Markedness bias in phonological learning: a Samoan case study and proposed experiment

February 17, 2023

Jennifer Kuo

1 Background

- How do learners resolve conflicting data patterns in morphophonological paradigms? English example from Berko (1958):

SG	PL
chie[f]	chie[f]s
lea[f]	lea[v]es
heaf	hea[f]s? hea[v]es?

  - Learner errors are informative.
  - Reanalysis: errors adopted into speech community, resulting in a type of language change.

- Malagasy: patterns of reanalysis suggest that learners are sensitive to both **frequency matching** and to a **markedness bias**.
  - ...where markedness is restricted to "active" markedness already present in stem **phonotactics**.

- **Goals for today:**
  1. Describe a Samoan case study showing the effect of markedness in reanalysis over time.
     - "replication" of Malagasy results
  2. Propose a nonce-word experiment that builds on these findings, by comparing the effect of learning biases in Samoan L1 vs. heritage speakers.

2 A Samoan case study: the reanalysis of ∅∼C alternations

- In the Polynesian languages (and many other Oceanic languages), a consonant of unpredictable quality may surface under suffixation (Pawley 2001)
- Samoan (Mosel and Hovdhaugen 1992): the -Cia 'ERGATIVE' suffix has various allomorphs, which mainly differ in their initial consonant (Table 1). Note: only /-a, -ina/, the vowel-initial allomorphs, are productive.[1]

ERG.	STEM	SUFFIXED	GLOSS	POc
a	rere	rere-a	to take	*rere
ina	iloa	iloa-ina	to see, perceive	*qilo
tia	pulu	pulu-**t**ia	to plug up	*bulu**t**
sia	laka	laka-**s**ia	to step over	*laka**s**
ŋia	tutu	tu-**ŋ**ia	to light a fire	*tutu**ŋ**
fia	utu	utu-**f**ia	to draw water	*qutu**p**
mia	inu	inu-**m**ia	to drink	*inu**m**
lia	tautau	tautau-**l**ia	to hang up	*sau**r**
na	ʔai	ʔai-**n**a	to eat	*kae**n**
ʔia	momo	momo-**ʔ**ia	to break in pieces	*mekme**k**

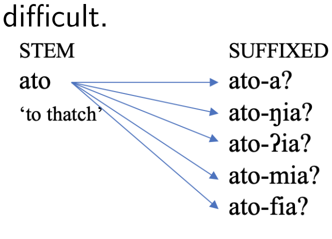
Table 1: Samoan ∅/C alternations

- Historical origin: word-final consonants of Proto-Oceanic (POc) were regularly lost.

POc	*inum	*inumia	*bulut	bulutia	
	inu	–	pulu	–	(C> ∅/__#)
Samoan	inu	inu-mia	pulu	pulu-tia	

- Ambiguity in the unsuffixed form can result in reanalysis (Hale 1973):

Conflicting patterns make learning difficult.



Reanalysis: errors adopted into speech community and passed through generations of speakers.

Samoan suffixed form			
POc	Expected	Actual	Reanalysis
*qatop	ato-fia	ato-a	f→ ∅
*akot	aʔo-tia	aʔo-ina	t→ ∅
*qulin	uli-na	uli-ŋia	n→ ŋ

[1] The relative distribution of a/ina is partially predictable from the length of a word.

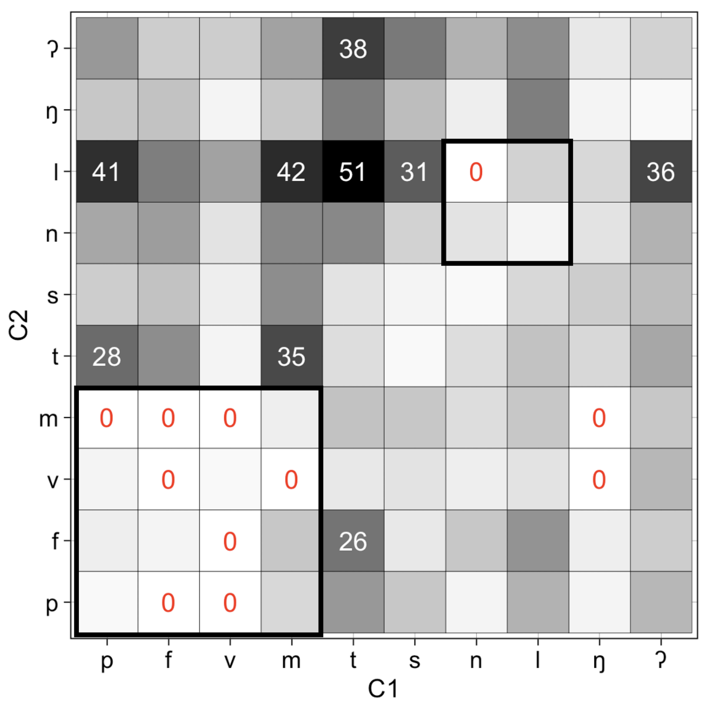
Preview of results:

- Reanalysis is towards the more frequent allomorph, but is sensitive to markedness effects.
  - Phonotactically marked outputs are more likely to be reanalyzed.
- How so? Samoan *roots* are subject to (non-local) consonant OCP effects. A model of reanalysis which incorporates these OCP effects outperforms a purely distributional model.

2.1 Phonotactic markedness: OCP effects in Samoan roots

- OCP: consecutive identical features are banned/dispreferred.
- Samoan is subject to transvocalic consonant OCP effects (Alderete and Bradshaw 2013)
  - Crosslinguistically well-attested (McCarthy 1988, 1994; Coetzee and Pater 2008; Wilson and Obdeyn 2009)
- OCP effects are particularly strong across:
  - Labials (*[+LABIAL][+LABIAL], e.g. *[fuma])
  - Coronal sonorants (*{n,l}{n,l}, e.g. *[nula])
- Visualized in (1) (data: corpus of 1,512 roots from Milner 1966, narrowed down from around 4200 headwords, and with pseudo-reduplicants removed).[2]

(1) *Transvocalic consonant co-occurences in modern Samoan roots*



[2]Samoan has two registers of speech, *tautala lelei* and *tautala leaga*. *Tautala lelei* preserves more segmental contrasts, but native speakers are generally cognizant of both levels. All data in this paper are from the *tautala lelei* register, as it is the one described in dictionaries and most scholarly work on Samoan.

- Patterns confirmed using the UCLA Phonotactic Learner (Hayes and Wilson 2008)
- Similar OCP effects are suggested to be a general property of Polynesian languages (Krupa 1971)

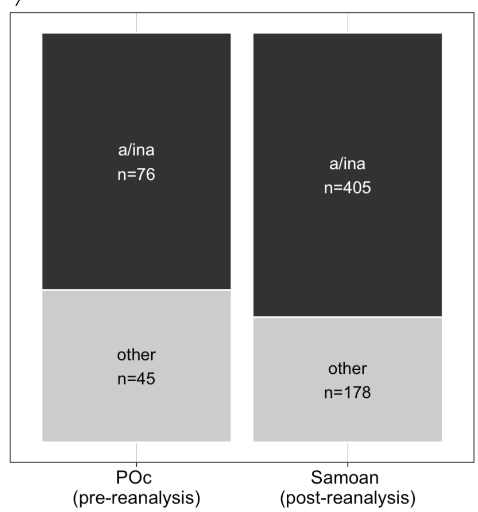
2.2 Reanalysis of ergative suffix allomorphs

- The allomorphs that existed prior to reanalysis are inferred by looking at the final consonants in Proto-Oceanic.
  - Data: 519 protoforms from the Austronesian Comparative Dictionary (Blust and Trussel 2010) and Polynesian Lexicon Project (Greenhill and Clark 2011)
- **Result 1:** Reanalysis is mostly towards /a, ina/, which is predictable from frequency distributions.

