Emergent Idiosyncrasy in English comparatives[*]

Brian W. Smith & Claire Moore-Cantwell

University of California Santa Cruz, University of Connecticut

1. Introduction

Speakers of a language have implicit knowledge not only of exceptionless phonological patterns in their language, but also variable patterns. Implicit knowledge has been observed of both **free variation**, or variation within lexical items, and **lexical variation** or variation across lexical items. An example of the former is t/d deletion in English (Guy 1994, Coetzee & Pater 2011). A single lexical item, say *west*, can be realized as [wɛst] or [wɛs] in the same phonological context. An example of lexical variation is stress in English. English speakers are aware of probabilistic trends across words, such as that heavy syllables attract stress (Chomsky & Halle 1968, Guion et al. 2003, Moore-Cantwell 2016). However, each individual word of English has just one stress, and does not vary (rélapse, never relápse; reláx, never *rélax). We examine a type of variation that lies in between these two extremes. Some, but not all, individual lexical items exhibit idiosyncratic preference for one variant or another, above and beyond the demands of the (variable) grammar. In particular, high-frequency lexical items exhibit idiosyncratic behvior, while the behavior of lower-frequency lexical items is probabilistic, and largely predictable from the preferences of the grammar. This work builds on Morgan & Levy (2016), which examines binomial expressions in English (e.g. *ladies and gentlemen*), finding idiosyncrasy in high-frequency forms, and grammatical behavior in low-frequency forms.

We examine the comparative in English, which can be realized two ways, either by the **morphological** (*happier*) or the **periphrastic** (*more happy*). For many adjectives (like *happy*) both options are quite good. The choice between *more* and *-er* is conditioned by a number of phonological factors, for example words three syllables long and longer very rarely take *-er* (**beautifuller*). However, many words exhibit idiosyncratic preferences that cannot be explained by phonology alone. For example *real* prefers to take *more* while *pale*, which has the same prosodic form and the same final consonant, prefers to take *-er*. Consider also *steadier* > *readier*, *sharper* > *apter*, and *sparser* > *falser*.

[*] For useful discussion, we would like to thank Ray Jackendoff, Emily Morgan, Aleksei Nazarov, Joe Pater, Stephanie Shih, Kie Zuraw, and audiences at at NELS 47, AMP 2016, USC, and the UConn ling lunch.

Given that adjectives of English do in fact exhibit idiosyncratic preferences for one form of comparative or the other, we take an expansive view of what is stored in a lexical entry. Specifically, we propose that at least some comparatives are stored as whole words (*happier*), or whole phrases (*more likely*), rather than being composed independently on each use. We implement this storage in the form of UR constraints (Zuraw 2000, Boersma 2001, Pater et al. 2012, Smith 2015), which demand that a particular adjective in the comparative form be realized in a specific way. UR constraints compete with grammatical constraints in a Maximum Entropy grammatical model (Goldwater & Johnson 2003). Both are learned together with a learning algorithm related to the Gradual Learning Algorithm (Boersma & Hayes 2001, Boersma & Pater 2016). We propose an induction and decay mechanism for UR constraints, so that they are induced only when needed, and decay when unused. Additionally, we train the model using data sampled based on lexical frequency, so that more frequent adjectives are experienced more often by the learner. With these two mechanics we find that at the end of learning (a) only some adjectives exhibit a strong preference which contradicts the grammar, and (b) more frequent adjectives are more likely to do so. Both of these results are consistent with observed corpus data.

2. List of factors in the literature and sketch of a phonological analysis

In this section, we review the phonological and frequency factors that condition comparative choice. In terms of syllable count, *-er* rarely occurs with stems longer than two syllables, and is favored in monosyllables. Finally stressed stems favor *more*, a preference that's especially clear in disyllabic stems (**robuster*). The fact that *-er* occurs with monosyllables and trochees, but not iambs or longer words, can be explained if *-er* selects for words with exactly one foot (McCarthy & Prince 1986/1996), namely a moraic trochee (H́ or ĹL).

(1) *Prosodic factors*
3+ syllables → *more* (Kytö & Romaine 1997, Mondorf 2003, Hilpert 2008)
monosyllables → *-er* (Kytö & Romaine 1997, Mondorf 2003, Hilpert 2008)
final stress → *more* (Mondorf 2003, Hilpert 2008)

The final segment of the stem also plays a role. Final [li] favors *more*, while final [i] (excluding [li]) favors *-er*. Final liquids, especially [r], disfavor *-er*. Elsewhere in English, near-adjacent liquids, especially identical ones, are avoided. Hall (2008) argues that avoidance of [rVr] drives local dissimilation ([ɪnfərɛd] for *infrared*), and prevents the suffix *-ery* from combining with r-final stems (*winery*,**beerery*), while Martin (2007) shows that [rVr] and [lVl] are underattested in the English lexicon, baby names, and neologisms.

