Beyond bigrams for surface-based phonotactic models: a case study of South Bolivian Quechua

Colin Wilson
JHU

Gillian Gallaghar
NYU

SIGMORPHON Berlin August 11, 2016

Some simple English phonotactics

* Onset (| ŋ) — dorsal nasal cannot appear in syllable onset (ex. *gnat* [næt])

* Coda (| h) — glottal fricative cannot appear in syllable coda (ex. *meh* [mɛ])

* $\begin{bmatrix} -\text{syll} \\ +\text{nasal} \\ +\text{coronal} \end{bmatrix} \begin{bmatrix} -\text{syll} \\ -\text{cont} \\ +\text{dorsal} \end{bmatrix}$ — coronal nasal cannot be followed by dorsal stop — subcase of nasal place assimilation (ex. sink [sɪŋk])

Are all phonotactic generalizations as 'simple' as these?

Simplicity is theory-dependent

Much of autosegmental, prosodic, metrical phonology reduces generalizations that appear complex and non-local to simpler rules or constraints
(e.g., Goldsmith 1990; see also notational conventions of Chomsky & Halle 1968, *SPE*)

- *CCC, *CC# (Kisseberth 1970; cf. syllable-based analyses of Halle & Vergnaud 1978, Selikirk 1981, Itô 1986/1989)

Coda Onset | | ʔi ɪ ⟨k⟩ h in	vowel epenthesis after stray consonants $\emptyset \rightarrow$ i / ⟨C⟩ __ (followed by resyllabification)

- 'Long-distance' sibilant harmony (ex. Aari, Hayward 1990)

ʒ ʃ s
\| \| \|
ʒa̤ːg-er-ʃ-e 'sew-PASS-PERF-3 sg' baʔ-s-e 'bring-PFV-3 seg'

Bigrams in surface-based phonotactics

Many computational models of phonotactics with one level (*surface forms*) have restricted generalizations to unigrams, bigrams [in practice, not in principle!]

- Peperkamp et al. (2006): phoneme learner, focused on rules A → B / _ D
  - see also Le Calvez et al. (2007), Boruta (2011, 2012), Fourtassi et al. (2014), Kempton & Moore (2014), Calamaro & Jarosz (2015)
  - cf. Martin et al. (2013), Learning Phonemes With a Proto-Lexicon, *Cognitive Science* (Experiment 3)
- Kirby & Yu (2007), Lexical and phonotactic effects on wordlikeness judgments in Cantonese, *ICPhS*: bigrams vs. neighborhood density
- Hayes & Wilson (2008): maxent learner with restricted use of trigrams
- Heinz (2010): precedence-based learning, focused on 2-local patterns
  - see also Heinz et al. (2011), Jardine & Heinz (2016), Learning Tier-based Strictly 2-Local languages, *ACL* (among many other papers by Heinz and collaborators)
- Daland & Pierrehumbert (2011), Learning Diphone-Based Segmentation, *Cognitive Science* (see Cairns et al. 1997, Hockema 2006; cf. Blanchard et al. 2010)

Bigrams in surface-based phonotactics

Other models of phonotactics or segmentation with bigram restriction:

- Vitevitch & Luce (2004), Marian et al. (2012)
- Adriaans & Kager (2009), Adriaans (2011)
- Mayer et al. (2010ab), Mayer (2012) on OCP and vowel harmony
- Räsänen (2011)
- McMullin & Hansson (2014), McMullin & Allen (2015), McMullin (2016)

Beyond bigrams

Previous research has identified some trigram constraints apparently not reducible to shorter generalizations, challenging the bigram limit.

- English *spVp, *spVb, *skVk, *skVg (cf. *pub*, *cake*); *ClVl (cf. *lilt*)
  (Fudge 1969, Clements & Keyser 1983, Davis 1984, 1989, Coetzee 2004, 2005)
- Pierrehumbert (1994), Syllable structure and word structure: a study of triconsonantal clusters in English (see also Pierrehumbert 1993 on Arabic roots)
  - "The assumption that the syllable grammar is stochastic, with the likelihood of medial clusters derived from the independent likelihoods [probabilities, CW] of the component codas and onsets, made an extremely successful contribution to the characterization of medial clusters" (p. 174).
  - Trigram constraints include *CT.C, *T.CC, *C_iCC_i (ex. *lfl, *lkl, *lpl)
- Kager & Pater (2012), Phonotactics as phonology: knowledge of a complex restriction in Dutch, *Phonology* *V:$CC_{[-coronal]})_\sigma$

Outline of the talk

Goals: motivate more formal work on complex phonotactic (incl. allophonic) distributions, raise issue of data sparseness, suggest feature-based approach

- Empirical case study
  - vowel allophony, other phonotactics of South Bolivian Quechua
- Beyond bigrams
  - surface-based analysis of Quechua high and mid vowels requires trigrams
- Surface frequencies, trigram sparseness, and generalization
  - compare learning with natural class sequences vs. segment sequences
- Beyond Quechua
  - many other attested patterns of allophonic distribution force us beyond bigrams, raising the same issues of sparseness and generalization

