Rules, Analogy, and Social Factors codetermine past-tense formation patterns in English

Péter Rácz
New Zealand Institute of Language Brain and Behaviour, University of Canterbury
peter.racz@

Clay Beckner
New Zealand Institute of Language Brain and Behaviour, University of Canterbury
clayton.beckner@canterbury.ac.nz

Jennifer B. Hay
New Zealand Institute of Language Brain and Behaviour, University of Canterbury
jen.hay@

Janet B. Pierrehumbert
Department of Linguistics / NICO Northwestern University.
New Zealand Institute of Language Brain and Behaviour, University of Canterbury
jbp@northwestern.edu

Abstract

We investigate past-tense formation preferences for five irregular English verb classes. We gathered data on a large scale using a nonce probe study implemented on Amazon Mechanical Turk. We compare a Minimal Generalization Learner (which infers stochastic rules) with a Generalized Context Model (which evaluates new items via analogy with existing items) as models of participant choices. Overall, the GCM is a better predictor, but the the MGL provides some additional predictive power. Because variation across speakers is greater than variation across items, we also explore individual-level factors as predictors. Females exhibited significantly more categorical choices than males, a finding that can be related to results in sociolinguistics.

1 Introduction

In this report, we present a psycholinguistic study of English past tense categories, using a nonce-probe experiment implemented on Amazon Mechanical Turk. The English past tense has been a testing-ground for a wide range of theories and predictions in psycholinguistics, including the processes of acquisition, the nature of lexical representation, and the representation of inflectional patterns as rules or as generalizations over specific items (Bybee and Slobin, 1982a; Rumelhart and McClelland, 1985; McClelland and Patterson, 2002; Albright and Hayes, 2003).

The present study investigates the factors influencing patterns of preferred past tense forms for particular verb classes. English past tenses are not merely a memorized list, but rather, verb categories can shrink, or expand to include new items. In everyday speech, there is evidence of ongoing influences from multiple verb classes, as verbs exhibit variation and slowly shift in their usage (*dived* vs. *dove*, *sneaked* vs. *snuck*), (Haber, 1976; Bybee and Moder, 1983).

Given that speakers can adapt their verbal categories to new situations, what is the best representation for the relevant morphological generalizations? In analogical models, the focus is on existing stored items in memory. The acceptability of a candidate past tense formation pattern for a particular candidate item is determined by patterns of similarity to stored items. Morphological innovation and productivity arises from generalizations over existing forms in the lexicon. To account for a speech error such as *glew* as the past tense of *glow* (Bybee and Slobin, 1982a), an analogical explanation would highlight the close similarity between *glow* and the present tense forms *blow, throw, know*, which provide the basis for an analogy with the past forms *blew, threw, knew*. Of particular interest is the Generalized Context Model (GCM) (Nosofsky, 1990; Albright and Hayes, 2003), an analogical model which assesses a category's suitability to a target item on the basis of feature-based similarities summed over category items, in addition to the category's size. It has already been successfully applied to model regular and irregular patterns in Arabic morphology (Dawdy-Hesterberg and Pierrehumbert, 2014).

Rule-based approaches propose more abstract representations of generalizations. Originally proposed to handle broadly applicable default patterns, (such as 'add *-ed* to express the past tense'), rule-based approaches have recently been extended to incorporate multiple stochastic rules. Albright and Hayes (2003) assign scores to morphological rules by training a Minimal Generalization Learner (MGL) over a dataset, an algorithm that iterates over pairs of words in the lexicon, hypothesizing generalizations conservatively on the basis of any phonological features that are

Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 55–63,
Baltimore, Maryland USA, June 27 2014.

shared across the words. A rule is scored according to how many items it applies to in the lexicon, weighted against cases in which the inferred phonological context is present but the rule fails to apply. The resulting system consists of a catalog of weighted natural class-based generalizations which compete with one another, and which are more or less likely to apply in various phonological contexts (for regular as well as irregular verbs). Albright and Hayes argue that the MGL outperforms the GCM in predicting participant behavior in a nonce-verb production task they conducted.

2 Experiment

We collected a large amount of data on irregular past tense formation in English with a nonce probe test, a classic method for exploring the productivity of inflectional morphology (Berko, 1958). Earlier studies used 30 or fewer participants per condition (Bybee and Slobin, 1982a; Albright and Hayes, 2003). By using Amazon Mechanical Turk, a burgeoning forum for psycholinguistic research (Munro et al., 2010), we were able to recruit a large number of participants and explore the role of individual-level factors in the choice of morphological patterns. Moreover, we tested participant preferences across a large dataset (316 nonce verbs) based on broad phonological sampling within verb classes, allowing for repeated trials across similar items for each participant.

