Modeling human-like morphological prediction

Modeling human-like morphological prediction

Eric Rosen
University of Leipzig
errosen@mail.ubc.ca

Abstract

We test a model of morphological prediction based on analogical deduction using phonemic similarity by applying it to German plural suffix prediction for a set of 24 nonce forms for which McCurdy et al. (2020) elicited human judgements, and which they found were poorly matched by productions of an encoder-decoder model of Kirov and Cotterell (2018). Their results raise the question of what kinds of models best mirror human judgements. We show that the predictions of the analogical models we tested mirror human judgements better than the encoder-decoder model.

1 Do neural models of morphological prediction emulate human behaviour?

Despite the recent success of neural models of morphological prediction such as the encoder-decoder (ED) model of Kirov and Cotterell (2018) (henceforth KC), two recent papers: Corkery et al. (2019) and McCurdy et al. (2020) (henceforth CMG and MGL) question how well these models' predictions of nonce forms match those of human judgements. Corkery et al. (2019) re-examine KC's application of their ED model to English past-tense nonce forms developed by Albright and Hayes (2003) (henceforth AH) through multiple random initializations of their model and find that KC's model predictions do not align with AH's results as well as reported by KC.

MGL pursue this question further by eliciting human judgements of possible German plural forms of 24 nonce words originally developed by Marcus et al. (1995) (henceforth M95): 12 'rhymes' with regular phonotactic patterns and 12 phonologically atypical 'non-rhymes', shown in table 1. As MGL put it, KC's claim, that "modern Encoder-Decoder (ED) architectures learn human-like behavior when inflecting English verbs, such as extending the regular past tense form to novel words" does not address a point made by M95: that neural models "may learn to extend not the regular, but the most frequent class – and thus fail on tasks like German number inflection, where infrequent suffixes like /s/ can still be productively generalized." As did CMG with AH's English nonce forms, MGL apply KC's ED model to M95's German nonce forms and compare them with their elicited human judgements. They find that the ED model fails to match human prediction in German plural formation, where, unlike in English, no class holds a majority.

Outline of the paper Here, we test to what extent an alternative model that predicts forms through analogical implicative relations can improve on an ED model for matching human prediction. In the rest of §1, we further discuss how MGL's wug test results compare with those of the ED model. In §2 we present variations on an alternative model of nonce word prediction. In §3 we compare the predictions of our model with MGL's human predictions. In §4 we compare our model with other models. In §5 we report tests made on real data. §6 concludes with a discussion.

Rhymes	Non-Rhymes
pind	fnahf
kach	pläk
spand	pnähf
spert	plaupf
klot	pröng
bral	fnöhk
raun	fneik
mur	bnöhk
vag	snauk
nuhl	pleik
pund	bnaupf
pisch	bneik

Table 1: 24 nonce forms developed by M95 and tested by MGL

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 133-142.
Held on-line February 7-9, 2022

MGL's wug test results Table 2, reproduced from MGL, shows MGL's wug-test results in percentages for each suffix. They find a high degree of variability among speaker data, where no plural class dominates, and /e/ is the most common suffix at around 45%. /en/ and /s/ are more common in non-rhymes than in rhymes. /er/ is less common in non-rhymes. Relatively low ratings for /s/ conflict with M95 who claim that /s/ is a default suffix that can apply in any environment.

Plural		Prod %
/-e/	R	45.3
	NR	44.4
/-(e)n/	R	25.0
	NR	34.7
/-er/	R	17.4
	NR	6.7
/-s/	R	4.2
	NR	6.4
/-∅/	R	2.7
	NR	2.7
other	R	5.4
	NR	4.8

Table 2: MGL's survey results (R=rhymes, NR=non-rhymes

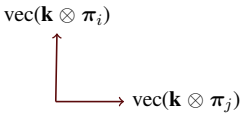
The coloured bar graphs in figure 1 (p. 5), reproduced from MGL, illustrate the differences in suffix prediction between the speaker data and their test of KC's ED model on the same nonce forms. The graphs show that the ED model predicts /en/ (purple) on the nonce forms way less than speakers. MGL suggest that the ED model over-predicts /e/ (blue) because of its frequency and does not capture minor patterns. They also observe that speaker production of /(e)n/ (purple) and /s/ (orange) is greater for Non-Rhymes relative to Rhymes. In the ED model, the tendency is reversed, where /e/ occurs for over 90% of Non-Rhymes.