South Bolivian Quechua phonotactics

Quechua languages

- Quechua (Kichwa, Runasimi) languages spoken by ~ 8 million people in South America: Peru, Ecuador, Bolivia, Argentina, ...
- Head-final (default SOV, modifiers precede N) but fairly free argument order
- Highly agglutinative and exclusively suffixing (N: number, seven cases, topic/focus, ...; V: subject person/number, tense, evidential, causative, reflexive, benefactive, incorporated object pronouns, ...)

wasi	-kuna	-pi	karuncha	-rqa	-ni
house	pl	locative	go away	past	1 s

- Extensive borrowing from and bilingualism in Spanish, bilingual education in Bolivia and Ecuador
- Dialect represented here is Cochabamba Quechua, variety of South Bolivian Quechua (group IIC of Quechua family, also includes Cuzco)
  (Gallagher 2010ab, 2011, 2012, 2013ab, 2014ab, 2015, 2016, to appear)

South Bolivian Quechua morpheme structure

1100+ native roots verified with a consultant

2σ (945, 86%)	CV.CV	(511)
	CVC.CV	(409)
	CV.C.VC	(19)
	CVC.CVC	(6)
3σ (125, 11%)	CV.CV.CV	(78)
	CVC.CV.CV	(23)
	CV.CVC.CV	(18)
	CVC.CVC.CV	(3)
	CV.CV.CVC	(2)
1σ, 4σ ($< 2\%$)		

Three representative suffixes

-nku '3 pl present' -spa 'gerund' -rqa '3 sg past'

South Bolivian Quechua consonants

	labial	dental	postalveolar	velar	uvular	glottal
plain	p	t	tʃ	k	q	(ʔ)
aspirate	p^h	t^h	$tʃ^h$	k^h	q^h	
ejective	p’	t’	tʃ’	k’	q’	
fricative		s	ʃ			h
nasal	m	n	ɲ			
liquid		l ɾ	ʎ			
glide		w		j		

transcription notes: ⟨r⟩ = [ɾ], ⟨y⟩ = [j]

- Plain uvular stop /q/ often realized as voiced sonorant [ʁ]
- Aspirated uvular stop /q^h/ often spirantized to [χ]
- Plain dorsal stops /k q/ spirantized to [x χ] preconsonantally, word-finally
- Glottal stop restricted to word-initial position, inserted to satisfy *[$_{PrWd}$ V (ex. /inti/ → [ʔinti] ‘sun’)
- Glottal fricative can occur unpredictably / contrastively (ex. [hatun] ‘big, tall’; [muhu] ‘seed’)

(Gallagher 2015; see also Rowe 1950, Cusihuamán 1976, Bills et al. 1969)

Laryngeal phonotactics

Plain stops/affricates and fricatives, sonorants are quite freely distributed

Aspirates and ejectives (together with [ʔ h]) are empirically quite restricted:

- Aspirates and ejectives occur only in roots and only prevocalically
- A root can contain maximally one aspirate or ejective

*$C^h \ldots C^h$	*C' ... C'
*$C^h \ldots$ C'	*C' ... C^h

  Even identical aspirates/ejectives are unattested (ex. *[tʃ'atʃ'a], cf. Bolivian Aymara - MacEachern 1999, Gallagher 2010)
- Aspirates and ejectives occur in medial position only in roots with initial fricatives, sonorants (ex. [satʃ'a] 'tree', [mat'i] 'forehead', [rak^{h}u] 'thick')
- Aspirates do not cooccur in roots with [h], ejectives do not with [ʔ]

(Gallagher 2013, 2014; see Parker & Weber 1996, Parker 1997 on Cuzco Quechua)

South Bolivian Quechua vowels

Typical three phoneme system /i ɑ u/ with allophonic variation

- Tense in open syllables ([i e ɑ o u]), lax in closed syllables ([ɪ ɛ ɐ ɔ ʊ])
- High vowels [i/ɪ u/ʊ] are in complementary distribution with mid vowels [e/ɛ o/ɔ]: mid occur in the context of uvular stops, high occur 'elsewhere'

[q'epij]	'to carry'
[q^{h}eʎu]	'lazy'
[qosɑ]	'husband'
[q'osɲi]	'garbage'
[leq'e]	'hat'
[seq'oj]	'to smack'
[hoq'o]	'damp'
[moq'ej]	'to love'
[erqe]	'son'
[p'esqo]	'bird'
[soNqo]	'heart'
[orqo]	'mountain'

no dorsals: [i u ɑ] *[e o]		
misi	'cat'	*mese
t'uru	'mud'	*t'oro
wɑsɑ	'back'	
velars: [i u ɑ] *[e o]		
siki	'behind'	*seke
ruku	'old'	*roko
pɑkɑ	'to cover'	
uvulars: [e o ɑ] *[i u]		
leq'e	'hat'	*liq'i
hoq'o	'damp'	*huq'u
ɑq^{h}ɑ	'corn beer'	