(2) *Final material*
final [r] → *more* (Mondorf 2003, Hilpert 2008, *pace* Kytö & Romaine 1997)
final [l] → *more* (Hilpert 2008, *pace* Kytö & Romaine 1997)
final [i] → *-er* (Kytö & Romaine 1997, Hilpert 2008)
final [li] → *more* (Mondorf 2003, Hilpert 2008)
final cluster → *more* (Mondorf 2003, Hilpert 2008)

While the effects of final [l]/[r] and stress can be explained by English phonology, we can't think of a reason why *-er* would be avoided with final clusters. One possibility is presented in Mondorf (2009), who explains the preference as a reflex of processing. Mondorf's hypothesis, called *more support*, is that *more* is used to ease the processing of complex structures, including phonologically complex ones. Another possibility is that the reported effect of clusters is an artifact of statistical modeling. In the logistic regression model of Hilpert (2008), words ending in a consonant followed by syllabic L (e.g., *humble*) are coded as ending in a final cluster (and not a final L).

A consistently strong predictor of comparative choice is adjective frequency. More frequent adjectives take *-er* more often, as do adjectives with a higher comparative to positive ratio - the frequency of the adjective in the comparative (*more*+*-er*) divided by the frequency of the positive form (not comparative or superlative). According to Mondorf (2009), frequent stems and comparatives are easier to access, so don't require *more*.

(3) *Frequency factors*
more frequent adjective → *-er* (Mondorf 2003, Hilpert 2008)
higher comparative-positive ratio → *-er* (Mondorf 2003, Hilpert 2008)

Boyd (2007) points out a more nuanced effect of frequency, related to the ones above. Adjectives that occur in the comparative more often tend to exhibit stronger preferences, not just for *-er* but also for *more*. We discuss this effect at some length below.

3. Corpus data and regression modeling

Our corpus study replicates many (but not all) of the phonological generalizations presented above and demonstrates that adjectives are sometimes idiosyncratic. Frequency and phonological factors alone aren't sufficient to fully account for comparative preference.

3.1 Corpus data

The data come from the Corpus of Contemporary American English (COCA: Davies 2008). COCA is a corpus of written and spoken English, containing 520 million words balanced across genres. To compile a list of comparatives, we searched for words tagged as comparative adjectives, and all *more* plus adjective sequences, where *more* was tagged as an adverb. To control for noise and mistagged words, we used the following exclusion criteria. We segmented *-er* forms by hand, and checked each stem against a list of adjectives, excluding stems not on the list. We also excluded any adjective that didn't occur at least 5 times with *more* and 5 times with *-er*. Although the data set doesn't control for syntactic context or semantics, we are confident that it contains only comparative adjectives. The remaining data – containing 313 different stems and 723,203 tokens – were annotated with IPA transcriptions from the CMU pronouncing dictionary (Weide 1994).

3.2 Logistic regression models

To test for lexical idiosyncrasy in comparatives, we fit two logistic regression models: one with random intercepts for each adjective and one without, in addition to the phonological and frequency factors. If the random intercept model provides a signficantly better fit to the data, we can conclude that lexical idiosyncrasy plays a role.

For the model with random intercepts, we fit a mixed effects logistic regression model using the package lme4 (Bates et al. 2013) in R (R Core Team 2017). The dependent variable was *more* vs. *-er*. The fixed effects were taken from the previous section, and were all dummy coded (1 if 'yes'; 0 if 'no'), except frequency measures. An explanatory variable for long stems (3+ syllables) was not included, since few long stems (just 4) survived the corpus exclusions. The mixed effects model also included a random intercept for adjective stem, while the fixed-effects-only model did not, but was otherwise the same.[1]

Some notes on coding. The comparative log frequency is the adjective's frequency in the comparative form (*more*+*-er*). The variables 'Final L' and 'Final syllabic-L' are exclusive – a stem with final [l] is coded as syllabic-L-final but not L-final – and so are the variables 'Final i' and 'Final li' – a stem with final [li] is coded as [li]-final but not [i]-final. The results for both models are summarized in the table in (4). A positive value indicates preference for *more*, and a negative value indicates preference for *-er*.

(4) *Results of regression models*

	Mixed effects model			Fixed effects model		
Variable	β	SE	p	β	SE	p
(Intercept)	3.235			4.418		
PROSODIC FACTORS						
One syllable	−4.426	0.092	***	−5.459	0.068	***
Final stress	1.890	0.093	***	1.913	0.070	***
FINAL SEGMENT FACTORS						
Final [i]	−2.243	0.208	***	−2.824	0.031	***
Final [li]	−0.562	0.385	n.s.	3.832	0.041	***
Final syllabic-L	0.310	0.824	n.s.	−0.598	0.032	***
Final L	0.561	0.186	***	1.137	0.031	***.
Final R	1.625	0.488	***	1.541	0.006	***
Final cluster	0.276	0.198	n.s.	1.016	0.020	***
FREQUENCY FACTORS						
Comp. log frequency	−0.487	0.026	***	−0.579	0.006	***
Comp.-pos. ratio	−3.960	0.090	***	−0.628	0.103	***
Log Likelihood	−44,281			−77,540		
(+ = *more*, − = *-er*)		* = p<0.05		** = p<0.01		*** = p<0.001

