5675.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3428–3448.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1745–1756.

Gereon Müller. 2015. *Remarks on nominal inflection in German*. Akademie Verlag.

Martin Neef. 2004. The relation of vowel letters to phonological syllables in English and German. *Written Language & Literacy*, 7(2):205–234.

Martin Neef. 2011. *Die Graphematik des Deutschen*, volume 500. Walter de Gruyter.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4609–4622.

Steven Pinker. 1998. Words and rules. *Lingua*, 106(1-4):219–242.

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

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.

D E Rumelhart and J McClelland. 1986. On Learning the Past Tenses of English Verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 216–271. MIT Press, Cambridge, MA.

Mark S. Seidenberg and David C. Plaut. 2014. Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate. *Cognitive Science*, 38(6):1190–1228.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *CoRR*, abs/2104.06644.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216.

Jochen Trommer. 2020. The subsegmental structure of German plural allomorphy. *Natural Language & Linguistic Theory*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 132–138, Online. Association for Computational Linguistics.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6682–6695.

Charles D. Yang. 2016. *The price of linguistic productivity: how children learn to break the rules of language*. The MIT Press, Cambridge, Massachusetts.

Eugen Zaretsky and Benjamin P Lange. 2016. 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 2015*, number Band 15 in Bamberger Beiträge zur Linguistik, pages 153–178. University of Bamberg Press, Bamberg.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

A Additional behavioural analyses

Here, we present two additional analyses for the nonce word stimuli from Marcus et al. (1995). Firstly, we present them to the model as compounds, to investigate whether these longer inputs change the model predictions too (see §3). We form a novel noun-noun compound with the nonce word in the second position while keeping the neuter tag (e.g. presenting `<n> t i e r b r a l </s>` to the model instead of `<n> b r a l </s>`), using three nouns of different genders ("der Zahn", "die Hand", "das Tier"). Generally, the plural class is determined by the second noun in a compound in German, but it is possible that our models might be biased by the first noun of the compound to emit a different plural class. There is only a small impact of the specific noun used to form a compound (Figure 11). A pattern that is more pronounced is that, overall, there are fewer *-s* predictions, and many more *-er* predictions. Considering that by creating a compound we increased the length of the nonce word (from 4.6 to 8.6), this suggests a correlation between input length and plural class emitted, as has been previously observed in the main paper.

Secondly, we investigate how models' predictions for nonce words change during training, as shown in Figure 12. Small fluctuations aside, the nonce predictions do not appear to change substantially after the point of overgeneralisation shown in Figure 3, even though the model's training accuracy increased until the end of training. This pattern is consistent in the remaining productions for which Figure 4 only showed the final epoch's predictions for brevity. It seems the predicted suffix classes have been decided for the majority of the inputs early on during training. Considering that German plural noun prediction suffers from a lack of generalisation for minority classes, this initial phase during training might be a crucial period to remedy this.

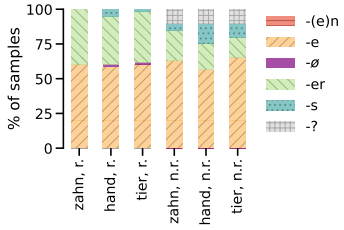
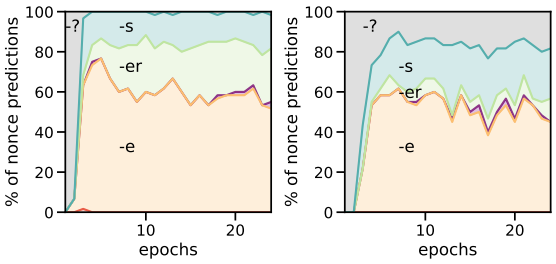
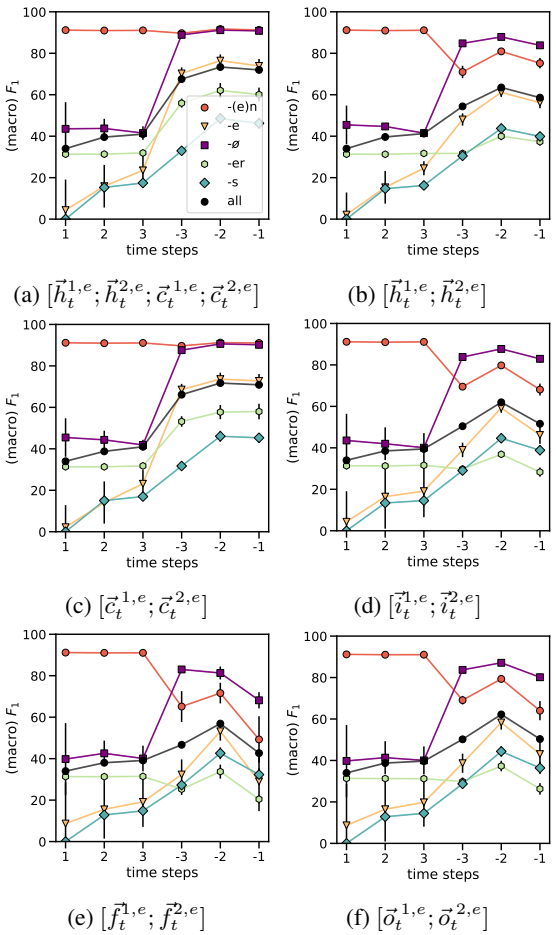