(Gallagher 2015; see also Bills et al. 1969, Laime Ajacopa 2007, Holliday 2014)

Complementary distribution of high and mid vowels

- Uvular stops condition mid vowels from both directions
  *IQ: [oqoj] ‘to swallow’ *QI: [q’epi] ‘to carry’
  - conditioning effect is not bounded by the syllable (ex. [o.qoj])
- Uvular stops condition mid vowels across an intervening consonant
  *I(C)Q: [p’esqo] ‘bird’, [orqo] ‘mountain’ (compare [soNqo] ‘heart’)
  - does not apply across a vowel (ex. [q’api]/*[q’ape] ‘bunch, bundle’)
  - data unclear for uvular fricatives, coda spirantization eliminates Q(C)V
- Uvular stops condition mid vowels across a morpheme boundary
  *I(C)Q: [take-rqa] ‘sing 3 sg past’, [hak’o-rqa] ‘grind 3 sg past’

Gallagher (2015) establishes that mid vowels are distinct from high vowels (approx. 100 Hz F1, 200 Hz F2) across these uvular environments and regardless of the conditioning uvular stop ([q q^h q’])

Key: K = velar stop, Q = uvular stop, I = high vowel, E = mid vowel, A = low vowel

Two-level analyses of high/mid vowel distribution

Distribution of non-low vowels in Quechua conforms to traditional 'conditioned allophone/elsewhere allophone' schema

Two-level frameworks can account for the pattern with simple (unigram, bigram) structural descriptions using a dorsal tier / projection

- *SPE*

  (i) Morpheme Structure Condition (MSC) *V[−high,−low]
  V[−low] ⇒ [+high] in underlying representations

  (ii) Vowel Lowering
  V[−low] → [−high] // __ [−continuant, +uvular] (applies on dorsal tier)

UR	/p'isqu/ 'bird'	/q'ipi/ 'to carry'	*/p'esqu/, */q'ipe/, etc.
Vowel Lowering	pesqo	q'epi	

If non-low vowels are instead [0high] in UR, need feature-filling Default High rule

Two-level analyses of high/mid vowel distribution

Distribution of non-low vowels in Quechua conforms to traditional 'conditioned allophone/elsewhere allophone' schema

Two-level frameworks can account for the pattern with simple (unigram, bigram) structural descriptions using a dorsal tier / projection

- *OT*

  Faith_C(high), *IQ, *QI $\gg$ *E $\gg$ Faith_V(high)

  (i) *E $\gg$ Faith_V(high) mid vowels are disallowed on the surface . . .

  (iii) *IQ, *QI $\gg$ *E except when immediately preceded / followed by a uvular (on the dorsal tier)

(hyp.) /q'ipe/	*IQ, *QI	*E	Faith_V(high)
[q'ipe]	* !	*	
[q'epe]		* * !	*
> [q'epi]		*	* *

Key: K = velar stop, Q = uvular stop, I = high vowel, E = mid vowel, A = low vowel

Single-level analysis of high/mid vowel distribution

Allophonic distribution is a type of phonotactic pattern, somewhat understudied from a computational perspective (compare syllable structure)

Recall that many computational models of phonotactics and phonotactic learning operate with a single (*surface*) level of representation

- cannot ban mid vowels from underlying representations
- cannot condition grammaticality of *E violation on input

A purely surface-based analysis of high/mid vowel distribution (constraints are inviolable and apply on the dorsal tier):

*QI, *IQ	high vowels are not allowed in uvular contexts
*¬Q E ¬Q	mid vowels are not allowed when preceded and followed by 'non-uvulars' i.e., *# E #, *# E K, *# E V, *K E #, *K E K, *K E V, *V E #, *V E K, *V E V

Beyond the bigram limit

Purely surface-based analysis of Quechua high/mid vowel distribution requires constraints more 'complex' than unigrams, bigrams

This expands the capacity of many phonotactic learning models, which must move beyond the bigram limit (and not just on the dorsal tier! see typology section)

Consequences of allowing 'complex' phonotactics:

- constraint search space becomes exponentially larger (e.g., Hayes & Wilson 2008)
- greater susceptibility to *accidental gaps* resulting from data *sparseness*
- differences emerge between feature- vs. segment- based learning models: *features provide a type of smoothing that counteracts sparseness*

Trigram sparseness in Quechua roots

Sparseness problem for language models

"Language model estimation is necessarily a sparse data problem, and so smoothing techniques become crucial to it."

"...any estimation of probabilities for a language model that depends on history necessarily suffers from the sparse data problem."