2 An alternative model

As an alternative to an ED model, we explore morphological prediction through *implicational relations* (Bonami and Beniamine, 2016; Ackerman and Malouf, 2016; Ackerman et al., 2009b,a) based on phonological similarity. Because we are trying to predict a plural from an affixless singular, we can't use principal parts and we can only guess an inflectional class through phonological clues and possibly what the phonology might suggest about semantics. If a speaker knows both the singular and plural of lexeme A, they can predict the plural of lexeme B from the singular if lexeme B is similar to A and forms the plural in the same way. e.g.: Fisch → Fische ('fish(es)'), Tisch → Tische ('table(s)')

We adopt a Vector Symbolic Architecture (VSA) model (Kanerva, 2009, 1988, 2017)[1] for representing sequences of phonemes, in which vectors are binary, with a typical dimension of 10,000. A phonological feature is represented by a randomly chosen sparse binary vector. The vector for each feature will be nearly orthogonal to all other features' vectors. A phoneme is represented by the sum of the feature vectors that compose it: for example, **k** = **cons** + **dorsal**, with features **sonor**, **voi**, **cont** at zero. (Bolded terms are vectors.) **g** differs from **k** just by the addition of feature **voi**. Each phoneme needs no more than 7 features to be represented. Basing phonemes on features means that the vectors of phonologically similar segments in the same position will be relatively close in the space (e.g., /k/ and /g/), if they differ by just one feature and relatively far (e.g., /k/ and /o/) if their features are mostly different.

To represent a *sequence* of phonemes, we superpose the encodings of all the phonemes, but each phoneme vector is cyclically permuted by one bit for each step in the sequence. Permutation moves a vector to a part of the space where it is nearly orthogonal to its non-permuted position and thus to where it will not interfere with other vectors as shown below. In this framework we can use phonological features in order to make deductions based on feature similarity.



Implicative relations (Ackerman and Malouf 2016, inter alia): e.g., *Bratsche* : *Bratschen* ::

[1] As noted by an anonymous reviewer, nothing in the analysis hinges on the particular model we are using for representing sequences of phonemes. We adopt the VSA model here for convenience, but what is crucial is the idea of predicting by feature-based similarity and (to be discussed below) word frequency.

Patsche : *Patschen* (‘viola’ sg : pl :: ‘paw’ sg : pl) are predicted by vector differences where $y_{pl} \simeq y_{sg} + x_{pl} - x_{sg}$ for lexemes x and y whose phonological-feature-based vector encodings are similar according to some similarity metric. Unlike conventional neural models, our model has no network and requires no training. Although the scores for choosing predictors have continuous values, the vector representations are effectively discrete.[2]

Nouns from the Unimorph dataset are used in conjunction with two frequency archives: Institut für Deutsche Sprache (2014) and Gambolputty. We convert both singular and plural forms to a phonemic representation using the German version of Bernard and eliminate a handful of words given non-German phonemes such as *psychothriller* (θ) or *chance* (ã) to end up with 36 phonemes, encodable with 16 phonological features.

Encoding German words To predict an unknown plural form[3] of lexeme A from its singular, we look for a lexeme B whose plural form is known and whose representation of the singular is close to lexeme A’s. For example, *Kind* ‘child’ is a possible candidate for predicting the plural of nonce *pind*. If the two singular forms being compared are unequal in length, we pad the left edge of the shorter one with dummy phonemes represented by zero vectors so that their right edges align.

