Supplementary Information – Morphological convergence as on-line lexical analogy

Péter Rácz, Clay Beckner, Jen Hay, Janet B./ Pierrehumbert

January 5, 2020

In this supplementary information, we detail (1) the structure of our nonce verb stimuli, the setup of (2) the Generalized Context Model (GCM) and (3) the Minimal Generalization Learner (MGL), and (4) how these models compare. We illustrate the models using fits on our *baseline* data.

We also discuss (5) the residualization method used to compare them, (6) how well they fit our ESP test, and (7) the regression methods used.

1 Regular and irregular verbs in English

Four irregular verb classes were defined for our stimuli, based on the vowel alternation and affixation processes that apply to the stem:

- DROVE ([aɪ]/[i] → [oʊ])
- SANG ([ɪ] → [æ])
- KEPT ([i] → [ɛ]Ct)
- BURNT ([ɜ]/[ɛ]/[ɪ] → [ɜ]/[ɛ]/[ɪ]Ct)

In the baseline experiment, participants are also tested on monosyllabic verbs that do not change in the irregular past tense (e.g. *cut*, *hit*). These forms are strong outliers and are not reported in the data.

The stimuli for the ESP experiment, 156 nonce verbs across four classes, are shown in Table 1 in the main text. Our formal criteria are based on the behavior of existing verbs in English and their categorization by Bybee & Slobin (1982), Moder (1992), and Albright & Hayes (2003). We made some adjustments to their categories. For instance, our DROVE class is a generalized version of Moder's class 4, described using the [aɪ]→ [oʊ] alternation but also including *weave*. (In contrast, Albright & Hayes restrict this class to this alternation.) The verb classes could be defined in a number of ways and still be mostly consistent with the work of Moder, for instance. We have trialed a set of possible small changes and found that they have no major effect on our overall results.

2 Implementation of the Generalized Context Model

2.1 Outline

Our implementation of the GCM evaluates the competition between two categories, *regular* and *irregular*, for each nonce verb base form. The framework of Nosofsky (1990) is adapted to morphophonology by using a segmental similarity calculation based on natural classes (Frisch et al., 2004). The same treatment of segmental similarity is used in the implementations of the GCM in Albright & Hayes (2003) and Dawdy-Hesterberg & Pierrehumbert (2014). We build on Dawdy-Hesterberg & Pierrehumbert (2014) in that we define our categories based on formal similarity.

2.2 Training data

Participants are presented with a sequence of nonce verb base forms, and have to pick either a regular or an irregular past tense form for each. The irregular past tense form is pre-determined by the class for the stem, so that, for a given verb, the participants can only choose between the regular past tense form and the irregular past tense form we assigned to the verb. (So, for instance, for *splive*, a verb in the DROVE class, they can choose either *splived* or *splove*, but not *splift* or *sploven*, etc.) For a given class (such as DROVE verbs), the GCM has a choice between two sets of verb types.

The irregular set consists of verb types in Celex that form their past tense according to the pattern captured by the class (such as an {[aɪ],[i]} → [oʊ] alternation).

The regular set consists of verb types that have base forms that are similar to these irregular forms but have a regular *-ed* past tense form as well as miscellaneous regular verbs – those that do not belong to our schemata . We narrow the regular set to monosyllabic forms. However, all polysyllabic irregular forms that could serve as a point of comparison are compounds based on monosyllabic forms (an example is *overwrite*, a compound form of irregular *write*). (A compound form might be more regular than a simplex form, but Celex will list both the regular and the irregular variant in both cases.) Table 1 shows the descriptions of the verb classes using regular expressions.

class	input regular expression	IPA	alternation
DROVE	`[i2][zvdltnk]$`	{i,aɪ} + {z,v,d,l,t,n,k} ##	{i,aɪ} → oʊ
SANG	`I(m|N|Nk)$`	{ɪ} + {m,ŋ,ŋk} ##	ɪ → æ
KEPT	`i[lpnm]$`	{i} + {l,p,n,m} ##	i → ɛCt
BURNT	`[3EI]nl$`	{ɜ,ɛ,ɪ} + {n,l} ##	{ɜ,ɛ,ɪ} → {ɜ,ɛ,ɪ}Ct

Table 1: Descriptions of verb classes in the GCM. 'C' marks any consonant.