Participants in our online study were presented with a forced choice task in which they had to pick either the regular or the irregular past tense form for an English nonce verb, presented in a carrier sentence. This was followed by a vocabulary task in which participants had to rate the familiarity of English nouns.

2.1 Stimuli

We set up five categories of irregular past tense formation based on phonological form of the present tense verb, and its corresponding candidate tense past forms. Each category exhibits phonological variability within the category, while also allowing for a specific phonological description. We avoided 'miscellaneous' verb classes, as well as wholly idiosyncratic patterns (such as *go–went*). Moreover, we are particularly interested in morphological classes which are known to display some indeterminacy (Haber, 1976), i.e., those classes which display some regular/irregular variation (*dived* vs. *dove*), due to the ready availability of multiple generalizations. The literature contains various taxonomies of English irregular verb classes (Bybee and Slobin, 1982a), but our current classification mostly represents a subset of the detailed verb classes outlined by Moder (1992).

The five categories of interest are as follows.

- **SANG**. Verbs that form the past tense with a vowel change from [ɪ] to [æ] (e.g. *sing–sang*, *sink–sank*, *swim–swam*).
- **BURNT**. Verbs that form the past tense by adding a [t], with no change in the stem vowel (e.g. *burn–burnt*, *spill–spilt*, *learn–learnt*). These items constitute a distinct set from regular English pasts such as *boss–bossed* which are articulated with a [t] allomorph, insofar as the **burnt** verb bases actually end in a voiced consonant but are nonetheless affixed with a voiceless stop.
- **KEPT**. Verbs that form the past tense by adding a final [t] and changing the stem vowel from [i] to [ɛ] (e.g. *keep–kept*, *mean–meant*, *feel–felt*).
- **DROVE**. Verbs that form the past tense with a vowel change from [aɪ] or [i] to [oʊ] (e.g. *drive–drove*, *weave–wove*, *ride–rode*).
- **CUT**. No-change past tense verbs, that is, verbs the past tense form of which is identical to their present tense form. (e.g. *cut–cut*, *cost–cost*, *hurt–hurt*). Verb bases in this class end in sounds that are already associated with the English past tense ([t] or [d]) (Bybee and Slobin, 1982a), although the nonce verb bases in the present study all end in [t].

We generated nonce verb forms by combining the category-specific restrictions spelled out above on the stem with a set of syllable onsets that occur in English. Using CELEX (Baayen et al., 1993), we then filtered the orthographic and phonetic transcriptions of the nonce stems, as well as the resulting past tense forms, to exclude real English words. Two native speakers checked the final list to remove additional real words that were not filtered out via the CELEX database (e.g., slang and informal terms). All our verb forms were monosyllabic– as are almost all English irregular verbs in general. The method used to generate the

stimuli means that some nonce forms looked more similar to real English verbs than others. This way we can tell whether similarities to a single form will strongly influence people's behavior in the case where the nonce form is highly similar to a single real form.

The **sang** and **cut** categories consist of 60 forms. The **burnt** category has 40, **drove** has 76, and **kept** has 80. The total number of nonce verbs is 316.

2.2 Setup

The experiment consisted of a forced choice task, in which participants had to pick a regular or irregular past tense form for each verb. Verbs were presented one at a time, visually, in a carrier sentence of the form 'I really like to VERB. Yesterday, I .'. Two buttons were presented under the carrier sentence, one with the regular past tense, adding *-ed*, and one with the irregular past tense. The irregular past tense was always the dominant pattern for the category. (So, for **cut**, it was identical to the present tense, etc.) The order of the two buttons was randomized for each verb. Each verb was presented once and the order of verbs was randomized for each participant.

The experiment was appended by a word familiarity rating task. The rating task was based on Frisch and Brea-Spahn (2010). It consisted of 50 nouns of varying familiarity, as well as 10 extremely common nouns and 10 nonce words. The 70 words were presented in a random order. The participant had to select, on a scale of 1-5, how familiar the given word was. Incorrect answers to the extremely common nouns and the nonce words were used as an exclusion criterion. Answers for the other items were used as an index of vocabulary level, which is predicted to affect morphological choices in both the GCM and MGL models.

