Over- and under-generalization in morphological learning[1]

Claire Moore-Cantwell
University of Massachusetts, Amherst
cmooreca@linguist.umass.edu

1 Overview

- In irregular morphological patterns, language learners must memorize the inflected forms for individual lexical items
- But they nevertheless form generalizations over those memorized forms (Zuraw, 2000, 2010; Albright and Hayes, 2003)
- The nature of these generalizations can be studied using a wug-test (Berko, 1958).
- I present evidence from a process of derivational morphology in Hebrew that langauge learners learn and employ multiple levels of generalization together
  - Conditional probabilities of derived form given phonological properties of the base
  - And type frequencies over derived form types, independently of any phonological properties of the base
- Lastly, I'll present a model that uses both types of knowledge to match speakers' performance on a wug-test

2 Levels of Generalization

1. Language learners' knowledge of their language involves extensive generalization as well as memorization.
   - In morphological patterns that show variability or exceptionality across but not within lexical items, derived forms must be memorized.
   - Several studies (Ernestus and Baayen, 2003; Hayes et al., 2009; Zuraw, 2000; Becker et al., 2011; Albright and Hayes, 2003) demonstrate that speakers have active knowledge of probabilistic trends across these memorized derived forms
   - Speakers can match these trends in a wug-test:

     (1) Law of Frequency Matching:(Hayes et al. (2009) p.826) Speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies.

[1]Materials can be found at `http://people.umass.edu/cmooreca/hebrewdvf/index.html`.
I wish to thank Joe Pater, John McCarthy, Lyn Frazier, and Kie Zuraw for lots of wonderful advice and direction. I also wish to thank Roy Becker-Kristal, Shmuel Bolozky, Rachel Borden, Will Quale, John Griffin, Chris Cantwell, and especially Michael Becker, Lena Feinleib and Aynat Rubinstein for help preparing and checking stimuli, and preparing the web interface for the experiment. Further thanks are due to Wendell Kimper, Robert Staubs, John Kingston, Anne Pycha, Tom Roeper, Iris Berent, Bruce Hayes, Anne-Michelle Tessier, audiences at UMMM and NELS 42, and especially the UMass 2nd year seminar for helpful discussion, comments, and questions. I am grateful to Michael Lavine and Caren Rotello for invaluable statistics help. This work was supported by a National Science Foundation Graduate Research Fellowship to the author, and is dedicated to the glory of God.

2. What types of generalizations do learners form?

- A lexicon allows for many possible generalizations over multiple lexical items.

  For example:

  – Nouns take the regular plural 90% of the time, the irregular 10% of the time

  – Nouns that end in [i] take the regular plural 75% of the time, and the irregular plural 25% of the time.

  – Nouns heard after 6:00pm take the regular 95% of the time, while nouns heard before 6:00pm take the regular only 85% of the time

- Learners must choose which characteristics of the base are important in calculating the probability of a derived form.

- I'll talk about the generalizations learners form in terms of *conditional probability* of a derived form given its base: P(Derived form | Base)

  – This is a number that is directly observable in a wug-test

  – And is also directly observable in the lexicon

  – So, these two numbers can be easily compared

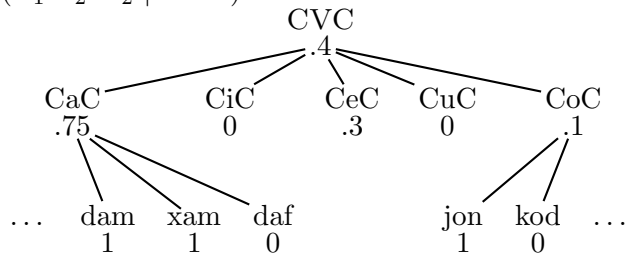
- Suppose some morphological process: Verbs are formed from nouns

  A brief preview of what's coming:

  – Verbs must fit a CVCVC template

  – CVC nouns can become verbs of the shape $C_1iC_2eC_2$ or of the shape CijeC

(2) $P(C_1iC_2eC_2 \mid \text{Noun})$



What do language learners learn about this situation?