Our starting point for the training set, following Albright & Hayes (2003), is the list of verbs in the Celex corpus (Baayen et al. 1993, based on Sinclair 1987) with a token frequency of 10 or above, encompassing 3257 forms. However, similarity requirements restrict the respective training sets. We use the *DISC* transcription in which each contrastive segment of English is represented by a unique character (`dr2v` equals [draɪv]).

Table 2 shows the number of verbs in Celex that were used as training sets for our verb classes. The irregular set consists of forms that match the schema and are irregular, whereas the regular set consists of such regular forms. The miscellaneous set is also restricted; it consists of monosyllabic verbs that do not belong to any of the schemata and are regular. These are included in the regular training set of *each* category.

The model calculates the similarity of a given nonce verb to the regular and the irregular set. Comparisons to stems in other classes are not calculated, as past tense markings for these classes were not available to participants in the forced-choice tasks.

verb class	irregular set	regular set	miscellaneous regular verbs
DROVE	14	83	1086
SANG	8	13	1086
KEPT	12	31	1086
BURNT	6	42	1086
sum	40	169	1086

Table 2: Number of forms in each verb class, GCM training data

2.3 Estimation

To calculate the similarity between two words, we first compute their dissimilarity. This is achieved using the string-edit (Levenshtein) distance, which is the smallest number of changes needed to transform one word into the other. Costs range from 0 (the corresponding segments are identical) to 1 (inserting or deleting an entire segment). Following Albright & Hayes (2003) and Dawdy-Hesterberg & Pierrehumbert (2014), costs between 0 and 1 are assigned to corresponding segments that are not identical, based on how much the segments differ.

All parts of the word are weighted equally, because although there is evidence that past tense formation in English is predominantly driven by overlaps 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. *stink*, *sink*, etc. – see Bybee & Moder 1983).

The following transformation, originating with Nosofsky (1990), is used to convert dissimilarity into similarity:

$$\eta_{ij} = \exp\left(-d_{ij}/s\right)^{p}$$

In the equation above, η_{ij} represents the similarity between form i and form j, while d_{ij} is the dissimilarity between the two forms. s and p are free parameters.

We explored a range of parameter settings and use $s = 0.9$ and $p = 1$ which provide the best model fit on the baseline data. (In contrast, Albright and Hayes use $s = 0.4$ and $p = 1$.)

When p is set to 1, as here, the similarity function is an exponential, rather than a Gaussian, function of the dissimilarity. The weighting parameter s controls how quickly the similarity decreases as the difference (or distance) between the forms increases. When s is small, the behavior of the model will be dominated by the small group of instances that differ very little from any given novel form. As it becomes larger, instances that differ more increase their influence on the overall model behavior (Nosofsky, 1990; Nakisa et al., 2001; Albright & Hayes, 2003; Dawdy-Hesterberg & Pierrehumbert, 2014). Thus, s effectively controls the size of the set of verbs that will be taken into account in determining the support for the regular versus the irregular outcome.

The overall similarity S_{iC_J} of a test form i to a set C_J is calculated by summing the similarity η_{ij} of each member j of class C_J to the test form i, and dividing by the summed similarity η_{ij} of each member k of class C_K (the class of all stored forms) to the test form i. This calculation is summarized in the following equation.

$$S_{iC_J} = \frac{\sum_{j \in C_J} \eta_{ij}}{\sum_{k \in C_K} \eta_{ik}}$$

2.4 Output format

The overall score used in our analyses is the *regularity score*, which is the complement to the *irregularity score* and reaches a maximum of 1.0 when the output is most likely to be regular. Unlike Dawdy-Hesterberg & Pierrehumbert (2014), there is no decision rule on top of the scoring, such that any form that is more likely than not to be regular is predicted to surface as regular all the time. This specific decision rule is statistically optimal, and was imposed in Dawdy-Hesterberg & Pierrehumbert (2014) in order to determine the ceiling performance for a computational model. The present paper, in contrast, analyzes data aggregated across human participants with differing decision thresholds. As discussed in Schumacher et al. (2014) and Schumacher & Pierrehumbert (2017), the input-output relationship in such aggregated data are typically reported to be nearly probability-matching.

