Computational Phonology, class 1: Introduction, and N-gram Models

Adam Albright

CreteLing 2022 — July 2022

SCAN ME

creteling2022.computational.phonology.party

Why computational phonology?

The many faces of computational phonology

- An engineering problem
- Distinct theoretical subfield
- Component of theoretical phonology

An engineering problem

As an engineering problem

- Encode allophony or co-occurrence restrictions, to improve accuracy or streamline TTS, speech recognition, spell-checking, etc.
- Encode variable processes (casual speech reduction, etc.)
- Characterize properties of different dialects and languages for automated language recognition

Theoretical questions

Provide guarantees about correctness or computational complexity of theories of phonological derivation or learning

- “The Recursive Constraint Demotion algorithm provably converges to a grammar that is consistent with the data” (Tesar & Smolensky 2000)
- “Gradual promotion and demotion of constraints provably converges to a grammar that is consistent with the data” (Magri 2013)
- “Phonological processes involve subregular computational complexity (Tier-based Strictly Local)” (Heinz, Chandlee, Jardine, etc.)

Theoretical questions

Highlight potential issues

- “Generating outputs in OT is NP-hard”

  (Eisner 1997; Idsardi 2006 *LI* squib and subsequent exchange with Kornai; followup by Riggle and colleagues)

- “The formal power needed to express correspondence constraints may ultimately make their evaluation intractable”

  (Potts and Pullum 2002 *Phonology* paper)

Theoretical questions

Or something in between…

- “Robust Interpretive Parsing with Resampling (RRIP) allows a learner to converge on the correct stress grammar 84% of the time using Optimality Theory with gradual constraint promotion/demotion” (Jarosz 2013)

The many faces of computational phonology

Such work is *computational* in the sense that it involves formalizing the theory in such a way that we can evaluate its computational properties. Actually *implementing* the algorithm is not necessarily instructive for these purposes.

The many faces of computational phonology

Find faster/more efficient ways to perform phonological derivations or learning

- Finite state formalization (e.g., Eisner 1999; Riggle 2004 UCLA dissertation)
- Heuristic search methods
  - Genetic algorithms (Belz and Eskikaya 1998)
  - Simulated annealing: (Bíró 2006 Groningen diss.)

The many faces of computational phonology

Find faster/more efficient ways to perform phonological derivations or learning

- Finite state formalization (e.g., Eisner 1999; Riggle 2004 UCLA dissertation)
- Heuristic search methods
  - Genetic algorithms (Belz and Eskikaya 1998)
  - Simulated annealing: (Bíró 2006 Groningen diss.)

General thrust: replace costly or uncertain procedures with ones that are more efficient, or whose properties can at least be evaluated. (And hope that the theory can still capture more or less the same facts.)

How do computational results inform theory?

- We better if we know we’re working with a system that makes phonology tractable
  - Prudent to start with mathematically well-defined models, keeping them as simple as possible
  - Elaborate slowly & cautiously to improve empirical coverage

How do computational results inform theory?

- We better if we know we're working with a system that makes phonology tractable
  - Prudent to start with mathematically well-defined models, keeping them as simple as possible
  - Elaborate slowly & cautiously to improve empirical coverage
- Yet the field "at large" often ignores such issues
  - Pragmatic approach: theory must also accommodate huge piles of complicated and diverse data
  - Often useful to pursue insights without knowing exactly how to formalize them $\Rightarrow$ preliminary sense of the fit to the data
  - If they work out well, hope that an elegant and tractable formalization can constructed post hoc

The many faces of computational phonology

Computational models as tools for developing theories

- Prosthetic thought[1]
  - Testing out ideas that would be impractical or impossible to check by hand
- Analytical hygiene
  - Verifying that a model derives the claimed output
- Baselines for evaluation
  - Qualitative and quantitative comparison of models with/without a given assumption
- Modular theorizing
  - Easily compare models that incorporate different assumptions while holding all else equal

[1] Term borrowed from Bruce Hayes.

Goals of this class

- Introduction to some basic computational concepts and approaches that are allowing phonologists to better understand humans and theories
- Demystify modeling
- I do not assume
  - Prior experience with programming or computational modeling
  - Prior knowledge of Optimality Theory, Harmonic Grammar, Maximum Entropy models, etc.
- I do assume
  - Some familiarity with phonological terminology and concepts (contrast, features, etc.)

Modeling humans

Hockett (1955) *Manual of Phonology*

- p. 147 “We know of no set of procedures by which a Martian, or a machine, could analyze a phonologic system—an entity, that is, to which even the basic biologic and cultural common denominator of humanness would be alien and would require specification. The only procedures which can be described are rules for a human investigator, and depend essentially on his ability to empathize.”

Hockett (1955) *Manual of Phonology*

- p. 147 “We know of no set of procedures by which a Martian, or a machine, could analyze a phonologic system—an entity, that is, to which even the basic biologic and cultural common denominator of humanness would be alien and would require specification. The only procedures which can be described are rules for a human investigator, and depend essentially on his ability to empathize.”
- The challenge: what do we need to build into a model in order to simulate phonological empathy?

Simulating humans

- Goal: use computational models to shed light on the phonological knowledge that humans have, and how they got it
  - Innate knowledge
  - Acquired knowledge
- Assessing our models
  - Ability to replicate the training data
  - Ability to simulate human judgments and productions (Turing test)

How do humans demonstrate phonological knowledge?

