Learner dynamics in a model of *wug* inflection: Integrating frequency and phonology

Stella Frank (stella.frank@ed.ac.uk), Kenny Smith (k.smith@ed.ac.uk)
Centre for Language Evolution, University of Edinburgh
Edinburgh, UK

Christine Cuskley (christine.cuskley@ncl.ac.uk)
Language Evolution, Variation, and Acquisition Group, Newcastle University
Newcastle upon Tyne, UK

Abstract

A recent large-scale *wug*-task study found that non-native speakers of English tend to produce fewer regular past-tense *-ed* inflections than native speakers (Cuskley et al., 2015). In this paper we present a model that can account for this difference in behaviour as resulting from a difference in input amounts and distributions. This model attends to both frequency, using Bayesian non-parametric methods, and phonological similarity between words, using a neural model of word forms, and unifies these factors within a single probabilistic framework. We show that the general pattern of over-use of irregular inflections in non-native speakers can result simply from exposure to a smaller amount of input and does not require any model-internal distinction of native and non-native speakers. Our model also captures the interaction between class frequency and phonological similarity that was evident across all participant productions.

Keywords: inflectional morphology, modelling, learning

Introduction

For decades the English past-tense has been the "fruit fly" of linguistic research on how morphological rules and exceptions are learned from limited exposure (see e.g., (Seidenberg & Plaut, 2014) for a review). The past tense *wug*-task (Berko, 1958), in which participants are asked to provide a past tense form of a novel verb, is the archetypal experiment for probing the ability of learners to generalise morphological patterns to new forms. It has classically been used to test children's knowledge of productive morphological patterns, but has also been applied to adults (Bybee & Moder, 1983; Prasada & Pinker, 1993; Albright & Hayes, 2003; Cuskley et al., 2015). Two factors have emerged as being crucial to patterns of morphological generalisation: frequency and phonology.

English regular and irregular verbs have different frequency distributions: irregular verbs tend to be more frequent than regular ones. This early observation (Bybee, 1985) has been confirmed and quantified by corpus studies: high frequency verbs are much more likely to have an irregular past tense form, and lower frequency irregular verbs are more likely to become regular (Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Cuskley et al., 2014; Newberry, Ahern, Clark, & Plotkin, 2017). This pattern is not limited to English: across languages, morphological irregularities are more frequently found among high frequency forms (Wu, Cotterell, & O'Donnell, 2019). As a result, language learners may thus initially be exposed to a disproportionate amount of irregularity early on in learning, compared to proficient learners who have access to a fuller picture of the language.

Phonology is also an important factor in morphological regularisation and generalisation. New verbs (e.g. novel forms in the context of a wug task) that are phonologically similar to existing irregulars are more likely to be inflected using a non-regular pattern (Bybee & Moder, 1983; Prasada & Pinker, 1993; Albright & Hayes, 2003). Phonology also interacts with frequency. The marked relationship between frequency and regularity exists alongside the presence of "phonological gangs" of irregular verbs (Bybee, 2003), such as *cling, fling, sling*, and *sting*, which all form the irregular past-tense in the same way. These "gangs" of lower frequency irregulars form higher frequency blocks of quasi-regularity, which then have the bulk to sustain irregularity over time (Bybee, 2003; Cuskley et al., 2014).

This quasi-regularity is also productive under certain conditions: participants in experimental contexts are willing to extend membership to novel forms, if they have sufficient phonological similarity to existing irregulars. In a recent large-scale study investigating phonological effects, Cuskley et al. (2015) asked adult native and non-native English speakers to inflect novel verbs in a wug-style task. The novel verbs were designed to be either phonologically close to existing irregulars, close to existing regulars, or equidistant from both. In both native and non-native speakers, the phonology of the novel verbs had a marked effect: participants were significantly more likely to provide non-*ed* forms for novel verbs that were phonologically close to existing irregular verbs than to frequent regulars. However, non-native speakers were significantly more likely than native speakers to produce non-*ed* forms across all novel verb types. Furthermore, among non-natives, age of acquisition and self-rated proficiency predicted irregularization rates: later and less proficient learners were more likely to provide non-*ed* forms.