Calculating the score of a predicted suffix for a given word We explored different possible combinations of hyperparameters for the model to see how well the results of each marched MGL’s human predictions. The hyperparameters included the following, where the hyperparameter choice for the results given below is starred:

1. The similarity metric for choosing predictive best neighbours of a nonce form. Calculating on raw cosine similarity between the vector for the nonce word and a candidate word did not spread out the values enough to sufficiently distinguish similar words from dissimilar ones.
   - Reciprocal of sum squared vector difference.
   - Further squaring the above value.
   - Reciprocal of sum of the absolute values of the difference of vectors.
   - *$\log \frac{1}{1-s}$, where s is the cosine similarity of the vectors.
2. The frequency score for each candidate word. We tried:
   - *Raw frequency.
   - Log frequency.
   - Squared raw frequency to spread the values out more and penalize infrequent words more as candidates.
3. The width of a beam search (*beam=6) among top-scoring neighbour candidates.
4. *Comparing the best candidate(s) for each possible suffix rather than just the suffixes that appear among the top candidates.[4]
5. *Scaling the similarity score to increase towards the end of the word. When taking the cosine distance between the vectors of a nonce word and a neighbour we take not the raw vectors but vectors where the values of the component for each phoneme in the string are boosted by factor s^i, where s is a scaling factor such as 1.2 and i is the ordinal position of the phoneme in the string. e.g., for nonce word *spand*, *Pfand* ‘pledge, deposit’ and *Brand* ‘fire’ would be better predictors than *Spalt* ‘crack’ or *Spatz* ‘sparrow’.
6. *Weighting the score by the negative exponential of the syllable count difference between the candidate and the source word so that analogies are biased to be based on prosodically similar words. (e.g., quadrisyllabic *Geburtstagskind* ‘birthday child’ is scored lower than *Kind* as a predictor for nonce *pind*.)
7. *Adding a score for each suffix based on the probability of each suffix in the candidate nonce word as assigned by a single-layer

[2]MGL trained the ED model on nouns in orthographic form and say “Unlike English, the phonological-orthographic mapping is straightforward in German, so we can use a written corpus for model training.” This isn’t quite true, given the non-negligible occurrence of foreign words in the corpus like *Babysitter, Boutique or Clique*, whose German pronunciations are idiosyncratic or mutually inconsistent.

[3]MGL abstract away from questions of umlaut. See Trommer (2021) for a detailed analysis of the interaction between gender, plural allomorphy and umlaut.

[4]The former option was suggested by Matías Guzmán Naranjo (p.c.) and yielded better results.

RNN trained on all the plurals of words in the database with frequencies above 100,000.

As noted by an anonymous reviewer, it is possible to engineer the choice of the above hyperparameters to make the results match MGL's human predictions as closely as possible. If it is the case that the mode of human prediction of nonce forms closely mirrors human prediction of real data, then these engineered choices are overfitting to the extent that they diverge from a model that is trained on and best predicts real data. On the other hand, it may not be the case that human speakers predict nonce words the same way they predict real words. The difficulty in getting a model to work equally well on prediction of real words and on nonce forms is noted by CMG (p. 3874 §5.1), who write: "It seems that the ED model displays a fundamental tension between correctly modelling humans on real words and nonce words." (p. 3874) A possible reason for this tension is given by Schmitz et al. (2021), who propose that nonce words are not "semantically empty shells" and that "[t]he resonance of morphologically simplex and complex pseudowords with the words in the mental lexicon influences the processing of these pseudowords." (slide 8) If their hypothesis is correct, then speakers judging these 24 nonce words may be using associations between these words and real words that are based not on the kinds of phonological similarities that our model measures, but instead on the kinds of onomatopoeic or phonaesthematic associations that Schmitz et al. (2021) suggest. In fact, we find that the hyperparameter choice that best predicts suffixes of real data does not necessarily best mirror MGL's human prediction results. As an illustration of the mismatch between nonce word and real data prediction, figure 3 graphs the nonce word predictions made by the same model that performed best (85% accuracy) on real data. This model over-predicts that /-s/, null and in some cases the 'other' and /-er/ suffixes. It should be understood then, that the results shown below illustrate how an ideal choice of hyperparameters can mirror MGL's nonce word predictions but they should not be considered as a held-out test set of real-data training.

Table 3 shows the normalized scores and top candidate for each suffix for the first nonce word *pind* under one hyperparameter combination.

[5]This word was incorrectly given a suffix [ee] instead of [ə] by the phonemizer.

Suffix	Best neighbour	Gloss	Score
er	Kind	'child'	0.474
e	Wind	'wind'	0.392
en	Mensch	'human'	0.081
s	Trend	'trend'	0.025
null	Cent	'cent'	0.020
oth	Konzern[5]	'corporation'	0.004

Table 3: Top neighbour and normalized score for each suffix for MGL's first nonce word *pind*.

