Learning a frequency-matching grammar together with lexical idiosyncrasy: MaxEnt versus Hierarchical Regression

Jesse Zymet
University of California, Berkeley

1 Introduction

Language learners can acquire principles that have differing levels of generalization. To take one example, a speaker might acquire a language with a variable phonological rule, applying across eligible words at some overall rate; yet any single word's rate of undergoing this rule might differ from the overall rate. What should our theory of learning and representation be, given that learners show an ability to acquire a system of *nested generalizations* such as this?

Here we take up the problem of learning such a nested generalization. In particular, this paper seeks a model that can accurately learn a statistical generalization across the lexicon, together with the idiosyncratic degree to which any particular lexical item departs from this overall generalization. Experimental research has uncovered language learners' ability to frequency-match to statistical generalizations across the lexicon, while also acquiring the idiosyncratic behavior of individual attested words. How can we model the learning of a frequency-matching grammar together with lexical idiosyncrasy? A recent approach based in the single-level regression model Maximum Entropy Harmonic Grammar (MaxEnt) makes use of general constraints that putatively capture statistical generalizations across the lexicon, as well as lexical constraints governing the behavior of individual words. I argue on the basis of learning simulations that the approach fails to learn statistical generalizations across the lexicon, running into what I call the GRAMMAR-LEXICON BALANCING PROBLEM. In MaxEnt, the general constraint and the set of lexical constraints are *a priori* equally viable hypotheses about data. Consequently, lexical constraints are too powerful: they come to learn each word's behavior, during which time frequency matching to the overall rate ceases and the general constraint becomes ineffective for modeling the grammar. Hence the cost of learning lexical idiosyncrasy is the inability to learn an accurate, frequency-matching grammar. A generality bias is therefore attributed to learners to the effect that general, grammatical constraints are privileged over lexical constraints, so that frequency matching using the former constraints is retained throughout the learning process. The model ultimately advanced here is rooted in HIERARCHICAL REGRESSION: multiple layers of regression structure, corresponding to different levels of a nested hierarchy of generalizations. Hierarchical regression is here shown to surmount the grammar-lexicon balancing problem—learning a frequency-matching grammar together with lexical idiosyncrasy—by encoding general constraints as fixed effects and lexical constraints as a random effect. The learner treats the grammar and lexicon differently, in that vocabulary effects are subordinated to broad, grammatical effects in the learning process. The model is applied to variable Slovenian palatalization, with promising preliminary results.

The paper is organized as follows. Section 2 summarizes some literature that supports the assumptions needed for the goal of our modeling task: to learn an accurate grammar together with lexical idiosyncrasies. Section 3 reviews the Maximum Entropy Harmonic Grammar approach to grammar and lexical idiosyncrasy, and provides the learning-theoretic argument against said approach. Section 4 advances hierarchical regression as a way to model grammar with lexical idiosyncrasy. Section 5 applies the hierarchical regression model to Slovenian palatalization. Section 6 concludes.

* Thanks to Bruce Hayes, Kie Zuraw, Robert Daland, Tim Hunter, Megha Sundara, and audiences at Stanford, UC Berkeley, UCLA, and AMP 2018 for invaluable feedback on this project.

Proceedings of *AMP 2018*

2 Speakers know aggregate generalizations across the lexicon together with lexical idiosyncrasies

A growing number of experiments suggest that language learners FREQUENCY-MATCH to statistical generalizations across the lexicon: during nonce probe studies, their responses in aggregate tend to match lexical frequencies (e.g., Zuraw 2000, Ernestus & Baayen 2003, Hayes & Londe 2006; see Hayes, Zuraw et al. 2009 for a more exhaustive list of citations). We briefly review a case from Hayes & Londe (2006).

In Hungarian vowel harmony, dative forms take one of two allomorphs of a suffix, *-nɛk* or *-nɔk*, depending on the backness of preceding stem vowel. Stems ending in front vowels tend to take *-nɛk* (e.g., [kɛ̠rt-nɛ̠k] 'garden'-DAT and [y̠ʃt-nɛ̠k] 'cauldron'-DAT), while stems ending in back vowels tend to take *-nɔk* ([ɔblɔ̠k-nɔ̠k] 'window'-DAT and [bi:rо̠:-nɔ̠k] 'judge'-DAT). According to the corpus study undertaken in Hayes & Londe (2006), 92% of monosyllabic stems ending in a front, unrounded vowel take *-nɛk*, whereas 8% of them exceptionally take *-nɔk*. In a wug test (Berko 1958), the investigators presented to Hungarian native speakers fake stems with the same shape, and asked whether they would inflect them with *-nɛk* or *-nɔk*. Their responses, in aggregate, matched the frequencies in the corpus very closely:

(1) Hayes & Londe (2006)

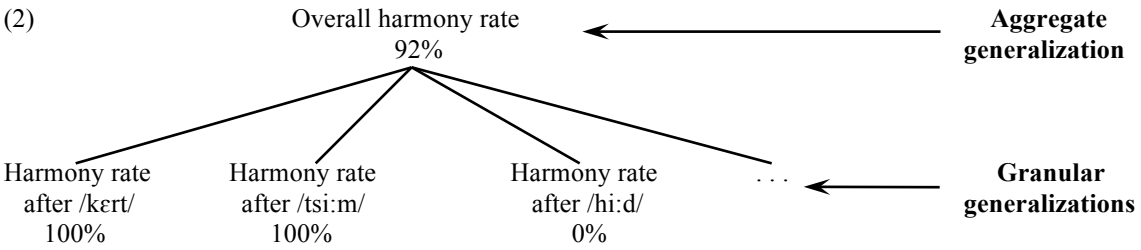
	-nɛk	*-nɔk*
Corpus rate:	92%	8%
Wug test rate:	93%	7%

While stems in the corpus take *-nɛk* 92% of the time, the wug test results show that speakers, in aggregate, inflect novel forms having the same shape with *-nɛk* at nearly the same rate of 93%.[1]

In addition to acquiring statistical generalizations across the lexicon, speakers must know which attested words are exceptional versus not (e.g., Ernestus & Baayen 2003; Zuraw 2000, 2010). There is evidence that speakers are even aware of gradient rates at which different words undergo a rule. In French, for example, different adverbs have differing propensities to undergo liaison before vowel-initial adjectives (Mallet 2008); moreover, a recent wug test suggests that speakers are sensitive to these distinctions when they are presented with these adverbs followed by novel vowel-initial adjectives (Zymet 2018).