Cuskley et al. (2015) suggested that the varying rates of irregularization may reflect differences in input: less proficient learners have had less exposure to the language. Comparing their performance to native speakers and corpora statistics indicated that non-natives may be over-estimating the productivity of quasi-regularity in English, generalising from the highly frequent irregular words in their limited input (see Figure 1). While there was considerable variety in the exact non-*ed* forms the participants provided, the forms were far from

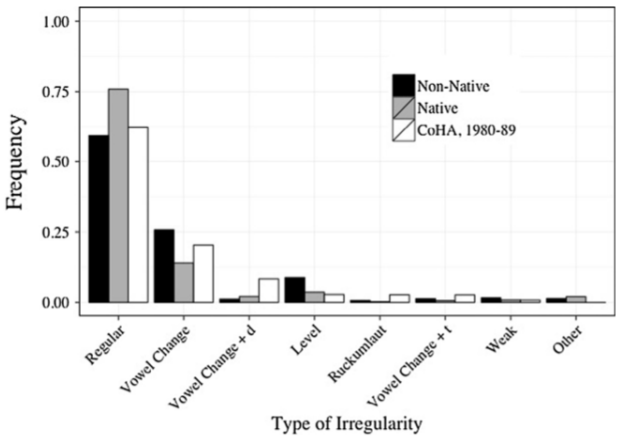

Figure 1: Reproduction of Figure 12 from Exp 2 in Cuskley et al. (2015), showing the distribution of past-tense forms produced by native and non-native English speakers, compared to the distribution of past-tense forms within a corpus (CoHA, Davies, 2010). (CC BY-NC-ND 4.0)

random, and could generally be tied to one of eight existing quasi-regular classes, shown in Figure 1, such as 'vowel change' (*spit–spat*) or 'level' (*cut–cut*).

These findings amount to a hypothesis not explicitly tested in the original study: learners with less input (later acquiring, lower proficiency non-natives) rely disproportionately on quasi-regularity found in non-*ed* forms with high *token* frequency. On the other hand, learners with more comprehensive input are influenced by the greater overall number of regular (*-ed* inflected) *types*.

In this paper we introduce a model of a Bayesian learner that aims to explicitly test this hypothesis. This model considers both frequency and phonological similarity as interacting but separable factors. Phonological similarity is evaluated using a neural network, allowing us to leverage their ability to detect patterns in high-dimensional spaces. This is somewhat reminiscent of the classic past-tense network of Rumelhart and McClelland (1986) and more recent versions (Kirov & Cotterell, 2018; Corkery, Matusevych, & Goldwater, 2019). However, those models generate past tense words forms directly, across the full vocabulary, while the neural networks here are used to generate probability distributions over forms within a given verb class, which are then incorporated within the Bayesian model.

We introduce the model incrementally, first describing the frequency component that assesses the relative importance of type and token frequency in verb class distributions. We then move on to describe the neural component responsible for estimating phonological similarity within a class. Finally, we evaluate the full model with both components against the patterns in the results from Exp 2 in Cuskley et al. (2015). The model captures the pattern of interaction between the amount of exposure (of which nativeness was a proxy in the original study), frequency, and phonology.

Modelling the Lexicon

Are type or token statistics more relevant to inferring the distribution of past tense forms in English? Arguably a cognitive model should capture both types of statistics, allowing the data, rather than the modeller, to indicate the relative importance of each. In this section we describe the Pitman-Yor process (PYP) (Buntine & Hutter, 2010; Goldwater, Griffiths, & Johnson, 2011), which can interpolate between types and tokens, depending on a particular parameter value: a learner who uses token vs. type-sensitive representations is thus, under such a model, merely tuning a continuous parameter, no different from all the other parameters that are updated as the learner sees data.

