Jesse Zymet
AMP ’18 Talk

Learning a frequency-matching grammar together with lexical idiosyncrasy: MaxEnt versus mixed-effects logistic regression

Jesse Zymet, UC Berkeley

SPEAKERS KNOW AGGREGATE GENERALIZATIONS *AND* IDIOSYNCRASIES

1. Language learners *frequency match* to statistical generalizations across the lexicon[1]

- E.g., Hungarian vowel harmony (Hayes & Londe 2006): dative forms takes *-nɛk* or *-nɔk*, depending on backness of preceding stem vowel. Stems ending in...
  - front V tend to take *-nɛk*: [kɛ̱rt-nɛ̱k] ‘garden’-DAT, [y̱ʃt-nɛ̱k] ‘cauldron’-DAT
  - back V tend to take *-nɔk*: [ɔblɔ̱k-nɔ̱k] ‘window’-DAT, [bi:ro̱:-nɔ̱k], ‘judge’-DAT

- Corpus study of monosyl. stems ending in front, unrounded V: **92%** take *-nɛk*; **8%** *-nɔk*.

- In wug tests, speakers presented with fake monosyllabic stems with a front unrounded vowel, in aggregate, ***closely frequency-matched to the 8% -nɔk rate***.

(1)

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

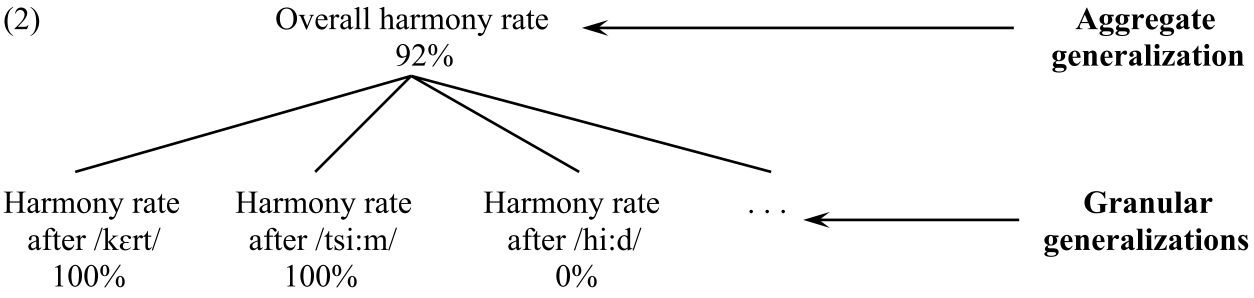
(Hayes & Londe 2006)

2. But language learners also know lexical idiosyncrasies

- Speakers know *which* attested words harmonize, versus not (Hayes & Londe 2006).
- French speakers even track *morpheme-specific rates* of liaison (Zymet 2018).

3. Language learners thus internalize *nested hierarchy of generalizations*:

(2)



Overall harmony rate
92%

Aggregate generalization

Harmony rate after /kɛrt/ 100%

Harmony rate after /tsi:m/ 100%

Harmony rate after /hi:d/ 0%

. . .

Granular generalizations

- Recent MaxEnt model (Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017):
  - General constraint to frequency match general trend across the lexicon (HARMONIZE)
  - Lexical constraints for specific attested words (HARMN(kɛrt), DON’T-HARMN(hi:d))

[1] Frisch, Broe, & Pierrehumbert 1996; Coleman & Pierrehumbert 1997; Eddington 1998, 2004; Berkley 2000; Zuraw 2000, 2010; Bailey & Hahn 2001; Frisch & Zawaydeh 2001; Albright 2002; Albright & Hayes 2003; Ernestus & Baayen 2003; Hayes & Londe 2006; Becker 2009; Hayes, Zuraw et al. 2009; *et seq*.

4. Today: Modeling learning of frequency-matching grammar with lexical idiosyncrasy

- Learning simulations reveal that **lexical constraints are too powerful in MaxEnt**:
  - o A priori, general constraint and set of lexical constraints considered *equally viable* hypotheses about the data in MaxEnt;
  - o at high levels of learning, lexical constraints come to explain every form in dataset, rendering the general constraint superfluous and ineffective.
  - o General constraint weight plummets to zero, failing to predict learners’ frequency-matching abilities in wug tests.

- **Solution**: Switch from MaxEnt—essentially single-level logistic regression model—to hierarchical **MIXED-EFFECTS LOGISTIC REGRESSION MODEL**.
  - o General/lexical constraints no longer equal: general constraints preserved as fixed effects; lexical constraints form random effect.
  - o Hierarchical model captures *hierarchy of generalizations*: aggregate trend + idiosyncrasies of individual words.
  - o We apply mixed model to variable Slovenian palatalization—with promising results.

MAXENT: THE GRAMMAR-LEXICON BALANCING PROBLEM

5. MaxEnt (Smolensky 1986, Goldwater & Johnson 2003, Hayes & Wilson 2008, *et seq*)

- Constraints have numerical weights instead of rankings;
- surface forms assigned probabilities as function of weights.
- Learning rooted in *accuracy* and *simplicity*: model takes constraints, finds best weights it can to fit overall rates in dataset; useless constraints discarded—weight set to zero.