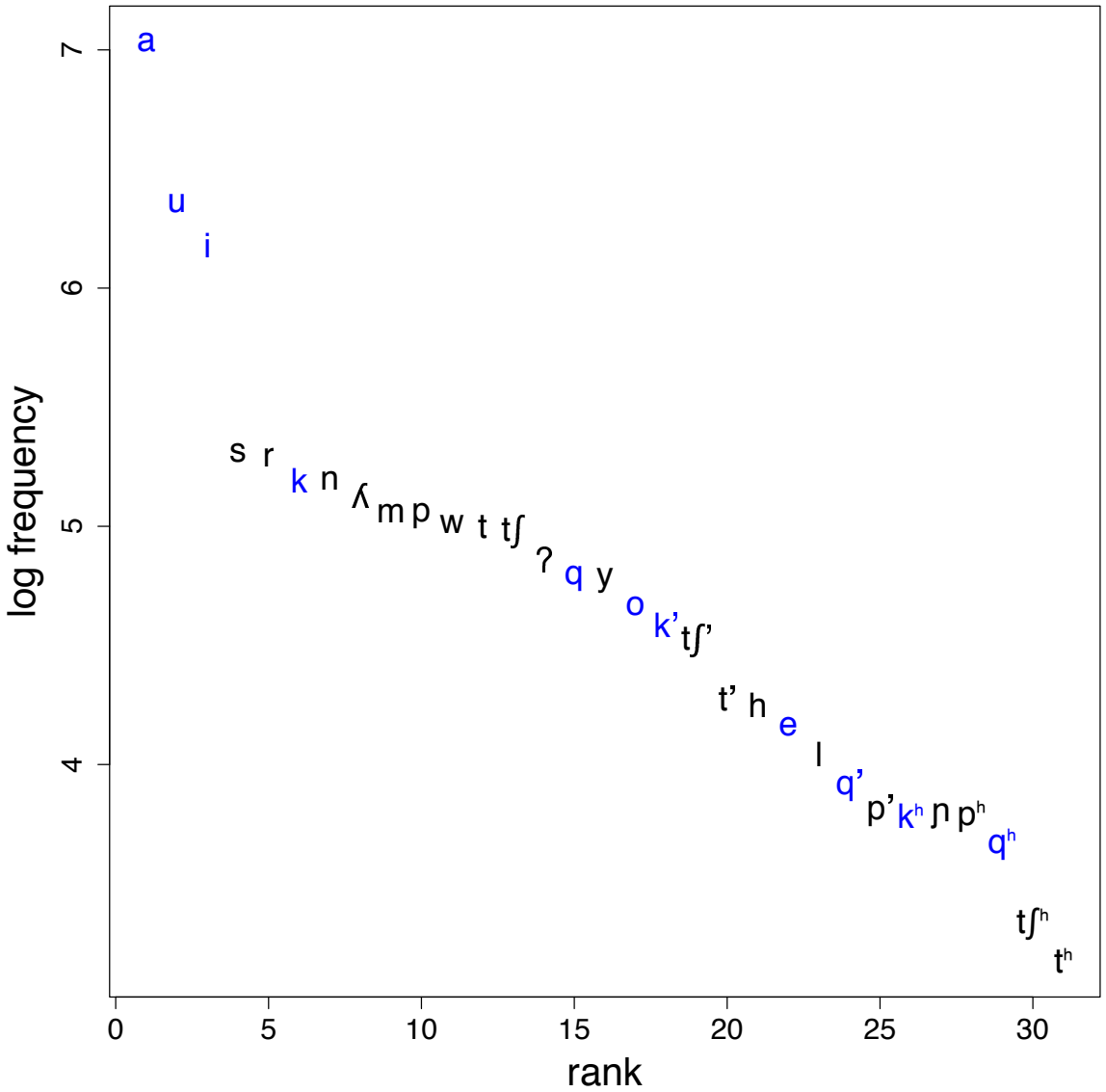
(Jelinek 1999: 75, 271)

Extensive discussion of smoothing approaches to sparseness, with primary references, in textbooks of Jelinek (1999), Manning & Schütze (1999, Ch. 6 'Statistical Inference: *n*-gram Models over Sparse Data'), Jurafsky & Martin (2009)

Does this problem arise in the phonotactic domain?

- Phoneme inventories are typically small (< 50; Maddieson 1984)
- But phoneme frequency distribution can be highly non-uniform
  (e.g., Pierrehumbert 1994; Kessler & Treiman 1997; Gorman 2011, 2013; Daland 2012)
- Allophonic conditioning can create a large set of surface segments

Segment distribution in roots (dorsal segments in blue)



(see e.g., Bengt 1968, Tambovtsev et al. 2007 on phoneme frequency distributions)

Aside: relating to Maddieson (1999), Phonetic Universals

- "...although /i/, /a/, and /u/ are all found in over 80% of the UPSID language sample, lexical counts show that, in many languages, /a/ is far more frequent than any other vowel (Maddieson 1992)."

  Quechua vowels: /a/ (1130) > /u/ (682) > /i/ (548)

- "In a count of lexical items in a sample of 25 languages of widely varied genetic and geographical groupings the most common individual syllable-onset consonants were /k/ (10 languages), /*t/ (5 languages), /l/, /s/ (2 languages each), and /m/, /n/, /d/, /h/, /r/ and /tɕ/ (1 language each)."