[1] R code for mixed effects model: glmer(More/Er ∼ oneSyll + moreThanTwoSylls + finalStress + final[i] + final[li] + final[el] + final[r] + finalCC + log(PosFreq) + compPosRatio + (1 | Stem), family="binomial")

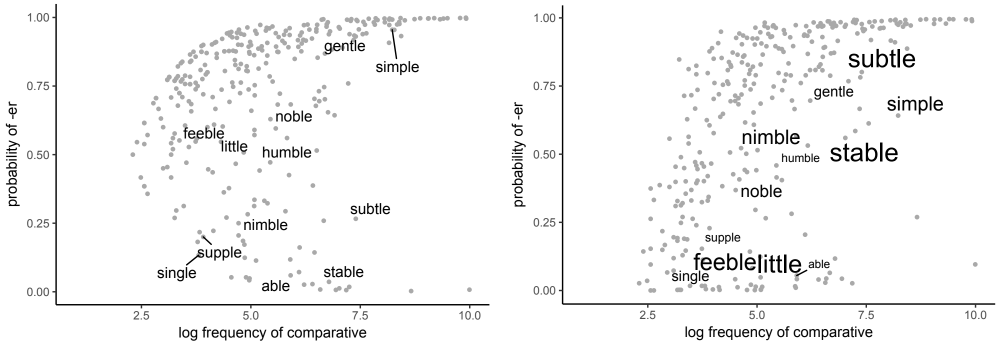
Among the prosodic and final segment factors, results are consistent with previous work except for final [li], final syllabic-L, and final cluster, none of which are significant in the mixed effects model. The difference between models with respect to final [li] is likely due to a single adjective: *likely*, which accounts for over 85% of tokens for [li]-final stems, and takes *more* more than 99% of the time. Among other [li]-final adjectives, the average rate of *more* is only 45%. The random intercept for *likely* in the mixed effects model means that it has less overall pull on the model. Similarly, the *-er*-favoring effect of syllabic-L in the fixed effects model likely comes from the adjective *simple*, which takes *-er* 95% of the time and makes up 40% of tokens for syllabic-L-final stems.

3.3 Idiosyncrasy and frequency

Phonology and frequency alone aren't enough to explain the behavior of some idiosyncratic adjectives, such as *likely* in the discussion above. This can be seen in the differences between the mixed and fixed effects models. The fixed effects model, with no means of capturing idiosyncrasy, has a much worse fit, with a log likelihood of -77,540, compared to the mixed effects model with -44,281. While testing for the significance of random effects is non-trivial and perilous, especially in GLMMs, we cautiously find a stastically significant difference between these models, even under very conservative assumptions. In a likelihood ratio test where we treat every stem as a degree of freedom, the mixed model provides a significantly better fit (χ^2=66,518, df=314, p<0.001).

The graphs below demonstrate idiosyncrasy in a different way. The graph on the left shows the probability of *-er* for each adjective, plotted against how often it occurs in the comparative. As frequency increases, the number of adjectives around 50% *-er* decreases, and adjectives become more categorical, moving toward *-er* or *more*. We've labeled disyllabic adjectives ending in syllabic L to demonstrate that adjectives with the same frequency and similar phonological shapes can demonstrate different comparative preferences. The graph on the right shows the predictions of the fixed-effects-only model, plotted against frequency. Adjectives with more error have larger labels. The fixed effects model, with phonological and frequency predictors, isn't sufficient to model the probabilities of some comparatives. While the prefences of some adjectives, like *supple* and *gentle*, can be explained by their frequency, others like *feeble* and *subtle* cannot.

(5) *Probability of* -er *for each adjective in the corpus (L) and model predictions (R)*



4. The model

In the previous sections, we demonstrate that in English, individual adjectives exhibit idiosyncratic preferences for either the morphological (*-er*) or periphrastic (*more*) comparative. Although general grammatical constraints govern words' preference to some extent, individual adjectives diverge from their grammatically dictated preferences. Furthermore, more frequent comparatives diverge more from the grammar's preferences. In this section, we propose a model of learning in which both grammatical trends and individual adjectives' preferences are acquired gradually, trained on the corpus data. We use Maximum Entropy grammar (Goldwater & Johnson 2003) to model gradient grammatical preferences, and we use UR constraints (Zuraw 2000, Boersma 2001, Pater et al. 2012, Smith 2015) to incorporate individual words' preferences into the system. In our model, UR constraints are induced only when needed, and decay when they are not used. Because of this, and because the learner learns gradually from corpus data, getting more opportunities to learn on more frequent items, our model predicts idiosyncratic behavior for high-frequency lexical items, and grammatically constrained behavior for low-frequency lexical items.