(2) *Historical distribution of passive allomorphs*

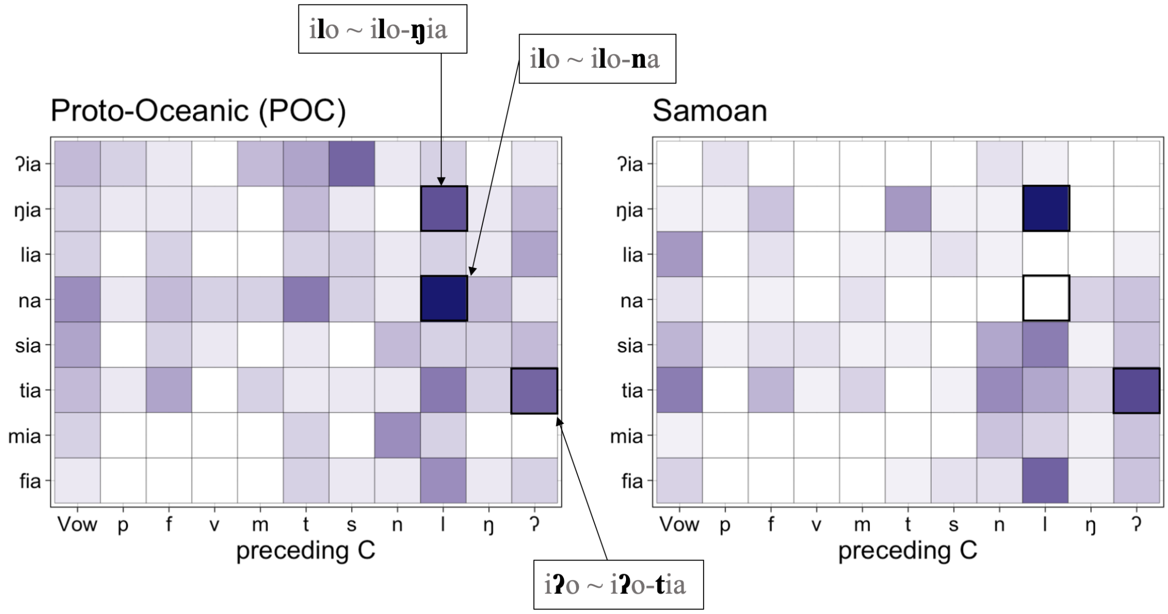
POc	Erg.	N	P	
vowel,*ʔ,*q	**a, ina**	**346**	**0.67**	⇐ ⋆
*p	fia	13	0.03	
*t,*c	tia	29	0.06	
*k	ʔia	24	0.05	
*s	sia	18	0.03	
*m	mia	11	0.02	
*n,*ɲ	na	41	0.08	
*ŋ	ŋia	21	0.04	
*r,*l	lia	16	0.03	

(3) *Reanalysis is generally towards a/ina*



- **Result 2:** phonotactically **marked** outputs are more likely to be reanalyzed.
- the Cia allomorph that surfaces is partially conditioned by the identity of the preceding consonant.
- Some regularities in POc are passed down to modern Samoan, while others are not. Figure (4) compares the distribution of Cia allomorphs in POc and Samoan.

(4) *Distribution of Cia allomorphs before and after reanalysis (/a, ina/ excluded)*



- Most notably, suffixed forms of the type [**il**o-**n**a] were relatively frequency pre-reanalysis, but never observed post-reanalysis.
  - motivated by coronal sonorant OCP (*{n,l}{n,l})

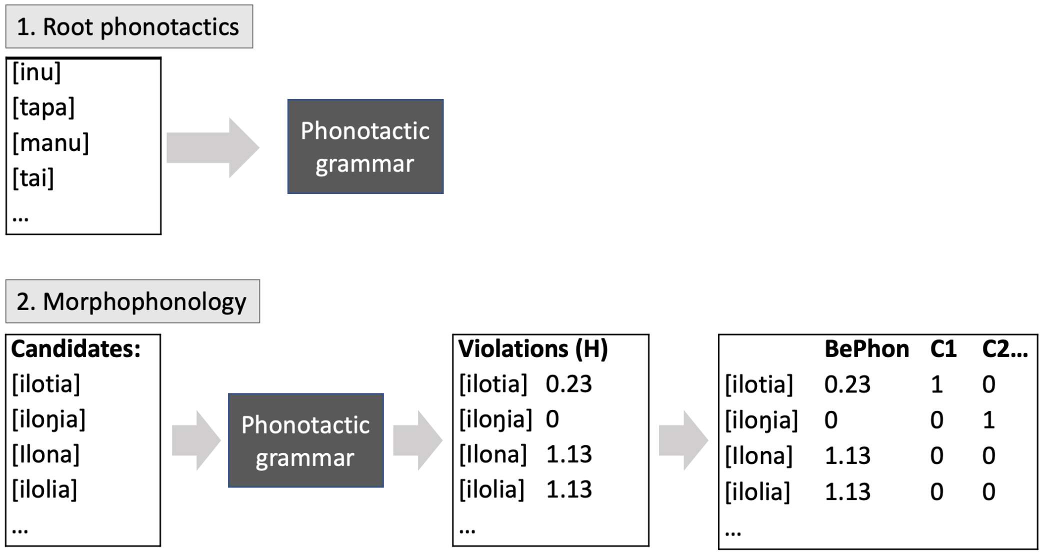
2.3 Modeling reanalysis of the Samoan ergative suffix

