A. Albright

Class 5: Refined statistical models for phonotactic probability

(1) (Virtually) no restrictions on initial CV sequences:

Vowel	/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

(2) Relatively more restrictions on VC combinations:

Vowel	/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

And compare also voiced:

Vowel	/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*

(3) CV co-occurrence for voiced stops

Vowel	/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*

And after sonorants:

Vowel	/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	*Europe*
[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	[ju]	— (yoink)

(4) Kessler & Treiman (1997)

Pearson's χ^2: tests whether relative frequencies of events match predicted (theoretical) frequencies

- In this case: is observed onset/coda asymmetry significantly different from the predicted (equal) distribution?

[k]	Onset	Coda
Observed	148	214
Predicted	181	181

(5) Calculation of χ^2:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

So for the [k] example:

$$\frac{(148-181)^2}{181} + \frac{(214-181)^2}{181} = 2 \times \frac{33^2}{181} = 12.033$$

(Incidentally: for most uses, Fisher's Exact Test is actually a more honest test)

(6) Nosofsky's GCM:

Similarity of *i* to existing items $j = \sum e^{-D \cdot d_{i,j}}$

Where

- $d_{i,j}$ = "psychological distance" between *i* and *j*
- D is a parameter (set to 1 or 2)
- e = 2.718281828

(7) Bailey and Hahn (2001): Adapting the GCM for neighborhood effects

- Similarity of items $d_{i,j}$ intuitively related to how differences they have
  - How many of their phonemes differ (*cat,cap* $>$ *cat,tap*)
  - How important those differences are (*cat, cap* $>$ *cat, cup*)
- Use *string edit distance* algorithm to calculate how many modifications are needed to transform one word into the other
- Use method devised by Broe (1993), Frisch (1996), and Frisch, Broe and Pierrehumbert (1997) to weight the relative cost of different modifications based on the similarity of the segments involved
- Also, want to let token frequency plays a role, but in a complex way: not only are low frequency words less important, but very high frequency words are also ignored
  - Implementation: add a quadratic weighting term, to allow greater influence of mid-range items (parabola-shaped function)

Similarity of $i = \sum (A{f_j}^2 + Bf_j + C) \cdot e^{-D \cdot d_{i,j}}$