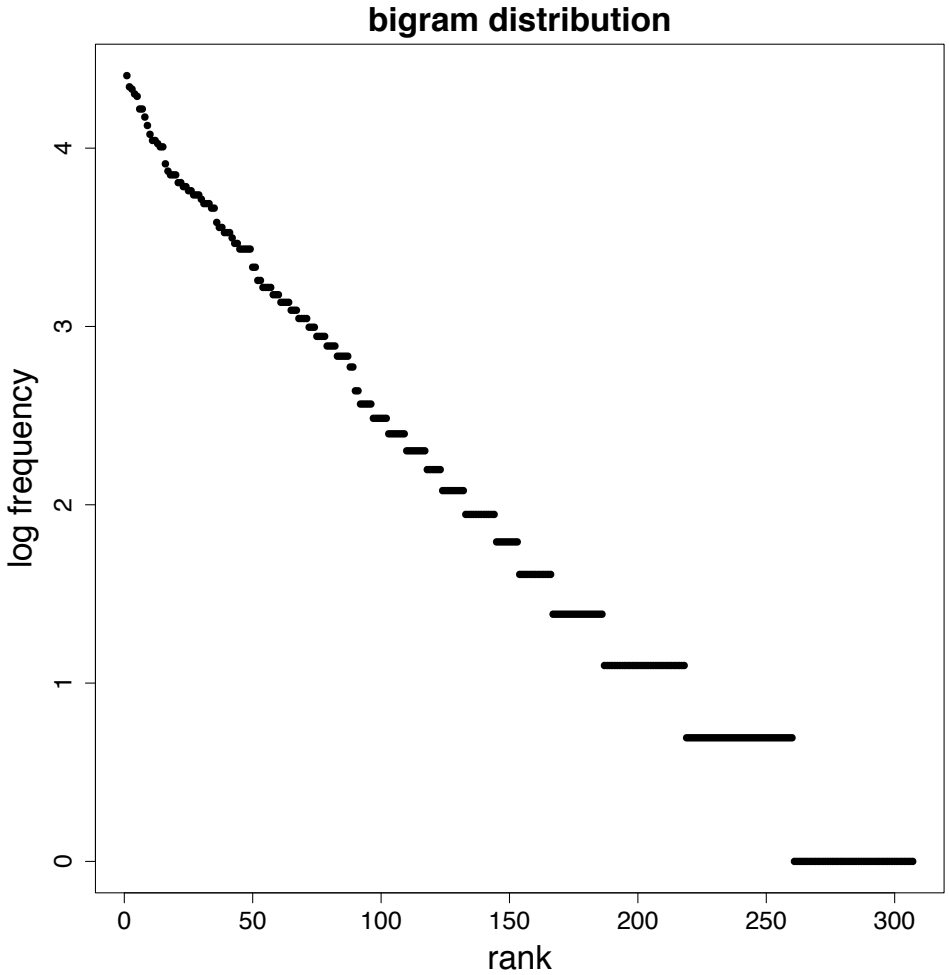
  Quechua most frequent consonants: /s/, /r/, /k/, /n/

- "There is generally a good correlation between the relative frequency of a segment type in terms of its appearance in the inventories of languages around the world and its frequency within the lexicon of particular languages."

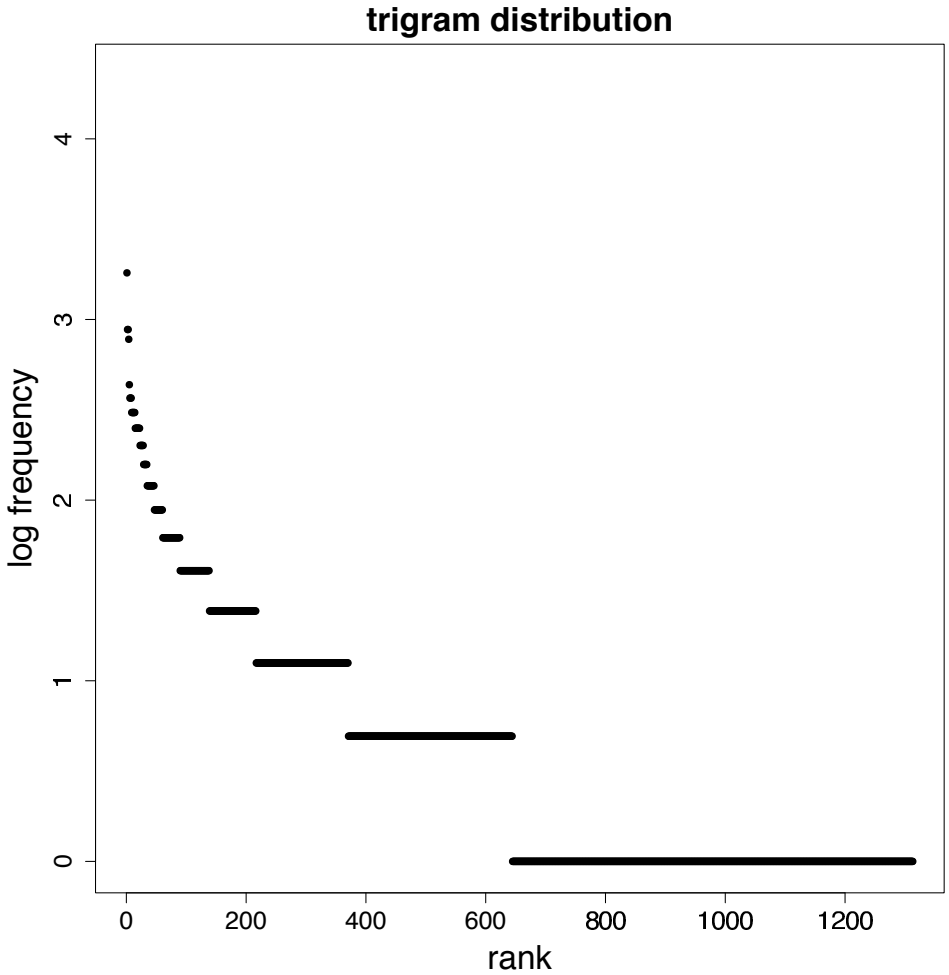
  Quechua dorsal consonants: /k/ > /q/ > /k'/ > /q'/ > /k^h/ > /q^h/