2.3 Participants

111 people took part in the experiment on Amazon Mechanical Turk during the course of two days. 51 were women, 60 were men, and 1 did not specify. The age range of the participants was 20-65, and the mean age was 34. All participants were native speakers of American English. Participants were paid three dollars. We excluded ten participants from the analysis because they failed to differentiate familiar from unfamiliar words in the vocabulary test.

Category	Experiment	Nonce Examples
drove	0.52	skride: skrode, skrided
sang	0.58	sking: skang, skinged
kept	0.59	skeep: skept, skeeped
burnt	0.67	skurn: skurnt, skurned
cut	0.83	skast: skast, skasted

Table 1: Categories and mean regularization ratings.

2.4 Results

The nonce verb categories have different rates of regular vs. irregular usage, as can be seen in Table 1. The *Experiment* column shows the mean regularization rates of the categories in our experiment. The **drove** class was regularized the least often, and the **cut** class the most often, with a considerable difference between the two.

The trends across verb classes are similar to those of Moder's (1992) nonce experiment. Note in particular the high regularization rate (83%) of the no-change class of verbs (**cut**). A search of CELEX indicates that no-change [t]-final verbs are quite widespread in English, represented by more than 30 types. Yet based on nonce responses, the English no-change pattern is not very prone to being applied to novel items. This finding matches observations by Bybee (1982b) that the no-change verb class has been on the decline in English, as evident from increasing regularization. One noteworthy feature of the **cut**-type verbs is that the phonological shape of the base is a quite unreliable indicator of verb class. That is to say, there are many [t]- final verb stems which typically take the regular -ed suffix (e.g., *gritted, salted, blasted*, and these provide counterexamples to the no-change pattern (cf. Moder (1992) on cue validity).

We fit a simple stepwise logistic mixed-effects regression model to the results with a maximal random effects structure, using regularization of individual verb form (yes or no) as an outcome variable and category as predictor. This model confirms the general finding that there is significant variation across the verb classes. (Significance values reported are based on difference with the **sang** class.) The **cut** class shows the highest rate of regularization ($p<0.001$), followed by the **burnt** class ($p<0.01$). It is followed by the **sang** and **kept** classes (these two do not differ significantly). The **drove** class shows the lowest rate of regularization ($p<0.01$).

Participant gender, age, and vocabulary size are not significant predictors of regularization in the simple logistic mixed effects model. However an examination of the data (Figure 1) reveals that for each verb class, variation across subjects is considerably greater than variation across items. This observation suggests that individual traits may play a role in morphological choices in a way that the simple model fails to capture. We will return to this issue after presenting the GCM and MGL model fits, and will find in the end that gender does affect response patterns.

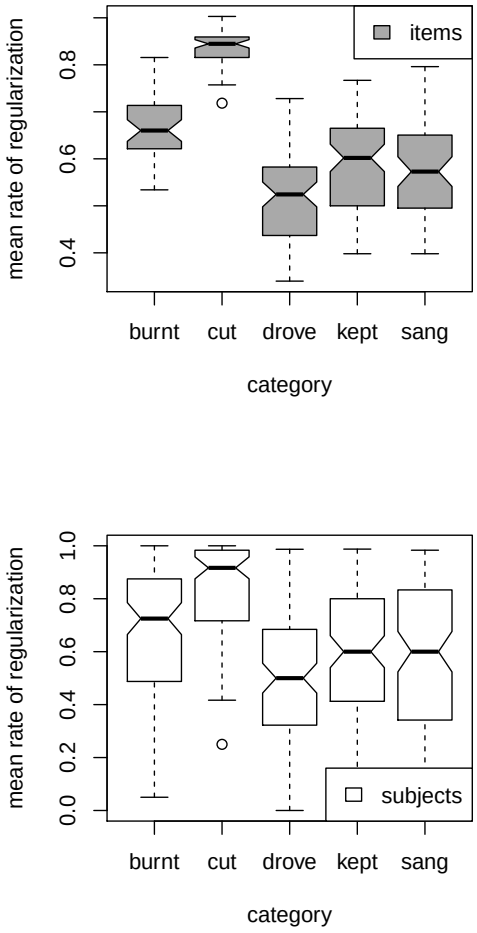

Figure 1: Across-item variation in regularization rates across category (above). Across-subject variation in regularization rates across category (below).

