Refined statistical models for phonotactic probability

24.964—Fall 2004
Modeling phonological learning

Class 5 (7 Oct, 2004)

Cooccurrence restrictions as evidence for the rime

Hammond (1999) *The Phonology of English*

- (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

Cooccurrence restrictions as evidence for the rime

Hammond (1999) *The Phonology of English*

- 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

Cooccurrence restrictions as evidence for the rime

Hammond (1999) *The Phonology of English*

- 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*

Cooccurrence restrictions as evidence for the rime

A few apparent restrictions:

- No [ʊ] before labials
- No [u], [ʊ] before [g] (and tense + [g] generally rare)
  - league, intrigue, fatigue, colleague
  - vague, plague, Hague, (Sprague)
  - vogue, rogue
- No [aʊ], [ɔɪ] before non-coronal stops

Cooccurrence restrictions as evidence for the rime

More on CV restrictions

- Equivalent for voiced stops (Hammond doesn't list)

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*

Cooccurrence restrictions as evidence for the rime

More on CV restrictions

- 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)

Cooccurrence restrictions as evidence for the rime

Equivalent VC restrictions

- Compare before sonorants:

Vowel	/m/	/n/	/ŋ/	/l/	/r/	/w/	/j/
[i]	team	mean	—	teal	tear	— (ew!)	—
[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	—	—

Some basic issues

- Are there in fact more nucleus-coda than onset-nucleus restrictions?
- What is the explanation for this?
- What about onset-coda restrictions (etc.)

Some basic issues

- What do we make of the few onset-nucleus restrictions that apparently exist? (are they accidental gaps?)

Kessler & Treiman (1997), p. 299: “Some phonemes are fairly uncommon in English, and the number of morphemes is finite, so some possible combinations may fail to exist just because they do not have a reasonable chance to occur. A count of zero cooccurrences does not mean there is a principled constraint against a sequence.”

Some basic issues

- What do we make of pseudo-restrictions, like the near lack of [ib], or the rarity of tense vowels + [g]? Kessler & Treiman (1997), p. 299: "On the other hand, finding a few cooccurrences does not mean that the phonemes combine freely. Some phonemes may be so common that one would expect them to appear together dozens or hundreds of times.

Kessler & Treiman (1997)

Basic idea:

- Statistical investigation of co-occurrence rates to reveal whether effects are (or tend to be) restricted to VC, as opposed to CV
- Strategies for overcoming the problem of interpreting small numbers and zeros:
  - Analyze by major classes (grouping together phonemes increases number of words in each count)
  - Statistical measures of association: provide statistical significance values

Kessler & Treiman (1997)

Strategy: statistical studies of a database of CVC words.

- Advantages of an in-depth study of just CVC's
  - Avoid needing to decide how to count complex onsets/codas
  - Avoid larger alignment problems when comparing words of different numbers of syllables
  - Smaller number of possible ONS-NUC and NUC-CODA possibilities (more compact probability mass); helps mitigate the problem of small numbers

Kessler & Treiman (1997)

Strategy: statistical studies of a database of CVC words.

- Caveat: CVC words are not “typical syllables” (p. 299)
  - They are stressed (so unable to tell us about stressless syls)
  - Minimal word effects exclude certain (otherwise legal) syllable types
  - CV and VC are also word edges (p. 299)

Kessler & Treiman (1997)

The training data:

- 2001 monomorphemic CVC words, from unabridged Random House Dictionary
- Started with full file: `dictAlign.txt` (in this week's `perlscripts`)
- Removed items with complex onsets and codas
- p. 299: "We were fervid in our zeal to eradicate polymorphemic words: A word was rejected if any part of it is used in the same sense in some other word, so that even words like *this* and *then* were omitted on the grounds that *th* may be a demonstrative morpheme."
- Also removed words "that the dictionary gave any reason to believe were not in current general use" (probably things marked *dial.* or *obs.*)

Kessler & Treiman (1997)

My attempt to simulate this:

- Started with `dictAlign.txt` (the Random House list)
- Perl script to convert transcription to something close to the CELEX one we've been using: `ConvertRHWordlist.pl`
- Script also looks up CELEX lemma frequency for each word
  - Lemma frequency = combined freqs of all inflect forms (*eat*, *eats*, *eating*, *ate*, ...)
  - Inflected forms don't occur in the list of lemmas; so if the RH list has inflected forms, they'll get freqs of 0)

Kessler & Treiman (1997)

My attempt to simulate this:

- Script also calculated CV template for each word
  - Syllabic [r̩] treated as V (doesn't occur in CELEX)
- Results: `RandomHouseMonosyllables.txt`
- (I get 2046 CVC monosyllables; I wasn't as fervid)

Kessler & Treiman (1997)

Study 1: relation between consonant type and syllable position

- Null hypothesis: every consonant can appear equally well as onset or coda
- Obvious falsification: *[ŋ] in onsets, *[h] in codas
- But beyond these categorical constraints, are there additional preferences?