Figure 11: Predicted plural classes for nonce words when presented as a compound with the neuter gender.



(a) Rhymes (b) Non-Rhymes

Figure 12: Predictions for the nonce nouns from Marcus et al. (1995), presented with the neuter gender tag.



(a) $[\vec{h}_t^{1,e}; \vec{h}_t^{2,e}; \vec{c}_t^{1,e}; \vec{c}_t^{2,e}]$ (b) $[\vec{h}_t^{1,e}; \vec{h}_t^{2,e}]$

(c) $[\vec{c}_t^{1,e}; \vec{c}_t^{2,e}]$ (d) $[\vec{i}_t^{1,e}; \vec{i}_t^{2,e}]$

(e) $[\vec{f}_t^{1,e}; \vec{f}_t^{2,e}]$ (f) $[\vec{o}_t^{1,e}; \vec{o}_t^{2,e}]$

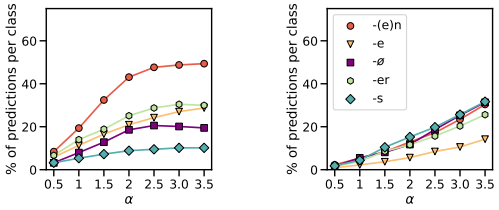
Figure 13: DCs trained over various model components extracted from the model's encoder. The DC is trained and evaluated with data per time step, separately. Negative time steps are relative to the position of the EOS token in position -1.

B Additional DC analyses

In Figure 13a, we visualise results for DCs trained on the concatenated hidden and memory cell states, for multiple time steps in the encoder. The remaining graphs present the same performance measures for DCs trained on (13b) the hidden states, (13c) memory cell states, and the (13d) input, (13e) forget and (13f) output gates.

C Additional analyses interventions

§5 presented causal interventions with $\alpha = 2.0$. That hyperparameter controls the size of the update of the hidden representation. A large α yields more successful interventions, but also more interventions that are not well-formed. As α increases, so does the frequency of these errors. We summarise this trend for each of the plural classes in Figure 14.



(a) Successful interventions (b) % not well-formed

Figure 14: Visualisation of the impact of the step size α on the percentage of changed predictions per target class, with correct predictions per class and predictions that are no longer well-formed.

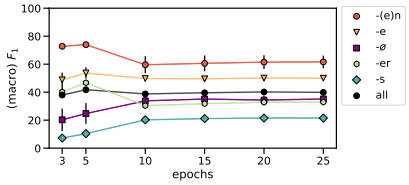


Figure 15: Performance (Macro F_1) over epochs for DCs trained on the control task.

D Control tasks

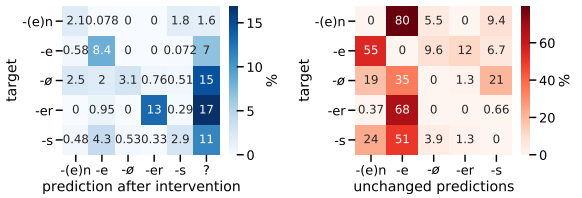
In order to assess the ability of our probes to learn a linguistically meaningless control mapping, we train a DC in a control setting (Hewitt and Liang, 2019), by randomly reassigning labels to each input. Note that we cannot use word identity, as is used by Hewitt and Liang (2019), as a basis for our label shuffling in our morphological task: words do not reappear. Instead, we randomly assign classes based on two features of the input word: gender tag and final two letters. For an example, see Table 4. Control classes are sampled based on their actual frequency as described in Table 1. We again train five DCs with different random initialisations, including different control labelings, and report their averages. The control DC can reach a 41% macro-averaged F_1-score. Figure 15 shows an overview of the accuracy of the control DCs per epoch.