- Basic idea: use **root phonotactics** to inform model of **morphophonology/reanalysis**.
- Phonotactic markedness constrains
  - A phonotactic model is trained on roots, and then used to assign violations to the candidate suffixed forms in the morphophonological grammar, as in (5).
  - UCLA Phonotactic Learner (Hayes and Wilson 2008), which is based in MaxEnt.
  - Two phonotactic grammars:
    - Bigram grammar: constraints are the set of all possible C_1C_2 combinations.
    - OCP grammar: constraints are restricted to OCP constraints that target natural classes,e.g. *[+CORONAL][+CORONAL], $*\begin{bmatrix}+\text{CORONAL}\\+\text{sonorant}\end{bmatrix}\begin{bmatrix}+\text{CORONAL}\\+\text{sonorant}\end{bmatrix}$
- Model of reanalysis:
  - MaxEnt Harmonic Grammar (Smolensky 1986; Goldwater and Johnson 2003)
  - Phonotactically motivated constraints (e.g. OCP-place) are biased to have high weight, following Wilson (2006), White (2013), etc.[3]

[3]Bias is implemented by giving the model a Gaussian prior, and assigning relevant constraints a higher μ.

– Iterated learning component (20 generations)

(5) *Incorporating phonotactic markedness into morphophonological grammar*

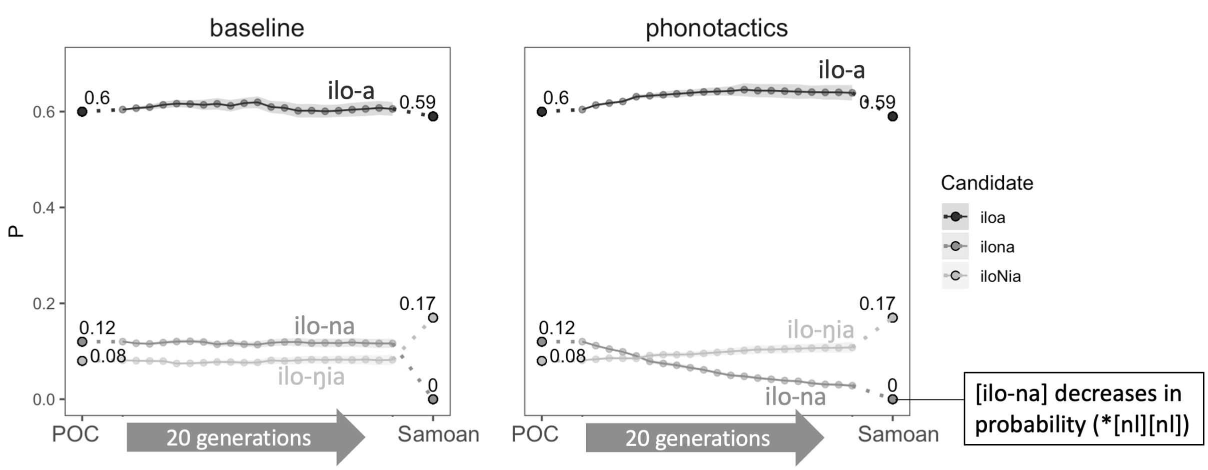


- **Result:** The models with a markedness bias perform better in terms of model fit to the modern Samoan distribution; see (6) for log-likelihood.
- In particular, the markedness models do better at predicting the output for words with coronal dissimilation; see (7).