In a PYP model over the vocabulary, each token is generated either directly from an existing cached value or by generating from the base distribution, responsible for the generation of individual types (which may, however, be repeatedly drawn). We will discuss the general effects of caching within the PYP first and then go on to describe the concrete base distribution we use in more detail.

Formally, the probability of token w_i taking on value v (i.e., being of type v), given seen data $\boldsymbol{w}$, hyperparameters a and b, and a base distribution G over word types, depends on K, the number of slots in the cache, each corresponding to a draw from the base distribution, and the number of tokens generated from that cached item, n_k:

$$p(w_i = v | w_1 \dots w_N, a, b, G) = \frac{b + Ka}{b + N} G(v) + \sum_{k=1}^{K} \frac{n_k - a}{b + N} [k = v]$$

The number of cached items is bounded by the number of types V and tokens N in the data ($V \leq K \leq N$). Importantly, as a non-parametric process, the vocabulary size is not predetermined: V is potentially as large as the number of unique draws from the base distribution, which may be infinite. However, for a given dataset, V (and K and N) is finite.

The degree of caching is governed by the model-level parameter $0 \leq a < 1$: lower values of a lead to more caching, resulting in fewer draws from the base distribution, whereas values of a near one result in very little caching, meaning nearly every token is independently generated from the base distribution. The posterior statistics of the base distribution will reflect these different usages. Identifying the base distribution with the 'mental lexicon', the no-caching setting results in a lexicon based on token statistics, while a heavy caching setting results in a lexicon reflecting type statistics (since at minimum each type must be drawn from the base distribution once, in order to be added to the cache). The optimal level of caching depends on the data distribution. Instead of fixing a, a is inferred together with the other parameters of the model (the cache behaviour, e.g. K). The other parameter b has relatively little influence and we set $b = 1$ throughout.

In the model, verb types are represented as a tuple of (class, form): $v = (c, f)$, where verb classes correspond to the inflec-

tion classes in the experimental data from Exp 2 in Cuskley et al. (2015), e.g. 'Regular' or 'Vowel Change', and the form is represented by the phonological form, as given by the CMU pronunciation dictionary. We use the frequencies of past tense verbs but represent them using their lemma/present tense form because the final goal is to estimate $P(c|\text{'wug'})$, i.e., to make predictions based on the lemma form. In the training phase, the verb classes are known, as are the underlying lemmas. This corresponds to estimating the state of the mental lexicon of a learner who knows the set of past tense forms, and thus can classify lemmas into the correct classes.

During testing, the model sees a novel verb form (e.g *'splink'*) without an accompanying class assignment, and the task is to calculate the probability distributions over classes given the form and the state of the Pitman-Yor process lexicon after training. Since the verb form is novel, it by definition has not been cached, and so it has to be generated by the base distribution directly. We define the base distribution over verb classes and forms as $G(c,f) = P(c)P(f|c)$, i.e. the form is conditioned on the verb class, while the class is generated independently. The first factor $P(c)$ captures the effect of relative frequency *across* classes, described in the section below, while $P(f|c)$ can capture phonological similarity *within* classes, described in the section thereafter.

The role of frequency

In this section we use the model described above to test the role of frequency across learners with differing amounts of experience. To do so, we estimate lexicons over different amounts of data, where the key parameters being inferred are a, controlling the degree of caching, and the base probability distribution, which depends on the degree of caching.