We standardize the regularity score to match the range of participant responses: [0,1]. The modified score is interpretable as the probability that the outcome will be regular in aggregated data. It is also appropriate to attribute this type of gradience to people's initial expectations about other people's behavior, on the assumption that people realistically encode the variability they have encountered.

2.5 Example: *splive*

The nonce form *splive* belongs to the DROVE class in our model. The two past forms of *splive* in the experiment are regular *splived* and irregular *splove*. It is compared to 1169 regular verbs – these are 83 verbs that match the DROVE schema (e.g. *side, hive, line*) and 1086 miscellaneous verbs. It is also compared to 14 irregular verbs (e.g. *drive, stride, smite*) in this class. Overall, it is more similar to the regular set: its *regularity score* is 0.56.

3 Implementation of the Minimal Generalization Learner

3.1 Outline

The Minimal Generalization Learner is an algorithm for forming input-output rules of varying generality, which then compete to generate the output. The Minimal Generalization Learner is implemented here from materials made available by Albright and Hayes (Albright & Hayes, 2003). These include their Segmental Similarity Calculator, implementing the natural class based similarity metric due to Frisch et al. (2004), also used in the GCM implementation[1].

3.2 Training data

For our model fitted on our baseline nonce word stimuli, the MGL is trained on regular and irregular English verbs with a minimum frequency cutoff of 10 in CELEX (Baayen et al., 1993), encompassing 4160 past/present verb transcriptions.

The MGL builds rules based on all verb forms in CELEX with a token frequency of 10 or above. However the structural descriptions of the resulting rules do not cover all these forms. Table 3 shows the number of unique forms covered by the structural descriptions of the 'regular' and 'irregular' rules that are relevant to each class.

The MGL generates multiple possible past tense forms for each nonce verb. We only consider those rules *relevant* that generate the past tense forms that appear in the experiment (e.g. *splive* : *splived* / *splove*). There is at most one relevant regular rule and one relevant irregular rule for one verb, but multiple rules can generate the (ir)regular forms for each verb class. We return to this in the next section.

Note that the sets of exceptions and related forms of each rule can overlap, both respectively and with each other. As a consequence, the MGL rules apply to fewer forms than apparent from the table: 456 (instead of 617) in total.

	regular rules		irregular rules	
verb class	related forms	exceptions	related forms	exceptions
DROVE	46	44	13	113
SANG	49	35	49	35
KEPT	60	32	10	9
BURNT	55	18	35	34

Table 3: Number of forms in each verb class, MGL training data

3.3 Estimation

The MGL begins by considering the relationship between each verb and its past tense as a 'rule'. For each pair of verbs in the training data, it then attempts to create a more general rule. It does so by aligning the wordforms and analyzing shared phonetic features. For example, merging the word-specific rules for *ring/rang* and for *stink/stank* yields a more general rule that expresses the information that they share: [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 structural description for each rule has a *scope*, which is the number of verbs conforming to the description, to which the rule might apply. The number of *hits* is the number of such verbs where the rule generates the correct output. In our example, *think*

[1]Due to issues with the MGL code, we had to fit the MGL separately for participants, and not in batches, in order to train it using segmental similarity.

and *blink* fall in the scope of the rule, but they are not hits, because their past tenses display other patterns (*thought* and *blinked*) The *raw confidence* of the rule is the ratio of hits to scope:

$$\textit{Raw confidence} = \frac{\textit{hits}}{\textit{scope}}$$

The raw confidence is 1.0 if the rule applies to all forms that meet its structural description. It is less than 1.0 if some forms meeting its structural description have past tenses other than that predicted by the structural change. Raw confidence values of 0 are not found, because a rule needs to apply to two or more examples to be posited in the first place.

The MGL raw confidence metric is adjusted on the basis of user-specified confidence limits, to generate an *adjusted confidence score* that takes into account the amount and distribution of available data. The MGL's lower limit affects how much confidence is assigned to rules that have a small number of instances; generalizations that are based on a smaller number of word types are penalized. The MGL's upper limit curtails the application of seemingly general rules which are in fact driven by a more specific rule (Albright & Hayes, 2002). The MGL is implemented here with its default settings, with the exception of the algorithm's confidence limits. We implement the MGL with lower and upper confidence limits of 55% and 95%, respectively, since these values afford the best fit to English verb data in Albright & Hayes (2003).