The accumulating evidence therefore suggests that speakers internalize a nested hierarchy of generalizations: the rate at which a process applies across words, together with the idiosyncratic behavior of individual words with regard to the process. Given in (2) is a schema for the hierarchy of harmony generalizations governing monosyllabic stems with a front vowel in Hungarian. The task at hand is to model the learning of a system of two levels of generalization such as this one.

(2)



3 The MaxEnt approach leads to the grammar-lexicon balancing problem

3.1 *MaxEnt background* In Maximum Entropy Harmonic Grammar (hereafter MaxEnt; Smolensky 1986, Goldwater & Johnson 2003, Hayes & Wilson 2008), constraints have numerical weights, and surface forms are assigned probabilities as a function of these weights. In this investigation, we only consider two

[1] Note that some experiments suggest that frequency matching can be overridden in cases where a generalization runs counter to a learning bias (e.g., Becker, Ketrez & Nevins 2011, Becker, Nevins & Levine 2012, Jarosz & Rysling 2017). For example, Becker, Nevins & Levine (2012) found that while a laryngeal alternation in the English plural (*leaf ~ leaves*) applies more regularly to monosyllables than polysyllables, a nonce probe study reveals no such preference, while a series of artificial language learning studies suggests a bias towards protecting initial syllables. Hence while frequency matching seems to be a general capability of learners, these studies suggest that patterns that contradict learning biases may be learned relatively poorly.

candidate surface forms for each input: one that undergoes some variable process, and another one that does not. Suppose we have *n* constraints and two candidate surface forms *x* and *y*. Let w_k be the weight of the *k*th constraint, and $C_k(x)$ be the number of times *x* violates constraint *k*. We define the *harmony* of candidate *x*, $H(x)$, and the *probability* of *x*, $P(x)$, as follows:

(3a) $H(x) = \sum_{k=1}^{n} w_k * C_k(x)$ (3b) $P(x) = e^{-H(x)}/(e^{-H(x)} + e^{-H(y)})$

The probability of the other candidate, *y*, is defined to be $P(y) = 1 - P(x)$.

The MaxEnt model has typically been used to learn constraint weights that result in the best possible fit to the frequency at which a variable process applies *across* the lexicon (e.g., Hayes, Zuraw et al. 2009). More recently, however, investigators have been using MaxEnt to additionally learn which words are exceptional versus not. A recent approach to capturing both kinds of knowledge uses general constraints to frequency-match to a trend across the lexicon (e.g., HARMONIZE in Hungarian, violated by the faithful candidate of every input eligible to undergo harmony), and lexically indexed constraints (hereafter lexical constraints) to capture the idiosyncratic behavior of particular morphemes (HARMONIZE(kɛrt), DON'THARMONIZE(hi:d), etc.) (Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017). Tanaka (2017) found that his MaxEnt model of Japanese surname data is characterized by a tradeoff: depending on the settings of the MaxEnt prior (Goldwater & Johnson 2003), the model either more accurately fits to overall rates in the data using general constraints, or it more accurately fits to lexical idiosyncrasies using lexical constraints; it thus requires identifying a balance between the role of lexical versus phonological factors (p. 197). In the following section, we assess whether this problem is general—i.e., we elucidate whether the MaxEnt approach can learn a frequency-matching grammar together with word-specific lexical idiosyncrasies from simple toy data displaying variation.

3.2 *The grammar-lexicon balancing problem* Suppose we have a dataset with some number of regulars and irregulars (e.g., 46 regulars and 4 irregulars) such that the irregularity rate across words is 8%, as in the Hungarian vowel harmony case from Section 2. Moreover, suppose that every word is consistent in its behavior: each regular behaves regularly (e.g., undergoes harmony) across essentially all of its tokens, while each irregular behaves irregularly (does not undergo harmony) across essentially all of its tokens. We use three constraints: BEREGULAR, BEREGULAR(regulars), BEIRREGULAR(irregulars). BEREGULAR is analogous to HARMONIZE in Hungarian, the general constraint demanding that every input eligible to undergo harmony do so; its weight should govern the behavior of the learner during wug tests. The other two constraints govern lexical knowledge: BEREGULAR(regulars) demands that words marked as regular in the underlying form behave regularly (much like HARMONIZE(kɛrt), etc., in Hungarian), while BEIRREGULAR(irregulars) demands that words marked as irregular behave irregularly (much like DON'THARMONIZE(hi:d)). The violation profiles for these constraints are given in Table 1 below.

A learning simulation was run to assess whether MaxEnt can accurately frequency-match to the 8% rate while learning the idiosyncratic behaviors of regulars and irregulars. The simulation was run using Excel Solver (Fylstra et al. 1998, Tay, Kek & Abdul-Kahar 2009), which can fit parameters of nonlinear models using either of the Conjugate Gradient Descent or Newton learning algorithms; in the following simulation, Conjugate Gradient Descent was used. The simulation consisted of several trials. In each trial: the frequencies were multiplied by some number, multiplier *m*; constraint weights were initiated at 0; the model was then set to learn constraint weights that predict the input frequencies; in the next trial, the weights would be reinitiated at 0, so that learning would restart from 50/50-rates. In early trials, the frequency multiplier *m* was set to be very small (e.g., *m* = 0.0001); but across trials the multiplier was gradually increased, so that learned parameter values would fit to the word frequencies with increasing accuracy.[2] The input to the various trials are given in the table below:

[2] So that the learner would not be required to learn infinite weights for the lexical constraints in the ideal, a small degree of token variation was introduced: 0.001% of the underlying regulars surface as irregular, while 0.001% underlying irregulars surface as regular.