3 Algorithmic Learning Models

We now turn our attention from the baseline effects of category variables, to investigate the predictions of particular algorithmic learning models that provide alternate representations for generalizations on the basis of similarity. Our analyses focus on the predictions of the Minimal Generalization Learner and the Generalized Context Model (Albright and Hayes, 2003; Nosofsky, 1990).

3.1 The two models

The Minimal Generalization Learner (MGL) (Albright and Hayes, 2002; Albright and Hayes, 2003) is an algorithm for inferring stochastic morphophonological generalizations over a set of training items (e.g., paired present and past tense forms). For each pair of items in the lexicon, the learner maximally aligns wordforms and analyzes shared phonetic features, thereby merging word-specific rules (*ring/rang* and *stink/stank*) into rules that express the most general applicable environment: [ɪ] → [æ] / [+coronal, + cont] ___ [ŋ].

Each rule inferred in this way is then further generalized on the basis of more comparisons; for instance, taking note of *swim/swam* expands the [ɪ] → [æ] rule to specify that it occurs before all [+nasal] consonants. The algorithm thus infers a set of natural-class based generalizations, which are weighted by comparing the number of hits for the past tense pattern (*ring/rang, drink/drank, sing/sang, stink/stank, swim/swam*, etc.) divided by the number of cases in which the alternation fails to apply although it *could* apply (thus tallying exceptions such as *think* and *blink*). This appproach favors generalizations that cover many cases, but penalizes those that are too broad because their phonetic environments encompass many exceptions. The MGL reliability metric is further adjusted to a confidence score, in which generalizations that apply to a smaller number of word types are penalized.

Note that the MGL algorithm automatically groups together items on the basis of shared phonological properties; thus, monosyllabic verbs are most likely to form strong generalizations with other monosyllabic verbs. Attempts to merge diverse wordforms under a single generalization would be more likely to incur penalties (i.e., exceptions). This feature of the MGL is important for comparing with the methods of the GCM (see below). Both algorithms allow for category-

specific similarities to play a role.

The Minimal Generalization Learner is implemented here from materials made available by Albright and Hayes (2003), including their Segmental Similarity Calculator based on Frisch et al. (2004). The MGL is trained on regular and irregular English verbs with a minimum frequency cutoff of 10 in COBUILD (Baayen et al., 1993), and excluding prefixed verb forms, thus encompassing 4253 past/present verb transcriptions. The MGL is implemented here with its default settings, which includes a lower 75% confidence interval for purposes of adjusting the reliability score.

The Generalized Context Model (GCM) is an instance-based model of categorization. To assign category membership to a novel instance, it first calculates its similarity to instances in preexisting categories. Then, it selects the category with members that are most similar to the novel instance (Nosofsky, 1990). Our implementation of the GCM has three notable aspects to it.

First, we used the GCM to categorize our nonce verb stimuli, basing the categories on real English verb types extracted from CELEX (as with the MGL). Second, we used the same segmental similarity calculator developed and used by Albright and Hayes and used by the Minimal Generalization Learner to calculate the similarity of phonetically transcribed word forms to each other, so that we could take the phonetic similarity of speech sounds into account instead of calculating similarity between word forms based on edit distance alone. We did not weight parts of the word forms differently, because there is evidence that although past tense formation in English is predominantly driven by similarities in word endings, onsets also play a role. (cf. the predominance of s+stop onsets in irregular verbs forming the past tense with a vowel change, e.g. *sing*, *sink*, etc.) (Bybee and Moder, 1983).

Third, our implementation of the GCM reflected the structure of the task. Recall from Section 2 that participants were presented with the stems of the nonce verbs in a sequence and had to pick either a regular or an irregular past tense form for them. The irregular past tense form was predetermined by category, so that, for a given verb, the participants could only choose between the regular past tense form or the irregular past tense form we assigned to the verb. (So, for instance, for *spling*, they could choose either *splinged* or *splang*, but not *splung* or *splingt*, etc.) For a given category (such as *sang* verbs), the GCM had a choice between two sets. The irregular set consisted of verb types in CELEX that form their past tense according to the pattern captured by the category (such as an [ɪ]–[æ] alternation). The regular set consisted of verb types that have a stem that matches the category (such as 'monosyllabic and stem vowel [ɪ]') but have a regular past tense. The model calculated the similarity of a given nonce verb to these two sets (depending on its category). In this paper, we report on category weights assigned to the *regular* category, which are comparable with both the results of the Minimal Generalization Learner and the rate of regularization in our experiment. We only used monosyllabic verbs in identifying relevant matches, for regular as well as irregular items.