Note that the MGL algorithm automatically groups together verbs on the basis of shared phonological properties; thus, verbs are most likely to form strong generalizations with other verbs that share the same onset or rhyme. 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. Both algorithms allow for category-specific similarities to play a role in rule formation.

3.4 Output format

Recall that in both our baseline and ESP experiment the trial task is prompted by a stem and offers a choice between a regular form and a specific irregular form, presented orthographically. In order to model this choice, we take the MGL rule for the stem that outputs the regular form (the *relevant regular rule*) and the rule that outputs the specific irregular form (the *relevant irregular rule*). If several regular / irregular rules generate the same form, we take the one with the highest *adjusted confidence*, following Albright & Hayes (2003). We use these rules to calculate the form's *relative (adjusted) confidence*.

Out of 156 test verbs in the ESP post-test, the Celex-trained MGL generates a *relevant regular rule* for every verb. It does not generate a *relevant irregular rule* for 28 verbs. These are all nonce verbs in the KEPT category (see Section **??**). In this category, irregular forms are derived from the stem through a vowel change (e.g. *greel* → *grelt*). The verbs missing the relevant irregular rule all have bases ending in <m>, <n>, or <l>. This is because, in our implementation, there is an insufficient number of verb types in the training set to support these irregular rules. Decreasing the cutoff criterion for the model leads to the generation of more of the currently 'missing' irregular rules, but the overall model fit becomes worse (cf. below). Therefore, we keep the cutoff criterion and assume that the adjusted confidence of the irregular rule for these 28 verbs is zero. We then take the relative confidence of the regular rule as compared to the regular *and* the irregular rule for each verb and took this as the adjusted regular confidence of the given verb. (If the irregular rule is missing, the value of this adjusted regular confidence is 1.) This is given by the following equation:

$$\textit{Relative (adjusted) confidence} = \frac{\textit{adjusted confidence of relevant regular rule}}{\textit{adj. conf. reg. rule + adj. conf. relevant irreg. rule}}$$

This relative adjusted confidence represents the MGL regularity score for an item, to be compared against the regularity score from the GCM (see Section 2).

3.5 Example: *splive*

The two past forms of *splive* in the experiment are regular *splived* and irregular *splove*. The *relevant regular rule* that generates the regular past tense is '∅ → [d] / {ð, ʃ, θ, ʒ, f, s, v, z} __'. The *structural description* indicates that this is a suffixation rule that can apply to forms that *end in an anterior fricative* (a natural class in our feature system). The *raw confidence* of this rule is 0.981. This is because this rule applies to most forms in its scope (530/540). The *adjusted confidence* is very similar: 0.973. This is because this rule applies to a large number of forms overall. The *relevant irregular rule* that generates the irregular form is '[aɪ] → [oʊ] / {ð, ʒ, dʒ, d, l, n, r, z} __v'. It applies to [aɪ] in the nucleus *preceded by voiced anterior consonant and followed by [v].* The rule applies to three forms (*drive, strive, dive*) and fails to apply to four (*thrive, contrive, rive, connive*). Its *raw confidence* is 0.428. Its *adjusted confidence* is slightly lower (0.412). This is because it applies to a smaller number of forms overall. The *relative (adjusted) confidence* of the predicted regularity of *splive* is $0.973/(0.973 + 0.412) = 0.702$.

4 Further notes on the GCM and the MGL

The training set for the GCM is focused initially; it is grouped into four verb classes and based on formal similarity with base forms of existing irregular verbs playing a role within each class. This, in effect, assumes lexical gangs as a starting point (Alegre & Gordon, 1999). In contrast, the MGL starts establishing rules across all forms in the starting dictionary.

As we see, however, the MGL also focuses the training set. The structural descriptions of the forms only cover a fraction of all verbs in the starting dictionary, organized by shared groups of segments. While disjunction increases the power of a theory (increasing the set of acceptable classes), disjunct classes frequently serve as input for phonological processes (Mielke, 2008) . In the MGL, the disjunction emerges despite the lack of an initial specification: while one rule accounts for the specific behavior of a given form, multiple rules cover a given class. The number of regular and irregular rules posited by the MGL for each verb class can be seen in Table 4.