4.1 Structure

Maximum Entropy grammar (MaxEnt) is a variety of Harmonic Grammar, in which constraint weights determine a probability distribution over output candidates. In (6), two candidate outputs are considered for the adjective *tall*. A harmony value ($\mathscr{H}$) is calculated for each candidate according to the equation given in (7), by multiplying the candidate's violations of each constraint by that constraint's weight, and summing over all constraints. That harmony value is then converted to a probability via the logit transform.

(6) *Maximum Entropy Tableau*

	p	$\mathscr{H}$	$1\sigma \rightarrow$*-er*	1→*more*	$\acute{\sigma}]_{\text{WD}}\rightarrow$*more*	
			6.96	3.41	4.27	← *weights*
TALL+COMP						
taller	**0.33**	-7.68		1	1	
more tall	**0.67**	-6.96	1			

(7) *MaxEnt equations*

$$\mathscr{H} = -\sum w_i * v_i \qquad\qquad p = \frac{e^{\mathscr{H}}}{\sum e^{\mathscr{H}}}$$

In the above tableau, the weights of $1\sigma \rightarrow$*-er*, 1→*more*, and $\acute{\sigma}]_{\text{WD}}\rightarrow$*-er* received their weights via training on a large data set of comparatives, discussed in detail below. In that data set, monosyllables tend to take *-er*, [l]-final adjectives tend to take *more*, and adjectives with final stress (including monosyllables) tend to take *more*. These competing pressures predict that *tall* should form the comparative with *-er* 33% of the time. In our simulation, we used a set of constraints corresponding to the phonological factors previously reported in the literature ((1),(2), and (3) above). They are repeated here, recast as OT-style constraints.

(8) *Constraints used in the simulation:*

Name	Assign a violation to	Weight
1σ →*-er*	monosyllabic adjectives with *more*	6.96
3+σ →*more*	three-syllable or longer adjectives with *-er*	6.62
l→*more*	l-final adjectives with *-er*	3.41
li→*more*	adjectives ending in [li] with *-er*	2.35
i→*-er*	i-final adjectives with *more*	3.15
r→*more*	r-final adjectives with *-er*	2.35
-CC→*more*	adjectives with a final cluster with *-er*	1.14
σ́]→*more*	adjectives with final stress with *-er*	4.27

We treat these constraints as placeholders for a more typologically-motivated constraint set (for example, OCP constraints for the liquids). Such a constraint-based model has not yet been developed for the comparatives, but we believe that the results we present here with respect to lexical idiosyncrasy would hold over a wide range of grammatical models.

In (6), the model predicts just 33% *taller* as the comparative of *tall*. In the corpus, however, *taller* is much more common than *more tall*. The latter is vanishingly rare (less than 1% of all occurrences). Mechanisms for including idiosyncratic information about a lexical item into a probabilistic analysis include high-weighted faithfulness (Zuraw 2000), lexical indexation of markedness constraints (Pater 2005, Becker 2009), lexically specific constraints (Moore-Cantwell & Pater 2016), and UR constraints (Zuraw 2000, Boersma 2001, Pater et al. 2012, Smith 2015). We use UR constraints.

UR constraints demand that a specific underlying form be used to produce a particular combination of morphological features. For example, such a constraint might demand that the underlying representation for TALL+COMPARATIVE be /moɹ+tɑl/. This constraint would be violated if the UR /tɑl+ɚ/, or any other UR, were used instead. In an analysis using UR constraints, candidates in a tableau are not just possible surface representations, but are instead UR-SR pairs. This is illustrated in (10).

(9) *Example UR constraints*

a. TALL+COMPARATIVE→ /moɹ+tɑl/ (→moɹ)
Assign a violation whenever the morpheme bundle of TALL and COMPARATIVE is not realized with the phonological underlying representation /moɹ+tɑl/

b. TALL+COMPARATIVE→ /tɑl+ɚ/ (→ -ɚ)
Assign a violation whenever the morpheme bundle of TALL and COMPARATIVE is not realized with the phonological underlying representation /tɑl+ɚ/

These UR constraints demand that specific sets of morphemes, or morphological features, be realized with a specific phonological UR. They are abbreviated in the following tableau as → -ɚ, and → moɹ, respectively. It is worth pointing out that these constraints are not restricted to *more* and *-er*, but rather determine the realization of the entire set of input morphological features for a language. For example, *good* in the comparative would have the UR constraint GOOD+COMPARATIVE → /bɛtɚ/, which would demand realization of the suppletive comparative form, and be violated by the URs /gʊd+ɚ/ or /moɹ+gʊd/.