UR	SR	Freq.	BEREG 0	BEREG(reg) 0	BEIRREG(irreg) 0
/Regular/	Regular:	≈ 46**m*			
	Irregular:	≈ 0	-1	-1	
/Irregular/	Regular:	≈ 0			-1
	Irregular:	≈ 4**m*	-1		

Table 1. MaxEnt input

We want MaxEnt to learn weights such that: in a wug test, an irregular form is picked roughly 8% of time; and the behavior of attested regulars and irregulars are predicted correctly roughly 100% of time. The settings *w*BEREG = 2.44, *w*BEREG(reg) = 4.5, *w*BEIRREG(irreg) = 9.4 yield great results for example, predicting that attested regulars surface as regular 99.9% of the time (plugging 2.44 + 4.5 into the inverse logit), attested irregulars surface as irregular 99.9% of the time (plugging in 9.4), and that irregulars would be selected roughly 8% of the time in a wug test (plugging in 2.44).

But does MaxEnt *learn* good weights from the input data? The results of the simulation are given in the figure and table below. In this simulation, the MaxEnt prior terms μ and σ were set to 0 and 100, respectively, for all constraints. With frequency multiplier set to 0, the model learns 0-valued weights, selecting irregulars at a 50/50-rate in wug tests. As we increase the frequency multiplier (e.g., to 0.0001), the model rapidly learns the behavior of regulars (predicting that they behave as regular most of the time even at $m = 0.0001$), but poorly predicts the behavior of the irregulars. At this stage, BEREG is used to explain much of the variation—its weight growing high at first—and so the model predicts a low nonce irregularity rate, as desired. But eventually the weights of the lexical constraints grow high, coming to explain increasingly more of attested data, while the weight of BEREG begins to drop, coming to explain increasingly less of data, eventually being rendered superfluous and ineffective. At that point, the learner selects regulars and irregulars, once again, at a 50/50-rate in wug tests, with no grammatical generalization learned. We observe that at no point in the learning process does the model yield an accurate, frequency-matching grammar for the 8% irregularity rate in the input data.

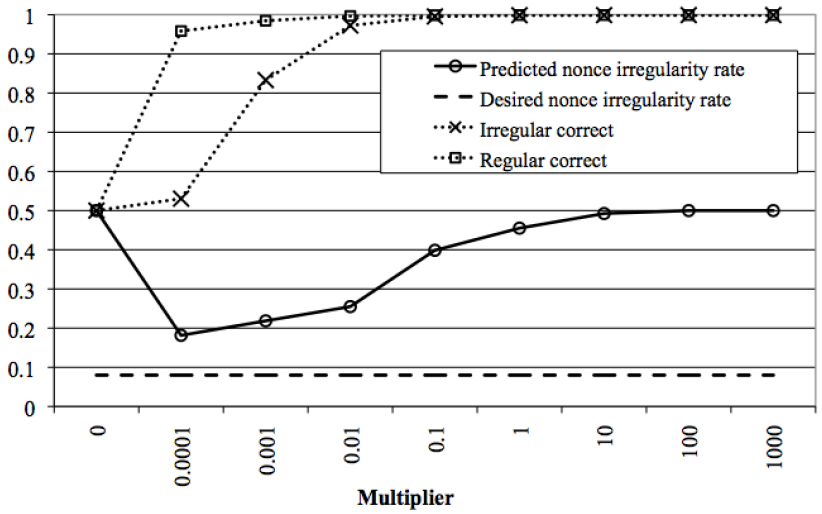


Figure 1. MaxEnt fails to learn generalization together with idiosyncrasy

Freq. multiplier	Be Reg	BeReg (reg)	BeIrreg (irreg)	Regular correct	Irreg. correct	Nonce irreg. rate
0	0	0	0	0.5000	0.5000	0.5000
0.0001	1.50	1.62	1.62	0.9582	0.5304	0.1815
0.001	1.27	2.88	2.88	0.9845	0.8331	0.2185
0.01	1.07	4.63	4.63	0.9966	0.9723	0.2548
0.1	0.41	6.22	5.82	0.9986	0.9955	0.3986
1	0.17	6.69	6.78	0.9989	0.9986	0.4551
10	0.02	6.87	6.89	0.9989	0.9989	0.4925
100	0	6.90	6.90	0.9989	0.9989	0.5000
1000	0	6.90	6.90	0.9990	0.9989	0.5000

Table 2. MaxEnt fails to learn generalization together with idiosyncrasy I

Note that the presence of the lexical constraints is a crucial prerequisite for model failure. To check that this was the case, I ran a learning simulation of the same dataset but with BEREG as the only constraint in the model. The model succeeded to frequency-match to the overall rate in this circumstance: at $m = 1$, for example, the model yielded wBEREG = 2.43, predicting that irregulars would be selected in wug tests at just about an 8% rate. The weight of BEREG did not shift nearly at all after this point in the simulation, remaining stable at least up to $m = 10$ million.