Table 5 shows the list of (irregular and regular) rules for one of the classes, the KEPT class. We choose this class as it provides a good illustration of the estimation process. An MGL rule describes an $A \rightarrow B$ *alternation* in an XAY *environment.* Since the KEPT rules

verb class	irregular rule	regular rule	sum
DROVE	18	4	22
SANG	10	3	13
KEPT	6	4	10
BURNT	1	3	4
sum	35	14	**49**

Table 4: The counts of unique rules posited by the MGL in each verb class

are all suffixing, the following environment (Y) is empty. For the preceding environment (X), we give a list of segments that are matched by the structural description if it applies to four segments or fewer. Otherwise, we give a phonological description. So, for instance, the first rule turns [il] to [ɛlt] following any obstruent. Note that in some cases, the MGL generates multiple rules for the same alternation (such as ip → ɛpt). As discussed in Section 3.4, in such instances, we choose the candidate rule with the highest adjusted confidence in order to calculate the MGL regularity score. (Fricatives and nasal consonants constitute a natural class in the segmental feature system used by Albright & Hayes 2003.)

Inspection of the table shows that the MGL fails to generate some expected irregular rules; in the KEPT class, the MGL omits rules for for any stems ending in [in] or [im]. Under the MGL, there is not enough lexical support for irregular rules of these types, and yet participants do irregularize nonce forms like *cheen* and *kleem* in the forced choice task. In this respect, the variegated similarity of the non-natural GCM class (which can be used to categorize *cheen*) allows for a better approximation of participant behavior. Note also that in some cases, the MGL generates an overly-specific rule; the rule for stems ending in [il] applies only after obstruents, whereas participants also irregularize KEPT stems with a sonorant in the onset (*greel*/*grelt*).

alternation	preceding environment
il → ɛlt	obstruent
ip → ɛpt	{j, l, r, w}
ip → ɛpt	dorsal consonant
ip → ɛpt	l
ip → ɛpt	any consonant
il → ɛlt	fricative or nasal consonant
∅ → d	alveolars
∅ → d	{b, m}
∅ → d	alveolar fricative or nasal consonant
∅ → d	{b, m, p}

Table 5: The unique rules posited for the KEPT class

The rule structure of the MGL likely follows from its parameter settings, which were tuned to optimize its performance on our fairly irregular-looking nonce word stimuli. It requires further work to determine the extent to which the MGL fit on individuals would improve if the MGL were tuned on individual participants.

5 Residualization of the two categorization models on the baseline data

Since the regularity scores of the GCM and the MGL for the nonce verbs are correlated ($r = 0.48$), we use residualization to see whether the two models make separate significant contributions to explaining variation in the baseline dataset (Section ??).

We fit two simple linear regression models; no random effects are explored in these models, since we are merely predicting algorithmic scores for each item. We rescale and center GCM and MGL scores for the model fits.

Model 1 predicts the GCM scores from the MGL scores (M1: GCM regular score $\sim$ MGL regular score). Model 2 predicts the MGL scores from the GCM scores (M2: MGL regular score $\sim$ GCM regular score). In two additional models, we then use the residuals from each model (M1, M2) as an estimate of the variation present in one score that cannot be attributed to the other. First, we use the residuals of Model 1 in another, mixed-effects logistic regression model along with the MGL scores to predict regular responses in the baseline task (M3: regular response $\sim$ MGL score + GCM residual + (1 + MGL score + GCM residual|participant) + (1 | verb)). We then residualize for the MGL. We use the residuals of Model 2 along with the GCM scores in a similar logistic regression model (M4: regular response $\sim$ GCM score + MGL residual + (1 + GCM score + MGL residual |participant)).

The residuals remain significant predictors in both models (M3: MGL score ($\beta = 1.27$, $p < 0.0001$), GCM residual ($\beta = 1.37$, $p < 0.0001$); M4: GCM score ($\beta = 2.29$, $p < 0.0001$), MGL residual ($\beta = 0.95$, $p < 0.0001$). Note that while this paper reports values fit to the baseline experiment data, the same pattern of results also holds for models fit to the pre-test data.