(6) *Model results*

model	**L**	**p**
Baseline	-2408.8	–
Phonotactics (OCP)	-2376.1	$p < 0.00001$
Phonotactics (bigrams)	-2400.746	$p<0.01$

(7) *Subset of model predictions for words with a preceding /l/*



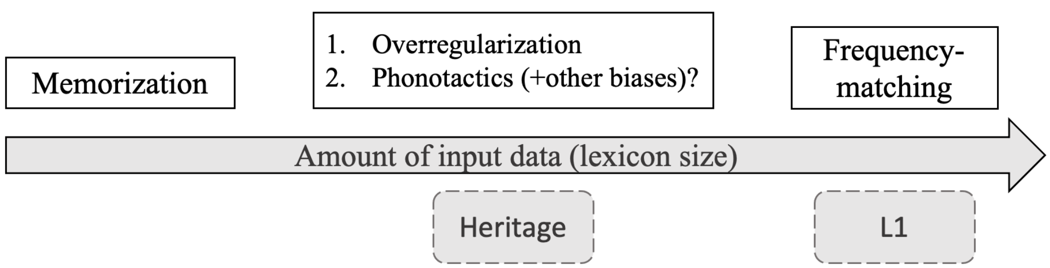
3 Next steps: markedness effects in the synchronic Samoan grammar

- Do Samoan speakers frequency-match when applying the ergative suffix to novel words, or are they also sensitive to markedness effects?
- Wug test on both fluent L1 and heritage speakers.
  - Both fluent and heritage speakers of Samoan are contactable.

3.1 Background: insights from heritage language phonology

- There is evidence that phonotactic learning is easier than alternation learning; even with limited input, adults possess sophisticated phonotactic intuitions (Oh et al. 2020).
- Many theories of acquisition and learning algorithms predict that learners will become better frequency-matchers as their lexicon increases in size.
  - Tolerance Principle (Yang 2016)
  - MaxEnt Harmonic Grammar (Zymet 2018)
  - Perceptron (online versions of MaxEnt; O'Hara 2020)
- **Prediction**: when learners have limited input, they will be more sensitive to markedness effects, and potentially other biases.

(8) *Relationship between lexicon size and morphophonological learning*



- **Heritage speaker**: a speaker who learns a (minority) language as their L1 at home, but then become dominant in the majority language spoken in the wider community (Polinsky and Kagan 2007).
  - Differs from L1 speakers in having less input data (Polinsky 2008)
  - Predicted to be **more sensitive** to markedness bias (and less prone to frequency-matching)
- In my own work on Seediq (Kuo, forthcoming), I wug-tested both middle-aged speakers (some heritage) and elderly speakers (truly native). Heritage speakers showed evidence of following (and exaggerating) a phonological generalization.

3.2 Methodology

- Nonce-word tests (wug tests; Berko 1958)
  - used to demonstrate speakers' knowledge about distributional facts of their lexicon (e.g. Zuraw 2000; Ernestus and Baayen 2003; Hayes and Londe 2006; Becker et al. 2011).
  - When speakers underlearn or overlearn patterns in the lexicon, this suggests the effect of a learning bias (e.g. Wilson 2006; Becker et al. 2011).
- Procedure:
  - Pretest: speakers provide phonological well-formedness ratings for the stimuli (i.e. novel stems).[4]
  - Open response: participants volunteer ergative forms for the wug verbs.
  - Rating task: speakers rate possible ergative forms for wug verbs; for each verb, 4 possible suffixed forms are provided:
    1. default ([ilo-a])
    2. OCP violation; same segment ([ilo-lia])
    3. OCP violation; same class ([ilo-na])
    4. no OCP violation (ilo-fia)
- Stimuli: words of the shape (C)V**C**V, conditioned by final consonant (in boldface).

Condition	**Example**	**Rating task**
labial	i**m**o	imo-a, imo-mia, imo-fia, imo-tia
coronal sonorant	i**l**o	ilo-a, ilo-lia, ilo-na, ilo-mia
coronal non-sonorant	i**t**o	ito-a, ito-tia, ito-na, ito-mia
filler	i**ʔ**o	iʔo-a, iʔo-ʔia, iʔo-ŋia, iʔo-tia

- Prediction: Both in production and goodness ratings, speakers will disprefer outputs which result in OCP violations. This effect is predicted to be stronger for heritage speakers.

4 Summing up (goals)

- I hope to obtain converging evidence from historical change (reanalysis), L1 speakers, and heritage speakers on how markedness bias interacts with frequency-matching.
- A corpus which aims to connect corpus results with experimental results, which can be used to test different models of learning.
  - Can be expanded to languages such as Tongan.
- Greater understanding of L2/heritage acquisition
  - Most existing work in heritage language phonology has focused on the learning of specific phonemes (e.g. Rao and Ronquest 2015; Kang and Nagy 2016), rather than the learning of morphophonological paradigms.