The distribution over classes $P(c)$ in the base distribution is modelled as a Dirichlet-Categorical, $c \sim DirCat(\alpha)$. Usefully, the posterior predictive probability of this compound distribution given data $\boldsymbol{k}$, needed for the wug task, has a closed form:

$$P(c|\boldsymbol{k}) = \int_\theta P(c|\theta)P(\theta|\boldsymbol{k})d\theta = \frac{n_c + \alpha}{N + C\alpha},$$

where n_c is the number of occurrences of c in $\boldsymbol{k}$, N is the total number of items in $\boldsymbol{k}$, and C is the number of categories. Importantly, since $P(c)$ is part of the base distribution, the data $\boldsymbol{k}$ consist only of the tokens drawn from the base distribution, *not* the cached tokens. In a setting in which tokens are cached aggressively, the base distribution will generate once for each type (in order to add it to the cache); $\boldsymbol{k}$ then consists of the set of types in the original data $\boldsymbol{w}$ and $P(c)$ is a (smoothed) estimator of class frequency in the vocabulary. Conversely, if tokens are not cached and instead always drawn from the base distribution, $\boldsymbol{k}$ will be identical to $\boldsymbol{w}$ and $P(c)$ is an estimator of token class frequency.

For now we set $P(f|c)$ to a simple distribution in which a form is generated as a draw from a uniform distribution over forms, $f \sim \frac{1}{F}$, where F is the number of unique forms in the dataset. Note that this distribution is not dependent on class: a form will have the same probability regardless of c. We will change this in the next section, and the full model includes a distinct distribution over forms for each class.

Inference We use Gibbs sampling to infer table configurations and slice sampling to infer values of a after every two iterations of Gibbs sampling. Forms and classes are observed in the data and not inferred. The hyperparameters apart from a are set to fixed values: $b = 1$ and $\alpha = 0.1$, implying a moderately sparse prior distribution over classes in the base distribution. The sampler is run for 100 iterations and converges quickly, after approximately 30 iterations in most cases; results are from the final sample.

Data Following Cuskley et al. (2015), we use the 1980s section of the Corpus of Historical American English (CoHA-1980) (Davies, 2010). We select the past tense verbs (tagged with `vvd`), filtering out suppletive verbs (forms of *be, do, have*, and *go*) and any types which occur fewer than three times in the entire CoHA-1980 corpus. Input datasets for the simulations are generated by sampling a given number of tokens from the set of remaining verbs according to their frequency. All samples will thus follow Zipf's law, with many tokens of high-frequency (possibly irregular) verbs, but also containing a large number of lower-frequency (possibly regular) forms. Each sample corresponds to a learner at a different stage of learning, under the simplifying assumption that learners are exposed to a non-biased sample of the language/corpus. (This clearly does not hold for child learners, inasmuch as child-directed speech is different from the written texts in CoHA; it may be a more valid assumption for literate adult language learners.)

Results Figure 2 shows the inferred probability distributions over classes across a set of simulations using datasets of different sizes, ranging from 100 to 1 million tokens. We see that distributions estimated on less data place less weight on the Regular class compared to distributions estimated over more data; this pattern corresponds to the difference in regular responses between native and non-natives (and among more and less proficient non-natives) in Exp 2 in Cuskley et al. (2015).

The values sampled for a varied with dataset size, with smaller datasets resulting in higher a: with $N = 100$, a fluctuated around .90 after convergence, while for the largest dataset, a was around 0.45. The caching behaviour reflected a, with proportionally more tokens being drawn from the cache in larger models. However, as the distributions in Figure 2 show, this did not lead to the expected type-vs-token statistics trade-off: all models inferred a distribution over classes that corresponded to the empirical frequency of word types in the data sample they were exposed to. However, smaller samples contained a higher proportion of certain irregular verb types, leading to a relative over-estimation of those classes (primarily those involving vowel change, 'VC').