- Judgments of identity and difference (contrast)
  - Hockett (1955): American English *wood* and *would* have the same vowel, *wood* and *wooed* have different vowels
- Phonotactic acceptability
  - Chomsky and Halle (1965): *blick* [blɪk] is a possible word of English, *bnick* [bnɪk] is not
- Alternations
  - Halle (1978): English plural *Bach*-[s] (cf. *Jarre*-[z])

We'll focus in this class on modeling phonotactics and alternations

The *blick* test

Halle (1978) “Knowledge unlearned and untaught: What speakers know about the sounds of their language.”

- Which of the following could be words of English?

ptæk	plæst	vlæs
θoʊl	sram	flɪtʃ
hlad	mgla	rtut

- What is the probability that [ptæk] could be a word of English?
  - Wordlikeness judgments
- Related, but distinct: how acceptable is [ptæk]?
  - Acceptability judgments

Evidence that infants learn about inventories

Jusczyk et al. (1993) Infants' sensitivity to the sound pattern of native language words

- Experiment
  - Lists: 15 abstract low frequency words each
  - Experiment: headturn preference procedure, American 6-month olds and 9-month olds
    - Infant seated on caregiver's lap in booth
    - Green light directly ahead, "centers" infant's attention
    - Then red light on left or right blinks, and sound comes from speaker on that side
    - Trial ends when infant looks away for more than 2 secs
- Result: 9-month olds (but not 6-month olds) listen longer to English words than to Dutch words
- Many possible ways of discriminating: sounds, sound combinations, words (unlikely)

N-grams

A very simple model

- As infants are exposed to speech, they accumulate evidence about the relative frequency of sounds
- By 9 months of age, infants have collected enough statistics about English to recognize that [θ] occurs, but [x] does not
- Task behavior: look at loudspeaker in proportion to frequency of sounds coming through (≈ ‘familiarity’)

Of course, there are many other possible interpretations, too! This is merely a simple statistical baseline model.

A warm-up model: unigrams

- Count the frequency of English phones
- Calculate probability of each phone in the English lexicon
- Generalization: probability of a word is joint probability (=product of probabilities) of its phones
- Is this enough to distinguish English from Dutch, for an English-learning child?

Bigram probability for sequences

Jusczyk et al. (1993), continued… (Exp 4)

- Are 9-month olds going beyond inventories, and also learning about combinatory possibilities?
- Test: lists of words restricted to sounds that (roughly) occur in both languages, but with combinatoric violations
  - English-only combinations: final voiced obstruents (*kudos*, *cubeb*, *aboard*), word-initial schwa (*astound*)
  - Dutch-only combinations: initial [kn] (*knoest*), initial [zw] (*zweten*), [lmp] (*zalmpjes*)

Bigram probability for sequences (*cont.*)

- Tested American and Dutch infants, 6 and 9 months old
- Results
  - American 9-mo. olds listen significantly longer to English words
  - Dutch infants listen slightly (but not significantly) longer to Dutch words
  - Post hoc survey revealed that Dutch infants hear (on avg.) 1.25 hrs of English per day
  - 6 month olds don’t show same preference

Distinguish low vs. high bigram frequency

Some possibly confirming evidence: Jusczyk et al. (1994) Infants' sensitivity to phonotactic patterns in the native language

- Refinement to previous finding: common vs. rare English sequences
- High probability words
  - [rɪs], [gɛn], [kæz], [ʃæn], [sɛtʃ]
- Low probability words
  - [jaʊdʒ], [ʃɔtʃ], [ðʌʃ], [fuv], [θɔʃ]
- Tested 6- and 9-month old, as before
- Results: 9-month olds (but not 6-month olds) attend longer to high probability words

Counting combinations

The simplest type of combinations: n-grams

- Substrings of letters/segments/etc.
  - N-grams = substrings of length N
  - E.g., bigrams: *aardvark* [aɹdvaɹk] → [aɹ], [ɹd], [dv], …

Bigram probability

- Bigram frequency $=$ Count of bigram in corpus
- Probability $= \dfrac{\text{Count of bigram in corpus}}{\text{Count of all bigrams in corpus}}$
- Problem: how to combine probabilities of bigrams within a word to yield a probability or score for the entire word?

Bigram probability

Combining bigram frequencies into scores for an entire word (*abcd*)

- Common move in the psycholinguistics literature: either *joint* or *average* bigram probability
  - Calculate prob. of component bigrams (*ab*, *bc*, *cd*)
  - Multiply them (joint probability) or average them (average probability)

Bigram probability

Combining bigram frequencies into scores for an entire word (*abcd*)

- Common move in the psycholinguistics literature: either *joint* or *average* bigram probability
  - Calculate prob. of component bigrams (*ab*, *bc*, *cd*)
  - Multiply them (joint probability) or average them (average probability)
- However, the bigrams of a word are not independent of one another!!
  - If bigram 1 *ab* ends in *b*, bigram 2 must begin with *b*
  - i.e., by the time we get to bigram 2, the *b* is “given”
  - It makes sense at each point to focus on probability, given the current segment, that the upcoming segment should occur next

Conditional/transitional probability

- Conditional probability of *b* given *a*
  - P(b|a) = P(ab) / P(a)
- N-gram probability of string *abcd* = joint conditional probability of each segment based on preceding N-1 segments
  - Bigrams: P(*abcd*) = P(a) × P(b|a) × P(c|b) × P(d|c)
- In practice, usually sensible to avoid a word/sentence including boundaries: *#abcd#*

Observing bigram probability

- Jusczyk et al (1994): “high” vs. “low” probability nonce words
  - Are they distinguished by unigram probability?
  - Are they distinguished by transitional bigram probability?