3 Graphical inter-model comparisons

Figure 1 compares predictions of MGL's human subjects with the ED model and Fig. 2 with one variation of the implicational model. The ED model greatly over-predicts the /e/ suffix at the expense of the other suffixes. The implicational model does not exactly match MGL's human predictions but we can see some patterns in common. A strong score of the /er/ suffix (green) in the first two nonce words occurs in both, but the implicational model is weaker on /er/ than human prediction on the third and fourth nonce words. The /e/ suffix (blue) is strong in the last two rhymes in both but is slightly weaker overall in the implicational model than in MGL's human judgements.

The /s/ suffix (orange) is somewhat over-predicted by the implicational model but mirrors the human judgements with a stronger overall prediction in non-rhymes than in rhymes. Overall over-prediction of /s/ by the implicational model is likely due to an abundance of foreign borrowings with this plural form in the Unimorph dataset. This can be seen if we calculate the frequencies of suffixes in the dataset and compare with the figures given by MGL, as shown in table 4. These calculations, using the frequency score from Institut für Deutsche Sprache (2014), are fairly close to the numbers given by MGL but /s/ is slightly higher.

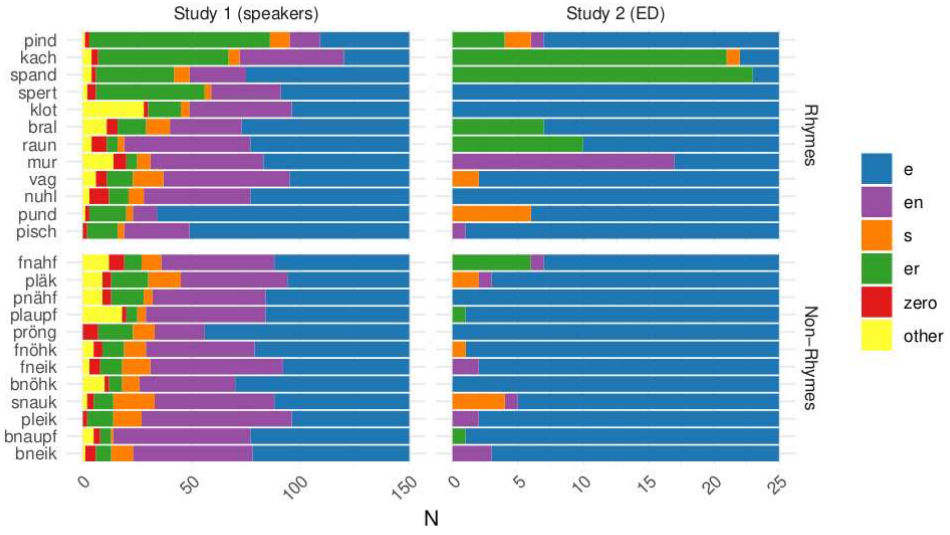


Figure 1: Plural class productions by item.