– In order to correctly speak the language, they have to learn that [dam] always becomes [dimem], but [daf] doesn't ever become *[difef]

– Do they learn that verbs of shape $C_1iC_2eC_2$ occur 75% of the time with nouns with [a], 30% of the time with nouns with [e], 10% of the time with nouns with [o], etc? – If yes, they would treat CaC nouns differently from CoC nouns

– Or do they learn that verbs of shape $C_1iC_2eC_2$ occur 40% of the time with nouns of shape CVC? – If yes, they would treat all CVC nouns the same

3 Hebrew Denominal Verb formation

3. The pattern:

- Semitic verbs consist of a three-consonant 'root' combined with a two-vowel vowel pattern which expresses derivational and inflectional morphology

(3) Vowel patterns:

g**a**d**a**l	g**a**d**e**l	g**i**d**e**l	meg**a**d**e**l
he grew	*he is growing*	*he raised*	*he is raising*

- This means verbs must be minimally of the shape CVCVC
- Nouns have no such structural requirement. When a verb is formed from a noun, the noun's consonants are simply fit into the verbal template

(4) Denominal Verbs (Bat-El, 1994, pg. 577-579)

Base		Derived verb	
praklit	*lawyer*	priklet	*practice law*
telegraf	*telegraph*	tilgref	*telegraph*
sandlar	*shoemaker*	sindler	*make shoes*
blof	*bluff*	bilef	*to bluff*

- But if the noun is too small (doesn't have enough consonants), then other strategies must be employed. Ussishkin (1999) identified five types of verb formed from two-consonant nouns

(5) Forms of verbs derived from CVC nouns:

Structure				Label
Doubling:				
1.	$\mathbf{C_1VC_2}$	→	$\mathbf{C_1iC_2eC_2}$	Plain Consonant Doubling (CD)
	dam	→	**dimem**	
	blood		*he bled*	
2.	$\mathbf{C_1VC_2}$	→	$\mathbf{C_1VC_2eC_2}$	Vowel Overwriting (Ov)
	kod	→	**koded**	
	code		*he encoded*	
Glide Formation:				
3.	$\mathbf{C_1VC_2}$	→	$\mathbf{C_1ijeC_2}$	Coronal Glide Formation (J)
	tik	→	**tijek**	
	file		*he filed*	
4.	$\mathbf{C_1VC_2}$	→	$\mathbf{C_1iveC_2}$	Labial Glide Formation[2](V)
	sug	→	**siveg**	
	type		*he sorted*	
Other:				
5.	$\mathbf{C_1VC_2}$	→	$\mathbf{C_1iC_2C_1eC_2}$	Reduplication (RED)
	daf	→	**difdef**	
	page		*he paged through*	

- Each noun has only one legal verbal counterpart
- But there are trends about which verbal form nouns of a given shape will take:
  - Nouns with high vowels take Glide Formation
  - Nouns with [a] and [e] take Consonant Doubling or Reduplication
  - Nouns with [o] take Vowel Overwriting

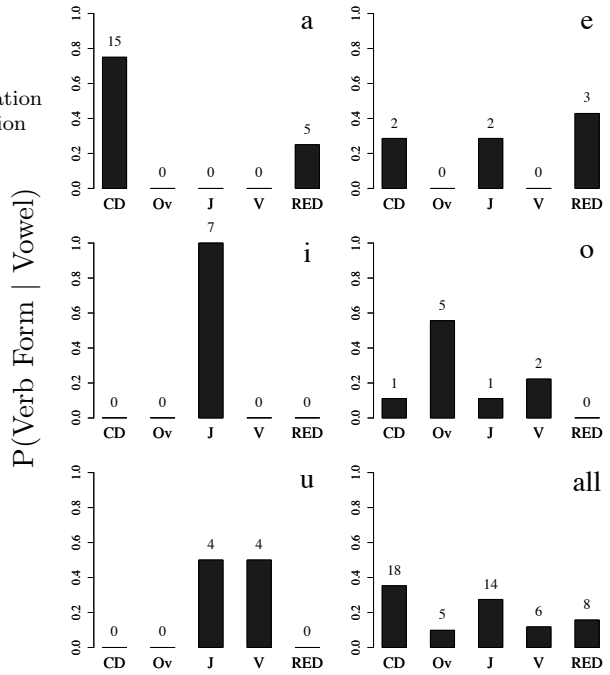
[2]Ussishkin (1999) argues that [v] 'counts' as the labial glide in Hebrew, even though it is not a sonorant. It is the consonantal counterpart of [u]

4. Corpus data

- A corpus of 52 noun-verb pairs was collected from a variety of sources: Ussishkin (1999); Bat-El (1994); Bolozky and Becker (2006), and native Hebrew speakers.
- The list is close to exhaustive - this pattern does not take up very much space in the Hebrew lexicon