(See Google Colab worksheet for a test)

The *blick* test in the lab

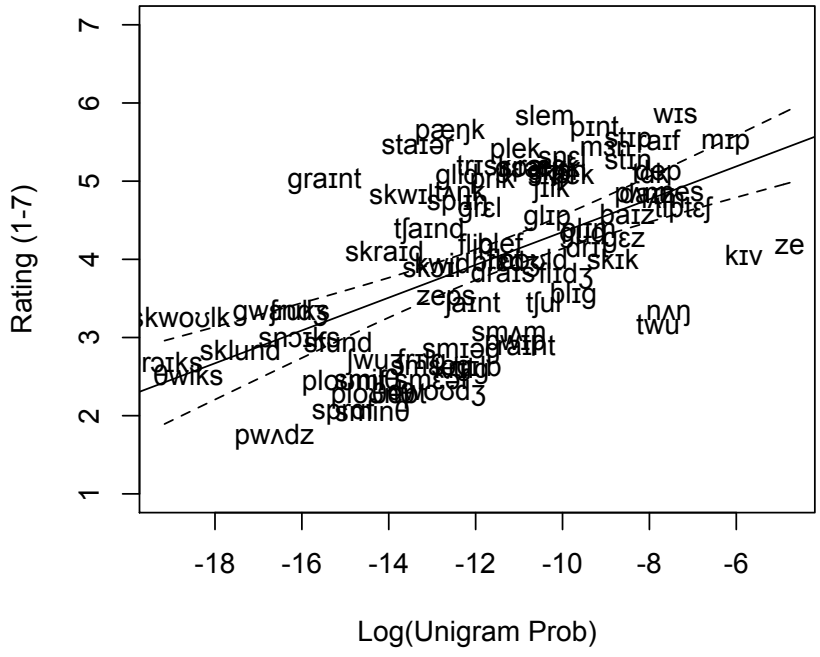
Albright and Hayes (2003): norming data for wug test of English past tenses

- Ultimate goal of study: assess acceptability of different past tense processes for nonce verbs of various shapes
  - *John likes to stin. Yesterday he stinned.*
- In theory, reactions to potential past tense forms could be influenced by two distinct factors
  - Morphological acceptability of regular -ed with stin (as opposed to stan, stun, …)
  - Phonotactic acceptability of the root (if stin is odd, then stinned will be too)
- Norming pre-test: collect ratings of phonotactic acceptability of the stems
  - How plausible would stin be as a word of English?

The *blick* test in the lab

slem	5.8	tɑk	5.1	gud	4.3	gwɛndʒ	3.3	smilθ	2.5
wɪs	5.8	tʃek	5.1	blef	4.2	ʃruks	3.3	ploʊmf	2.4
pɪnt	5.7	glid	5.1	gɛz	4.2	nʌŋ	3.3	dwoʊdʒ	2.3
pæŋk	5.6	graɪnt	5.0	drɪt	4.2	skwoʊlk	3.3	ploʊnθ	2.3
raɪf	5.5	prik	5.0	flip	4.2	twu	3.2	θept	2.3
stɪp	5.5	ʃɪlk	4.9	ze	4.2	smʌm	3.1	sminθ	2.1
mɪp	5.5	daɪz	4.8	skraɪd	4.1	snɔɪks	3.0	sprɑf	2.1
staɪər	5.5	nes	4.8	kɪv	4.1	sfund	2.9	pwʌdz	1.7
mɜn	5.4	tʌŋk	4.8	flɛt	4.0	pwɪp	2.9		
plek	5.4	skwɪl	4.8	noʊld	4.0	raɪnt	2.9		
snɛl	5.3	lʌm	4.8	skɪk	4.0	sklund	2.8		
stɪn	5.3	pʌm	4.8	brɛdʒ	3.9	smɪəg	2.8		
ræsk	5.2	splɪŋ	4.7	kwid	3.9	frɪlg	2.7		
trɪsk	5.2	grɛl	4.6	skɔɪl	3.9	ʃwuʒ	2.7		
spæk	5.2	tip	4.6	draɪs	3.8	θrɔɪks	2.7		
dep	5.1	tɛʃ	4.6	flɪdʒ	3.8	trɪlb	2.6		
gɛər	5.1	baɪz	4.6	blɪg	3.5	krɪlg	2.6		
glɪt	5.1	glɪp	4.5	zeps	3.5	smɛəg	2.6		
ʃɜn	5.1	tʃaɪnd	4.4	tʃul	3.4	θwiks	2.5		
skɛl	5.1	plɪm	4.4	ʃaɪnt	3.4	smɛrf	2.5		

Unigram probabilities (r = .575)