See also Lindblom & Maddieson (1988) on *basic* vs. *elaborated* articulations

Bigram, trigram distribution in roots



$> 15\%$ have frequency of 1



$> 50\%$ have frequency of 1

These figures exclude $\#x$ and $x\#$ grams, some of which are highly frequent

Non-occurring legal trigrams: VCV

Non-occurring trigrams $V_iC_jV_k$ that are composed of attested bigrams (V_iC_j, C_jV_k) and satisfy known phonotactic restrictions of Quechua roots

ikʰi ik'a ile ilo ilu iʎi iʎo ime imo imu ini ino inu ipe ipo ip'e ip'i ip'o ise iso ite ito itʰi itʰo itʃo itʃʰa itʃʰe itʃʰi itʃʰo itʃ'e itʃ'i itʃ'o it'e it'o it'u

eʎi eʎo ema eme emi emo emu eni eno enu epe epo eqʰa eqʰo eq'e era eri esa ese esi eso esu etʃa etʃo

akʰi ale ali alo aʎo ame amo ano ape apo ap'e ap'o ap'u aq'e ase aso ato atʰa atʰo atʃo atʃʰa atʃʰe atʃʰo atʃʰu atʃ'e atʃ'o at'e at'o

ola ole olo olu oʎa oʎi oʎo oma ome omo omu ona oni ono onu oɲu ope opi opo oqʰe oqʰo ose oso osu ota ote oto otʃi otʃo otʃu owe owu oyo oyu

ukʰi ule uli ulo uʎo ume umo uni uno upe upo upʰu use ute utʰa utʰi utʰo utʃo utʃʰe utʃʰi utʃʰo utʃʰu utʃ'e utʃ'o ut'e ut'o

N = 140+

Non-occurring legal trigrams

Other common root sequence types:

- CVC
- VCC
- CCV

Expect trigram distribution to be sparse, lacking many phonotactically legal sequences, in these substructures as well

Ssparseness on dorsal and other tiers will depend on (i) how many segments are projected onto the tier and (ii) whether segments are projected as individual units or merged into classes (ex. dorsal tier Q, K, I, E, A)

Learning Quechua phonotactics

Whole-language analysis

Attempt to learn the phonotactics of the entire language, both roots and inflected forms, on several appropriate tiers (see Hayes & Wilson 2008 on Wargamay)

tier	projected segments	phonotactic generalizations
dorsal	dorsal consonants (K Q), vowels (I E A)	high/mid complementarity
C-dorsal	dorsal consonants	*K...Q, *Q...K (morpheme-bounded)
laryngeal	stops, affricates, h ʔ	laryngeal cooccurrence restrictions
C/V	(all)	*#V, *VV, *CCC
segmental	(all)	consonant cluster restrictions, etc.