My investigation into adjusting MaxEnt's L2 prior in response to this problem has failed to surmount the problem. Assuming the three constraints have the same value for σ, varying σ merely varies the time at which the learner's predicted nonce irregularity rate begins to drift away from the overall irregularity rate: for example, if σ were set to be smaller, then the no-grammar outcome would be achieved at a higher frequency multiplier than if σ were set to be higher. Setting a stronger prior for the lexical constraints and a weaker one for the general constraint (in particular, σ = 10 for the lexical constraints, σ = 1,000 for BEREG) also did not affect the outcome meaningfully: the model still arrives at the same outcome, but at a high frequency multiplier. For more discussion on these points, see Zymet (2018).[3]

I also checked to see if the results would be different if we had constraints for each word. Instead of inputting two groups of data (one group of regulars, one group of irregulars), I inputted a table with 50 groups: 46 regular words surfacing as regular 99.9% of the time, and 4 irregular words surfacing as irregular 99.9% of the time—for an overall irregularity rate of 8%. I coded 51 constraints: BEREG, BEREG(reg1), ..., BEREG(reg46), BEIRREG(irreg1), ..., BEIRREG(irreg4). The results are similar to those above—see bolded predictions in the table below: very early in learning, the model yielded a close prediction of the overall irregularity rate, but poor predictions of word-specific rates (in particular, the irregulars); as the lexical weights grow and the model comes to learn idiosyncrasy with greater accuracy, the predicted nonce irregularity rate drifts away from the overall irregularity rate in the learning data.[4]

[3] It is unclear as to whether the same outcome would obtain if we were to use other priors (*cf.* Hughto, Lamont, Prickett & Jarosz 2019). This is something I plan to investigate in the future.

[4] I do not address here why the model behaves differently when regulars and irregulars are grouped together (wBEREG plummets with lexical learning) versus treated separately (wBEREG increases with lexical learning, failing to mimic nonce exceptionality). Regardless, in both simulations we find that better lexical learning comes with poorer general learning.

	$m = 0.0001$	$m = 0.01$	$m = 1$	$m = 100$
*w*BEREG	2.44	3.87	5.61	5.70
Predicted irregularity rate across nonces	**0.08**	**0.02**	**≈ 0.00**	**≈ 0.00**
*w*BEREG(reg) constraints	0.03	0.50	1.09	1.20
For each reg.-marked form, pred. % tokens behaving as reg.	**0.92**	0.99	0.99	0.99
*w*BEIRREG(irreg) constraints	0.38	5.74	11.23	12.58
For each irreg.-marked form, pred. % tokens behaving as irreg.	**0.12**	0.87	0.99	0.99

Table 3. MaxEnt fails to learn generalization together with idiosyncrasy II

The model frequency-matches to the overall rate across a wide range of values for *m* only when it is tasked to fit the weight of BEREG exclusively: *w*BEREG = 2.43, so that the predicted irregularity rate for words—attested or novel—is 0.08.

These results were checked using the learning algorithms provided in Excel Solver: in particular, the Conjugate Gradient Descent method and the Newton method. Other algorithms may yield different results; but recent research generally casts doubt on single-level regression models' ability to accurately model nested generalizations (see Johnson 2013 for more discussion). To take the results given in Moore-Cantwell & Pater (2017), their learning model was able to learn lexical trends together with idiosyncrasy in a variable voicing alternation in Dutch, but with predicted overall rates exaggerated toward the poles (p. 63).

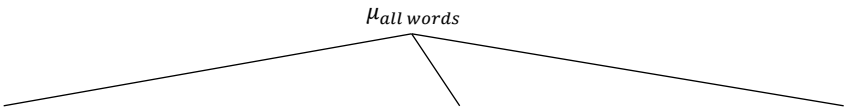
Hence, at the very least, the results taken together suggest that single-level regression models like MaxEnt are challenged by the task of learning systems of nested generalizations such as the one defined by lexically variable processes. I call this the GRAMMAR-LEXICON BALANCING PROBLEM. In MaxEnt, the general constraint and the set of lexical constraints are *a priori* equally viable hypotheses about data. Consequently, lexical constraints are too powerful: they come to learn each word's behavior, during which time frequency matching to the overall rate ceases and the general constraint becomes ineffective for modeling the grammar. Hence the cost of learning lexical idiosyncrasy is the inability to learn an accurate, frequency-matching grammar. We therefore search for a theory that can learn and sustain a frequency-matching grammar even while accurately learning lexical idiosyncrasies.

4 The hierarchical solution to grammar and lexicon

4.1 *Hierarchical regression background* We seek in particular a model that incorporates a GENERALITY BIAS to the effect that general, grammatical constraints are privileged over lexical constraints, so that frequency matching with the former constraints is retained throughout the learning process. In light of the fact that we are modeling rates at two distinct levels of generality—word-specific rate, and overall rate across words—it would seem natural here to invoke a theory rooted in hierarchical regression. How does the hierarchical mixed-effects logistic regression model fare? Similar to binomial logistic regression, hierarchical regression models arrange constraints hierarchically into fixed effects and random effects. Here fixed effects are taken to govern the generalizations on the higher level of the hierarchy—for our purposes, the statistical phonological generalizations present across the lexicon, i.e. the rate of regularity across words. Random effects govern the generalizations lower in the hierarchy—here, the idiosyncratic rates of individual words. I call this model HIERARCHICAL MAXENT. Hierarchical regression is used widely to capture statistical generalizations together with idiosyncrasies in variable datasets: linguists in particular have employed random intercepts to measure by-word/lexical class idiosyncrasy (Fruehwald 2012, Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017, inter alia); Shih & Inkelas (2016) and Shih (2018) even adopt the hierarchical model as a theory of learning and competence for their data.