❖ But MaxEnt fits to *overall rates*; investigators hadn’t tried to get MaxEnt to also learn which words are un/exceptional until recently. The new approach:
  - o General constraints for overall trend, lexical constraints for specific-word behavior
  - o Moore-Cantwell & Pater (2016), Zuraw & Hayes (2017), Tanaka (2017), *inter alia.*

6. Does the MaxEnt approach to learning frequency matching & idiosyncrasy work?

- Suppose we have 46 regulars, 4 irregulars—irregularity rate of **8%**.
- 3 constraints: BEREG, BELEX(regulars), BELEX(irregulars) initiated at 0 weight
- If we want to learn the dataset better? Multiply frequencies by 10. Worse? By 0.1.
- (Caveat: introduced *a little* variability: 0.001% /regs/ surface [irreg]; 0.001% /irregs/ as [reg])

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

Table 1: *MaxEnt input*

- We want MaxEnt to learn weights such that:
  - in wug test, irregular form picked **~8%** of time;
  - *attested* words (=words in learner input) are pronounced correctly ~100% of time.
  - *w*(BEREG) = 4, *w*(BELEX-reg) = 3, *w*(BELEX-irreg) = 11 gives great results.

- But does MaxEnt *learn* good weights from the input? Let's run learning simulation using Excel Solver, which can fit parameters of nonlinear models (Fylstra et al. 1998, Harris 1998):
  - Trial run by using data in Table 1,
  - and multiplying frequencies of the dataset by a small factor (0, 0.001, etc.)—we call this "childhood". We learn poor weights (*w*(BELEX-irreg)=0.5) that don't fit the data.
  - After each trial, increase frequency factor slightly, get new weights—"adolescence".
  - When frequencies get large and we think we have final weights—"adulthood".

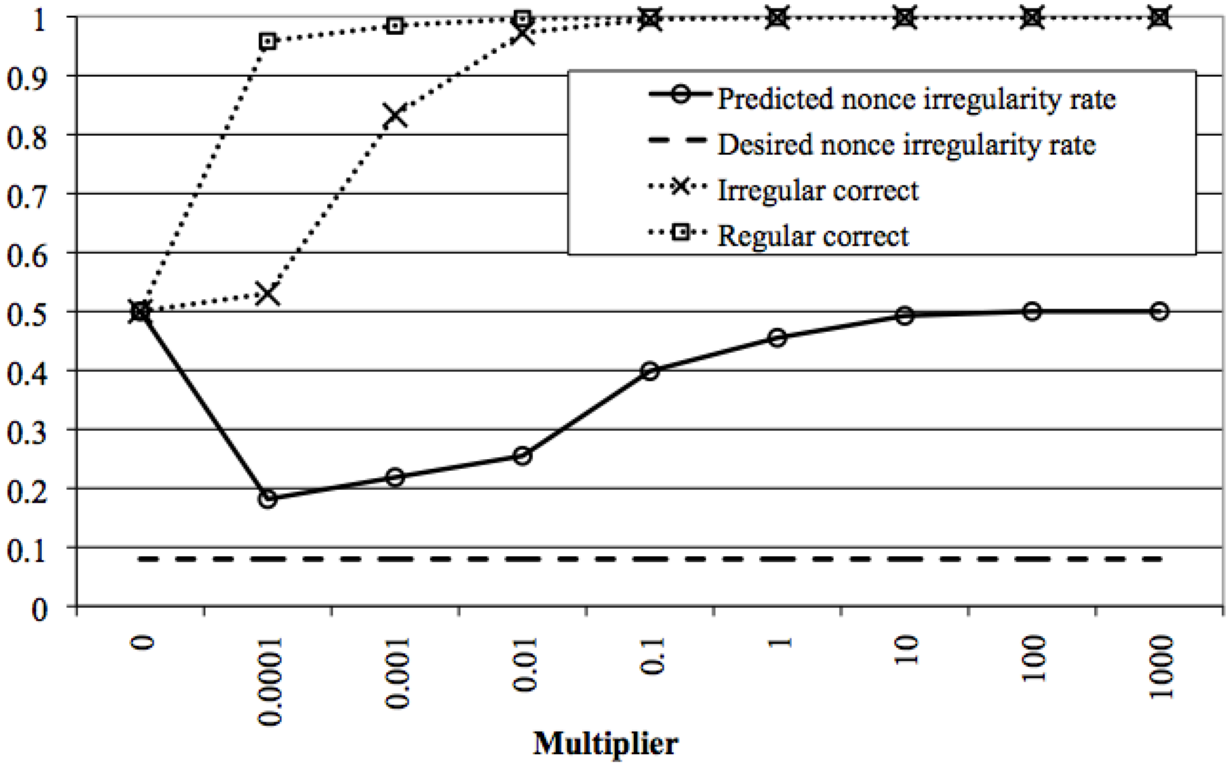


Figure 1: *MaxEnt fails to learn generalization together with idiosyncrasy* (σ = 100)

- With frequency factor 0, baby learns 0-valued weights, prefers 50/50 regular/irregular.
- As child grows (freq. factor 0.0001), rapidly starts to learn regulars, slowly tackling irregulars.
- Early in learning, BEREG is used to explain much of the variation—we see that with low nonce irregularity rate.
- But eventually BELEX constraints grow very high, coming to explain entire set of attested data. BEREG comes to explain increasingly *less* of data, eventually *perishing*.
- By adulthood (freq. mult. 1000), **BEREG sinks to 0, rendered superfluous/ineffective.**
- At that point, the learner selects regulars/irregulars at 50/50 rate in wug tests—forgetting the grammar entirely. See Appendix for simulation output numbers.

7. Hence, the GRAMMAR-LEXICON BALANCING PROBLEM. In MaxEnt...

- A priori, general constraint/set of lexical constraints *equally viable hypotheses* about data,
- **consequently lexical constraints too powerful**: lexical constraints learn each word's behavior before general constraint matches overall 8% trend, at which point frequency matching ceases and the general constraint becomes ineffective.
- No phonological learning; just lexical learning. Implausible that speakers fail wug tests once they learn lexicon (see Shademan 2007 for learning in elderly).
- We need a theory that, while accounting for idiosyncrasies, ***preserves grammar***.