Figure 1: MGL's plural class productions compared to those of the ED model

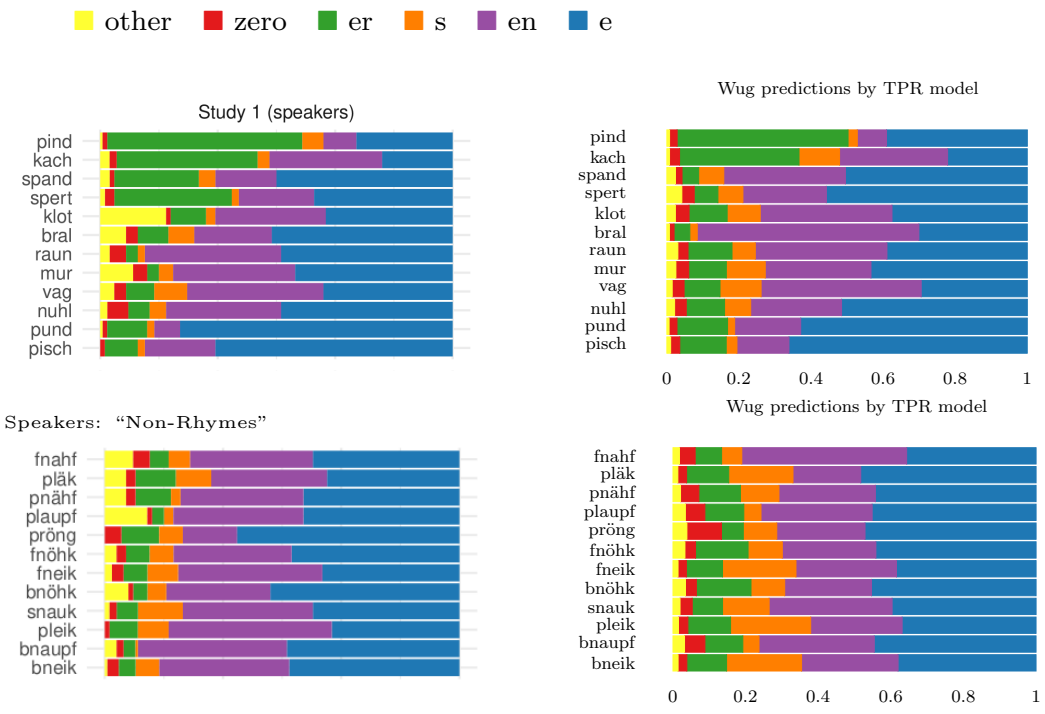


Figure 2: MGL's plural class productions compared to those of the implicational model

Suffix	Type (MGL)	Token (MGL)	Calculated here
/-(e)n/	.48	.45	.450
/-e/	.27	.21	.265
/-∅/	.17	.29	.189
/-er/	.04	.03	.035
/-s/	.04	.02	.048
other	–	–	.013

Table 4: Frequencies of suffixes in MGL's results and those produced by the current model

Calculating Spearman rank correlations MGL calculate the Spearman rank correlations between ED model production ranks and those of human speakers for each suffix across all nonce forms. They conclude from their results that there is no "statistically significant difference from the null hypothesis of no correlation." Following their approach, we perform a similar calculation to compare one set of implicational model results with MGL's speaker judgements. Table 5 shows the rank of each suffix for each nonce word for the implicational model's predictions and MGL's speaker judgements (IMP:MGL).

Nonce \ Suffix	oth	∅	er	s	en	e
pind	6:6	5:5	1:1	4:4	3:3	2:2
kach	6:5	5:6	1:1	4:4	2:2	3:3
spand	5:5	6:6	4:2	3:4	2:3	1:1
spert	5:6	6:4	4:2	3:5	2:3	1:1
klot	6:3	5:6	3:4	2:5	2:2	1:1
bral	6:5	5:6	3:3	4:4	1:2	2:1
raun	5:5	6:3	3:4	4:6	2:2	1:1
mur	6:3	5:5	4:6	3:4	2:2	1:1
vag	6:5	5:6	4:4	3:3	1:1	2:2
nuhl	6:6	5:4	3:3	4:5	2:2	1:1
pind	6:6	4:5	3:2	5:4	2:3	1:1
pisch	6:6	5:5	3:3	4:4	2:2	1:1
fnahf	6:3	5:6	3:5	4:4	1:2	2:1
pläk	6:5	5:6	4:3	3:4	2:2	1:1
pnähf	6:4	5:6	3:3	4:5	2:2	1:1
plaupf	6:3	4:6	3:4	5:5	2:2	1:1
pröng	5:6	4:5	6:3	3:4	2:2	1:1
fnöhk	6:5	5:6	3:3	4:4	2:2	1:1
fneik	5:6	6:5	4:4	3:3	2:1	1:2
bnöhk	5:3	6:6	3:5	4:4	2:2	1:1
snauk	6:6	5:5	4:4	3:3	2:2	1:1
pleik	6:6	5:5	4:4	3:3	2:1	1:2
bnaupf	6:3	4:5	3:4	5:6	2:2	1:1
bneik	6:6	5:5	4:4	3:3	2:2	1:1
$\sum d^s$	59	28	37	19	7	4

Table 5: Rank comparisons: speaker judgements and implicational model

Calculating the Spearman rank correlation between MGL's speaker judgements and model productions for each suffix, we get the correlations shown in table 6 for those calculated by MGL (ρ_{ED}) and for the implicational productions (ρ_{IMP}). The results show that the relative ranks for each suffix for each nonce word mirror those of MGL's wug tests fairly closely.