(6) Distribution of verbal forms in the corpus

CD Consonant Doubling
Ov Vowel Overwriting
J Coronal Glide Formation
V Labial Glide Formation
RED Reduplication



In (6), the y axis represents the probability of occurrence of that verbal form with that vowel, and the numbers above the bars represent raw counts

- Nouns with high vowels always take some kind of glide formation
- Nouns with [a] typically take consonant doubling
- Nouns with [o] tend to take vowel overwriting
- But glide formation happens with nonhigh vowels too, there is no way to determine which kind of glide formation will happen with [u], and nouns with [o] are highly variable.

5. OT-style analysis

- For now, I treat these tendencies as categorical patterns, and follow Ussishkin (1999)
- In Consonant Doubling forms, the verbal vowel pattern must overwrite the noun's vowel

(7) Emergence of the vowel pattern

a. MAX-V-STEM: *Assign a violation to every input stem vowel without an output correspondent*

b. MAX-V-AFFIX: *Assign a violation to every input affix vowel without an output correspondent*

c. MAX-V-AFFIX ≫ MAX-V-STEM

/dam + ie/	MAX-V-AFFIX	MAX-V-STEM
a. damem	*!	
→ b. dimem		*

- But high vowels are preserved in the form of a glide

(8) High vowels are preserved

a. ID-μ: *Assign a violation to every output segment whose value for moraicity (syllabicity) does not match its input counterpart*

b. MAX-V-STEM $\gg$ ID-μ

/ti$_1$k/	MAX-V-STEM	ID-μ
a. tikek	*!	
→ b. tij$_1$ek		*
/su$_1$g/		
a. sigeg	*!	
→ b. siv$_1$eg		*

- In the rest of Hebrew, there is lots of evidence for MAX-V-STEM being very low ranked (input vowels are overwritten all the time), but to get the difference between verbs derived from different vowels, it must outrank some constraints

4 The wug-test

6. Methods

- This is a free-form production study (fill-in-the-blank)
- 20 nonce words were constructed, of the form CVC, 4 with each vowel (a,e,i,o,u)
- The nonce nouns were presented aurally in tiny stories.
- Stories were recorded by a native speaker of Hebrew who had phonetic training

(9) Example item

> *Spoken:* In Bat-Jupiter, fruit can't be grown because it takes up too much space. Fruit has to be shipped from Earth. In order to keep the fruit from weighing too much and taking up space on the ship from Earth, the fruit has to be compressed somehow. A machine called a *mok* first removes all the water from the fruit, then removes the skin, seeds, and any other part that won't be eaten, and finally vacuum-packs it.
>
> *Written:* When they're in the middle of preparing a shipment, Earth technicians _ many kilograms of fruit per day.

- 27 native speakers were recruited through the internet

7. Results

- Subjects responded with a fairly wide variety of verbal form types
  - Sometimes with an existing Hebrew verb rather than a novel one
  - Sometimes with a derived (nonce) verbs of types which do not exist as denominal verbs in Hebrew

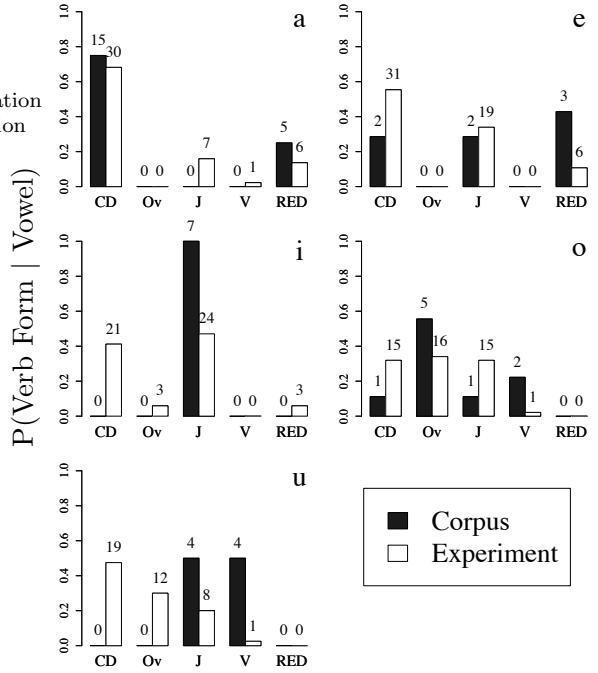
(10) Variety of Responses

Response	Count	Percentage
Denominal verb forms:	240	44%
Other verb forms:	170	31%
Existing verbs:	85	16%
Other:	45	8%

I will analyze only the denominal verb form responses.