(10) *Tableau illustrating UR-SR pairs*

			p	$\mathscr{H}$	FAITH	→ -ɚ	→ moɹ	1σ →-*er*	l→*more*	σ́]→*more*
					10	7.1	1.1	6.96	3.41	4.27
TALL+COMP										
a.	/moɹ+tɑl/	→ moɹtɑl	0.005	-14.06		1		1		
b.	/moɹ+tɑl/	→ tɑlɚ	0.00	-24.78	1	1			1	1
c.	/tɑl+ɚ/	→ tɑlɚ	**0.995**	-8.78			1		1	1
d.	/tɑl+ɚ/	→ moɹtɑl	0.00	-18.06	1		1	1		
e.	/tɑl/	→ tɑl	0.00	-22.84		1	1	1	1	1
f.	/tɑl/	→ tɑlɚ	0.00	-25.87	1	1	1		1	1

In (10), the high-weighted faithfulness constraint forces each UR to surface faithfully, so that the choice between forms of the comparative is a choice between URs, rather than a choice between phonological operations. We do not include faithfulness, or unfaithful candidates in the simulation. Rather, we assume that the relevant faithfulness constraints would get a high weight in English since m's and o's typically don't get epenthesized or deleted, suffixes don't become clitics, etc. Our simulation will focus on candidates like a. and c. In (10), the markedness constraints alone would predict a 33% probability on *taller*, just like in (6), but the UR constraints adjust this probability. Since the UR constraint demanding *-er* has a very high weight, *taller* gets nearly 100% probability. Different adjectives, with different UR constraints, or different weights on them, would have different predicted distributions across the two forms of the comparative. In the next section, we present a learning model which learns the weights of the grammatical constraints in (8) as well as inducing UR constraints for some adjectives, and learning weights for them.

4.2 Learning

We use a variation of the Perceptron learning algorithm (Rosenblatt 1958), which has previously been adapted to the learning of OT-like grammars (Boersma & Hayes 2001, Pater 2005). Perceptron learning is a form of error-driven learning, in which parameters are updated whenever the learner makes a mistake on a learning datum. At each learning 'timestep' a learning datum is sampled, and the grammar is updated based on that datum. Here, each learning datum is an adjective plus its comparative form, for example (*tall*, *taller*), and adjectives are sampled based on frequency – higher-frequency comparatives are used more frequently as learning data, and less frequent comparatives are used less frequently. The learner uses the bare adjective as input to its current grammar, and predicts a comparative. If the current grammar were the one in (6), there would be a 33% chance that the learner would predict *taller*, and a 66% chance of predicting *more tall*. If the learner happened to select *more tall* this would count as an error, and an update would occur.

In perceptron-based OT-learning models, the weights of all relevant constraints are updated by the same amount, the learning rate δ. Constraints which prefer the actual output of the learning datum are promoted, while constraints which prefer the error are demoted. A single update is illustrated below, with *more tall* as the error, and *tall* as the correct form.

(11) *Perceptron update:*

a. $\overrightarrow{w}_{t+1} = \overrightarrow{w}_t + (\overrightarrow{v}_{err} - \overrightarrow{v}_{correct}) * \delta$

b. Example update: δ=0.01

	1σ→*-er*	3+σ→*more*	l→*more*	li→*more*	i→*-er*	r→*more*	r→*-er*	-CC→*more*	$\acute{\sigma}$]→*more*
$\overrightarrow{v}_{taller}$	0	0	1	0	0	0	0	0	1
$\overrightarrow{v}_{more\ tall}$	1	0	0	0	0	0	0	0	0
$\overrightarrow{v}_{err} - \overrightarrow{v}_{correct}$	1	0	-1	0	0	0	0	0	-1
$\overrightarrow{w}_t$	6.96	6.62	3.41	2.35	3.15	2.35	0.58	1.14	4.27
$\overrightarrow{w}_{t+1}$	**6.97**	6.62	**3.40**	2.35	3.15	2.35	0.58	1.14	**4.26**

1σ →*-er*, l→*more*, and $\acute{\sigma}$]→*more* are updated, while the other constraints are unchanged, since they cannot adjudicate between *taller* and *more tall*. The weight of 1σ →*-er* is moved up, since the error violates this constraint, while the weights of the other two are moved down, since the correct form violates them.

In the learning model we present here, we introduce two additional mechanics which govern the learning of UR constraints. Following work such as (Pater 2010, Nazarov 2016), our learner induces UR constraints during learning. Specifically, UR constraints are introduced whenever the model makes an error on a learning datum. In the above hypothetical case, where *taller* is the correct output, and *more tall* is the error, the UR constraint Tall+Comparative→ /tɑl+ɚ/ would be created.

(12) **UR constraint induction:** On error, create a UR constraint of the form Adj+Comparative → *correct output* (if it does not already exist)