Suffix	ρ_{ED}	ρ_{IMP}	p_{IMP}
oth	n.a.	0.972	< 0.001
∅	n.a.	0.987	< 0.001
er	0.05	0.983	< 0.001
s	0.33	0.991	< 0.001
en	0.28	0.997	< 0.001
e	0.13	0.998	< 0.001

Table 6: Rank correlations for each suffix

Pearson correlations Table 7 shows the calculated Pearson correlation between MGL's production scores and the implicational model's for each of the six suffixes. Calculated individually, 3 of the 6 suffixes show significant correlation. But calculated across all suffixes we see strong correlation.

Suffix	r	p-value	significant
oth	0.159	.458	no
∅	0.318	.096	no
er	0.748	.00001	yes
s	0.578	.003	yes
en	0.340	.103	no
e	0.713	.0001	yes
all suffs.	0.902	$\simeq 0$	yes

Table 7: Pearson correlations

A possible reason for the model's poorer correlation for suffix ∅ is corpus noise. For the first six nonce words the highest scoring predictive words are dominated by spurious referents: (a) non-nouns such as *zweifel* 'second' or *drittel* 'third' with null plurals, (b) words given an incorrect null plural such as *cent* 'cent', or (c) proper names like *Siemens* or *Lutz*. And /en/'s poorer correlation is due to nonce words that got low scores for /en/ in MGL's wug tests in which our model over-predicts /en/ as a result of measuring similarity by featural closeness rather than an exact structural description. For example, *spand* gets a low score for /en/ from speakers but a high score from top candidate with an /en/ plural *Mensch* 'human', whose vowel and final consonant differ minimally from those of *spand* in their features.

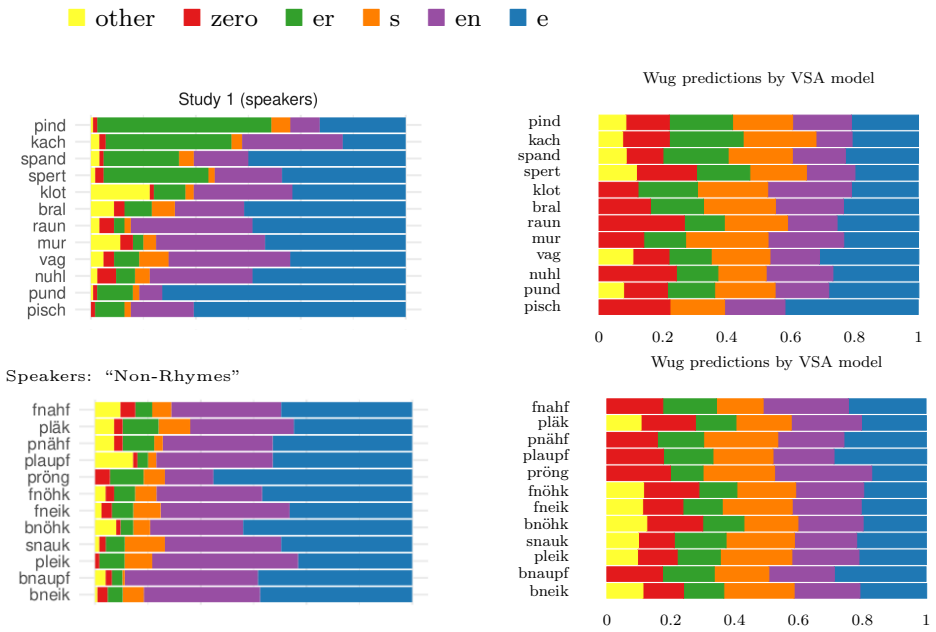


Figure 3: MGL's plural class productions compared to those of a variation implicational model that worked well on real data

Figure 4: MGL's plural class productions compared to those of Albright and Hayes' rule-based model

In summary, our Spearman and Pearson correlation results indicate that the implicational model aligns moderately well with the rankings of suffixes for each nonce word but doesn't always reflect the differences in suffix preferences by speakers among the nonce words – especially for the suffixes 'oth', 'null' and /en/.