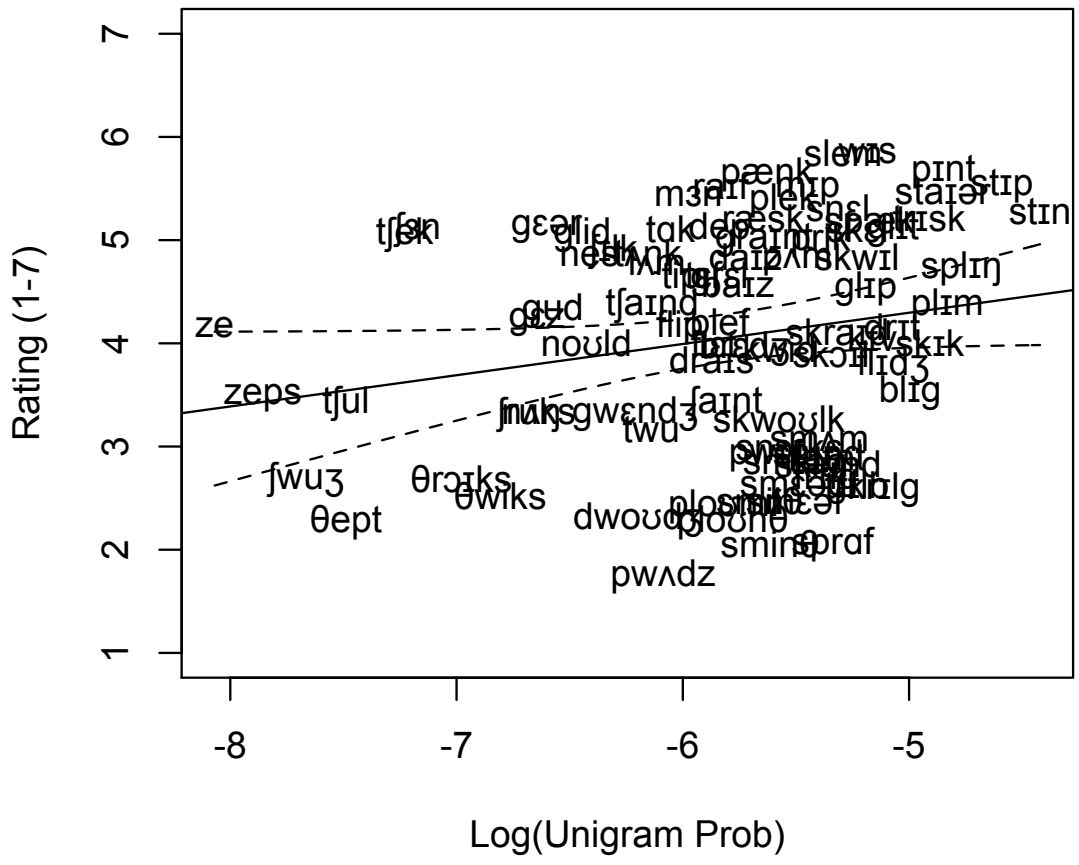


- Unigram probabilities alone already do fairly well
- Phonotactic restrictions ⇒ rare/unattested combinations ⇒ lower segment frequency

Thought experiment

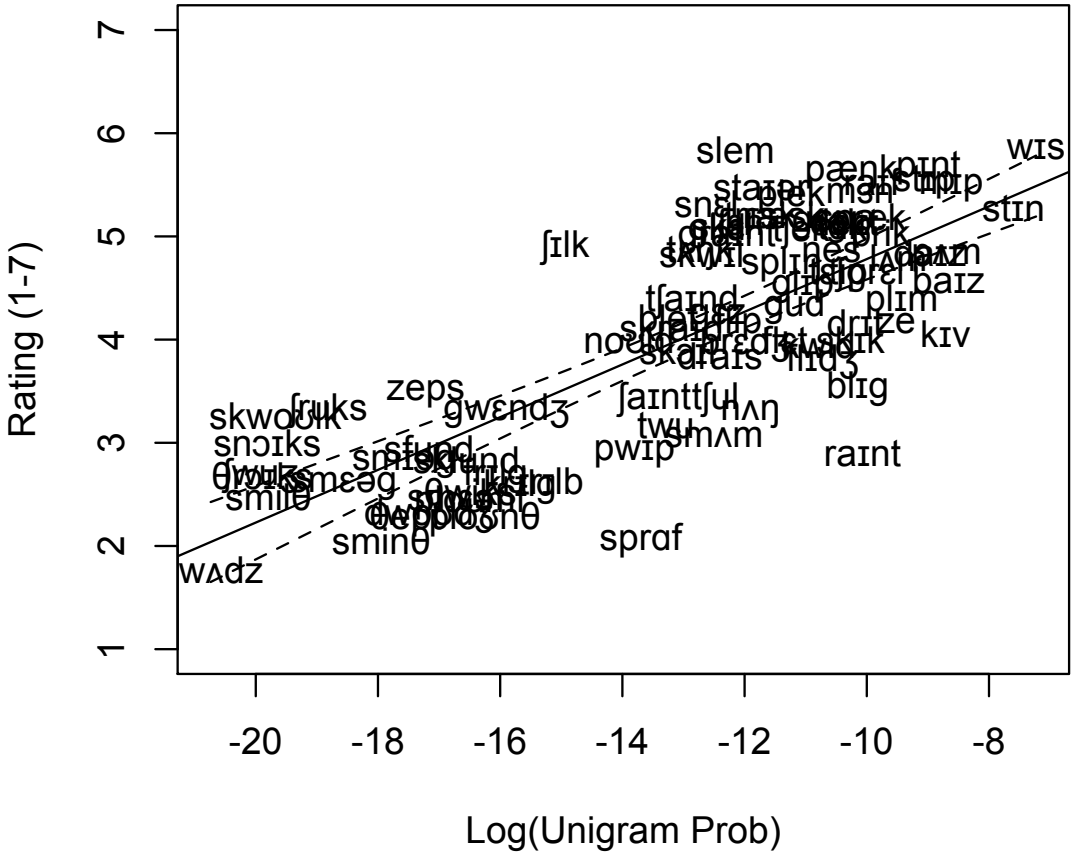
- Propose an experiment to show that unigram frequency is insufficient as a model of human judgments