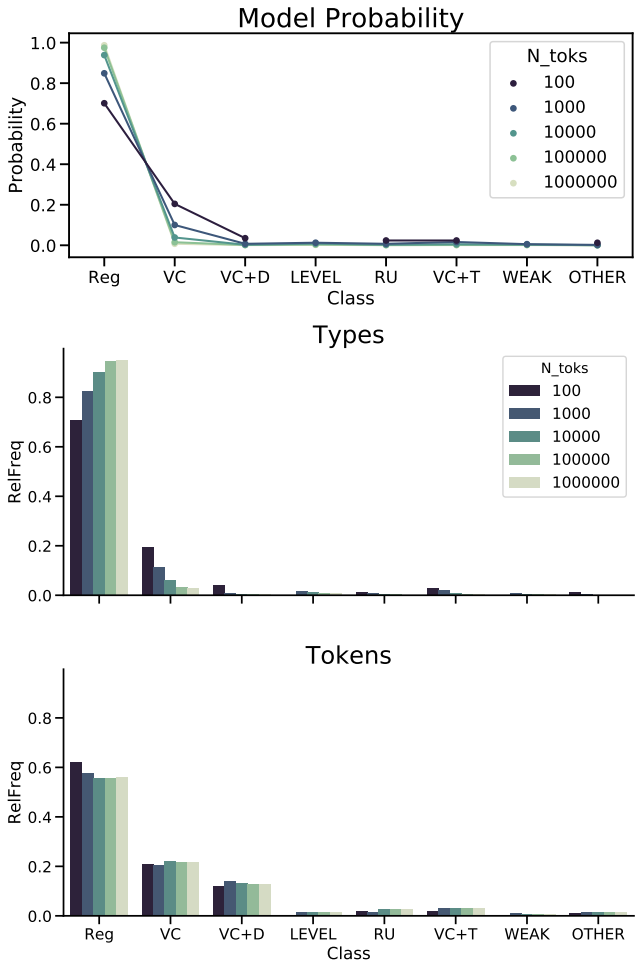

Figure 2: Top: Probability distribution over classes after training on different amounts of data (N_toks). Learners with more exposure prefer the regular class. (The line corresponding to $N = 100$ is discontinuous because some classes were not seen.) Below: Empirical distribution of types and tokens in differently sized samples. Class labels correspond to those in Figure 1.

These results support the hypothesis that non-native speakers may be over estimating quasi-regularity on the basis of their exposure to proportionally more irregular word types, compared to native speakers. However, it does not seem to be the case that estimating verb class distributions requires attention to token statistics, counter to the original hypothesis.

Phonological similarity

Inflection patterns are influenced by phonological similarity as well as frequency patterns investigated in the previous section. In this section, we focus on modelling $P(f|c)$, the distribution of forms within a class. Roughly, forms known to be within a class should receive high probability, along with novel forms that are plausible members of that class. Plausibility in this case is based on phonological similarity: does the novel form contain similar sequences of sounds as other members of the class?

We structure $P(f|c)$ as a language model, in which the probability of the form is the product of the probability of the characters, generated in sequence: $P(f|c) = \prod_i^{|f|} P(f_i|f_{0\ldots i-1}, c)$. We depart from the Bayesian methodology and use a neural network to estimate these distributions. Neural networks are able to capture rich non-linear patterns in data that would be hard, if not impossible, to define explicitly: the potential is that, in order to accurately predict the next character in a single form, the network will make generalisations across all the forms in the class. Since forms and classes are known during the training phase, we only have to train each class-specific model once, keeping inference tractable.

Within the neural language model, the model predicts each character in the word based on the history (all preceding characters in the word). Forms are predicted as a sequence of characters corresponding to the phonological representation of the word (i.e., the model predicts pronunciation, not orthography). Language models are cluster specific, corresponding to $P(f|c) = P_{lm_c}(f)$, and thus capture the probability distribution over forms in that class. Ideally, this distribution could capture the phonetic similarity and variability of members of the class: intuitively, a distribution over forms that has seen words like *bend, send, lend*, and thus assigns them high probability, would also give high probability to similar but unseen words like *nend*.