Although residualization is sometimes critiqued when used as a stopgap for collinearity or sign changes in a regression model (Wurm & Fisicaro, 2014), these concerns do not apply in the present case. We use residualization as a tool to test whether one metric contributes beyond what is covered by a different metric of the same attribute, as in Baayen et al. (2006), and as also acknowledged in Wurm & Fisicaro (2014).

6 Additional information regarding the role of the GCM and the MGL in the post-test

6.1 Concordance Indices

Table 6 summarizes the concordance indices between our data and the various categorization models. The individual-GCM improves on the Celex-GCM in terms of explaining the post-test data. The individual-MGL does not improve on the Celex-MGL.

Our concordance indices are calculated using the somers2 function in R's `Hmisc` package (Harrell Jr, 2008).

Note that the indices in Table 6 directly calculate C between the relevant algorithm score (MGL or GCM) and the binary post-test outcome (regular or irregular). As a diagnostic for our regression models, we also calculated concordance indices between model predictions and participants' regular/irregular responses. The concordance indices for multiple regression models are, unsurprisingly, considerably higher than the indices calculated on the basis of a single variable.

	training data	method	test data	C (test data)
Celex-GCM	real verbs	'analogy'	verbs in baseline	**0.58**
Celex-MGL	real verbs	'rules'	verbs in baseline	**0.59**
Celex-GCM	real verbs	'analogy'	verbs in post-test	**0.6**
individual-GCM	real verbs + verbs seen in ESP	'analogy'	verbs in post-test	**0.68**
Celex-MGL	real verbs	'rules'	verbs in post-test	**0.61**
individual-MGL	real verbs + verbs seen in ESP	'rules'	verbs in post-test	**0.55**

Table 6: Summary of concordance indices between our data and the categorization models. The values above the line show the concordance indices for the baseline data, as reported in Section **??**. The values below the line show the indices for the ESP post-test data.

The models in Table **??** and Table **??** both have an index of 0.84, indicating an acceptable level of predictive ability (Harrell, 2013).

Since the random effects likely account for a large amount of this variance, we calculated the marginal r^2 values for both models: these are 0.27 and 0.22, respectively (the conditional values are 0.45 and 0.47). Conditions and learning models explain some amount of the variation, but the data are noisy.

6.2 Assessing separate contribution of individual models over Celex models

The text (Section **??**) explains that we calculate the extra information given by the individual models by subtracting the Celex model prediction from the individual model prediction for each item for each participant, and that we find that the extra information from the individual-GCM models is predictive over and above the Celex-GCM model.

We do this for both the GCM (Model 5: regular response $\sim$ Celex GCM score + extra GCM information + participant pre-test mean + (1 + Celex GCM + extra GCM information|participant) + (1|verb) and the MGL (Model 6: regular response $\sim$ Celex MGL score + extra MGL information + participant pre-test mean + (1 + Celex MGL + extra MGL information|participant)+ (1|verb)).

In Model 5, the extra information from the individual-GCM is a significant predictor of post-test responses ($\beta = 2.61$, $p < 0.001$), above and beyond the Celex-GCM contribution. In similar Model 6, however, the extra information from the individual-MGL is not a significant predictor of post-test responses ($\beta = 0.21$, $p = 0.23$). This means that the individual-GCM has some success in modeling post-test behavior when trained with real English verbs and information from the ESP test, whereas adjustments made to the individual-MGL do not improve predictive power.

7 Regression modeling procedures

Our regression model selection in this article proceeded as follows. Except where stated otherwise, we fit logistic mixed-effects regression models, using R's lme4 package (Bates et al., 2015). Non-significant predictors are removed from our models, as determined by a series of comparisons of nested models, using likelihood ratio tests in R's anova function. However, our model tables (e.g., Tables **??** and 6) include the significance levels generated by lme4 summaries.

We investigate maximal random effects structures, allowing for random intercepts for participants and stimulus items (verbs), as well as random slopes for all within-unit factors. We compare nested random effects structures as above, and retain any random slopes and intercepts which are supported by likelihood ratio tests (Baayen et al., 2008).