Average bigram probability (r = .203)



- Averaging bigram probabilities is a much worse model
- Word length effects
- “Bottleneck” effects

Transitional bigram probability (r = .771)



- Transitional bigram probability substantially better
- Initial/final restrictions, and rare vs. common bigrams

A recurring finding

- For attested (common vs. rare) sequences, transitional bigram probability is hard to beat as a model of acceptability judgments
  - Baseline: a simple model that sets the bar for more sophisticated models
- Important to remember: most acceptability experiments focus on acceptable or nearly-acceptable items

Positional bigram probability

Position matters

- Often, the probability of a given sequence depends on its position within the word
  - E.g., [ʒɻ] is much more common word-finally than word-initially
- A certain amount of positional dependence is already dealt with by using transitional probabilities
  - P(ɻ|ʒ) may be sort of high, but P(ʒ|#) is very low
- Another strategy in the literature: separate counts depending on position

Positional bigram probability

Jusczyk et al. (1994), p.

- “We operationally defined phonotactic probability based on two measures: (1) positional phoneme frequency (i.e., how often a given segment occurs in a position with a word) and (2) biphone frequency (i.e., the phoneme-to-phoneme cooccurrence probability)…All probabilities were computed based on log frequency-weighted values. The average summed phoneme probability was .1926 for the high-probability pattern list and .0543 for the low-probability pattern list.”

Positional bigram probability (*cont.*)

- “A high-probability phonotactic pattern also consisted of frequent segment-to-segment cooccurrence probabilities. In particular, we chose CVC phonetic patterns whose initial consonant-to-vowel cooccurrences and vowel-tofinal consonant cooccurrences had high probabilities of occurrence in the computerized database. For example, for the pattern /ɹɪs/, the probability of the cooccurrence /ɹ/ to /ɪ/ was high, as was the cooccurrence of /ɹ/ to /s/”

Vitevitch and Luce's Phonotactic Probability Calculator

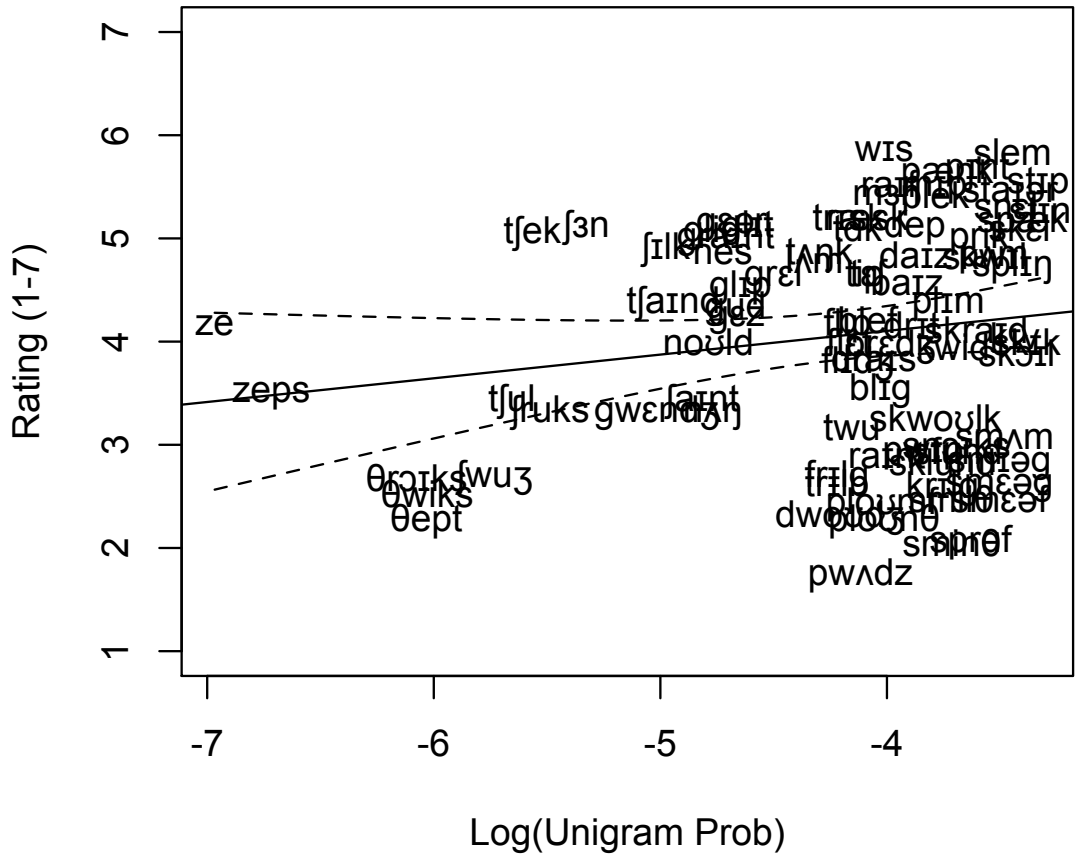
Vitevitch & Luce (2004) Phonotactic probability calculator[1]