- We search for model possessing **GENERALITY BIAS**: general, grammatical constraints must be privileged to lexical constraints in the learning process.
  - Adjusting MaxEnt penalty term does not work: dividing σ's by 10 = dividing freq. multiplier by 100—merely *delays* overfitting (see Appendix).
  - High σ(BEREG)/low σ(BELEX) so far does not work; overfits at higher multiplier.

LEARNING LEXICAL VARIATION WITH MIXED-EFFECTS LOGISTIC REGRESSION

8. What about the hierarchical MIXED-EFFECTS LOGISTIC REGRESSION model?

- Similar to binomial logistic regression, except constraints hierarchically arranged as follows:
  - **Fixed effects**: those constraints that we are actually interested in—e.g., phonological constraints, yielding the statistical generalizations in the dataset
  - **Random effects**: constraints that capture the idiosyncrasies in the data—deviations from generalizations captured by fixed effects.
  - We might call this **Mixed Effects Maximum Entropy Harmonic Grammar**.
- Used widely in science to capture trends & idiosyncrasies in variable datasets;
- Linguists employ random intercepts to measure by-word/lexical class idiosyncrasy (Fruehwald 2012, Zuraw & Hayes 2017, Smith & Moore-Cantwell 2017, *inter alia*);
- Shih & Inkelas (2016)/Shih (2018) even adopt multilevel model as theory of learner.

9. Mixed models *hierarchical*: random effects "depreciated" relative to fixed effects

We have a fixed effect—a general constraint—BEREGULAR, whose weight is estimated based on average harmony rate across the entire dataset—**92%**.

(3a) *w*(BEREG): $\mu_{all\ words}$

- We want this weight to accurately estimate the average rate across all words, as that would be a **frequency-matching grammar**, mimicking human behavior in wug tests.

We have a random effect (random intercept) consisting of weights for lexical constraints:

- *w*BELEX–irreg1, for example, estimated by rate irregular1 (0.001) ...
- ***and* by overall rate across dataset**:

(3b) *w*(BeLex–irreg1): $\lambda_{irregular1} * \mu_{irregular1} + \left(1 - \lambda_{irregular1}\right) * \mu_{all\ words}$

Raudenbush & Bryk (2012), Snijders & Bosker (2012)

- λ: value between 0 and 1, depends on size of the group: *w*(BeLex–irreg1) will be determined more by $\mu_{irregular1}$ if data have lots of irreg1 tokens rather than few.
    - Predicts more idiosyncrasy with frequent forms, but more grammatical behavior with infrequent forms (Morgan & Levy 2016, Moore-Cantwell & Smith 2016).
- **Think of mixed models as follows**: fixed effect weights predicts overall rate, and random effect weights predict ***word-specific offsets*** from overall rate.
- Source of the generality bias: lexical constraint weights *depend* on overall average rate.

10. Mixed model performs well on strict exceptionality dataset

We want the learning model to predict:

- With BeReg, the average rate across all Words in the dataset—hence a frequency-matching grammar
- With BeLex–reg/BeLex–irreg, the specific rates for every word.

We run a model of the dataset using the `glmer` function of the *lme4* package R.

- weight of BeReg is the general intercept
- weight of BeLex constraints are the coefficients of the levels of the random intercept.

To extract predicted nonce rate from model, you cannot simply plug *w*BeReg into inverse logit—rather, you must "average" over the levels of the random intercept (Pavlou et al. 2015).

- This involves a complex integral that cannot be calculated analytically;
- Zeger et al. (1998) provide a good approximation:
    - *c* is constant equal to $\frac{16\sqrt{3}}{15\pi}$
    - τ^2 is variance of random intercept (14.77)

(4)

$$\frac{\exp\left(\frac{w\text{BEREG}}{\sqrt{c^2\tau^2+1}}\right)}{1+\exp\left(\frac{w\text{BEREG}}{\sqrt{c^2\tau^2+1}}\right)}$$

(5) **Results**:

Word	*w*BELEX	***Actual rate***	***Predicted rate***
reg1	0.69	**0.999**	**0.999**
reg2	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**

***w*BEREG = 6.167**

OVERALL IRREGULARITY RATE: 8%
PREDICTED NONCE IRREGULARITY RATE: 7.4%

This model:

- Predicts word-specific rates—learns lexical effects.
- Frequency-matches overall rate—mimicking subjects in wug tests—without lexical constraints *starving* general constraint. ***Grammar sustained after lexical learning.***

11. Mixed model performs well on dataset with different lexical rates

- Consider the following: twelve words, each with 1000 tokens, with the different tokens undergoing, say, harmony, at different rates.

Word	Rate	Word	Rate	Word	Rate
1	0.00	5	0.30	9	1.00
2	0.00	6	0.80	10	1.00
3	0.10	7	0.90	11	1.00
4	0.20	8	1.00	12	1.00

Average over all rates: **0.61**

Table 2: *propensity dataset*

- Two kinds of constraints:
  - APPLY (HARMONIZE), whose weight should frequency match to **61%** overall rate
  - APPLY-Word1, ..., APPLY-Word12, assists with specific rates

(6) **Results:**

Word	*wLex constr.*	***Actual rate***	***Predicted rate***
Word1	-16.56	**0.000**	**0.000**
Word2	-16.56	**0.000**	**0.000**
Word3	-7.32	**0.100**	**0.100**
Word4	-6.51	**0.200**	**0.200**
Word5	-5.97	**0.300**	**0.300**
Word6	-3.74	**0.800**	**0.800**
Word7	-2.93	**0.900**	**0.900**
Word8	7.14	**1.000**	**0.999**
Word9	7.14	**1.000**	**0.999**
Word10	7.14	**1.000**	**0.999**
Word11	7.14	**1.000**	**0.999**
Word12	7.14	**1.000**	**0.999**

***w*HARMONIZE = 5.130**

OVERALL AVERAGE APPLICATION RATE: 0.61
PREDICTED APPLICATION RATE TO NONCE WORDS: 0.66

- I tried MaxEnt on this dataset:
  - Outcomes similar to other dataset, except *w*APPLY vacillates/plummets to 0 at high levels of lexical learning—see Appendix.
  - See Zymet (2018) for further details.