Random effects are depreciated relative to fixed effects. In particular, as it pertains to the problem at hand, fixed effects are designed to predict the overall rate across words, and random effects are designed to predict *word-specific offsets* from the overall rate. To see how this works, we apply the model to the dataset posed in Section 3.2. We have a fixed effect, the general constraint BEREG, whose weight is estimated by the overall regularity rate across the entire dataset, 92%. With this weight we want our model to accurately predict the rate across all words, as that would be a frequency-matching grammar, mimicking human behavior in wug tests. We also have a random intercept, whose coefficients correspond to the weights of the lexical constraints—in the simulation below, constraints for each of the fifty words. As these word-specific coefficients are embedded in a random intercept, any coefficient is estimated by a weighted average between the rate across all words and the word-specific rate. Given below is a hierarchical scheme for how fixed coefficients are estimated (4a) and how random coefficients are estimated (4b) (Raudenbush & Bryk 2002, Snijders & Bosker 2012). *w*BEREG(reg1), for example, is estimated by a weighted average between the rate across all fifty words in our data ($\mu_{all\ words}$) and the rate at which Regular1 behaves regularly (μ_{reg1}). In this scheme, λ_i, called the reliability of Word*i*, is a group-specific constant taking a value between 0 and 1, and depends on size of the group (Raudenbush & Bryk 2002, Snijders & Bosker 2012). For example, *w*BEREG(reg1) will be determined more by μ_{reg1} if the data have more Regular1 tokens rather than fewer.

(4a)



(4b) $\lambda_{reg1} * \mu_{reg1} + (1 - \lambda_{reg1}) * \mu_{all\ words}$ $\lambda_{reg2} * \mu_{reg2} + (1 - \lambda_{reg2}) * \mu_{all\ words}$ . . .

The parameters of a mixed model are the coefficients of the fixed effects, the variances of the random effects (for our purposes, one per intercept), and residual variance (Raudenbush & Bryk 2002, Snijders & Bosker 2012). The objective function of a mixed-effects logistic regression model—the likelihood of the observed dataset given the fixed-effect coefficients and the variances (leaving aside a regularization term)—does not possess a closed form, and so these parameters are typically estimated by maximizing a Laplace approximation of this likelihood. This is achieved by applying the penalized iterative reweighted least squares algorithm, which performs batch gradient descent first on the fixed effect coefficients and then on the random variances—iteratively—until the relative change in predictors has fallen below a threshold value, at which point the iterates are said to have converged (Bates 2009, pp. 28-31).

The correct way of extracting predicted nonce rates from model is to average over the levels of the random intercept (rather than, say, simply plugging *w*BEREG into the inverse logit formula, which produces exaggerated overall rates) (Skrondal et al. 2009, Pavlou et al. 2015). This averaging involves an integral that cannot be calculated analytically; R's *predict()* function gives an approximation to this, and was used to extract predictions in the simulation given in the following section.[5]

4.2 *Hierarchical regression learns a frequency-matching grammar together with idiosyncrasy*

We want the model to learn with the weight of BEREG a grammar that accurately frequency matches to the overall rate, and with the weights of the lexical constraints the specific rates for every word. We achieve this using mixed-effects logistic regression. We run a model of the dataset using the `glmer` function of the *lme4* package R (Bates & Maechler 2011), taking *w*BEREG to be the coefficient of the fixed-effect intercept and the weights of the lexical constraints to be the coefficients of the levels of a random intercept. The predictions of this model are given in the table below. This model succeeded in learning a grammar that frequency matches to the overall rate, mimicking participant behavior in wug tests. The model learned a general constraint weight of *w*BEREG = 6.17 and a random variance of τ^2 = 14.77, and so the predicted nonce irregularity rate is 0.074. Moreover, the model accurately predicts lexical idiosyncrasies: regulars behave as regular, and irregulars behave as irregular, the vast majority of the time.

[5] Zeger et al. (1998) provides a commonly used approximation to the estimated general rate. For our model, the approximation would be $exp\left(\frac{wBEREG}{\sqrt{c^2*\tau^2+1}}\right)/(1+exp\left(\frac{wBEREG}{\sqrt{c^2*\tau^2+1}}\right))$, where *c* is constant equal to $16\sqrt{3}/(15\pi)$, and τ^2 is the variance of random intercept.

Word	***w* lex. constr.'s**	**Observed regularity rate**	**Predicted regularity rate**
reg1	0.69	0.999	0.999
...			
reg46	0.69	0.999	0.999
irreg47	-12.46	0.001	0.002
...			
irreg50	-12.46	0.001	0.002
Observed overall irregularity rate:			**0.080**
Predicted nonce irregularity rate (*w*BEREG = 6.17, τ^2 = 14.77):			**0.074**

Table 4. Hierarchical regression learns overall generalizations together with idiosyncrasy

Before moving forward, I note that learning simulations in Section 6.3 of Zymet (2018) suggest that hierarchical regression can accurately learn overall generalizations together with lexical idiosyncrasy even when the idiosyncrasy manifests as lexical propensities—gradient word-specific rates patterning across the complete spectrum. Even more simulations suggest that MaxEnt, on the other hand, is challenged by the task of learning overall generalizations together with lexical propensities (Zymet 2018, Section 6.2.2).

5 Application of hierarchical regression to lexically specific Slovenian palatalization

5.1 *Lexical specificity in Slovenian palatalization* Here we apply hierarchical regression to the case of variable velar palatalization in Slovenian, in which stems ending with a velar (*k*, *g*, *x*) palatalize before particular suffixes (obla**k**-a 'cloud'-GEN, obla**tʃ**-itsa 'cloud'-DIM; dow**g**-a 'long'-GEN, dow**ʒ**-ina 'length'). Only some suffixes trigger the process: Jurgec (2016), citing Toporišič (1976/2000), reports that only a handful of (common) suffixes of the 200 suffixes in the language trigger palatalization.