In the model in Table **??**, we elect to mean center continuous variables to counteract collinearity that becomes evident when including the interaction between and *lexical typicality* . Although some debate exists regarding the practice of transforming variables (Belsley, 1991; Echambadi & Hess, 2007), centering variables is common practice to remove nonessential ill conditioning. That is, relationships that inevitably exist between main and product terms (Aiken et al., 1991; Jaccard & Turrisi, 2003; Jaeger, 2008). We note that if *verb baseline mean* is *not* centered, collinearity is high (VIF=15.79) in this model. A non-centered model also yields a significant main effect for *lexical typicality*, with increased post-test regularization in the random and reversed conditions. However, we elect to take a more conservative approach by not including a potentially spurious main effect. Note that other than this main effect difference, models with and without centering yield results that are qualitatively indistinguishable; all other main effects and interactions are present in either case. The collinearity for the centered model (Table **??**) is acceptable, with a maximum VIF score of 2.89.

Since it is unclear whether the absence of a difference between random and reversed conditions is due to a lack of data, we refit the model with the baseline average : typicality interaction on the random and reversed conditions only. Then we calculated the Bayes Factor of the null model (with no interaction) over the full model (with interaction) using Bayesian Information Criteria calculated by lme4. The result strongly supports the null model over the full model, indicating that the preference for the null model is *not* due to the lack of evidence to reject it.

References

Aiken, Leona S.; Stephen G. West; and Raymond R. Reno. 1991. *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Albright, Adam, and Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, vol. 6, 58–69. Stroudsburg, PA.

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90.119–161.

Alegre, Maria, and Peter Gordon. 1999. Rule-based versus associative processes in derivational morphology. *Brain and Language* 68.347–354.

Baayen, R. Harald; Douglas J. Davidson; and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

Baayen, R. Harald; Laurie B. Feldman; and Robert Schreuder. 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 55.290–313.

Baayen, R. Harald; Richard Piepenbrock; and Hedderik van Rijn. 1993. *The CELEX lexical database on CD-ROM.* Philadelphia: Linguistic Data Consortium.

Bates, Douglas; Martin Mächler; Ben Bolker; and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67.1–48, DOI: 10.18637/jss.v067.i01.

Belsley, David A. 1991. *Conditioning diagnostics: Collinearity and weak data in regression.* New York: Wiley.

Bybee, Joan L., and Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language* 59.251–270.

Bybee, Joan L., and Dan I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58.265–289.

Dawdy-Hesterberg, and Janet B. Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. *Language, Cognition and Neuroscience* 29.1268–1282.

Echambadi, Raj, and James D. Hess. 2007. Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science* 26.438–445.

Frisch, Stefan A.; Janet B. Pierrehumbert; and Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228.

Harrell, Frank E. 2013. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* New York: Springer Science & Business Media.

Harrell Jr, et al, Frank E. 2008. Hmisc: A package of miscellaneous R functions. *R package version* 3. Online: `http://biostat.mc.vanderbilt.edu/Hmisc`.

Jaccard, James, and Robert Turrisi. 2003. *Interaction effects in multiple regression.* Thousand Oaks, CA: Sage.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59.434–446.

Mielke, Jeff. 2008. *The emergence of distinctive features.* Oxford: Oxford University Press.

Moder, Carol Lynn. 1992. *Productivity and categorization in morphological classes.* Buffalo, NY: State University of New York dissertation.

Nakisa, Ramin C.; Kim Plunkett; and Ulrike Hahn. 2001. A cross-linguistic comparison of single and dual-route models of inflectional morphology. *Models of language acquisition: Inductive and deductive approaches*, ed. by Peter Broeder and Jaap Murre, 201–222. Cambrdge, MA: MIT Press.

Nosofsky, Robert M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* 34.393–418.

Schumacher, R. Alexander, and Janet B. Pierrehumbert. 2017. Prior expectations in linguistic learning: A stochastic model of individual differences. *Proceedings of the 39th Meeting of the Cognitive Science Society (CogSci2017)*. London, UK.

Schumacher, R. Alexander; Janet B. Pierrehumbert; and Patrick LaShell. 2014. Reconciling inconsistency in encoded morphological distinctions in an artificial language. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci2014)*, 2895–2900. Quebec City,CA.

Sinclair, John M. (ed.) 1987. *Looking up: An account of the COBUILD project in lexical computing*. London: Collins.

Wurm, Lee H., and Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language* 72.37–48.