APPLYING THE MIXED MODEL TO VARIABLE SLOVENIAN PALATALIZATION

- For example, only *some* suffixes trigger it.

(7a)

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

  - From Toporišič (1997/2000): Of 200 suffixes, only a handful trigger palatalization.

- Different palatalizing suffixes trigger at different rates, suggesting suffix identity plays role:

(7b)

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

- *Stems* undergo at different rates before same suffix, suggesting stem identity plays role.

(7c)	Stem		Stem before diminutive -it͡sa		
	obla**k**-a	'cloud'-GEN	obla**tʃ**-it͡sa	'cloud'-DIM	*Undergoer*
	nɔ**g**-a	'leg'-GEN	nɔ**g**-it͡sa ~ nɔ**ʒ**-it͡sa	'leg'-DIM	*Vacillator*
	ja**k**-a	'yak'-GEN	ja**k**-it͡sa	'yak'-GEN	*Non-undergoer*

10. Jurgec (2016) on Slovenian palatalization

- Jurgec extracted words with velar-final stem + palatalizing suffix from two dictionaries:
  - *Dictionary of Standard Slovenian* (Bajec 2000; 110,000 word types)
  - *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):
  - Text corpus w/ ~1.2 billion tokens from written sources ca. 1990–2011.
  - His resulting data set included ~5.7 million tokens.

- Jurgec suggests phonological factors condition variation in his data:
  - Suffixes with front vocoids trigger more regularly
  - Velars *k*, *g* undergo more regularly than *x*.
  - Suffixes with *t͡s* trigger less regularly.
  - Palatalization regularly applies to avoid geminate in /...{k, g}+k/ (-*k* = -DIM)
  - Palatalization blocked by distant postalveolars earlier in the stem.

- Jurgec gives MaxEnt account of *phonological* conditioning; suffix idiosyncrasy encoded with **[+/- Pal'n]**—only picks out suffixes with *any degree* of palatalization.
  - But he *does* observe suffix-specific rates in his study—lexical propensities left to further research.

12. Building upon Jurgec (2016): a corpus investigation into lexical propensities

- I show that:
  - Morphemes have **LEXICAL PROPENSITIES**: suffixes trigger at different rates, and stems undergo at different rates, patterning across an entire spectrum (**[0.7 Pal'n]**).
  - Mixed model encodes propensities while frequency matching to trends.

- Extraction method similar to Jurgec:
  - Words consisting of velar-final stems + palatalizing suffix extracted from *Dictionary of Standard Slovenian*.
  - Each extracted stem concatenated with each of nine suffixes, creating hypoth. words
  - Fed each word into *Gigafida*, extracting frequencies/token rates
  - Yielded ~3 million tokens of words either undergoing/not undergoing palatalization

- I calculated palatalization rates for each suffix. /ag/ undergoes 22% of time before -/je/, /kak/ 99% of time; average rate before -/je/ is 88%.

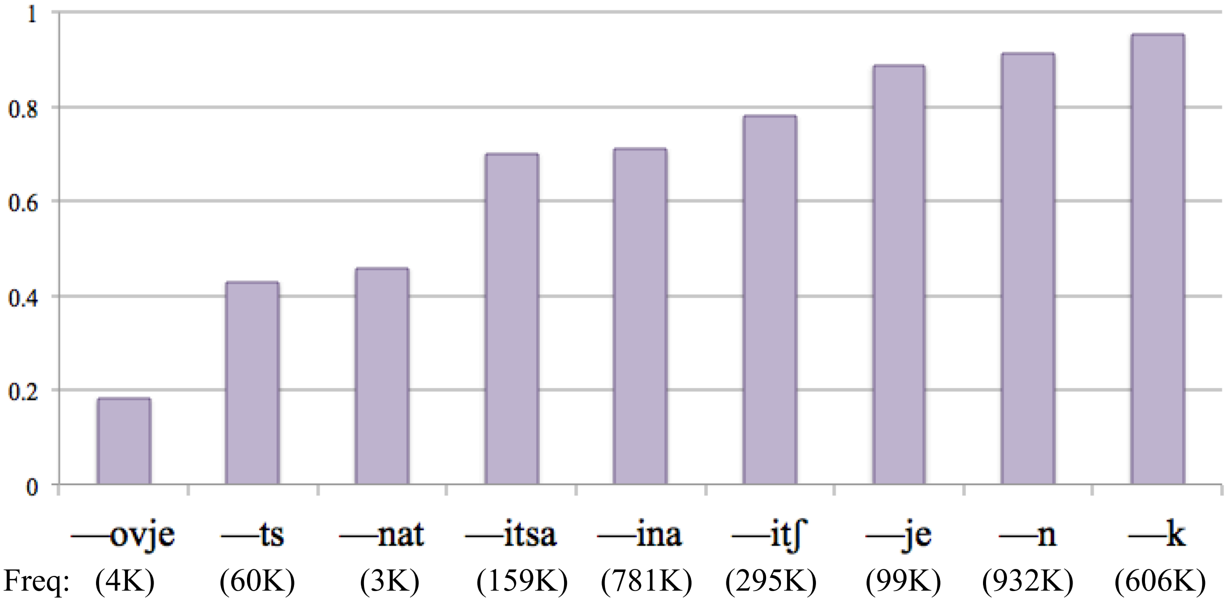


Figure 2a: *palatalization rates for each suffix*

What about stems? A histogram of rates across 246 stems occurring before at least four suffixes:

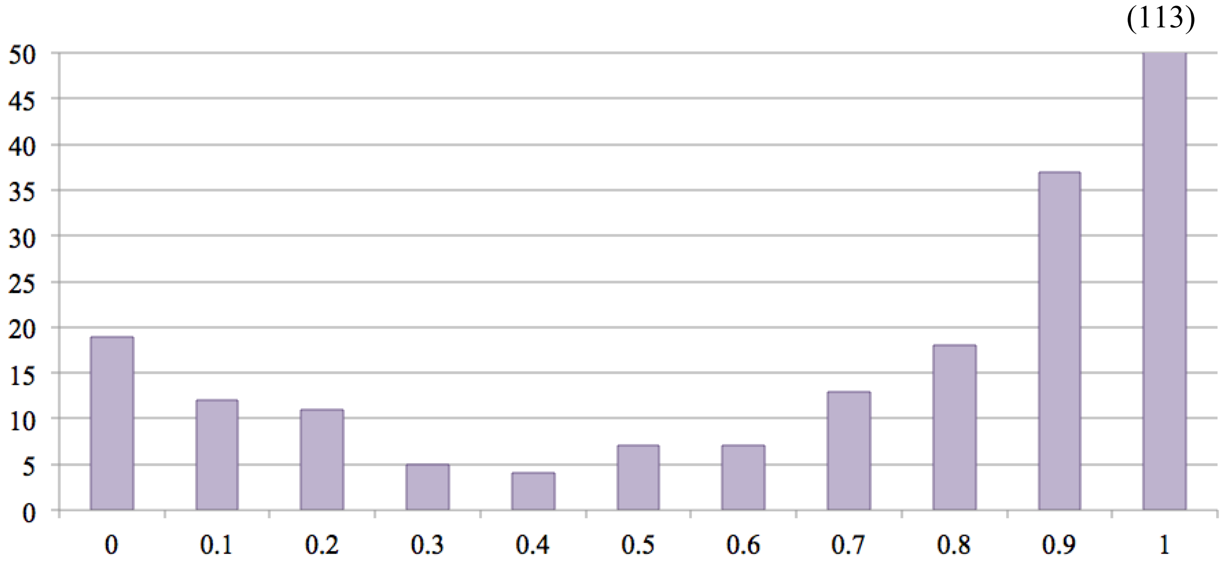


Figure 2b: *Histogram of stem palatalization rate frequencies*

- Results suggest morphemes have **LEXICAL PROPENSITIES**: suffixes trigger at different rates, and stems undergo at different rates, patterning across an entire spectrum.

- We use mixed-effects logistic regression to encode morphemes on a spectrum (**[0.7 Pal’n]**)—significantly improves model fit relative to binary scale (**[+/- Pal’n]**).
  - Models run using `glmer` functions of *lme4* package (Bates & Maechler 2011) in R.

- In this handout, we focus on/compare performance of following logistic models:
  - **Baseline Model**, containing fixed effects for:
    - Stem-final velar identity (*k, g, x*)
    - Whether suffix begins with a front vocoid
    - Whether stem contains an earlier post-alveolar
    - Whether the suffix contains a post-alveolar affricate
    - (Contains random effect for whole word; thus we’re regressing over frequency-weighted types.)
  - **Stem+Suffix Model**, containing:
    - all factors in Baseline Model
    - plus stem identity and suffix identity, ***encoded as random intercepts***.

- Models compared using Akaike Information Criteria (AIC; Akaike 1973), which scores models based on number of parameters and fit to the data: **lower score = better**.
  - See Bolker et al. (2009) for justification on using this to compare mixed models.

13. Results of Baseline Model

- Stem-final velar identity significant: *k* > *g* (seemingly a faith effect: k → tʃ but g → ʒ)
- Geminate avoidance significant
- ʃ...tʃ+ avoidance significant
- Suffix with *ʦ* significantly associated with lower rates
- `frontvocoid` not significant
- AIC: **8767.8**

```
             Estimate Std. Err.   z value       p
Intercept      3.95    0.47         8.29     <0.001 ***
ref: consg
consx          1.59    0.66         2.41      0.015 *
consk          1.96    0.47         4.11     <0.001 ***
kk             4.94    0.80         6.12     <0.001 ***
frontvocoid   -0.52    0.39        -1.32      0.183 (n.s.)
S...S         -1.67    0.78        -2.12      0.033 *
suff.with.ts  -3.53    0.46        -7.55     <0.001 ***
```

Output 1a: *Baseline Model results for Slovenian palatalization*

13. Results of Stem+Suffix Model

- Significant *k* > *g* effect and geminate effect, but no ʃ...tʃ+ effect or suffix-with-*ʦ* effect
- **Stem and suffix variances highly positive**—suggest stem and suffix condition variation
- Stem variance bigger than suffix variance: maybe undergoers louder than triggers; or linearly-first-element bias; or just relative morpheme counts. Feel free to ask in Q&A.
- AIC value: **7801.5** — substantial reduction from Baseline Model’s **8767.8**;

```
Random effects:
 Groups Name        Variance Std.Dev.
 stem   (Intercept) 68.06    8.25
 suffix (Intercept) 19.54    4.42
Number of obs: 2940918; words: 4822; stems: 2720; suffixes: 9

Fixed effects:
              Estimate Std. err. z value   p
Intercept:     1.15      2.24       0.51   0.60
ref: consg
-------------------------------------------------------
consx          2.36      1.00       2.35   0.02     *
consk          2.59      0.69       3.75  <0.001    ***
-------------------------------------------------------
k+k            7.94      1.32       6.01  <0.001    ***
-------------------------------------------------------
frontvocoid    2.72      3.01       0.90   0.366
S...S         -1.20      1.12      -1.06   0.284
suff.with.ts  -1.88      3.58      -0.52   0.598
```

Output 1b: *Stem+Suffix Model results for Slovenian palatalization*

14. AICs suggest suffix *and* stem identities matter ($p < 0.001$ by likelihood ratio test)

- Baseline Model AIC: **8767.8**
- Suffix-Only Model AIC: **8283.7**
- Stem-Only Model AIC: **8128.9**
- Stem+Suffix Model AIC: **7801.5**
- See Zymet (2018) for further elaboration on all these models.

