Learning Opaque Generalizations: The Case of Samala (Chumash)

Jeffrey Heinz* William Idsardi**

*University of Delaware

**University of Maryland

LSA 84th Annual Meeting
Baltimore, MD
January 9, 2010

Learning opaque generalizations in phonology

1. How can phonological generalizations be automatically discovered from surface forms when they are obscured by others?
2. Discuss 2 different UG-based proposals which shuffle the data in principled ways to reveal obscured generalization
3. Case Study: Samala (Chumash) (Applegate 1972, 2007), simplified into a phonotactic learning problem
   - Correct misconceptions about the phonology of Samala
   - Study interaction between long-distance and local processes

Learning opaque generalizations in phonology

1. How can phonological generalizations be automatically discovered from surface forms when they are obscured by others?
2. Discuss 2 different UG-based proposals which shuffle the data in principled ways to reveal obscured generalization
3. Case Study: Samala (Chumash) (Applegate 1972, 2007), simplified into a phonotactic learning problem
    - Correct misconceptions about the phonology of Samala
    - Study interaction between long-distance and local processes

Learning opaque generalizations in phonology

1. How can phonological generalizations be automatically discovered from surface forms when they are obscured by others?
2. Discuss 2 different UG-based proposals which shuffle the data in principled ways to reveal obscured generalization
3. Case Study: Samala (Chumash) (Applegate 1972, 2007), simplified into a phonotactic learning problem
   - Correct misconceptions about the phonology of Samala
   - Study interaction between long-distance and local processes

Samala (Inezeño Chumash)

Maria Solares (1842-1923)

↓

John Peabody Harrington (1884-1961)

↓

Dr. Richard Applegate

`www.chumashlanguage.com`

The Corpus

- 4800 words drawn from Applegate 2007, generously provided in electronic form by Applegate (p.c).

35 Consonants

	labial	coronal	a.palatal	velar	uvular	glottal
stop	p pˀ pʰ	t tˀ tʰ		k kˀ kʰ	q qˀ qʰ	ʔ
affricates		t͡s t͡sˀ t͡sʰ	t͡ʃ t͡ʃˀ t͡ʃʰ			
fricatives		s sˀ sʰ	ʃ ʃˀ ʃʰ	x xˀ		h
nasal	m	n nˀ				
lateral		l lˀ				
approx.	w	y				

6 Vowels

i	ɨ	u
e		o
	a	

(Applegate 1972, 2007)

Opaque generalizations in Samala

Consider these processes in Samala (Applegate 1972):

1. **Local Assimilation**: [s] becomes [ʃ] before adjacent coronals [t,l,n] only across morpheme boundaries
2. **Sibilant Harmony**: the rightmost sibilant causes sibilants to the left to agree in anteriority

/s-ti-jep-us/ ‘3s tells 3s’

Local Assimilation
predicts [ʃtijepus]

Sibilant Harmony
predicts [stijepus]

/s-ti-jep-us/ ‘3s tells 3s’

Local Assimilation
predicts [ʃtijepus]

which is evidence against sibilant harmony!

Sibilant Harmony
predicts [stijepus]

/s-ti-jep-us/ ‘3s tells 3s’

Local Assimilation
predicts [ʃtijepus]

which is evidence against sibilant harmony!

Sibilant Harmony
predicts [stijepus]

which is evidence against local assimilation!

The facts of Samala

Local Assimilation
predicts [ʃtijepus]

Sibilant Harmony
predicts [stijepus]

/s-ti-jep-us/→ stijepus

(Applegate 1972, 2007; texts at `www.chumashlanguage.com`)

Contra much of the secondary phonological literature!

(Poser, 1982, 1993; Hansson, 2001; McCarthy, 2007)

The misreading

- Applegate (1972:119-120) states that the harmony process has some exceptions, such as when the local process can apply and gives /s-ti-jep-us/→[ʃtijepus] as an example.
- BUT Applegate meant these were *token* exceptions, not *type* ones. (Applegate p.c.)
- Applegate estimates 95% of the forms like /s-ti-jep-us/ are pronounced like [stijepus] in Harringtons copious notes of Samala (p.c).

The misreading

- Applegate (1972:119-120) states that the harmony process has some exceptions, such as when the local process can apply and gives /s-ti-jep-us/→[ʃtijepus] as an example.
- BUT Applegate meant these were *token* exceptions, not *type* ones. (Applegate p.c.)
- Applegate estimates 95% of the forms like /s-ti-jep-us/ are pronounced like [stijepus] in Harringtons copious notes of Samala (p.c).