Values reported here were generated with no frequency cutoff. Alternate runs with the frequency threshold enforced produce no change in the model. The model is run with the default parameter settings of $s = 0.3$, $p = 1$ with respect to calculating the weighted similarities between items. When p is set to 1, as here, the similarity function is exponential, rather than Gaussian. The weighting parameter s controls the tradeoff in the relative importance of the size of the verb category (the 'gang size') vs. the amount of similarity (measured via edit distance between phonological forms) (Nosofsky, 1990; Nakisa et al., 2001; Albright and Hayes, 2003; Dawdy-Hesterberg and Pierrehumbert, 2014).

Figure 2 shows three plots. The first one depicts the relationship between the predictions of the GCM (regular category weight) and experimental ratings (mean participant regularization) for individual verb types used in the experiment. The Spearman rank correlation is highly significant ($rho = 0.497$, $p < 0.001$). The second one depicts the relationship between the MGL model predictions (reliability rating of the regular form) and mean participant regularization in the experiment. The Spearman rank correlation between these variables is highly significant ($rho = 0.393$, $p < 0.001$). The predictions of the two models are z-scored to allow for comparability. The third plot shows the relationship between the predictions of the GCM and the MGL for individual verb types in the experiment. The Spearman rank correlation between these variables is highly significant

CATEGORY	GCM	MGL
SANG	**0.65**	0.55
CUT	**0.18**	-0.19
DROVE	0.37	**0.64**
KEPT	**0.52**	0.18
BURNT	**0.48**	0.24
ALL	***0.5***	0.39

Table 2: Correlations table: Spearman's rank correlations between mean regularization in the experiment and the predictions of the two models

(rho $= 0.347$, p < 0.001), but the correlation is far from perfect. Comparing the overall correlations and patterns in Figure 2, it appears that the GCM is doing a better job of predicting the variation across items than the MGL is. We now turn to an examination of the predictions within our verb classes.

3.2 Model comparisons within verb class

Table 2 shows Spearman rank correlations between mean regularization in the experiment and the predictions of the two models for the five verb categories. Overall, GCM does a better job. The no-change (*cut*) verb class is especially illustrative of the differences between the two models. Note that the MGL is negatively correlated with our experimental data for this category. As noted above, this verb class appears to be strikingly nonproductive; participants display a strong preference for regularizing a wide range of t-final forms. The MGL underestimates the regularization of nonce verbs that resemble *cut* and *hit*, while overestimating the regularization of forms like *vurt, slurt, plurt*. The no-change irregular form of such verbs must be modeled on a pattern with a sole English exemplar (*hurt–hurt*), and the Minimal Generalization model (in contrast with the GCM) is swayed very little in such cases. This is one of several cases where the GCM predicts subject preferences better than the MGL does, seemingly because the irregular form requires modeling a response on a sole exemplar.

There is one verb category where the MGL outperforms the GCM: the **drove** class. Here, the MGL does especially well because it makes an accurate prediction about one subcategory of items: nonce verbs like *quine* and *sline* are regularized by participants (*quined,slined*) more often than other members of the **drove** class. Here, it seems that