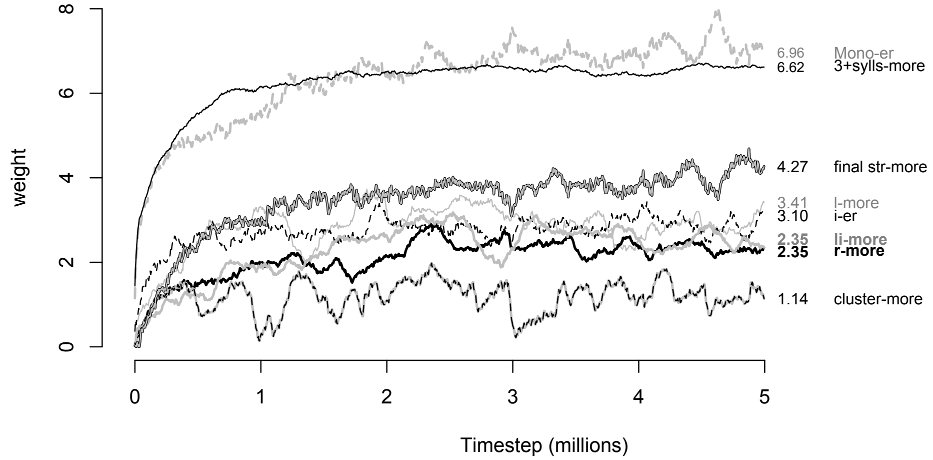
UR constraints are induced with an initial weight of 10 (a high weight in this system – see the faithfulness constraint in (10)). If one or more UR constraints already exist for the adjective, their weights are also updated according to the perceptron update mechanic. UR constraints preferring the correct output will be promoted, and UR constraints preferring the error will be demoted. They also decay as learning progresses. All UR constraints in the system decay at the same rate of 0.0001 per timestep, and if a UR constraint decays to zero it is removed from the system. This can happen for two reasons: either the adjective to which the UR constraint belongs is not sampled again as a learning datum in the 10,000 timesteps it takes for the constraint to decay away, or the adjective is sampled, but no error is made. Low-frequency adjectives will usually wind up without UR constraints because they will not be sampled often enough, and adjectives whose behavior is completely predictable based on the weights of the grammatical constraints will also wind up without UR constraints, because no errors will be made on them.

This decay mechanic allows for simple but constrained induction of UR constraints – UR constraints can be freely induced, but those which are less useful in predicting the overall behavior of the training data will decay away. As we will discuss below, UR constraints for high-frequency forms, and forms which diverge from the predictions of the grammar (like *tall* above) acquire very high weights, and do not decay away.

4.3 Results

We trained the learner on 1.1 million comparatives gathered from COCA. The data set is similar to the one used to fit the regression models, but we include rare adjectives and adjectives with categorical preferences for *-er* or *more*. The data consisted of about 4600 distinct adjectives, the majority of which occurred only once. The model was trained for 5 million timesteps, which was enough for the weights of the general grammatical constraints to stabilize, as shown in (13). As mentioned above, adjectives were sampled according to frequency, so that high-frequency forms are trained on more often than low-frequency forms. In (14), the probability of each adjective taking *-er* as its comparative is plotted against the log frequency in COCA of that adjective in the comparative.

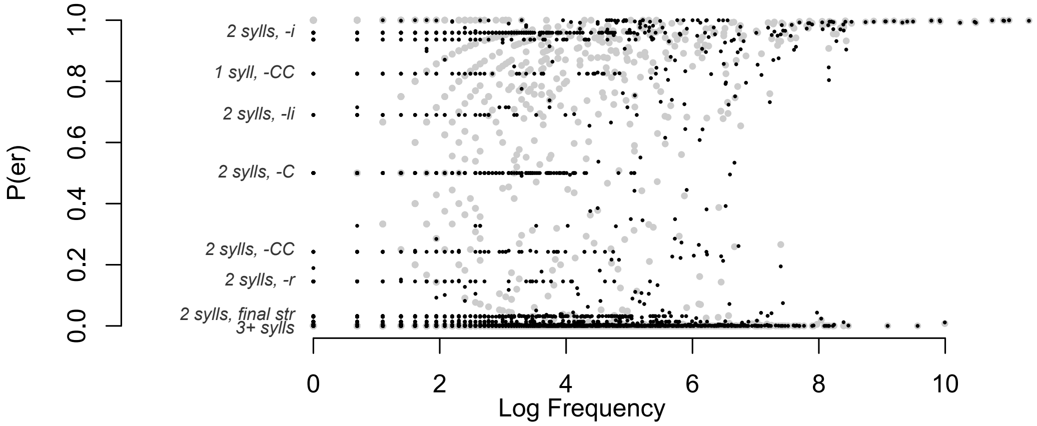
(13) *Weights over time during training.*



Adjectives that are high-frequency are matched very closely by the model, while low-frequency adjectives are matched less closely. Adjectives with a log frequency below about 5 tend to fall into probability bins - this is because they do not have UR constraints, so their probability is entirely determined by the weights of the grammatical constraints. Each probability bin is determined by a particular violation profile, which many adjectives in the training set share. For example, monosyllables ending in a cluster, such as *vast*, *crisp*,

lax all violate 1σ →*-er* when they appear with *more* and all violate -CC→*more* when they appear with *-er*. Several of these bins are labeled in (14).