(5a)

Stem		Triggering suffix /-itsa/		Non-triggering suffix /-inja/	
lu**k**-a	port-GEN	lu**tʃ**-itsa	port-DIM	lu**k**-inja	port-DIM
bo**g**-a	god-GEN	bo**ʒ**-itsa	god-DIM	bo**g**-inja	god-DIM

The variation is very fine-grained. Different palatalizing suffixes carry lexical propensities—they trigger palatalization at different rates (Jurgec 2016, Zymet 2018), suggesting that suffix identity conditions variation. Given below are data from my Slovenian corpus (see Section 5.3) on the idiosyncratic propensities of different suffixes after the stem /luk/:

(5b)

/luk-itʃ/ port-DIM	/luk-ina/ port-ABS	/luk-itsa/, port-DIM
lu**tʃ**-itʃ, **18%** (558/3147)	lu**tʃ**-ina, **50%** (50/100)	lu**tʃ**-itsa, **98%** (39/40)
luk-itʃ, 82% (2589/3147)	luk-ina, 50% (50/100)	luk-itsa, 2% (1/40)

In addition, stems undergo at different rates before the same suffix, suggesting stem identity plays a role (Jurgec 2016, Zymet 2018). This can be observed below:

(6c)

Stem		Stem before diminutive -itsa		Stem status
obla**k**-a	'cloud'-GEN	obla**tʃ**-itsa	'cloud'-DIM	*Undergoer*
nɔ**g**-a	'leg'-GEN	nɔ**g**-itsa ~ nɔ**ʒ**-itsa	'leg'-DIM	*Vacillator*
ja**k**-a	'yak'-GEN	ja**k**-itsa	'yak'-GEN	*Non-undergoer*

The task taken up here is to capture general phonological trends emerging from these data, together with the kinds of lexical idiosyncrasy present therein.

5.2 *Jurgec (2016) on Slovenian palatalization* In his pioneering study of the palatalizing process, Jurgec (2016) extracted words with a velar-final stem and one of nine palatalizing suffixes from two dictionaries: the Dictionary of Standard Slovenian (Bajec 2000; 110,000 word types) and the Slovenian Orthographic Dictionary (Toporišič 2001; 130,000 word types). To obtain token rates for each word, he fed

them into Gigafida (Logar-Berginc et al. 2012), a text corpus with around 1.2 billion tokens from written sources ca. 1990–2011. Obtaining palatalization rates for his purposes was possible because palatalization is spelled in the Slovenian orthography. His resulting data set included ~5.7 million tokens—0.5% of all the tokens in Gigafida, a substantial portion of it, suggesting that palatalizing suffixes are frequent. Jurgec suggests that phonological factors condition variation in his data: suffixes with front vocoids trigger more regularly than those without; stems with final *k* and *g* undergo more regularly than those with final *x*; suffixes with ʦ trigger less regularly; palatalization regularly applies to avoid geminates (i.e., to /{k, g}+k/, where *-k* = -DIM); and palatalization can be blocked by distant postalveolars earlier in the stem. Jurgec indeed observes he suffix-specific rates in his study, but the account of them is left to further research. Jurgec gives a MaxEnt account of phonological conditioning, while suffix idiosyncrasy is accounted for with the binary feature [+/- Palatalization], picking out suffixes with any degree of palatalization.

5.3 *Building upon Jurgec (2016): a corpus investigation into Slovenian palatalization* This section gives preliminary results of a hierarchical regression model of the palatalization grammar together with the different kinds of lexical idiosyncrasy observed. It will be shown that the model learns phonological trends across the data together with lexical distinctions. The model results further suggest that stems and suffixes have lexical propensities: suffixes trigger at different rates, and stems undergo at different rates, patterning across an entire spectrum (e.g., [0.7 Palatalization]).

My data collection methods were similar to Jurgec's. Words with a velar-final stem and a palatalizing suffix were extracted from *Dictionary of Standard Slovenian*. Each extracted stem was concatenated with each of the nine suffixes, creating a set of hypothetically existent words to be fed into Gigafida. Extracting the frequencies for these words from Gigafida yielded a corpus totaling to around 3 million tokens, with each word having some palatalization rate.

Some basic calculations over the data give us the sense that stems and suffixes have idiosyncratic propensities to participate in palatalization. For each suffix *-X*, a suffix-specific palatalization rate was calculated by averaging over token rates for every *Y-X*, *Y* being a stem (e.g., /ag/ undergoes palatalization 22% of the time before /-je/, but /kak/, 99% of the time before /-je/; averaging over rates, we obtain 88% for /-je/). The figure below gives suffix-specific rates for the nine suffixes, together with their frequencies.

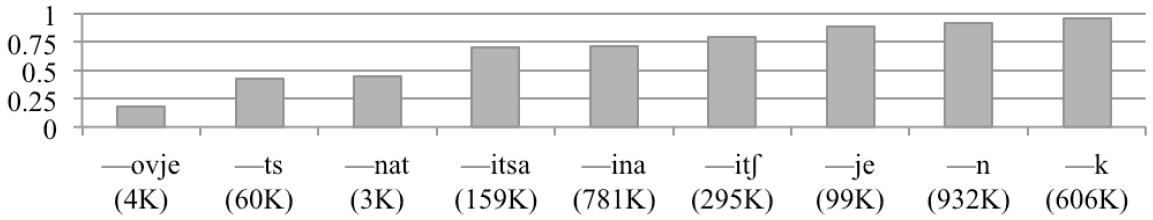


Figure 2a. Palatalization rates for each suffix

Stem-specific rates were calculated for 246 stems occurring before at least four suffixes. The figure below presents a histogram that plots palatalization rate against the frequency of stems with that rate.

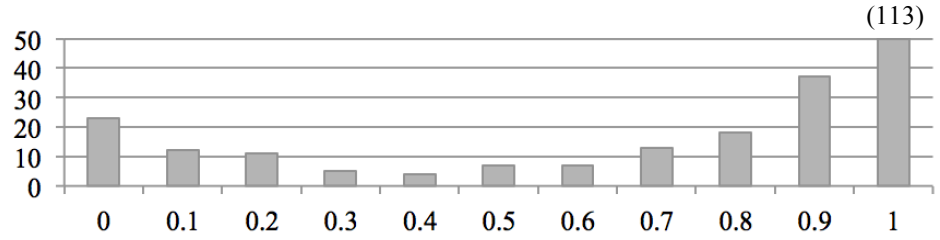


Figure 2b. Histogram of stem palatalization rate frequencies

In anticipation of the discussion below, I mention here two major phonological trends that emerged from my data: first, *k*-final and *x*-final stems palatalized at an overall higher rate than *g*-final stems (85%,

80%, and 70%, averaging over types);[6] and palatalization applied essentially categorically in the data to avoid a geminate (/{*k*, *g*}+*k*/).