(11) Conditional probability of output given input: comparing corpus and experiment.

CD Consonant Doubling
Ov Vowel Overwriting
J Coronal Glide Formation
V Labial Glide Formation
RED Reduplication



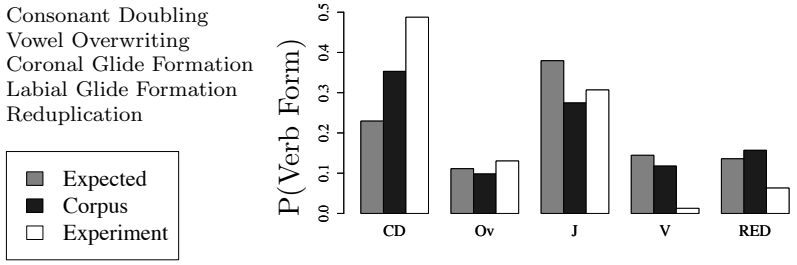
8. Conditional probability of verb given noun's vowel is an imperfect match between the lexicon and the experiment
   - Subjects have overgeneralized Consonant Doubling: it occurs more often and in more contexts than in the lexicon
   - Coronal glide formation and vowel overwriting also occur in more contexts than they do in the lexicon
   - Labial glide formation has been undergeneralized. It occurs only three times throughout the whole experiment
   - There are generally more exceptions, and less clear generalizations in the experiment - the distributions are 'smoothed'
9. The type frequencies of verbal forms in the experiment match the type frequencies in the lexicon better than expected by chance

   Verbal form type frequencies in the lexicon are a direct result of vowel frequencies in CVC nouns

   - Consonant doubling is common in the lexicon because nouns with [a] are common, and coronal glide formation is less common simply because nouns with [i] are less common
   - (12) compares the output type frequencies in the corpus and the lexicon to what would be predicted if speakers matched the conditional probabilities of verb given vowel on an equal number of words with each vowel.

(12) Comparing the experimental results to the corpus.

CD Consonant Doubling
Ov Vowel Overwriting
J Coronal Glide Formation
V Labial Glide Formation
RED Reduplication



10. Statistical analysis: A Poisson Regression

    - Tests whether verbal forms occur equally often
    - And whether the noun's vowel affects the frequency of occurrence of each verbal form

      Result: Consonant doubling is significantly more frequent than any other verbal form in the experiment (p< .002), but not in the lexicon (p= .48)

      Also: Adding noun's vowel as a predictor of frequency of occurrence of verbal form significantly improves the model's fit (p< .001)

5 Local Summary

11. Consonant Doubling has been overgeneralized

12. Verbal form type frequencies in the experiment match verbal form type frequencies in the lexicon

13. The distribution produced by experimental subjects is smoothed with respect to the lexical distributions

6 Modeling

My goal is to produce a model that takes as input the Hebrew lexicon (of denominal verbs) and predicts the experimental results

14. Maximum Entropy grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008; Wilson, 2006)

    - A subspecies of Harmonic Grammar (Smolensky and Legendre, 2006; Pater, 2009)
    - Uses harmony scores calculated from weighted constraints to assign probabilities to output candidates

      (13) a. $H = \sum w_i C_i(x)$
      Where w is a vector of constraint weights, C is the set of constraints, so that $C_i(x)$ is the number of times x violates constraint C_i.

      b. $P_x = e^{(-H_x)} / \sum_i e^{(-H_i)}$

    - I use a batch learning algorithm (Wilson and George, 2009) to find a set of weights

15. The constraint set

    - A simplified version of the constraint set in Ussishkin (1999)[3]

      (14) Constraints for consideration

a.	MAX-V-STEM	*Don't delete a vowel from the stem*
b.	MAX-V-AFFIX	*Don't delete a vowel from the affix (the piʕel vowel pattern must surface intact)*
c.	ID-μ	*Don't change the moraicity (syllabicity) of a segment*
d.	ID-V-LO	*Don't change the lowness of a vowel*
e.	ID-V-HI	*Don't change the height value of a vowel*
f.	MAX-LAB	*Don't delete a labial feature*
g.	ID-V-SON	*Don't change the sonority of a vowel*
h.	*REDUPLICATION	*Don't be reduplicated*

[3] 'Simplified' here means that I've left out a set of prosodic markedness constraints which force the verb to be bisyllabic and prevent illegal clusters and vowel hiatus.