15. Mixed model learns *the phonology*, frequency matching to statistical trends

- Matching to overall palatalization rate for *k*-final stems, and *g*-final stems
- Predicts *k* > *g* effect
- Predicts geminate-avoiding palatalization

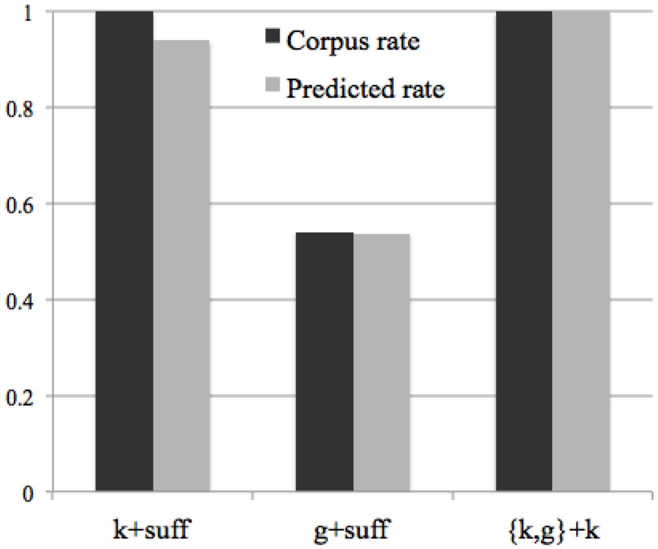


Figure 3a: *model succeeds in predicting phonological trends*

16. The mixed model learns lexical propensities

- Fares well in predicting suffix-specific rates:

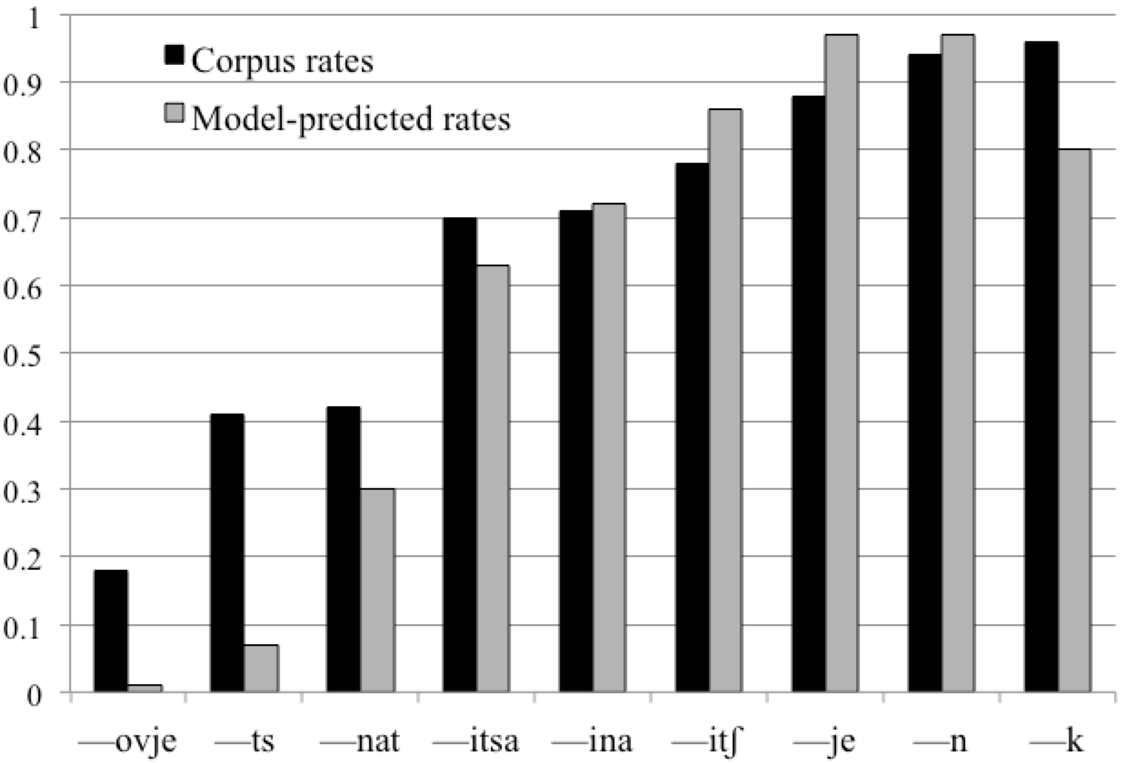


Figure 3b: *model-predicted suffix rates generally match corpus rates*

- I submit **mixed-effects logistic regression** as viable approach to modeling lexical variation—to learning of frequency-matching grammar with lexical propensities.

CONCLUSION

- Language learners internalize nested hierarchy of generalizations:
  - they can frequency match to aggregate statistical generalizations across the lexicon,
  - but also know which words are idiosyncratically exceptional, and which are not.