4 Other models

Testing with Albright and Hayes' rule-based model We also ran the nonce words through the rule-based model of Albright and Hayes (2003) trained on the Unimorph corpus. Graphical results are shown in figure 4. We found that the model over-predicts the null suffix at the expense of the /er/ suffix, which hardly occurs at all. A possible reason is that the rule-based model requires an exact structural description to trigger a rule. The null suffix occurs abundantly, e.g., in nonce word *pind* because of a default rule with strength 0.396 that makes no changes to a stem ending in one of {d, l, n, r, s, t, z} if no other rules are triggered. For the /er/ suffix to occur requires a very specific rule or else a default rule with strength 0.001.

Other rule-based and symbolic models Further testing with other rule-based models could determine how well rule-based models can model MGL's human wug prediction. Payne et al. (2021) test a rule-based model based on Yang (2016)'s Tolerance Principle using morphosemantic and phonological features that include gender features when tested on German plural formation. They test stochastically sampled nouns from German CELEX, so it remains to be tested what their model would predict for MGL's wug forms.

Beniamine and Naranjo (2021) take an approach to morphological prediction that shares some common elements with ours. They use multiple alignments of forms in inflectional paradigms. Versions of our model that do not truncate a candidate word to equalize its length with the nonce word do predict stem changes such as umlaut[6], but because we are comparing with MGL's results, which abstract away from stem changes, we do not allow for possible gaps in alignment. On the other hand, because our VSA model uses binary vectors in a distributed representation, graded, continuous degrees of similarity can be used in a way that is not possible with purely symbolic models.

[6]For example, some predictions for nonce word *kach* with the /-er/ suffix produce /kɛçɛr/ with umlauted /a/.

Calderone et al. (2021) report on their morphological prediction experiments on nonce verb forms in English, German and Dutch, using several variations of a model that combines a bidirectional LSTM with 'fine alternation patterns' that figure in analogical deduction of word forms. They report Pearson correlations for regular and irregular verbs for their best-performing model along with Albright and Hayes' Minimal Generalization Learner and a purely analogical model of Nosofsky (1990). It is difficult to compare their results with ours because they are dealing with verb inflection rather than noun plurals and systems that have a clear regular/irregular split. They report ratings of 0.583 and 0.595 for regulars and irregulars respectively, which roughly compares with our results for the /s/ suffix and are lower than our result for /e/ and /er/. The results of their model, which, like ours, uses analogical deduction, but in a different way, provides further support for the role of analogical deduction in morphological prediction.

One approach that we did not take was to present all the nonce forms as neuter as MGL appear to have done. Among the 17,488 neuter nouns in the dataset, only 83, or 0.48% have an /(e)n/ suffix. Given the relatively strong presence of this suffix in MGL's wug predictions, it is not clear how presenting each nonce form as neuter would produce such results.

5 Testing on real data

The Unimorph dataset As mentioned above, we found that there was an inconsistency between model variations that best predicted the suffixes of real words in the Unimorph dataset and those that best matched MGL's results for nonce word prediction. We found that using log frequencies rather than raw frequencies gave better results for real-word prediction, so that infrequent words would have more weight, since many infrequent words in the dataset that we are testing will also have infrequent phonological neighbours. For real-word prediction, we also did not use RNN-generated perplexities which were specifically tailored to the wug words and whose main purpose was to allow suffixes that were a not a first choice for a wug word to have non-zero scores. The model achieves 85.6% accuracy on a sample of 3,390 items as compared with 88.8% reported by MGL with the ED model. The ED model arguably has an advantage in identifying foreign words in that it used

orthographic rather than phonemic input, which gives clues to a word's foreignness. For example a word spelled with c followed by a letter other than h or k is likely foreign. Foreign words make up a sizeable portion of words our model misses: for example *mustang*, *body*, *kanu* 'canoe', *strip*, *gun*, *overtime*.