We use a mixed-effects logistic regression model to learn overall generalizations in the data together with lexical idiosyncrasy.[7] Zymet (2018) (Section 3.2) used the Akaike Information Criterion to compare the performance of four mixed-effects logistic regression models of the data. The baseline model included Jurgec's factors as fixed effects and whole word as a random intercept. A second model additionally included stem identity as a random intercept, a third model additionally included suffix identity as a random intercept, and a fourth model additionally included both stem identity and suffix identity as random intercepts. The fourth model—which I call the Stem+Suffix Model—performed the best according to the criterion. Here I summarize the results of the Stem+Suffix Model (rerun in somewhat simplified form—see Footnote 9 below). The goal here is not to compare predicted rates against observed rates for the different subpopulations of the data, but simply to show that this model is promising for learning phonological trends together with lexical idiosyncrasy.

The model was run using the glmer functions of the *lme4* package in R. We include fixed effects for Jurgec's phonological factors: stem-final velar identity (*k, g, x*); whether the suffix begins with a front vocoid; whether the stem contains an earlier post-alveolar; and whether the suffix contains an alveolar affricate. Moreover, we include random intercepts for stem, suffix, and whole word.

Before proceeding, I briefly mention that the baseline model, which lacked random intercepts for stem and suffix, revealed that all of Jurgec's factors were significant predictors of variation with the exception of suffix-initial front vocoids. Coefficients of the significant predictors went in the direction anticipated in Jurgec (2016).

The results of the Stem+Suffix Model are given below. Stem-final velar identity was a significant predictor, with coefficients indicating that *k*, *x* > *g*; in addition, underlying geminates were a significant predictor, with the coefficient indicating that palatalization applies more consistently to avoid them. The other predictors, not shown, were not significant.[8] The model was rerun with non-significant predictors removed one by one in descending order of *p*-value, until only the two significant predictors mentioned above remained, as shown below. Remarkably, all three random intercepts had positive variance, suggesting that lexical variability is attributed to individual morphemes and whole words simultaneously; in particular, some of it is attributable to stems, some if it to suffixes, and some of it to whole words.[9]

Random int.	**Variance**	**Fixed effects**	**Estimate**	**Std. err.**	***z*-value**	***p***
Word (4822)	43.31	Intercept (ref.: stem-final *g*)	1.58	1.83	0.86	0.679
Stem (2720)	62.20	Stem-final *x*	2.26	0.95	2.39	0.017
Suffix (9)	26.93	Stem-final *k*	2.52	0.65	3.85	< 0.001
		/{*k*, *g*}+*k*/	7.57	1.24	6.09	< 0.001

Table 5. Stem+Suffix Model results for Slovenian palatalization

This model learned the two major phonological trends previously aforementioned: namely, that *k*- and *x*-final stems palatalize frequently relative to *g*-final stems; and that palatalization applies more often where a geminate would be avoided. In addition, the model learned lexical distinctions. In particular, the coefficients of the random intercept for suffix track the coarse suffix rates calculated in Table 2a. This can be observed in the table below.

[6] This can be seen as an effect of faithfulness to [continuant]—whereas *k* maps to *ʧ* and *x* to *ʃ*, *g* maps to *ʒ*.

[7] These models were run using glmer's nloptwrap optimizer and with nAGQ = 0 to speed up the estimation process (Bates & Maechler 2011).

[8] However, the factor for earlier postalveolar in the stem was close-to-significant in my simulations (roughly $p = 0.1$). It had a negative coefficient, predicting that earlier postalveolars inhibit palatalization, in line with Jurgec (2016).

[9] The models run in my prior studies included log(word frequency) as a fixed effect. The effect was statistically significant, but with an extremely small positive coefficient; rounding the log values to the nearest hundredths place resulted in an extremely small negative coefficient. Word frequency was thus excluded from the models run here.

	-ovje	**-ts**	**-nat**	**-itsa**	**-ina**	**-itʃ**	**-je**	**-n**	**-k**
Suff. rate	0.18	0.42	0.44	0.70	0.71	0.79	0.88	0.91	0.95
Suff. coef.	-4.82	-2.75	-1.02	0.22	0.66	1.47	3.31	3.11	-0.19

Table 6. Suffix coefficients track coarse suffix rates

Note that the high observed rate for /-*k*/ is predicted primarily by the geminate avoidance coefficient, rather than by the suffix coefficient for /-*k*/ alone.

My modeling of the Slovenian data is presently a work in progress, and so these results should be taken as preliminary. Future investigations will yield how closely the mixed model predicts rates in various subpopulations of data delimited by phonological factors, lexical factors, or both. Nonetheless, the results are promising, suggesting that the hierarchical regression model can learn phonological generalizations across the Slovenian data, as well as capture lexical distinctions therein.

6 Conclusion

The literature suggests that speakers internalize a nested hierarchy of generalizations: they generally can frequency-match to aggregate statistical generalizations across the lexicon, but also know which words are idiosyncratically exceptional, and which are not. MaxEnt does not recognize the concept of hierarchicality of generalizations, as it is rooted in single-level regression, and is therefore challenged by the task of accurately learning multiple levels of generalization. Group-specific constraints come to explain the entire dataset, rendering the general constraint ineffective for modeling the overall statistical generalizations. Hence the cost of learning lexical idiosyncrasy is an inability to learn a frequency-matching grammar, leading to the grammar-lexicon balancing problem. I suspect this problem is broader than just lexical variation: if the learner knows two groups of data have different rate, and averages over these rates when encountering novel data lying outside both groups, then how could we model this averaging if we have an accurate model of the group rates? For a hierarchy of generalizations we can implement a hierarchical theory, rooted in mixed-effects logistic regression, which surmounts the balancing problem. As it pertains to the relationship between grammar and lexicon, this model dictates that idiosyncratic effects of vocabulary are subordinated to the broader effects of grammar. Prior studies into the phonology and the lexicon have suggested hierarchical regression as potential model; here I give an argument for why it should be our theory of language learning and competence.