[4]This step, following Albright and Hayes (2003), is included to address the possible confounding influence of stem well-formedness on morphological intuitions

References

Adam Albright and Bruce Hayes. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161, 2003. doi: 10.1016/S0010-0277(03)00146-X.

John Alderete and Mark Bradshaw. Samoan root phonotactics: Digging deeper into the data. *Linguistic Discovery*, 11(1), 2013.

Michael Becker, Nihan Ketrez, and Andrew Nevins. The surfeit of the stimulus: analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language (Baltimore)*, 87(1):84–125, 2011. doi: 10.1353/lan.2011.0016.

Jean Berko. The child's learning of English morphology. *Word*, 14(2-3):150–177, 1958.

Robert Blust and Stephen Trussel. Austronesian comparative dictionary, web edition. *Blust's Austronesian Comparative Dictionary Website*, 2010.

Andries W Coetzee and Joe Pater. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory*, 26:289–337, 2008.

Mirjam Ernestus and R Harald Baayen. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79(1):5–38, 2003. doi: 10.1353/lan.2003.0076.

Sharon Goldwater and Mark Johnson. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl, editors, *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pages 111–120. Stockholm: Stockholm University, Department of Linguistics, 2003.

Simon J Greenhill and Ross Clark. Pollex-online: The Polynesian lexicon project online. *Oceanic Linguistics*, pages 551–559, 2011.

Kenneth Hale. Deep-surface canonical disparities in relation to analysis and change: An Australian example. *Current trends in linguistics*, 11(19731):401–458, 1973.

Bruce Hayes and Zsuzsa Cziráky Londe. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology*, 23(1):59–104, 2006. doi: 10.1017/S0952675706000765.

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

Yoonjung Kang and Naomi Nagy. VOT merger in Heritage Korean in Toronto. *Language Variation and Change*, 28(2):249–272, 2016.

Viktor Krupa. The phonotactic structure of the morph in Polynesian languages. *Language*, pages 668–684, 1971.

Jennifer Kuo. Evidence for prosodic correspondence in the vowel alternations of Tgdaya Seediq. *Phonological Data and Analysis*, forthcoming.

John J McCarthy. Feature geometry and dependency: A review. *Phonetica*, 45(2-4):84–108, 1988.

John J McCarthy. The phonetics and phonology of Semitic pharyngeals. In Patricia Keating, editor, *Phonological structure and phonetic form*, pages 191–233. Cambridge University Press, 1994.

George Bertram Milner. *Samoan Dictionary; Samoan-English, English-Samoan*. ERIC, 1966.

Ulrike Mosel and Even Hovdhaugen. *Samoan reference grammar*. Scandinavian Univ. Press, 1992.

Y Oh, S Todd, C Beckner, J Hay, J King, and J Needle. Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific reports*, 10(1):1–9, 2020.

Charlie O'Hara. Frequency matching behavior in on-line MaxEnt learners. *Proceedings of the Society for Computation in Linguistics*, 3(1):463–465, 2020.

Andrew Pawley. Proto polynesian *-CIA. In *Issues in Austronesian Morphology: A festschrift for Byron W. Bender*. Pacific Linguistics, 2001.

Maria Polinsky. Gender under incomplete acquisition: Heritage speakers' knowledge of noun categorization. *Heritage language journal*, 6(1):40–71, 2008.

Maria Polinsky and Olga Kagan. Heritage languages: In the 'wild' and in the classroom. *Language and linguistics compass*, 1(5):368–395, 2007.

Rajiv Rao and Rebecca Ronquest. The heritage Spanish phonetic/phonological system: Looking back and moving forward. *Studies in Hispanic and Lusophone Linguistics*, 8(2):403–414, 2015. doi: doi:10.1515/shll-2015-0016. URL `https://doi.org/10.1515/shll-2015-0016`.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, chapter 6, pages 194–281. Cambridge, MA: MIT Press, 1986.

James C. White. *Bias in phonological learning: Evidence from saltation*. PhD thesis, UCLA, 2013.

Colin Wilson. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982, 2006. doi: 10.1207/s15516709cog0000_89.

Colin Wilson and Marieke Obdeyn. Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. *Ms, Johns Hopkins University*, 2009.

Charles Yang. *The price of linguistic productivity: How children learn to break the rules of language*. MIT press, 2016.

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

Jesse Zymet. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. PhD thesis, University of California, Los Angeles, 2018.