- MaxEnt/single-level regression doesn't recognize *hierarchicality of generalities*. I suspect problem is broader than just lexical variation:
  - If learner knows two groups of data have different rates,
  - and averages over rates when encountering novel data lying outside both groups,
  - then how could we model this averaging if we have accurate model of group rates?
  - MaxEnt: specific constraints enough to explain data, general constraint superfluous.

- Mixed-effects logit/mixed-effects MaxEnt surmounts balancing problem:
    - *Hierarchical* theory for a hierarchy of generalizations
    - Idiosyncratic effects of vocabulary subordinated to broad effects of grammar
    - Prior studies suggest it as potential model; today I give reason *why* this should be our theory of language competence.

- Future questions I hope to work on:
    - How should hierarchical theory look—e.g., how to plug random intercept into theory?
    - Exactly what constraints/kinds of constraints should be considered fixed vs. random?
    - How to expand mixed-effects MaxEnt to cover more than just the binomial case?
    - How to get >2 levels of generalization?

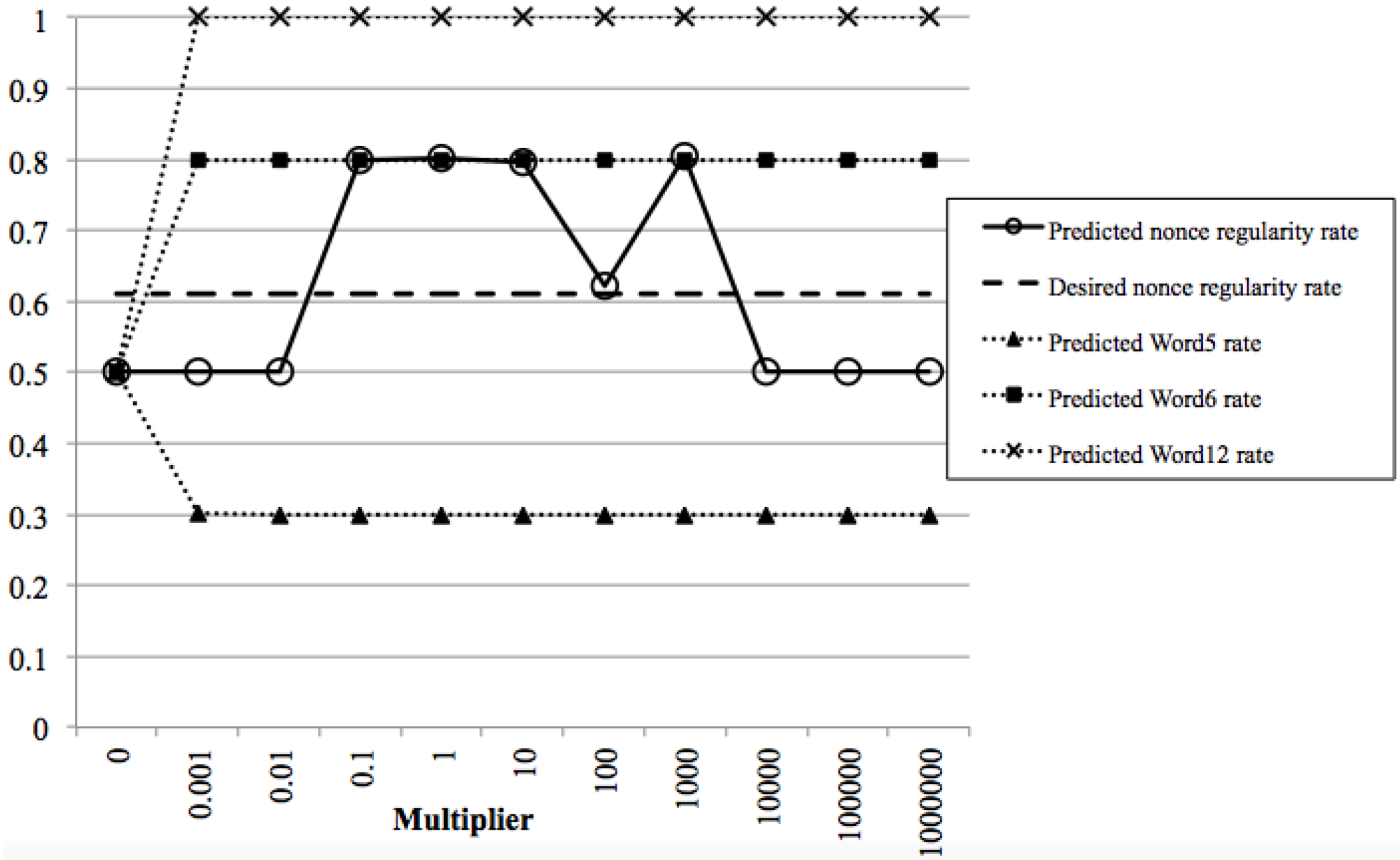
Appendix

FAILED LEARNING SIMULATION IN MAXENT (strict exceptionality, 8% irregularity rate)

Freq. multiplier	**Be Reg**	BeLex (regs)	BeLex (irregs)	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**	0.5304	0.9582	0.1815
0.001	**1.27**	**2.88**	**2.88**	0.8331	0.9845	0.2185
0.01	**1.07**	**4.63**	**4.63**	0.9723	0.9966	0.2548
0.1	**0.41**	**6.22**	**5.82**	0.9955	0.9986	0.3986
1	**0.17**	**6.69**	**6.78**	0.9986	0.9989	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**	0.9989	0.9990	0.5000

Table A1: *MaxEnt learning simulation output numbers for strict exceptionality data*

FAILED LEARNING SIMULATION IN MAXENT (propensities dataset, 61% applic’n rate)