References

Bajec, Anton. 2000. *Slovar slovenskega knjižnega jezika: Electronic edition.* Ljubljana: SAZU and Fran Ramovš Institute for the Slovenian Language.

Bates, Douglas. 2009. *Linear mixed model implementation in lme4*. Ms., University of Wisconson—Madison.

Bates, Douglas & Martin Maechler. 2011. Package 'lme4'. R.

Baayen, R. Harald, Dogulas J. Davison, & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4): 390-412.

Becker, Michael, Andrew Nevins & Nihan Ketrez. 2011. The Surfeit of the Stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87(1): 84-125.

Becker, Michael, Andrew Nevins & Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88(2): 231-268.

Ernestus, Mirjam & R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1): 5-38.

Fruehwald, Josef T. 2012. Redevelopment of a Morphological Class. In Mao-Hsu Chen, Aaron Ecay, Sabriya Fisher, Aaron Freeman, Lauren Friedman, Kyle Gorman, Anton Ingason, Marielle Lerner, Laurel MacKenzie, Hilary Prichard, & Kobey Schwayder (eds.) *University of Pennsylvania Working Papers in Linguistics* 18(1).

Fylstra, Daniel, Leon Lasdon, John Watson & Allan Waren. 1998. Design and use of the Microsoft Excel solver. *Interfaces* 28(5): 29-55.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*.

Hayes, Bruce & Zsuzsa Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23(1): 59-104.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3): 379-440.

Hayes, Bruce, Kie Zuraw, Peter Siptar & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4): 822-863.

Hughto, Coral, Andrew Lamont, Brandon Prickett, & Gaja Jarosz. 2019. Learning Exceptionality and Variation with Lexically Scaled MaxEnt. In Gaja Jarosz, Max Nelson, Brendan O'Connor & Joe Pater (eds.) *Proceedings of the Society for Computation in Linguistics* 2(11).

Jarosz, Gaja & Amanda Rysling. 2017. Sonority sequencing in Polish: The combined roles of prior bias and experience. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert Staubs (eds.) *Proceedings of 2016 Meetings on Phonology*. Washington, DC: Linguistic Society of America.

Johnson, Daniel Ezra. 2013. Progress in regression: Why sociolinguistic data calls for mixed-effects models. Ms.

Jurgec, Peter. 2016. Velar palatalization in Slovenian: Local and long-distance interactions in a derived environment effect. *Glossa* 1(1): 24.

Logar-Berginc, Nataša, Simon Krek, Tomaž Erjavec, Miha Grčar, Peter Halozan & Simon Šuster. 2012. Gigafida corpus.

Mallet, Géraldine. 2008. *La liaison en français: Descriptions et analyses dans le corpus PFC.* Doctoral dissertation, Université Paris Ouest, Nanterre La Défense.

Moore-Cantwell, Claire & Joe Pater. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics* 15: 53-66.

Pavlou, Menelaos, Gareth Ambler, Shaun Seaman & Rumana Z. Omar. 2015. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Medical Research Methodology* 15:59.

Raudenbush, Stephen W., & Anthony S. Bryk, 2012. *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Thousand Oaks: Sage Publications.

Shih, Stephanie. 2018. Learning lexical classes from variable phonology. In Yuki Seo and Haruya Ogawa (eds.) *Selected Papers from Asian Junior Linguists Conference* 2: 1-15. ICUWPL.

Shih, Stephanie & Sharon Inkelas. 2016. Morphologically-conditioned tonotactics in multilevel Maximum Entropy grammar. In Gunnar Hansson, Ashley Farris-Trimble, Kevin McMullin, & Douglas Pulleyblank (eds.) *Proceedings of the 2015 Annual Meeting on Phonology*. Washington, DC: Linguistic Society of America.

Skrondal Anders & Rabe-Hesketh, Sophia. 2009. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Statistics in Society, Series A* 172(3): 659-687.

Smith, Brian W. & Claire Moore-Cantwell. 2017. Emergent idiosyncrasy in English comparatives. In Andrew Lamont & Katie Tetzloff (eds.) *NELS 47: Proceedings of the 47th meeting of the North East Linguistic Society*. Amherst: Graduate Linguistic Student Association.

Snijders, Tom & Roel Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis, 2nd edition*. Sage.

Tanaka, Yu. 2017. *The sound pattern of Japanese surnames.* Doctoral dissertation, UCLA.

Tay, Kim Gaik, Sie Long Kek, & Rosmila Abdul-Kahar. 2012. A Spreadsheet Solution of a System of Ordinary Differential Equations Using the Fourth-Order Runge-Kutta Method. *Spreadsheets in Education (eJSiE)* 5(2): 5.

Toporišič, Jože. 1976/2000. *Slovenska slovnica*. Maribor: Obzorja.

Toporišič, Jože (ed.). 2001. *Slovenski pravopis*. Ljubljana: SAZU.

Zeger, Scott L., Kung-Yee Liang, & Paul S. Albert. 1998. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44(4): 1049-1060.

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology.* Doctoral dissertation, UCLA.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28(2): 417-472.

Zuraw, Kie & Hayes, Bruce. 2017. Intersecting constraint families: An argument for harmonic grammar. *Language* 93(3): 497-548.

Zymet, Jesse. 2018. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning.* Doctoral dissertation, UCLA.