- Position-independent bigram probability: P(*xy*) = $\frac{\text{(Frequency-weighted) count of } xy}{\text{(Frequency-weighted) count of all bigrams}}$
- Positional bigram probability: P(*xy*@position *n*) = $\frac{\text{(Frequency-weighted) count of } xy \text{ at position } n}{\text{(Frequency-weighted) count of all bigrams at position } n}$

- Positions: 1st, 2nd, 3rd, … phoneme of word
  - E.g., [st] in 3rd position: *best*, *abstain*, *austere*, …
- In point of fact, this (rather than average biphone frequency) is what Jusczyk et al. use

[1] http://www.people.ku.edu/~mvitevit/PhonoProbHome.html

Average positional bigram probability (r = .162)



- Calculations here don’t incorporate token frequency
- However: this usually makes little difference

Other concepts of position

Sensitivity to syllable structure: Pierrehumbert (1994 Labphon III): what determines probability of attestation of CCC clusters, such as [lfr] in *belfry*?

- Possibility 1: prob([lf], [fr])
- Possibility 2: argmax(prob(coda [l], onset [fr]), prob(coda [lf], onset [r]))
- Claim: metric sensitive to syllable-position works best
  - Best predictor of which CCC clusters are tested (best linear separation)
  - Also best model of data from experiment on low frequency clusters

Another syllable structure effect

A longstanding observation: onset-rhyme dissociation

- Categorical phonotactic restrictions tend to target VC, not CV combinations (Fudge 1969)

CV	/p/	/t/	/k/
[i]	peel	teal	keel
[ɪ]	pick	tick	kick
[e]	pale	tale	kale
[ɛ]	pen	ten	Ken
[æ]	pan	tan	can
[u]	pool	tool	cool
[ʊ]	put	took	cook
[o]	poke	toke	coke
[ɔ]	Paul	tall	call
[ʌ]	puff	tough	cuff
[a]	pot	tot	cot
[aɪ]	pine	tine	kine
[aʊ]	pout	tout	cow
[ɔɪ]	poise	toys	coin
[ju]	puke	—	cute

VC	/p/	/t/	/k/
[i]	leap	neat	leek
[ɪ]	lip	lit	lick
[e]	rape	rate	rake
[ɛ]	pep	pet	peck
[æ]	rap	rat	rack
[u]	coop	coot	kook
[ʊ]	—	put	book
[o]	soap	coat	soak
[ɔ]	—	taught	walk
[ʌ]	cup	cut	tuck
[a]	top	tot	lock
[aɪ]	ripe	right	like
[aʊ]	—	bout	—
[ɔɪ]	—	(a)droit	—
[ju]	—	butte	puke

Another syllable structure effect

A longstanding observation: onset-rhyme dissociation

- Categorical phonotactic restrictions tend to target VC, not CV combinations (Fudge 1969)

CV	/b/	/d/	/g/
[i]	beep	deep	geek
[ɪ]	bin	din	gill
[e]	bait	date	gait
[ɛ]	bet	deck	get
[æ]	back	Dan	gap
[u]	boon	dune	goon
[ʊ]	book	—	good
[o]	boat	dote	goat
[ɔ]	ball	doll	gall
[ʌ]	bun	done	gun
[a]	bot	dot	got
[aɪ]	buy	dine	guy
[aʊ]	bout	doubt	gout
[ɔɪ]	boy	doi(ly)	goi(ter)
[ju]	butte	—	(ar)gue

VC	/b/	/d/	/g/
[i]	grebe	lead	league
[ɪ]	bib	bid	big
[e]	babe	fade	vague
[ɛ]	Deb	bed	beg
[æ]	tab	tad	tag
[u]	tube	food	—
[ʊ]	—	could	—
[o]	robe	road	rogue
[ɔ]	daub	laud	log
[ʌ]	rub	bud	rug
[a]	cob	cod	cog
[aɪ]	bribe	ride	—
[aʊ]	—	loud	—
[ɔɪ]	—	void	—
[ju]	cube	feud	fugue

This is a grebe.

Another syllable structure effect

A longstanding observation: onset-rhyme dissociation

- Categorical phonotactic restrictions tend to target VC, not CV combinations (Fudge 1969)

CV	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
[i]	meat	neat	—	leap	reap	weep	yeast
[ɪ]	mitt	nip	—	lip	rip	whip	yip
[e]	mate	Nate	—	late	rate	wait	yay
[ɛ]	met	net	—	let	wreck	wet	yet
[æ]	mat	nap	—	lap	rap	wax	yak
[u]	moot	newt	—	lute	route	woo	you
[ʊ]	Muslim	nook	—	look	rook	wood	you'll(?)
[o]	moat	note	—	lope	rope	woke	yoke
[ɔ]	moss	naught	—	log	Ross	walk	yawn
[ʌ]	mutt	nut	—	luck	rut	what	young
[a]	mock	knock	—	lock	rock	wand	yard
[aɪ]	mine	nine	—	line	rhyme	whine	—
[aʊ]	mouse	now	—	lout	route	wound	(yowl)
[ɔɪ]	moist	noise	—	loin	Roy	—	(yoink)
[ju]	music	—	—	—	—	—	—

Another syllable structure effect

A longstanding observation: onset-rhyme dissociation

- Categorical phonotactic restrictions tend to target VC, not CV combinations (Fudge 1969)