Considering additional allophonic detail would only increase sparseness (ex. many instances of [e o] would become [ɛ ɔ] due to closed-syllable laxing)

Two learning models

(1) Feature-based maximum entropy (Maxent; version of Hayes & Wilson 2008)

- Natural classes derived from feature matrix
  (ex. Q = [−continuant, +dorsal, −high] = {q, q^{h} q'})

- Negative phonotactic constraints stated over classes, on tiers
  (ex. dorsal:*QI = *[−continuant, +dorsal, −high] [+syllabic, +high])

- Probability of surface form given by weighted constraint violations
  $p(x) \propto \exp\left(-\sum_{k=1}^{M} w_k \cdot c_k(x)\right)$ each $w_k \geq 0$

- Greedy constraint induction according to *gain* heuristic
  (Della Pietra et al. 1997, Perkins et al. 2003)
  - estimate improvement to $\log p(data)$ from adding each constraint c
  - add constraint c^* with highest gain above threshold γ
  - halt when no constraint has gain above threshold

Initialization: one constraint per surface segment, one per word length

Two learning models

(2) Segment-based sequence model

(TSL; version of Heinz et al. 2011 with multiple tiers as in McMullin & Allen 2015)

- Learner initialized with $G_t = \emptyset$ for each tier t
- For each word x in learning data, for each tier t
  add sequences of length $1 \leq m \leq 3$ present in x on tier t to G_t

ex. if [q'epi] is first word processed in learning, grammar becomes

tier (t)	G_t
dorsal	{ #, q', e, i; #q', q'e, ei, i#; #q'e, q'ei, ei# }
C-dorsal	{ #, q'; #q', q'#; #q'# }
laryngeal	{ #, q', p; #q', q'p, p#; #q'p, q'p#}
CV = segmental	{ #, q', e, p, i; #q', q'e, ep, pi, i#; #q'e, . . . , pi# }

- After learning, a word x is grammatical iff, for every tier t, all of its sequences of length $1 \leq m \leq 3$ are present in G_t

Model similarities

- Provided with same segments, tiers, and maximum constraint length
- Feature-based constraints in Maxent are equivalent to segment-based constraints with tied weights (see class-to-segment compilation in FSA toolkits)
  ex. *QI $= \{$*qi, *qu, *q$^{\text{h}}$i, *q$^{\text{h}}$u, *q'i, *qu $\}$
- 'Positive' constraints in TSL can be equivalently (though not as efficiently) expressed with negative constraints: $G_t \rightarrow \bar{G}_t$

Model differences

- Maxent constraints can be violable, unlike those of TSL — for purpose of comparison we required all learned constraints to be surface-true
- Maxent constraints are weighted, unlike those of TSL — we defined a word to be grammatical iff it violates no learned constraint
- Maxent grammars naturally generalize beyond the attested sequences (n-grams), unlike TSL — this is the important difference
  - TSL effectively learns a constraint *xyz for every trigram not observed in the learning data (and similarly for bigrams)
  - Maxent learns only constraints that have *observed* violations significantly smaller than would be *expected by chance*
    - chance calculation takes into account unigram frequencies via the baseline constraints, other info from previously learned constraints

See Pierrehumbert (1994), Gorman (2013) on assessing 'complex' phonotactics w.r.t. chance

Model evaluation

Models were evaluated with 5-fold cross-validation on the corpus of root and inflected forms, with additional tests for restrictiveness (e.g., laryngeal coccurrence) and generalization (e.g., exhaustive testing on VCV sequences)

		Maxent	TSL
sim1	held-out	.999 (0.002)	.954 (0.010)
	legal test	1.00	.546 (0.025)
	illegal test	0	0
sim2	held-out	.999 (0.002)	.959 (0.095)
	legal test	1.00	.562 (0.055)
	illegal test	0	0
sim3	held-out	.999 (0.001)	.962 (0.004)
	legal test	1.00	.569 (0.026)
	illegal test	0	0

0-2 of 880+ held-out forms rejected by Maxent (cf. 30-39 by TSL)

Other models under evaluation: Precedence Learner (Heinz 2010), segment-based Maxent (see also Heinz & Rodgers 2010)

Learned Maxent constraints

- Dorsal tier

*[−high][+high]	(generalization of *QI)
*[+high][−high]	(generalization of *IQ)
*#[−high,−low][+syllabic]	(subcases of *¬ Q E ¬ Q)
*#[−high,−low][+high]	
*[+syllabic][−high,−low]#	
*[+syllabic][−high,−low][+high]	
*[+syllabic][−high,−low][+syllabic]	
*[+high][−high,−low]#	
*[+high][−high,−low][+syllabic]	
*[+high][−high,−low][+high]	

These constraints generalize to accept dorsal-tier trigrams such as [oq^{h}e] that do not occur in the lexicon

Learned Maxent constraints

- C-dorsal tier
  - *[+high][−high] (*K . . . Q)
  - *[−high][+high] (*Q . . . K)
- Laryngeal tier
  - *[−continuant][+constricted glottis] (no stops before ejectives in roots)
  - *[−spread glottis][+constricted glottis] (accounts for *ʔ . . . C')
  - *[−sonorant][+spread glottis] (generalization of *C'/C^h . . . C^h)
  - *[−son, +c.g.][−son][−son] ?
  - *[−s.g.][−c.g.][−son] ?
- C/V tier
  - *[+syllabic][+syllabic] (*VV)
  - *#[+syllabic] (*#V)
  - *#[][−syllabic] (effectively *#CC)
  - *[−syllabic][−syllabic][−syllabic] (*CCC)
  - *[−syllabic][−syllabic]- (*CC at morpheme edge)