Conclusions:

1. The canonical pronunciation is [stijepus].
2. Sibilant Harmony has priority over Local Assimilation.

Which process has priority is learned

- In Canadian French (Poliquin, 2006), pre-fricative tensing has priority over [ATR] harmony.
- Also, in Shimakonde, two harmony processes interact opaquely (Ettlinger, Bradlow and Wong 2010).
- There is no principle of UG which requires harmony patterns to have greater priority; which generalization obscures the other must be learned.

The Problem

- Given [stijepus] ‘3s tells 3s’, how do we conclude *st is active in the language?
- How can generalizations be learned in the face of regular exceptions?

The Problem

- Given [stijepus] '3s tells 3s', how do we conclude *st is active in the language?
- How can generalizations be learned in the face of regular exceptions?

The Problem

- Given [stijepus] '3s tells 3s', how do we conclude *st is active in the language?

- How can generalizations be learned in the face of regular exceptions?

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions)

Translating Samala into a phonotactic learning problem

Local Assimilation

*s[+coronal]

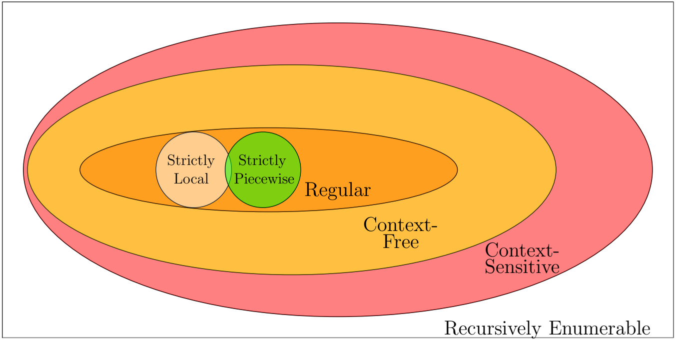
abbreviated ***st**

Sibilant Harmony

$$*\begin{bmatrix} +\text{strident} \\ \alpha\text{anterior} \end{bmatrix} \ldots \begin{bmatrix} +\text{strident} \\ -\alpha\text{anterior} \end{bmatrix}$$

abbreviated ***s. . . ʃ**

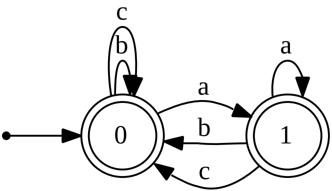
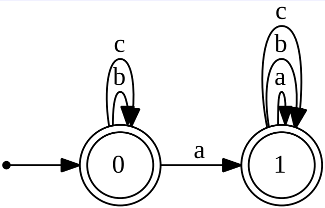
Learning local and long-distance phonotactic constraints



- Strictly 2-Local (SL) grammars describe constraints like **st*
- Strictly 2-Piecewise (SP) grammars describe constraints like **s. . . ʃ*
- SL-k and SP-k constraints are provably efficiently learnable from distribution-free, positive evidence
- SL-k and SP-k distributions are provably efficiently estimable

(McNaughton and Papert 1971, Rogers and Pullum 2007, Heinz 2007, Rogers et. al to appear, Garcia et. al 1991, Jurafsky and Martin 2008, Heinz and Rogers in prep, Vidal et. al 2005a,b)

Strictly Local and Strictly Piecewise

Strictly 2-Local (e.g. *st)	Strictly 2-Piecewise (e.g. *s. . . ʃ)
Contiguous subsequences	Subsequences (discontiguous OK)
Immediate Predecessor	Predecessor
Concatenation (·)	Less than (<)
	
0 = have not just seen an [a]	0 = have never seen an [a]
1 = have just seen an [a]	1 = have seen an [a] earlier

(McNaughton and Papert 1971, Simon 1975, Rogers and Pullum 2007, Rogers et. al. 2009, Heinz and Rogers in prep)

The Estimation of SL_2 Distributions (bigram model)

x	p	t	k	q	x$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions)

(Garcia et. al 1991, Jurafsky and Martin 2008)

The Estimation of SL_2 Distributions (bigram model)

x	p	t	k	q	x$\notin \Sigma - \{p,t,k,q\}$
Counts(sx)	29	29	37	20	728
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions)

(Garcia et. al 1991, Jurafsky and Martin 2008)