VC	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
[i]	team	mean	—	teal	tear	(ewww!)	—
[ɪ]	Tim	tin	sing	till	—	—	
[e]	tame	pane	—	tale	tear	—	—
[ɛ]	hem	ten	—	tell	—	—	—
[æ]	ham	tan	tang	pal	—	—	—
[u]	tomb	tune	—	tool	tour	—	—
[ʊ]	—	—	—	full	—	—	—
[o]	tome	tone	—	toll	tore	—	—
[ɔ]	—	lawn	long	tall	—	—	—
[ʌ]	hum	ton	tongue	(skull)	—	—	—
[a]	Tom	con	—	doll ???	tar	—	—
[aɪ]	time	tine	—	tile	tire	—	—
[aʊ]	—	town	—	scowl	hour	—	—
[ɔɪ]	—	coin	—	toil	—	—	—
[ju]	fume	(im)mune	—	fuel	pure	—	—

A more systematic demonstration

Kessler and Treiman (1997)

- Statistical analysis of all CVC monosyllables in Random House dictionary
- Attempted to assess degree of over/underrepresentation of CV, VC C-C combinations
- Result in a nutshell: many VC and C-C restrictions, virtually no statistically significant CV restrictions

Making counts sensitive to syllabic position

- Pierrehumbert (1994): biphone probability across a syllable boundary may not matter
  - Or, may have less importance, at least
- Kessler and Treiman (1997): biphone probability across onset-nucleus boundary may not matter

Another way to get some syllable effects

Sequences larger than biphones

- l.f vs. lf. $=$ lf$\begin{bmatrix} +\text{son} \\ +\text{cont} \end{bmatrix}$ vs. lf[$-$son], lf[$-$cont]
- Or more generally: gradient *lf] decomposed
  - High: lfa, lfi, lfu, … lfr, lfl
  - Low: lfm, lfn, lfs, lft, …
- An immediate problem: sparse data
  - Even if […l][f…] is “perfect”, it’s not guaranteed to co-occur with all surrounding vowel contexts
  - Accidental gaps: not ruled out by any principle, just happen to be unattested given finite set of words
  - N-grams of larger length have many more possibilities $\rightarrow$ lower individual probabilities $\rightarrow$ often 0 attested occurrences

Sparse data at larger value of n

Strategies for coping with this (smoothing)

- Discounting: steal some probability mass from more frequent items, to give to unattested items
- Deleted interpolation: combine longer and shorter *n*-grams
  - $\mathrm{P(c|ab)} = \omega_1 \mathrm{P(c|ab)} + \omega_2 \mathrm{P(c|b)} + \omega_3 \mathrm{P(c)}$
- Back-off: combine information from lower orders only if higher order is unattested
  - $\mathrm{P(c|ab)} =$

$$\left\{ \begin{array}{ll} \mathrm{P(c|ab)} & \text{if count(abc)} > 0 \\ \alpha_1 \mathrm{P(c|b)} & \text{if count(abc)} = 0 \text{ and count(bc)} > 0 \\ \alpha_2 \mathrm{P(c)} & \text{otherwise} \end{array} \right\}$$

☞ See Jurafsky and Martin chapter for an overview

Non-local dependencies

Newport and Aslin (2004) Learning at a distance I.

- Constructed 3-syl stimuli: 5 fixed σ_1-σ_3 frames, variable σ_1
  - bi $\left\{\begin{array}{l}\text{di}\\ \text{ku}\\ \text{to}\\ \text{pa}\end{array}\right\}$ te, gu $\left\{\begin{array}{l}\text{di}\\ \text{ku}\\ \text{to}\\ \text{pa}\end{array}\right\}$ do, pi $\left\{\begin{array}{l}\text{di}\\ \text{ku}\\ \text{to}\\ \text{pa}\end{array}\right\}$ ra, etc.
  - Transitional prob from σ_1 to $\sigma_2 = .25$
  - Transitional prob from σ_2 to $\sigma_3 = \sigma_3$ to $\sigma_1 = .20$
  - Transitional prob. from σ_1 to $\sigma_3 = 1.00$
  - See Table 1, p. 132

Non-local dependencies (*cont.*)

- Adults exposed for 21 mins, tested on real vs. part-words
  - Adults: can just ask in randomized list: word or not?
- Result: at or below chance performance!
  - Even with more training, simpler languages, etc.
  - Only evidence of learning is ability to distinguish real words from non-words (syls in completely non-occurring orders), which can be distinguished using local transitional probabilities

Non-local dependencies

Comparison: non-local consonant or vowel dependencies

- Constructed languages with $C_1C_2C_3$ frames, variable vowels—or vice versa
  - $p\begin{Bmatrix} a \\ o \end{Bmatrix}g\begin{Bmatrix} i \\ u \end{Bmatrix}t\begin{Bmatrix} æ \\ e \end{Bmatrix}$, $d\begin{Bmatrix} a \\ o \end{Bmatrix}k\begin{Bmatrix} i \\ u \end{Bmatrix}b\begin{Bmatrix} æ \\ e \end{Bmatrix}$
  - $\begin{Bmatrix} p \\ d \end{Bmatrix}a\begin{Bmatrix} g \\ k \end{Bmatrix}u\begin{Bmatrix} t \\ b \end{Bmatrix}e$, $\begin{Bmatrix} p \\ d \end{Bmatrix}o\begin{Bmatrix} g \\ k \end{Bmatrix}i\begin{Bmatrix} t \\ b \end{Bmatrix}æ$
- Frames don't exhibit harmony or any such phonological affinity; merely a statistical reliability
- As before, test words vs. part-words
- Result: good discrimination in both conditions!