6 Discussion

As discussed above, we tested many variations of the model, for which there is not space to list the details of each one's result. The variation that made the most difference to the results was the inclusion of frequency scores. A further step with this model is introduce learning, so that instead of having positional vectors intentionally orthogonal, they are allowed to move together closer in the space so that a phoneme in one position can have some measured similarity with the same phoneme in a nearby position.

Given that right-aligned edge calculation, features of segments and prosodic shape of implicational candidates were all found to contribute to predicting plurals of nonce forms based on word similarity, it is notable, as observed by an anonymous reviewer, that word similarity appears to be a multidimensional calculation that involves all of these properties.

The fact that MGL's wug tests results give non-negligible scores to all the suffix classes for most of the 24 items suggests that each suffix has found some niche or set of niches in the sense of Aronoff (2021), who gives the example of the ongoing niche competition between English /-er/ $\sim$ /-est/ and adverbs *more* $\sim$ *most* as a very complex one in its distribution. Moreover, the fact that no suffix behaves like an overpowering default choice in MGL's results suggests that the niche distribution of the German plural suffixes is also complex. Further tests will help determine to what extent this implicational model may have an advantage over purely symbolic models by being able to capture subtle distinctions between niches through its distributed representations of word forms.

Acknowledgements

This research was originally inspired by discussions of the Corkery et al. (2019) and McCurdy et al. (2020) papers at Colin Wilson's 2021 morphology seminar and by a presentation by Coleman Haley on those two papers at Paul Smolensky's Neurosymbolic Computation Lab group last spring, both at Johns Hopkins. Thanks to three anonymous reviewers and to all the members of the above two groups for helpful comments and suggestions. Thanks also to audiences at the Fifth American International Morphology Meeting where an earlier version of this research was presented. All errors are my own.

References

Farrell Ackerman, James Blevins, and Robert Malouf. 2009a. *Analogy in Grammar: Form and Acquisition*, chapter Implicative Relations in Word-Based Morphological Systems. Oxford University Press.

Farrell Ackerman, James Blevins, and Robert Malouf. 2009b. *The Oxford Handbook of Morphological Theory*, chapter Word and Paradigm Morphology. Oxford University Press.

Farrell Ackerman and Robert Malouf. 2016. *Cambridge Handbook of Morphology*, chapter Implicative Relations in Word-Based Morphological Systems. Cambridge University Press, Cambridge.

Adam Albright and Bruce Hayes. 2003. Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. *Cognition*, 90:119–161.

Mark Aronoff. 2021. Three ways of looking at morphological rivalry. Keynote talk at the 5th American International Morphology Meeting.

Sacha Beniamine and Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. In *Proceedings of the Society for Computation in Linguistics*, volume 4.

Mathieu Bernard. phonemizer.

Olivier Bonami and Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure*, 9(2):156–182.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. Technical report, arXiv.

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

Gambolputty. dewiki-wordrank.

Institut für Deutsche Sprache. 2014. Korpusbasierte Wortformenliste DeReWo, DeReKo-2014-II-MainArchive-STT.100000. www.ids-mannheim.de/derewo.

Pennti Kanerva. 1988. *Sparse Distributed Memory*. MIT Press, Cambridge, MA.

Pentti Kanerva. 2009. Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. *Cognitive Computation*, 1(2):139–159.

Pentti Kanerva. 2017. Stanford Seminar - Computing with High-Dimensional Vectors. Youtube talk at https://www.youtube.com/watch?v=zUCoxhExe0o.

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.

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

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. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756.

Robert M. Nosofsky. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34:393–418.

Sarah Payne, Caleb Belth, Jordan Kodner, and Charles Yang. 2021. The Recursive Search for Morphological Productivity. Poster presented at the 5th American International Morphology Meeting.

Dominic Schmitz, Ingo Plag, and Dinah Baer-Henney. 2021. Reconsidering pseudowords in morphological research. Slides for talk at the 5th American International Morphology Meeting.

Paul Smolensky. 1990. Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems. *Artifical Intelligence*, 46:159–216.

Jochen Trommer. 2021. The subsegmental structure of German plural allomorphy. *Natural Language and Linguistic Theory*, 39:601–656.

Charles Yang. 2016. *The price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT Press.