16. Regularization

- MaxEnt models employ regularization terms that penalize high constraint weights
- Lower constraint weights result in smoother probability distributions
- In order to model the 'smoothing' that happens in the experimental data with respect to the lexicon, I'll employ a very strong regularization term, which restricts constraint weights to have a variance of 5
- Further: MAX-V-STEM will get an especially strong bias towards zero (σ^2=0.5), to reflect the fact that it is violated frequently in the rest of the language

(15) Weights learned from training on the corpus with a strong regularization term

17. MaxEnt at work

- This tableau illustrates the weights learned when the regularization discussed above is applied
- It also illustrates each candidate's harmony score (**H**), and predicted probability (**P**).
- Note that the predicted probability distribution is 'smoothed' with respect to the lexical distribution in this example - this is because the weights are forced to remain low during the weight-finding process
- Also note that it's still not a great match for the experimental frequencies (.09 vs. .48 for consonant doubling)

	Corpus probs	MAX-V-STEM	MAX-V-AFF	IDENT-μ	IDENT-V-LOW	IDENT-V-HIGH	MAX-LAB	IDENT-V-SON	*REDUPLICATED			*Experiment probs*
Weights$_{\text{corpus}}$		**.8**	**3.0**	**0**	**2**	**1.6**	**2.6**	**1.8**	**0.8**	**H**	**P**	*p*
/$C_1u_3C_2$/												
a. $C_1iC_2eC_2$	**0**	1					1			**-3.4**	**0.09**	*.48*
a. $C_1u_3C_2eC_2$	**0**		1							**-3**	**0.14**	*.30*
a. $C_1ij_3eC_2$	**.5**			1			1			**-2.6**	**0.23**	*.2*
a. $C_1iv_3eC_2$	**.5**			1				1		**-1.8**	**0.49**	*.03*
a. $C_1iC_2C_1eC_2$	**0**	1					1		1	**-4.2**	**0.04**	*0*

18. Incorporating output frequency

- The probabilities predicted by the MaxEnt model are scaled by the overall output type frequencies

(16) $P(Verb \mid Vowel) = P_{MaxEnt}(Verb \mid Vowel) * P(Verb)$

- The numbers produced by this equation are then scaled so that the total probability for each vowel sums to 1