Perhaps a substantive bias to attend to certain co-occurrences, rather than others?

References

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

Chomsky, Noam and Morris Halle. 1965. Some controversial questions in phonological theory. Journal of Linguistics 1, 97–138.

Halle, Morris. 1978. Knowledge unlearned and untaught: What speakers know about the sounds of their language. In Linguistic Theory and Psychological Reality., ed. by Morris Halle; Joan Bresnan; and George Miller, 294–303. MIT Press.

Hockett, Charles F. 1955. A Manual of Phonology. Baltimore: Waverly Press.

References (*cont.*)

Jusczyk, Peter W.; Angela Friderici; Jeanine M.I. Wessels; Vigdis Y. Svenkerud; and Ann Marie Jusczyk. 1993. Infants' sensitivity to the sound patterns of native language words. Journal of Memory and Language 32, 402–420.

Jusczyk, Peter W.; Paul A. Luce; and Jan Charles-Luce. 1994. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language 33, 630–645.

Kessler, Brett and Rebecca Treiman. 1997. Syllable structure and the distribution of phonemes in English syllables. Journal of Memory and Language 37, 295–311.

References (*cont.*)

Pierrehumbert, Janet. 1994. Syllable structure and word structure. In Papers in laboratory phonology III: Phonological structure and phonetic form, ed. by P. Keating, 168–190. Cambridge: Cambridge University Press.

Vitevitch, Michael S. and Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in english. Behavior Research Methods, Instruments, & Computers 36, 481–487.

Transitional probability

Saffran, Aslin, and Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* 274, pp. 1926-1928.

- Infants trained on text of 3-syllable “words”
  - E.g., *pabiku*, *tibudo*, *golatu*, and *daropi*

Transitional probability

Saffran, Aslin, and Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* 274, pp. 1926-1928.

- Infants trained on text of 3-syllable “words”
  - E.g., *pabiku*, *tibudo*, *golatu*, and *daropi*

Example

pabikugolatupabikudaropitibudogolatudaropipabikutibudopabiku

ropigolatupabikugolatutibudodaropitibudo…

Transitional probability

Saffran, Aslin, and Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* 274, pp. 1926-1928.

- Infants trained on text of 3-syllable "words"
  - E.g., *pabiku*, *tibudo*, *golatu*, and *daropi*

Example

pabikugolatupabikudaropitibudogolatudaropipabikutibudopabiku
ropigolatupabikugolatutibudodaropitibudo…

- Items controlled so no syllable was used in more than one word
- Non-final syllables always followed by the same syllable (*pa* can only come before *bi*, etc.)
- Final syllables could be followed by any other word (*ku* could be followed by *ti*, *go*, or *da*)

Transitional probability

Saffran, Aslin, and Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* 274, pp. 1926-1928.

- Infants trained on text of 3-syllable “words”
  - E.g., *pabiku*, *tibudo*, *golatu*, and *daropi*

Example

pabikugolatupabikudaropitibudogolatudaropipabikutibudopabiku
ropigolatupabikugolatutibudodaropitibudo…

- Items controlled so no syllable was used in more than one word
- Non-final syllables always followed by the same syllable (*pa* can only come before *bi*, etc.)
- Final syllables could be followed by any other word (*ku* could be followed by *ti*, *go*, or *da*)

Transitional probability

Saffran, Aslin, and Newport (1996) Statistical Learning by 8-Month-Old Infants. *Science* 274, pp. 1926-1928.

- Infants trained on text of 3-syllable “words”
  - E.g., *pabiku*, *tibudo*, *golatu*, and *daropi*

Example

pabikugolatupabikudaropitibudogolatudaropipabikutibudopabiku
ropigolatupabikugolatutibudodaropitibudo…

- Items controlled so no syllable was used in more than one word
- Non-final syllables always followed by the same syllable (*pa* can only come before *bi*, etc.)
- Final syllables could be followed by any other word (*ku* could be followed by *ti*, *go*, or *da*)

Transitional bigram probability

Example

Words: *pabiku*, *tibudo*, *golatu*, and *daropi*

pabikugolatupabikudaropitibudogolatudaropipabikutibudopabiku

ropigolatupabikugolatutibudodaropitibudo…

- Results: after two minutes of exposure, infants can reliably distinguish “words” from “part-words”
  - “Words”: strings of syls that always occur together, like *pabiku*
  - “Part-words”: strings of syls that occur together, but only occasionally (e.g., *kudaro*)
  - Distinguish: prefer to look longer at a speaker playing part-words
- Claim: in order to do this, they must be able to track (somehow) the sequencing of syllable combinations like *pa* and *bi*

Evidence that humans care about transitional probability

Aslin, Saffran & Newport (1998) Computation of conditional probability statistics by 8-month-old infants. *Psych Sci* 9, pp. 321–324.

- Same set-up as before: four words (*pabiku*, *tibudo*, *golatu*, and *daropi*), no repeated syllables
- Small change: two words occurred twice as often as the other two in the training
  - How does this change the syllable bigram probabilities? how about the transitional probabilities?
- Test: tested "part-words" vs. low frequency "words"
- Result: infants still distinguished between the two
  - As before, preferred to look longer at part-words

☞ What statistical differences might they (in principle) have been responding to?