The Estimation of SL_2 Distributions (bigram model)

x	p	t	k	q	x$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions)

Chi-squared test not significant, p=0.264

(Garcia et. al 1991, Jurafsky and Martin 2008)

The Estimation of SP_2 Distributions

$P(x \mid b <)$		x			
		s	t͡s	ʃ	t͡ʃ
b	s	0.0325	0.0051	0.0013	0.0002
	t͡s	0.0212	0.0114	0.0008	0.
	ʃ	0.0011	0.	0.067	0.0359
	t͡ʃ	0.0006	0.	0.0458	0.0314

Table: SP2 probabililties of sibilant occuring sometime after another one (collapsing laryngeal distinctions)

(Rogers et. al to appear, Heinz and Rogers in prep)

Proposal #1

Remove data points confounded by the obscuring generalization and re-estimate

- Since Sibilant Harmony has priority over Local Assimilation, we'd like to remove words with sibilant harmony since they lead us to overestimate *st*.

1. Identify the obscuring generalization through correlation
2. Remove all data points which conform to the obscuring generalization
3. Re-estimate

Proposal #1 (detail)

s t i j e p u s	u s t a s ɨ n
s u m u ʔ	p a l u w o y o t͡ʃ
a s p a x a n u s	n i p o w
ʔ o x p o n u ʃ	s e
t͡s a y a ʔ	m a n ɪ s u s
q a l i w i l p i	s u s t a k u y u s
e x q e n	n i p a t
u s w a w a n u s	i ʃ o y

Table: Example words illustrating proposal #1

Proposal #1 (detail)

	st	s. . . s		st	s. . . s
s t i j e p u s	1	1	u s t a s ɨ n	1	1
s u m u ʔ	0	0	p a l u w o y o t͡ʃ	0	0
a s p a x a n u s	0	1	n i p o w	0	0
ʔ o x p o n u ʃ	0	0	s e	0	0
t͡s a y a ʔ	0	0	m a n ɪ s u s	0	1
q a l i w i l p i	0	0	s u s t a k u y u s	1	1
e x q e n	0	0	n i p a t	0	0
u s w a w a n u s	0	1	i ʃ o y	0	0

Table: Example subset of words illustrating proposal #1.
Check for correlation.

Proposal #1 (detail)

	st	s...s		st	s...s
s t i j e p u s	1	1	u s t a s ɨ n	1	1
s u m u ʔ	0	0	p a l u w o y o t͡ʃ	0	0
a s p a x a n u s	0	1	n i p o w	0	0
ʔ o x p o n u ʃ	0	0	s e	0	0
t͡s a y a ʔ	0	0	m a n ɪ s u s	0	1
q a l i w i l p i	0	0	s u s t a k u y u s	1	1
e x q e n	0	0	n i p a t	0	0
u s w a w a n u s	0	1	i ʃ o y	0	0

Table: Example subset of words illustrating proposal #1.
Check for correlation.

Proposal #1 (detail)

	st	s…s		st	s…s
~~s t i j e p u s~~	1	1	~~u s t a s i n~~	1	1
s u m u ʔ	0	0	p a l u w o y o t͡ʃ	0	0
~~a s p a x a n u s~~	0	1	n i p o w	0	0
ʔ o x p o n u ʃ	0	0	s e	0	0
t͡s a y a ʔ	0	0	~~m a n ɪ s u s~~	0	1
q a l i w i l p i	0	0	~~s u s t a k u y u s~~	1	1
e x q e n	0	0	n i p a t	0	0
~~u s w a w a n u s~~	0	1	i ʃ o y	0	0

Table: Example subset of words illustrating proposal #1.
Remove s…s words.

Proposal #1 (detail)

	st	s. . . s		st	s. . . s
~~s t i j e p u s~~	1	1	~~u s t a s i n~~	1	1
s u m u ʔ	0	0	p a l u w o y o t͡ʃ	0	0
~~a s p a x a n u s~~	0	1	n i p o w	0	0
ʔ o x p o n u ʃ	0	0	s e	0	0
t͡s a y a ʔ	0	0	~~m a n ɪ s u s~~	0	1
q a l i w i l p i	0	0	~~s u s t a k u y u s~~	1	1
e x q e n	0	0	n i p a t	0	0
~~u s w a w a n u s~~	0	1	i ʃ o y	0	0

Table: Example subset of words illustrating proposal #1.
Estimate SL2 again.