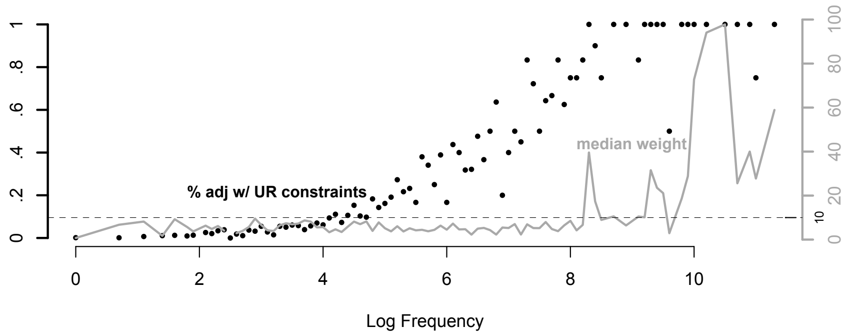
(14) *Observed (grey) and predicted (black) P(er) for all words in the training set.*



Adjectives that are high-frequency are matched very closely by the model, while low-frequency adjectives are matched less closely. Adjectives with a log frequency below about 5 tend to fall into probability bins - this is because they do not have UR constraints, so their probability is entirely determined by the weights of the grammatical constraints. Each probability bin is determined by a particular violation profile, which many adjectives in the training set share. For example, monosyllables ending in a cluster, such as *vast*, *crisp*, *lax* all violate 1σ →*-er* when they appear with *more* and all violate -CC→*more* when they appear with *-er*. Several of these bins are labeled in (14).

Higher-frequency comparatives are more likely to have one or more UR constraints associated with them, and also tend to have higher weights on those UR constraints. This is illustrated in (15) with 89 frequency bins, of width 0.1 in log space. Weights above 10 (the starting weight for newly induced UR constraints) are achieved via update.

(15) *% forms with a UR constraint (black) and median weight of UR constraints (grey)*

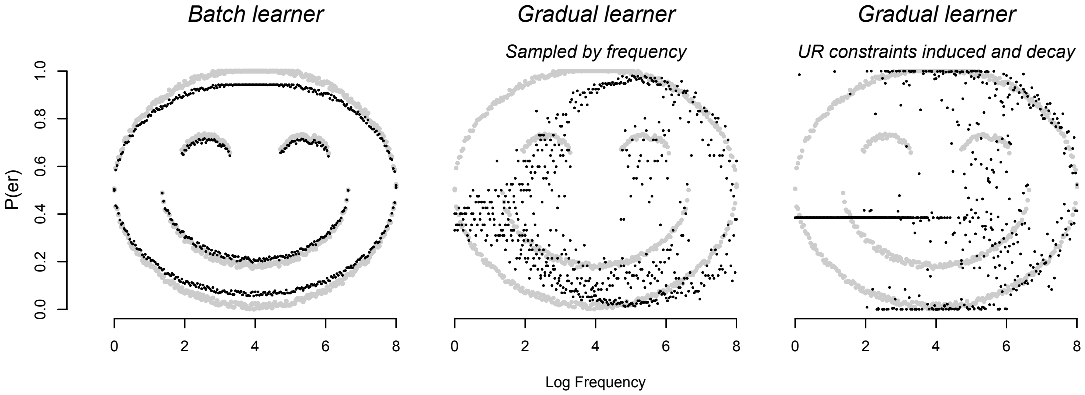


In this simulation, higher-frequency forms acquire high-weighted UR constraints, which allows them to diverge from the general grammatical trends in the system. UR constraints are induced for low-frequency forms, but they decay away since those forms are seen only rarely. The behavior of low-frequency forms is therefore determined mostly by the weights of the general grammatical constraints, while the behavior of high-frequency forms is determined by their individual UR constraints.

The online nature of learning in our model, as well as the induction and decay of UR constraints, are crucial to achieving this result. To illustrate this, we compared the behavior of our model to the behavior of two simpler models on the toy dataset in (16). In all models, two simplistic grammatical constraints were used: one preferring *more* and one preferring *-er*. First, we fit a 'batch' learning model, in which constraint weights are fit for the entire dataset at once without sampling. We used the MaxEnt Grammar Tool (Wilson & George 2009), and gave every adjective both possible UR constraints (demanding *more* and demanding *-er*). Because frequency is not incorporated into this model in any way, both high-frequency and low-frequency forms are fit equally well, and are in fact fit very closely, as can be seen in the leftmost panel of (16). The predictions only diverge from the training data at all because we used a very strict (L2) regularization term: σ^2=100. Similar closeness of fit would obtain if this model were trained on the actual comparatives in (14).

Next, we trained a gradual learner which sampled each form according to frequency, but which did not incorporate UR constraint induction/decay. Instead, like the batch learner every adjective had both UR constraints. The results of this model are given in the middle panel of (16). Because higher frequency forms are sampled more during training, they are fit better, but all forms diverge from the grammar somewhat (which predicts about 40% *-er*). Given enough training iterations, the learner will eventually encounter low-frequency adjectives enough to match the toy data. The rightmost panel is the results of training our learner, with UR induction and decay. Despite being trained on data very unlike natural languages, this learner still arrives at a final state very much like the English comparatives, as well as English binomial expressions (Morgan & Levy 2016), in which high-frequency forms are idiosyncratic and low-frequency forms are governed by the grammar.