Model parameters The architecture of the character language models is intentionally kept simple, in order to avoid overfitting on small datasets. All models have the same architecture: input is given to a single LSTM layer with 128 dimensions, followed by a learned linear layer with softmax normalisation on the output, giving a probability distribution over the next character. Learning is done using RMSProp with default parameters. The input to the model is the phonological word form in which each segment is mapped to a feature vector over 13 phonological features (e.g. the vowel in *wug* would be represented as a vector corresponding to features such as -consonant, +mid). Output prediction is over the 35 XSampa symbols used to represent pronunciation. Representing the input using features enables the model to generalise over segment types (e.g. vowels) more easily than with categorical ('one-hot') representations; conversely, on the output side, the model is constrained to producing legal symbols, instead of potentially generating unattested combinations of features.

Training All training is done over word types, not tokens. First, a general English character language model is trained using early stopping on a validation set. The training data for this dataset is sampled from the set of words in CoHA-1980s that are not tagged as a verb (any tag not starting with v) and are shorter than a maximum length of 10. The dataset is ten times the size of corresponding verb dataset, roughly approxi-

mating the fact that speaker's vocabularies contain more non-verbs than verbs, and that small-*N* learners will have smaller vocabularies. The validation dataset includes only word types that do not appear in training. For each class, the initial model is further fine-tuned for another 50 epochs on the word types in the class (most classes do not have enough types to do early stopping on a separate validation set.) As before, verbs are represented with the present tense/lemma form from the CMU pronunciation dictionary.

We leave more comprehensive testing of the full model for the next section, but we first confirm that the model is working as a similarity metric, assigning higher probability to members of its class than to members of other classes. Note that during training there is no pressure for the model to be discriminative between classes: each class language model is trained on positive evidence only, and never sees negative evidence in the form of examples from other classes.

As an initial check that the language models assign more probability to verbs within their class over verbs outside their class, we separately test held-out verb forms from the two classes that are sufficiently large: 'Regular' and 'Vowel Change' (VC). For each class, we train a language model as described above, on the approximately 100 verb types found when sampling 200 tokens (systematically slightly fewer types for VC than Regular). We separately take 30 unseen verb forms from each class as a test set. We can then compare within-class probability (testing on the same verb class as the model was trained on) against across-class probability (the probability assigned by the class-specific model to words from a different class). Over ten repetitions, the Regular language models on average assigned probability to Regular test verbs that was 1.5 times as high as that assigned to VC test items; the difference was even higher for VC language models, with within-class (but unseen) VC verbs assigned 2.9 times higher probabilities than outside-class Regular verbs. The fact that the VC class has fewer types compared to Regular seems to have led to a tighter distribution, but it is striking that even the quite diverse Regular class assigns lower probability to the (relatively short and common) forms in the VC class than unseen Regulars.

The Full Model

In this section we evaluate the complete model, where we use the character language model from the previous section as the distribution over forms $P(f|c)$ within the base distribution of the PYP model introduced earlier. This model incorporates both class frequency biases and biases towards phonological similarity: the model prefers high-(type)-frequency classes, due to $P(c)$, but this preference can be overruled by similarity, if another class assigns higher probability to the particular form being assessed in $P(f|c)$.

We test the full model on the set of 15 nonce words used in Exp 2 in Cuskley et al. (2015), in order to evaluate whether it is sensitive to the same factors (frequency and phonology) as human participants. We also again examine the effect of learning from different amounts of data. Our evaluation assesses two hypotheses, corresponding to the factors of frequency and similarity:

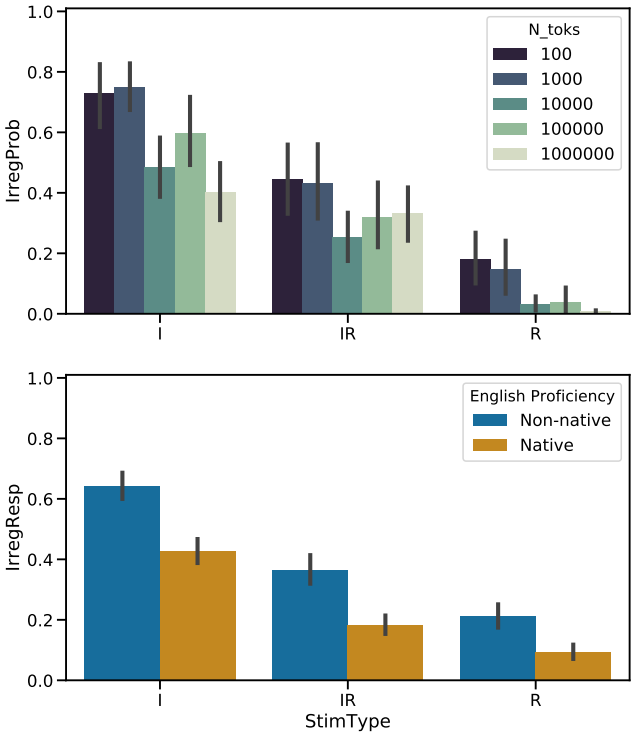

Figure 3: A comparison of the model's predictions of irregular classes (top) to the irregularisation rates of participants in Exp 2 in Cuskley et al. (2015) for the same stimuli (bottom). The bottom sub-figure replicates Figure 6 in Cuskley et al. (2015), in which all non-regular forms are counted together. The different models correspond to different dataset sizes: values of N_toks are the number of tokens sampled for training. Model results show an average of ten runs for each setting; error bars show 95% CIs. StimType corresponds to the verb type of the nearest neighbour of the novel verb: `I`: Irrregular, `R`: Regular, `IR`: novel verb is equidistant between a regular and irregular verb.

- That differences between full models trained on small amounts of data and larger amounts of data will qualitatively correspond to the differences between non-native and native speakers;
- That the full model will be sensitive to differences in phonological similarity of test words, in the same way as participants were.

For each word, we calculate $P(c|f) = \frac{P(c)P(f|c)}{\sum_{c'} P(c')P(f|c')}$ using the posterior predictive base distributions from the PYP. We compare the total model irregularization probability (i.e., the probability of assigning a given form to a non-Reg class) against irregularization rates in Exp 2 in Cuskley et al. (2015),

in Figure 3. The test forms are grouped into one of three classes, depending on their phonological nearest neighbour: `I` forms are closest to an existing irregular, `R` forms are closest to a regular, and `IR` are midway between a regular and irregular form. Nearest neighbours were matched for frequency; see Cuskley et al. (2015) for details on stimuli construction.

Figure 3 shows that firstly, all models are sensitive to phonological similarity in the same way as the human participants were: forms that are phonologically closer to irregulars assigned more probability of belonging to irregular classes, while test forms more similar to existing regulars have higher probability of belonging to the Reg class (i.e., lower probability of being irregular).

Secondly, like non-native speakers, models trained on a smaller sample of data assign relatively higher probability to irregular classes, compared to models trained on larger datasets. This is the effect of $P(c)$: the distribution over classes is more weighted towards irregular forms in a small-data model (recall Figure 2) for each of the stimuli types.

Conceivably, phonological similarity and dataset size could also interact, since smaller models estimate phonological similarity on a different vocabulary than larger models, containing fewer low-frequency forms. If these low-frequency forms are systematically distributed (e.g., they are likely to be longer) then the presence or absence of these forms in the dataset could affect phonological similarity judgements. We found that larger models generally judged the stimuli to have lower phonological similarity to the Regular class than smaller models, that is, $P(f|Reg)$ was larger in models trained on smaller dataset than in models trained on more word forms. However, differences in phonological similarity were very small compared to differences in cluster probability across models trained on different dataset sizes.