Results

- Only 14 of the 29 *st* words are in *s. . . s* words!
- The other 15 are within morphemes.

x	p	t	k	q	x$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions).

Results

- Only 14 of the 29 *st* words are in $s \ldots s$ words!
- The other 15 are within morphemes.

x	p	t	k	q	x$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	24	15	28	16	511

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions). **Results after removing s. . . s words.**

Results

- Only 14 of the 29 *st* words are in $s \ldots s$ words!
- The other 15 are within morphemes.

x	p	t	k	q	x$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	24	0	28	16	511

Table: Counts of s-stop pairs in the corpus (collapsing laryngeal distinctions). **Desired Results!**

Summary of Proposal #1

- Check for an interaction between two different initial estimations of probability distributions and then revise.
- This procedure fails here because of another confound: morphological context.
- If we had a way of detecting this (e.g. Goldsmith 2001), it too could be subject to the above procedure.

Proposal #2

Search for SL2 constraints via comparison to similar sounds

- Prior knowledge of where to search can provide direct evidence not only of *st, but also of the repair (s→ʃ).
- To illustrate, compare *sx* and *ʃx* counts with a chi-squared analysis.

Searching for *st despite the confound

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

Searching for *st despite the confound

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

Searching for *st despite the confound

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

Searching for *st despite the confound

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

Chi-squared Test

x	p	t	k	q	$X \notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

x	p	t	k	q	$x \notin \Sigma - \{p, t, k, q\}$
Counts(sx)	0.106	-5.292	-0.318	0.616	1.706
Counts(ʃx)	-0.097	4.871	0.293	-0.567	-1.571

Table: Residuals from χ^2 test on counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions). $\chi^2 = 58.0274$, df $= 4$, p-value $=$ 7.53e-12.

Chi-squared Test

x	p	t	k	q	X$\notin \Sigma - \{p,t,k,q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

x	p	t	k	q	x$\notin \Sigma - \{p,t,k,q\}$
Counts(sx)	0.106	-5.292*	-0.318	0.616	1.706
Counts(ʃx)	-0.097	4.871*	0.293	-0.567	-1.571

Table: Residuals from χ^2 test on counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions). $\chi^2 = 58.0274$, df = 4, p-value = 7.53e-12. Highlighted cells $p < 0.05$ (critical value=3.84)

Unigram counts of C2 are misleading

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

x	p	t	k	q
Count(ʃx)	-3.006	7.058*	-1.265	-4.183*
Count(x)	0.618	-1.451	0.260	0.860

Table: Residuals from χ^2 test on counts of *ʃx* pairs with counts of x in the corpus (collapsing laryngeal distinctions). $\chi^2 = 1.2497$, df = 3, p-value < 2.2e-16. Highlighted cells $p < 0.05$ (critical value=3.84)

Unigram counts of C2 are misleading

x	p	t	k	q	X$\notin \Sigma - \{p, t, k, q\}$
Counts(sx)	29	29	37	20	728
Counts(ʃx)	33	134	48	18	762
Counts(x)	1333	1679	1373	1130	28029

Table: Counts of s-stop and ʃ-stop pairs in the corpus (collapsing laryngeal distinctions).

Would we conclude that **q→t/ʃ____**?

x	p	t	k	q
Count(ʃx)	-3.006	7.058*	-1.265	-4.183*
Count(x)	0.618	-1.451	0.260	0.860

Table: Residuals from χ^2 test on counts of $ʃx$ pairs with counts of x in the corpus (collapsing laryngeal distinctions). $\chi^2 = 1.2497$, df = 3, p-value < 2.2e-16. Highlighted cells $p < 0.05$ (critical value=3.84)

Proposal #2 Summary

- Prior knowledge guides the right comparisons to make correct inferences despite confounded data
- Generally, the idea is to compare ax sequences (SL or SP) with bx sequences where a and b are similar.

Conclusion

1. We corrected a misreading in earlier literature

   /s-ti-jep-us/→[stijepus], not *[ʃtijepus]

2. We identified a new well-defined learning problem and explored two different approaches
3. Correct statistical inference is possible, but only with the right model, i.e. structured probabilistic models

   (Yang 2000, Goldwater 2006, Hayes and Wilson 2008, and many others)

Acknowledgements

Thanks to

- Dr. Richard Applegate
- 2008-2009 U. of Delaware Research Fund Grant
- National Institutes of Health #7R01DC005660-07