Freq. multiplier	APPLY	FAITH$_5$	APPLY$_6$	APPLY$_{12}$	**Pred. nonce rate**	Pred. Word5 rate	Pred. Word6 rate	Pred. Word12 rate
0	0.00	0.00	0.00	0.00	**0.50**	0.50	0.50	0.50
0.001	0.00	0.84	1.39	11.19	**0.50**	0.30	0.80	1.00
0.01	0.00	0.85	1.39	10.64	**0.50**	0.30	0.80	1.00
0.1	1.39	2.23	0.00	8.77	**0.80**	0.30	0.80	1.00
1	1.40	2.25	0.00	11.85	**0.80**	0.30	0.80	1.00
10	1.37	2.22	0.02	16.17	**0.80**	0.30	0.80	1.00
100	0.50	1.34	0.89	8.25	**0.62**	0.30	0.80	1.00
1000	1.41	2.27	0.00	6.16	**0.80**	0.30	0.80	1.00
10000	0.00	0.85	1.39	12.60	**0.50**	0.30	0.80	1.00
100000	0.00	0.85	1.39	11.84	**0.50**	0.30	0.80	1.00
1000000	0.00	0.85	1.39	11.84	**0.50**	0.30	0.80	1.00

Table A2: *MaxEnt learning simulation output numbers for propensity data*

OVERFITTING OUTCOME GENERAL ACROSS MAXENT PENALTY SETTINGS

- E.g., multiplying σ's by 10 yields same result as multiplying frequency multiplier by 100.
- Evident in the table below, which presents results of a series of learning simulations of the strict exceptionality dataset from above (but only fitting the weight of BeReg to it).
- Hence decreasing σ merely has the effect of delaying learner overfitting

	σ = 1		**σ = 10**		**σ = 100**	
	irreg. rate	weight	irreg. rate	weight	irreg. rate	weight
***m* = 0.01**	0.4748	0.1008	0.1127	2.0629	0.0213	3.8258
***m* = 1**	0.1127	2.0629	0.0213	3.8258		
***m* = 100**	0.0213	3.8258				

Table : *identical learning outcomes across different values of m and* σ

- Manipulating μ also has no effect—yields same learning outcome as if we set μ = 0.
- What about high σ(BeReg) and low σ(BeLex)?
- I tried it on a few strict exceptionality datasets (but not including the one given in this handout...), and so far the results are negative:
- Setting σ = 1,000 for BeRegular and σ = 10 for the lexical constraints, for example, still yielded overfitting, albeit at a high frequency multiplier.

COEFFICIENTS FOR STEMS AND SUFFIXES IN SLOVENIAN

Suffix	Rate
-ovje	-4.05
-ina	-1.27
-nat	-0.40
-itʃ	-0.38
-ts	-0.16
-itsa	0.16
-k	0.58
-je	1.48
-n	4.03

Stems (sample)	Rate
trak-	-5.34
tramik-	0.00
tradicionalistik-	0.55
tragikomik-	1.14
travmatik-	1.30
tragik-	2.31

- Coefficients run the gamut, suggesting gradience.
- Suffix coefficients generally track suffix rates we saw toward the beginning.

References

Akaike, Hirotugu. 1973. Information theory and an extension of the maximum likelihood principle. in Petrov, B. N.; Csáki, F., 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, 267–281.

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. Language 78: 684-709.

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

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? Journal of Memory and Language 44, 568–591.

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

Bates, Douglas & Martin Maechler. 2011. Package ‘lme4’. R.

Becker, Michael. 2009. Phonological Trends in the Lexicon: The Role of Constraints. Doctoral Dissertation, University of Massachusetts, Amherst.

Berkley, Deborah Milam. 2000. Gradient obligatory contour principle effects. Doctoral dissertation, Northwestern University.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution, 24(3): 127–135.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop, ed. by John Coleman, 49–56. East Stroudsburg, PA: Association for Computational Linguistics.

Eddington, David. 1998. Spanish diphthongization as a non-derivational phenomenon, Rivista di Linguistica 10: 335-354.

Eddington, David. 2004. Spanish Phonology and Morphology: Experimental and Quantitative Perspectives. Amsterdam: John Benjamins.

Ernestus, Mirjam and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. Language 79, 5–38.

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

Frisch, Stefan A. & Zawaydeh, Bushra. 2001. The psychological reality of OCP-Place in Arabic. Language 77, 91-106.

Fruehwald, Josef T. 2012. Redevelopment of a Morphological Class. University of Pennsylvania Working Papers in Linguistics 18(1). Available at: https://repository.upenn.edu/pwpl/vol18/iss1/10.

Fylstra, D.; Lasdon, L.; Watson, J.; and Waren, A. 1998. Design and use of the Microsoft Excel solver. Interfaces, Vol. 28, No. 5, 29-55.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In the Proceedings of the Stockholm workshop on variation within Optimality Theory.

Harris, Daniel. 1998. Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver. Journal of Chemical Education 75(1).

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

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry, 39, 379–440.

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

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

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

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. http://www.gigafida.net: Amebis.

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, 2002. Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks: Sage Publications.

Shademan, Shabnam. 2007. Grammar and Analogy in Phonotactic Well-formedness Judgments. Ph.D. dissertation, University of California, Los Angeles.

Shih, Stephanie. 2018. Learning lexical classes from variable phonology. In Proceedings of AJL2.

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

Smith, Brian W. & Claire Moore-Cantwell (2017). Emergent idiosyncrasy in English comparatives. In Andrew Lamont and Katie Tetzloff, eds., NELS 47: Proceedings of the 47th meeting of the North East linguistic Society. Amherst: Graduate Linguistic Student Association. pp. 127-140.

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.

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, University of California, Los Angeles. ROA-788.

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

Zymet, Jesse. 2018. Lexical propensities in phonology: corpus and experimental studies, grammar, and learning. Ph.D. Dissertation, University of California, Los Angeles.