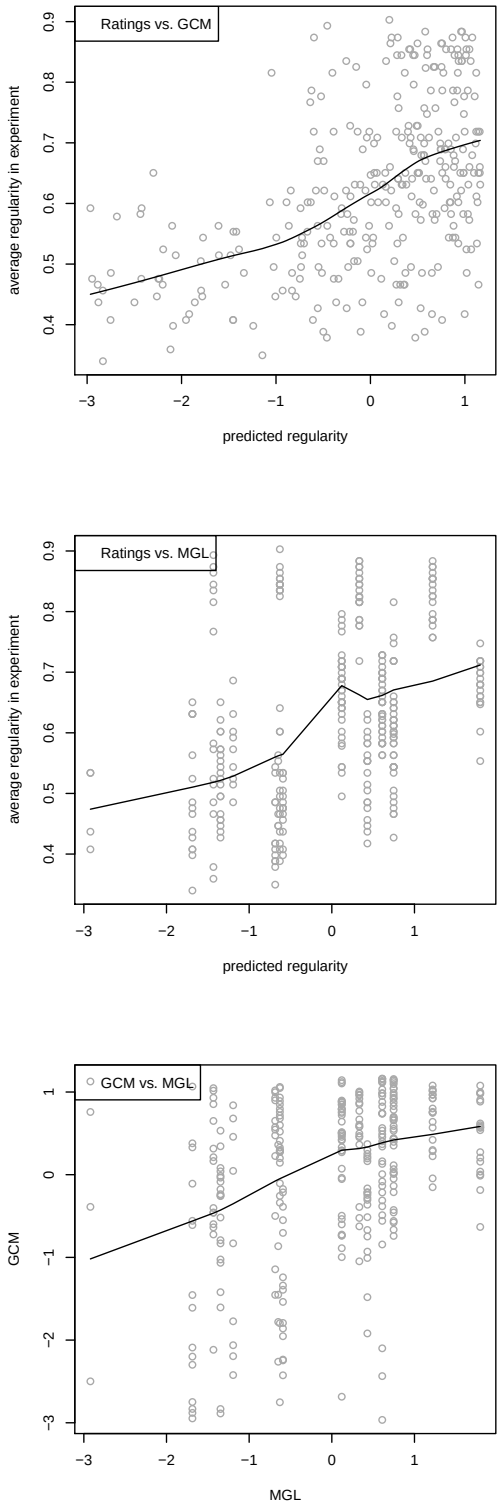

Figure 2: Above: experimental ratings versus GCM predictions. Middle: experimental ratings versus MGL predictions. Below: MGL predictions vs. GCM predictions. (With lowess lines added.)

the irregular past would need to be modeled on one closely-related English item (*shine–shone*), but similar English verbs offer many exceptions to any abstract generalization (*line–lined, mine–mined, whine–whined*, not to mention the transitive verb *shine–shined*). Such a situation causes the MGL to correctly classify all *-ine* final verbs as highly prone to regularization, because *-ine/-one* type irregulars are all dispreferred in the experiment. However, the GCM makes a wide range of predictions for these stimuli on the basis of different segmental similarities with training items (e.g., based on the syllable onsets).

On the whole, comparing the two models on the verb classes suggests that analogy to individual instances is a better approximation of the behavior of our subjects than recourse to abstract generalizations. It is true, however, that both the GCM and the MGL each only explain a part of the observed variance. In order to test whether the two models contribute differently to explaining participant behavior in our dataset, we fitted a simple stepwise logistic mixed-effects regression model on the results with maximal random effects structure, using regularization on the individual verb form (yes or no) as an outcome variable. Instead of verb category, we used the GCM and the MGL regularization rates as predictors. Both predictors are significant. An analysis of variance test reveals that the regression model that includes the predictions of both categorization models provides a significantly better fit than the models including either alone. We tested nonlinear effects of MGL and GCM, using restricted cubic splines, but nonlinearity did not significantly improve the model. Participant age and gender are not significant. Vocabulary size explains some variation, though does not quite meet the threshold of .05 for significance. The interaction of GCM predictions and participant gender, however, is significant. The model coefficients can be seen in Table 3.

Predictor	b	z	sig.
(Intercept)	0.71	4.5	***
MGL	12.4	3.38	***
GCM	1.11	5.05	***
gender (male)	-0.02	-0.07	(n.s.)
vocabulary	-0.25	-1.77	.
GCM : gender (male)	0.47	2.12	*

Table 3: Effects of rules vs. analogy in the regression model

3.3 Individual-level factors

As both MGL and GCM make reference to existing patterns in the lexicon, we hypothesized that the precise size and contents of an individual's vocabulary is likely to produce individual variation in terms of the lexical support available for certain patterns. Individuals with higher vocabulary scores may be more likely to have robust stored instances of irregular, lower frequency, minority past tense patterns. We might therefore predict that they are more accepting of irregular realizations. This is, to some degree, confirmed by the strength of vocabulary as a predictor of regularization in our final model. A potential interaction of vocabulary size and the two models of past tense formation is that these models likely have different predictions when trained on vocabulary sets of various sizes – this is a clear direction of future research.