19. Does adding output type frequencies to the model help?

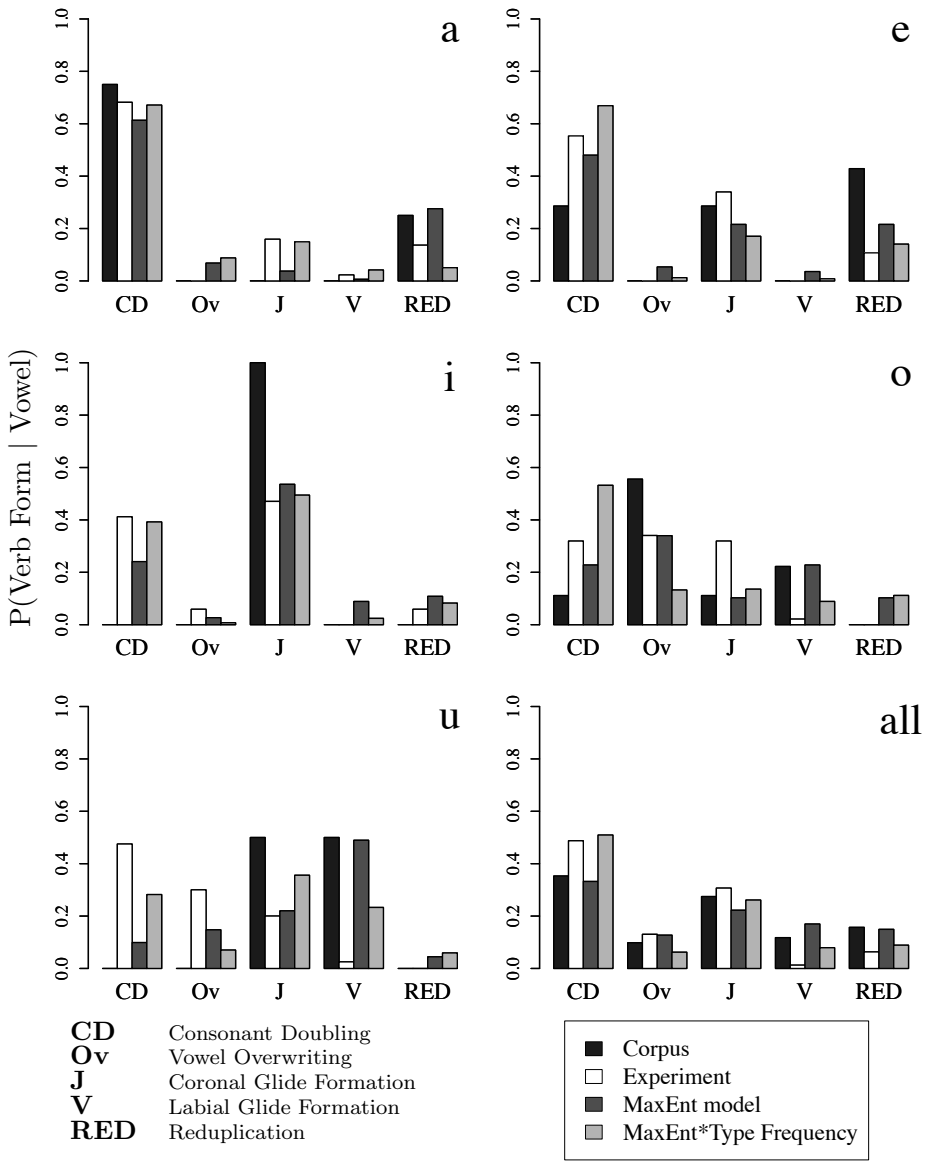
- I'll use the Chi-squared difference test to test whether the addition of output frequencies to the model significantly improves its performance

(17) Chi square value for various models of the experimental data

Model	χ^2	df	χ^2_{diff}	df_{diff}	p
MaxEnt	163.5	8			
MaxEnt					
+ Output type frequency	127.7	12	35.8	4	$< .001$

- Scaling the expected probability distributions by the output type frequencies significantly improves the model's fit to the experimental data.

(18) The final model



7 Conclusions

20. Hebrew speakers do form generalizations over the lexically variable pattern of denominal verb formation
    - But not just over phonemically-detailed levels of generalization like that which formed the basis of the analysis in Ussishkin (1999)
    - In this experiment, speakers demonstrated knowledge of a level of generalization which took into account no phonemic details, and which was not very 'useful' in that it led to a low degree of certainty regarding the outcome.
21. Speakers demonstrate knowledge of two different levels of generalization
    - Conditional probability of verb type given noun's vowel
    - And type frequency of verbs
22. I've presented a model that integrates these two types of knowledge, and produces a decent match to the experimental data.

References

Albright, A. and Hayes, B. (2003). Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition*, 90:119–161.

Bat-El, O. (1994). Stem modification and cluster transfer in modern hebrew. *Natural Language and Linguistic Theory*, 12:571–596.

Becker, M., Ketrez, N., and Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 87(1):84–125.

Berko, J. (1958). The child's learning of english morphology. *Word*, 14:150–77.

Bolozky, S. and Becker, M. (2006). Living lexicon of hebrew nouns. ms. UMass Amherst.

Ernestus, M. and Baayen, H. (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language*, 79(1):5–38.

Goldwater, S. and Johnson, M. (2003). Learning ot constraint rankings using a maximum entropy model. In Spenader, J., Eriksson, A., and Dahl, O., editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Hayes, B., Zuraw, K., Siptár, P., and Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85(4):822–863.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, 33:999–1035.

Smolensky, P. and Legendre, G. (2006). *The harmonic mind: from neural computation to optimality-theoretic grammar*. MIT Press, Cambridge, Massachusetts.

Ussishkin, A. (1999). The inadequacy of the consonantal root: Modern Hebrew denominal verbs and output-output correspondence. *Phonology*, 16(3):410–442.

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30:945–982.

Wilson, C. and George, B. (2009). The maxent grammar tool. Department of Cognitive Science, Johns Hopkins University and Department of Linguistics, UCLA (http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/).

Zuraw, K. (2000). *Patterned Exceptions in Phonology*. PhD thesis, University of California, Los Angeles.

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