Typology of ‘complex’ phonotactics

Typology of surface trigram phonotactics

Vowel lowering / retraction by consonants

- Many other Quechua languages (e.g., Cuzco)

- Chilcotin (Northern Athabaskan; Cook 1992)
  Vowels have 'flattened' allophones after / before 'flat' (pharyngealized, [+RTR]) consonants, 'sharp' allophones elsewhere
  i → ᵊi/e, u → o, æ → a
  ɪ → ᵊɪ, ʊ → ŏ, ɛ → ə

"Similar patterns are seen in many languages with uvulars or other post-velar consonants, including Eskimo-Aleut languages (Rischel 1972; Dorais 1986), Interior Salish languages (Bessell 1998), Nuu-chah-nulth (Wakashan) (Wilson 2007), Chilcotin (Athabaskan) (Cook 1983, 1993; Bird 2014), Aymara (de Lucca 1987; Adelaar with Muysken 2004) and many varieties of Arabic (McCarthy 1994; Shahin 2002; Zawaydeh 1998; Al-Ani 1970; Butcher and Ahmad 1987)" (Gallagher 2015: 2)

Typology of surface trigram phonotactics

Tongue root (ATR / RTR) vowel harmony systems (e.g., Casali 2003, 2014)

- Luo (Dhulou; Western Nilotic; Swenson 2015)
  [+ATR] low vowel [ʌ] occurs only after / before [+ATR] vowels,
  [−ATR] low vowel [a] elsewhere

- Mundari (Eastern Nilotic; Stirtz 2014)
  [+ATR] mid vowels [e̤ o̤] ("about half way between [ɛ] and [e], [ɔ] and [o]") occur only after / before [+ATR] vowels, [−ATR] [ɛ ɔ] elsewhere

"In a good number of nine-vowel /i ɪ e ɛ a ɔ o ʊ u/ or seven-vowel /i ɪ ɛ a ɔ ʊ u/ languages (e.g., Ahanta, Akan, Anum, Chumburung, Dagara, Didinga, Efutu, Gichode, Gonja, Kinande, Krachi, Lama, Lugbara, Nawuri, Nkonya, Talinga-Bwisi and Waja), for example, the low vowel /a/ has a [+ATR] allophone derived via assimilation to neighboring [+ATR] vowels" (Casali 2014: 7)

"A good number of /2IU/ languages which have seven-vowel /i ɪ ɛ a ɔ ʊ u/ languages (or the same basic system plus an additional central vowel)... disallow mid [-ATR] vowels phonetically either preceding or following high [+ATR] vowels, due to a process that realizes /ɛ/, /ɔ/ as [+ATR] allophones [e], [o] in the relevant context" (Casali 2014: 10)

Typology of surface trigram (morpho)phonotactics

Nasal harmony frequently creates non-contrastive segments and can be bidirectional (e.g., Walker 2000)

Conjunctive (both-side) allophone conditioning

- tut $\rightarrow$ tyt (e.g., Cantonese; Flemming 1995 / 2002)
- intervocalic consonant voicing / lenition (e.g., Kirchner 1998, Kaplan 2010)

Gradient intervention effects and double triggers in vowel harmony

- BN vs. BNN Hungarian stems (e.g., Ringen & Kontra 1989, Hayes & Londe 2006)
- $V_{[+rnd]}$ $V_{[+rnd]}$ vs. $V_{[+rnd]}$ Oroqen stems (e.g., Zhang 1996, Walker 2001)

(See also Frisch et al. 2004, Zymet 2014, McPherson & Hayes 2016 on distance effects)

Open question: are there effects beyond the trigram window (on appropriate tiers)?

Summary

- Several computational models of phonotactics have restricted attention to 'simple' surface generalizations
- Phonotactic patterns such as Quechua high/mid vowel distribution (and many more in the typology) require 'complex' surface constraints
- Complexity raises the texbook problem of sparseness (here, under-attestation of surface trigrams) at least for low-resource lexicons
- Feature-based probabilistic models, which learn constraints only against statistically significant 'gaps', are a promising approach to this problem

Open questions

- Degree of restrictiveness / generalization in Maxent model depends on one parameter (gain threshold) — how should it be estimated from the data?

- Generalization ability of feature-based statistical models remains a matter of simulation, calling for proofs — are there existing results?

- Does the sparseness problem for phonotactic learning remain with a much larger learning corpus?

  ex. Quechua lexicon of AntiMorfo (Gasser 2009, 2011), others?

- Should we reject one-level phonotactic models in favor of traditional, two-level approaches? How many 'complex' constraints would remain?

- More generally, what are the costs and benefits of underlying representations (a kind of hidden structure) for phonotactic learning?

Please let us know about other phonotactic models of interest to you!

Thank you!

Thanks to Cochabamba Quechua speakers and experimental participants.
This work was partially supported by NSF grant BCS 1222700 to GG.