We further use the control DCs to perform causal interventions. As described in §5, we only perform interventions on those predictions that are well-formed, but mapped to the incorrect class. The

Gender & Singular form	Features	Control class
<m> Strauch	m, ch	*-(e)n*
<m> Tisch	m, ch	*-(e)n*
<m> Wagen	m, en	*-e*

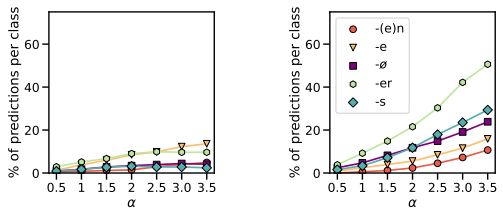
Table 4: An excerpt from a possible control label shuffling. The correct plural forms emitted by the recurrent model are Sträucher (*-er*), Tische (*-e*) and Wagen (*-ø*) respectively. The control mapping instead picks a random class based on two features: gender and final two letters. "Strauch" and "Tisch" are assigned to the same randomly chosen class *-(e)n*, such that the mapping is deterministic, but linguistically meaningless.

gradient produced by the control DC, with which we update the activations is, unlike the original setup, not informed by the actual class outputted by the recurrent model. This "class" in our control task is instead a random set of words, that do not necessarily correspond to the plural form emitted. Successful interventions are therefore coincidental. The results for all interventions are listed in Figure 16. The impact of the hyperparameter α on the control DC is visualised in Figure 17. Some outputs can be corrected using an effectively meaningless hidden state update: the maximum percentage of corrected predictions never reaches above 20% for any particular plural class, as can be seen in Figure 17a (compare Figure 14a).



(a) Changed predictions (b) Unchanged predictions

Figure 16: The results of performing interventions ($\alpha = 2.0$) using the DC trained on the control task. (a) Target class distribution for interventions that changed the prediction's suffix. (b) Target class distribution for interventions that did not change the suffix.



(a) Successful interventions (b) % not well-formed

Figure 17: The impact of α on the changed predictions per target class for DCs trained on the control task.

E Rule-based model

We train rule-based models in the manner suggested by Belth et al. (2021). A part of the model can be seen in Figure 18. We make visualisations of the full models available here.

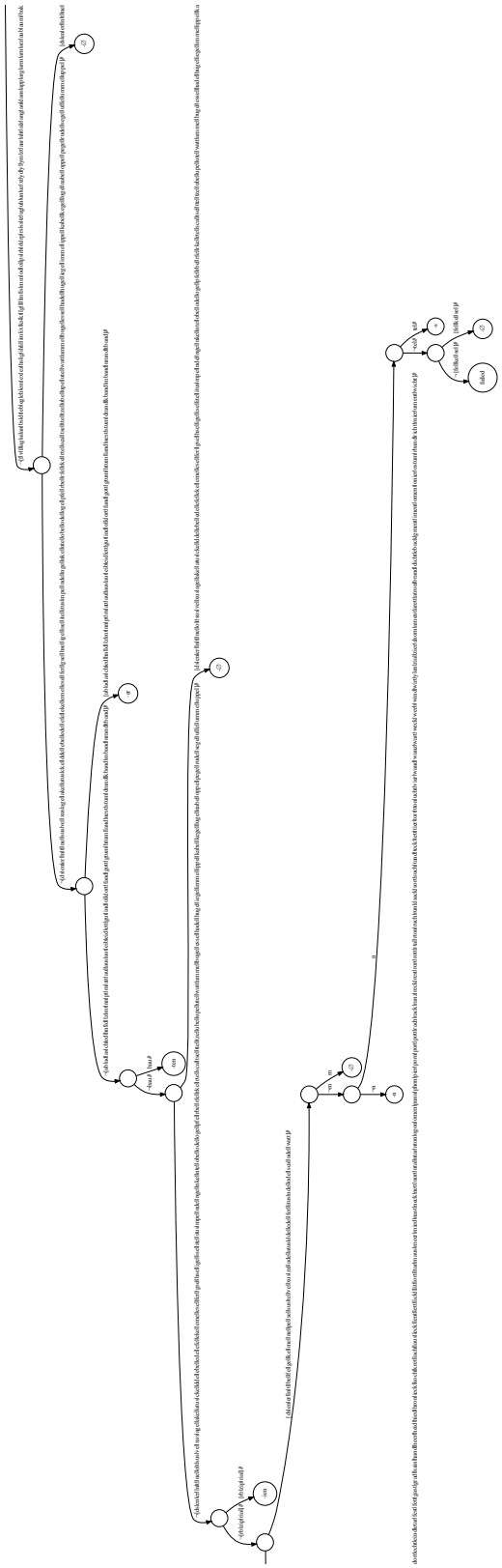


Figure 18: A part of a rule-based cognitive model (Belth et al., 2021), trained using the Wiktionary training set.