We also tested the effects of participant gender, as women have been reported to be more biased towards more standard language (Labov, 2001). This would mean that conformity to speech community standards in whether a form is irregular or regular (essentially, getting it 'right') could be highly valued by women. Consistent with this observation, we find a significant interaction between GCM and participant gender. Females show a steeper slope for the GCM than the males do. When there is low analogical support for regularization, females have a tendency to prefer irregular forms more than males do, but this difference is reversed for items where the GCM provides strong support for the regular. In that case, females prefer regular forms more than males do. To put it differently, females categorize the verb forms more in our dataset than the males do.

It is interesting to note that our results differ from Hartshorne and Ullman's (2006) child data on real English verbs. They found more overregularization for girls than for boys. The mechanism they suggest relies on girls having more precocious verbal ability, as is commonly reported. These results may seem hard to reconcile, since the adult women in our study did not regularize more than men (there was no significant overall effect of gender), nor did they have larger vocabularies, as measured by our vocabulary inventory. However, they are compatible if we assume that

the real verbal lexicon is rather well learned by adulthood (as reflected in the weakness of vocabulary level as a statistical predictor in our model) and that the gender difference we observed taps the social factors mentioned by Labov, which are learned gradually during childhood and adolescence.

4 Conclusions

Our results suggest that both the GCM and MGL models contribute important insights into factors underpinning perceived wellformedness. Individuals are heavily influenced by the combined analogical force of existing lexical forms. They generalize over items. However, they also, it appears, generalize over these generalizations - forming more abstract 'rules' or associations that operate in parallel with the token-based analogical processes. While this seems to be the interpretation that is pointed to by this current data set, verification of the joint role of these types of processes clearly requires a lot more explicit testing in different and varied data sets, including real verbs in addition to nonce forms. Recent models in phonological processing and speech perception certainly point to a hybrid model, in which instance-based processing and reasoning sits alongside more abstract structures, and in which both types of processes may be jointly operative – with the balance affected by many factors including the particular nature of the task at hand (Pierrehumbert, 2006). Indeed, we would predict that it should be possible to design morphological tasks which more readily tap into purely analogical processes, or into more abstract generalizations.

Acknowledgments

This project was made possible through a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. Hay and Beckner were also supported by a Rutherford Discovery Fellowship awarded to Hay. The authors would like to thank Adam Albright, Patrick LaShell, Chun Liang Chan, and Lisa Garnard Dawdy-Hesterberg. All faults remain ours.

References

Adam Albright and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pages 58–69. Association for Computational Linguistics.

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

R Harald Baayen, Richard Piepenbrock, and Rijn van H. 1993. The CELEX lexical data base on CD-ROM.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14:150–177.

Joan L Bybee and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language*, pages 251–270.

Joan L Bybee and Dan I Slobin. 1982a. Rules and schemas in the development and use of the English past tense. *Language*, pages 265–289.

Joan L Bybee and Dan I Slobin. 1982b. Why small children cannot change language on their own: Suggestions from the English past tense. In *Papers from the 5th international conference on historical linguistics*, volume 21.

Lisa Garnand Dawdy-Hesterberg and Janet B Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience*, (ahead-of-print):1–15.

Stefan A Frisch and Maria R Brea-Spahn. 2010. Metalinguistic judgments of phonotactics by monolinguals and bilinguals. *Laboratory Phonology*, 1(2):345–360.

Stefan Frisch, Michael Broe, and Janet Pierrehumbert. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22:179–228.

Lyn R Haber. 1976. Leaped and leapt: a theoretical account of linguistic variation. *Foundations of Language*, pages 211–238.

Joshua K Hartshorne and Michael T Ullman. 2006. Why girls say holded more than boys. *Developmental Science*, 9(1):21–32.

William Labov. 2001. *Principles of linguistic change Volume 2: Social factors*. Blackwell.

James L McClelland and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11):465–472.

Carol Lynn Moder. 1992. *Productivity and categorization in morphological classes*. Ph.D. thesis, State University of New York at Buffalo.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

Ramin C. Nakisa, Kim Plunkett, and Ulrike Hahn. 2001. A cross-linguistic comparison of single and dual-route models of inflectional morphology. *Peter Broeder, & Jaap Murre, Models of Language Acquisition: Inductive and Deductive Approaches*, pages 201–222.

Robert M Nosofsky. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34(4):393–418.

Janet B Pierrehumbert. 2006. The next toolkit. *Journal of Phonetics*, 34(4):516–530.

David E Rumelhart and James L McClelland. 1985. *On learning the past tenses of English verbs*. Institute for Cognitive Science, University of California, San Diego.