Conclusion

Frequency and phonological similarity are the key factors in morphological generalisation to new forms, and the model presented in this paper captures them in a single framework. In doing so, it is able to capture the three-fold interaction between language proficiency (or exposure), class frequency, and phonological similarity to existing verbs that was present in the data from Cuskley et al. (2015). Remember that this model was not fit to participant productions, and even the input data (sampled from a large corpus) is a very rough approximation of what learners of English are likely to be exposed to. Nevertheless, we found that the model displayed the same pattern of behaviour as participants, indicating both that the interacting phenomena are quite robust and that model can reliably capture them.

We find that our small-data learners behave similarly to non-native speakers, in that they are more likely (than large-data models and native speakers) to generate irregular forms, due to the proportionately higher frequency of irregular verb types in their limited input. A question remains of whether this model can also explain the behaviour of native-speaker child learners, who famously go through a phase of regularising irregular forms (e.g., *'goed'*, Marcus et al., 1992). Investigating child acquisition would require training the model on varying amounts of child-directed speech input and evaluating against age-appropriate nonce-word productions. Note that this model is able to assign a known word form to a class other than the one seen in training (i.e., by generating a new entry in the PYP and assigning higher probability to a different class than the training class); this would allow for over-regularisation behaviour.

The combination of a non-parametric process and neural models presented here is powerful, since it allows us to make use of the different strengths of each methodology. However, it also comes with limitations. It is currently not tractable to infer the set of verb classes, i.e., to do unsupervised clustering of the verbs, in the current framework, which would also require additional learning algorithms (e.g. to sample class membership). Moreover the likelihood $P(f|c)$ given by the neural network is a point estimate dependent on network parameters, instead of the posterior marginalised over parameters that a full Bayesian approach would require. Bayesian neural networks could remedy this issue in the future (e.g. Fortunato, Blundell, & Vinyals, 2017).

This model is structured to support multiple productive verb classes. Since irregularization rates of novel verbs in English are non-zero (and occasionally quite high) in experimental settings, a model should be able to predict productive use of non-regular verb inflection. However, it remains a fact that the English past tense is extremely biased towards a single regular form, and the set of irregular verbs is limited. Many other languages have much richer inflection systems, with multiple productive high-frequency classes. In future work we plan to test the model on data from such a language (e.g. Polish noun inflection, Dąbrowska, 2008), in order to test the cross-linguistic viability of the model's underlying assumptions about the importance of frequency and phonological similarity.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 681942).

References

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

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

Buntine, W., & Hutter, M. (2010). A Bayesian view of the Poisson-Dirichlet process.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). John Benjamins Publishing.

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.

Bybee, J., & Moder, C. L. (1983). Morphological categories as natural categories. *Language*, *59*(2), 251-270.

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the association for computational linguistics.*

Cuskley, C., Colaiori, F., Castellano, C., Loreto, V., Pugliese, M., & Tria, F. (2015). The adoption of linguistic rules in native and non-native speakers: Evidence from a wug task. *Journal of Memory and Language*, *84*, 205–223.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PloS One*, *9*(8), e102882.

Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, *58*(4), 931–951. doi: 10.1016/j.jml.2007.11.005

Davies, M. (2010). *The corpus of Historical American English (COHA): 400 million words, 1810-2009.* Retrieved from `https://www.english-corpora.org/coha/`

Fortunato, M., Blundell, C., & Vinyals, O. (2017). Bayesian recurrent neural networks.. doi: arXiv:1704.02798

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, *12(Jul)*, 2335–2382.

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), i. doi: 10.2307/1166115

Newberry, M. G., Ahern, C. A., Clark, R., & Plotkin, J. B. (2017). Detecting evolutionary forces in language change. *Nature*, *551*(7679), 223.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*(1), 1–56. doi: 10.1080/01690969308406948

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of english verbs. In J. L. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216–271). MIT Press.

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190–1228.

Wu, S., Cotterell, R., & O'Donnell, T. J. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the association for computational linguistics.*