(16) *Results on toy data comparing three models*



5. Conclusion

In this paper, we present a corpus analysis of the English comparative, together with a learning model in which UR constraints affiliated with individual adjectives determine each adjective's idiosyncratic preferences for *more* or *-er*. These UR constraints are induced during gradual learning in a MaxEnt system, and decay when they are not used. Because of this induction and decay, the model predicts a trade-off between lexical variation and idiosyncratic behavior at high frequencies, and within-item variation following the predictions of the probabilistic grammar at lower frequencies.

References

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. 2013. *lme4: Linear mixed-effects models using eigen and s4*. R package version 1.0-5.

Becker, Michael. 2009. Phonological trends in the lexicon: the role of constraints. Doctoral dissertation, University of Massachusetts Amherst.

Boersma, Paul. 2001. Phonology-semantics interaction in OT, and its acquisition. In *Papers in experimental and theoretical linguistics*, ed. Kirchner et al, volume 6. University of Alberta, Edmonton.

Boersma, Paul, & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul, & Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In *Harmonic grammar and harmonic serialism*, ed. John McCarthy & Joe Pater. London: Equinox Press.

Boyd, Jeremy Kenyon. 2007. Comparatively speaking: a psycholinguistic study of optionality in grammar. Doctoral dissertation, UCSD.

Chomsky, Noam, & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Coetzee, Andries W., & Joe Pater. 2011. The place of variation in phonological theory. In *Handbook of phonological theory*, ed. John A. Goldsmith, Jason Riggle, & Alan C. Yu, 401–434. Wiley-Blackwell, 2 edition.

Davies, Mark. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Goldwater, Sharon, & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, & Osten Dahl, 111–120.

Guion, Susan G., J.J. Clark, Tetsuo Harada, & Ratree P. Wayland. 2003. Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech* 46:403–427.

Guy, Gregory R. 1994. The phonology of variation. In *CLS 30: Papers from the 30th Regional Meeting of the Chicago Linguistic Society. Volume 2: The Parasession on Variation in Linguistic Theory*, 133–149.

Hall, Nancy. 2008. Perceptual errors or deliberate avoidance? Types of English /r/-dissimilation. In *Annual Meeting of the BLS*, volume 34, 133–144.

Hilpert, Martin. 2008. The english comparative–language structure and language use. *English language and linguistics* 12:395–417.

Kytö, Merja, & Suzanne Romaine. 1997. Competing forms of adjective comparison in Modern English: What could be more quicker and easier and more effective? In *To explain the present: studies in the changing English language in honour of Matti Rissanenissan*, ed. Terttu Nevalainen & Leena Kahlas-Tarkka. Helsinki: Societe Neophilologique.

Martin, Andy. 2007. The evolving lexicon. Doctoral dissertation, UCLA.

McCarthy, John J., & Alan Prince. 1986/1996. Prosodic morphology 1986. Technical Report RuCCS-TR-32, Rutgers University Center for Cognitive Science.

Mondorf, Britta. 2003. Support for more-support. *Topics in English Linguistics* 43:251–304.

Mondorf, Britta. 2009. *More support for more-support: The role of processing constraints on the choice between synthetic and analytic comparative forms*. Studies in language variation. John Benjamins Publishing Company.

Moore-Cantwell, Claire. 2016. The representation of probabilistic phonological patterns: neurological, behavioral, and computational evidence from the English stress system. Doctoral dissertation, University of Massachusetts Amherst.

Moore-Cantwell, Claire, & Joe Pater. 2016. Gradient exceptionality in maximum entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15:53–66.

Morgan, Emily, & Roger Levy. 2016. Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition* 157:382–402.

Nazarov, Aleksei. 2016. Extending hidden structure learning: features, opacity, and exceptionality. Doctoral dissertation, University of Massachusetts Amherst.

Pater, Joe. 2005. Learning a stratified grammar. In *Proceedings of the 29th Boston University Conference on Languega Development*, ed. Alejna Brugos, Manuella R. Clark-Cotton, & Seungwan Ha, 482–492. Somerville, MA: Cascadilla Press.

Pater, Joe. 2010. Morpheme-specific phonology: constraint indexation and the inconsistency resolution. In *Phonological argumentation: essays on evidence and motivation*, ed. Steve Parker, 123–154. Equinox.

Pater, Joe, Robert Staubs, Karen Jesney, & Brian Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of SIGMORPHON2012*, 62–71. Montréal, Canada: Association of computational Linguistics.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenblatt, Frank. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65:386.

Smith, Brian W. 2015. Phonologically-conditioned allomorphy and UR constraints. Doctoral dissertation, University of Massachusetts Amherst.

Weide, Robert L. 1994. CMU pronouncing dictionary.

Wilson, Colin, & Benjamin George. 2009. The maxent grammar tool. Department of Cognitive Science, Johns Hopkins University and Department of Linguistics, UCLA.

Zuraw, Kie. 2000. Patterned exceptions in phonology. Doctoral dissertation, UCLA.

Brian W. Smith, Claire Moore-Cantwell
b@phrenology.biz, c@phrenology.biz