Kessler & Treiman (1997)

Study 1 approach:

p. 300: "To determine whether the frequencies are affected by syllable position, we performed for each consonant type separate two-cell goodness-of-fit tests with Pearson's χ^2, computing the expected frequencies under the null hypothesis that consonants would be evenly distributed between onset and coda. Because all words had exactly one onset and one coda consonant, this means that each consonant should occur half the time in an onset and half the time in a coda."

Kessler & Treiman (1997)

Example: distribution of a few phonemes

Consonant	Total occurences	Onset	Coda
p	308	151	157
t	404	133	271
k	362	148	214
f	201	123	78
θ	66	22	44
ʃ	132	67	65
h	143	143	0
ŋ	46	0	46

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

Kessler & Treiman (1997)

Calculation of χ^2:

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

So for the [k] example:

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

Kessler & Treiman (1997)

Side-note:

- If you ever find yourself need to compare proportions like this, a more widely accepted procedure is Fisher's Exact Test
- This is especially true when dealing with smallish numbers (as we have here)
- For more information, see http://www.tufts.edu/~gdallal/ctab.htm

Kessler & Treiman (1997)

Comparing the relative strengths of these preferences:

- Obviously, some are categorical, and important ([h], [ŋ], [j], [w])
- Also, for their set, [ʒ] and [ð]
- G^2 test on remaining phonemes, broken down by place of articulation, shows additional effects (Table 4, p. 303)

Kessler & Treiman (1997)

The novel finding of study 1:

- Prefer onsets: /b, d͡ʒ, ʃ, r/
- Prefer codas: /z, θ, n, t, l, k/

(My own counts don't replicate every aspect of this. For example, [ʃ] came out very evenly in my counts; things apparently change when you include also the [ʃVCC] and [CCV ʃ] words (clash, trash, fresh, etc.))

Kessler & Treiman (1997)

Study 2: combinatorial tendencies (co-occurrence restrictions)

- Three-way contingency tables
- {Place, Manner, Voice} $\times$ {Height, Backness, Tense} $\times$ {Place, Manner, Voice}
- Tested all three as independent factors (27 comparisons), then each as independent of the other two, then each pair independent of the third
- Tried to compare magnitude of interactions by rescaling

Kessler & Treiman (1997)

Results, part 1:

- p. 305, Table 5
- Preponderance of Ons-Coda and Nucleus-Coda associations; virtually no Onset-Nucleus associations
  - OC and VC associations are more often significant, and account for a large proportion of the effect

Kessler & Treiman (1997)

Results, part 2:

- Guided by results in Table 5, look more carefully at interactions between particular features in particular slots
- Example: in top left of Table 5, onset place and coda place are associated (regardless of vowel quality)
  - Two-way comparisons of onset and coda place interactions are given in Table 6
- Similarly, coda manner, voice, and place all show VC associations (at least with some vowel qualities (Table 7)

Kessler & Treiman (1997)

Phoneme-by-phoneme comparison (Table 8)

- Generally more VC than CV sequences associated in this way (both positively and negatively)
- Many "less frequent" cases appear to be motivated by markedness (*[ki], *[wæ], *[lVl], *[rVl], effects of coda [r]), though many are not
- Under a view in which explanation lies in explaining what is dispreferred, we may not need to worry so much about why particular cases are *more* frequent than expected.

Kessler & Treiman (1997)

Calculating co-occurrence: `CooccurrenceCharts.pl`

- Program outputs observed counts of co-occurrence
- Also, *expected* counts based on joint probability of the two subparts
- Demo: using Excel to calculate observed over expected (O/E) values

Do these results replicate Kessler and Treiman? What are some reasons why they might differ a bit?

Kessler & Treiman (1997)

An obvious next question: are speakers aware of any of this?

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000) English speakers' sensitivity to phonotactic patterns. In M. B. Broe & J. Pierrehumbert, eds., *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, pp. 269-282. Cambridge University Press.

- Created a set of novel words with rimes of low and high cooccurrence probability

- Presented them to speakers, and asked them to rate from 1 ("doesn't sound at all like an English word") to 7 ("sounds very much like an English word")

- Result: items whose rimes have high cooccurrence probability are rated higher. (And ditto for higher CV probability!)

- Also did a study with children, using a simpler task of choosing the more English-sounding item
- Also did a study asking (adult) speakers to blend nonsense monosyllables with likely rimes ([həɹk] + [jɪg] → [hɪg] or [həɹg]) and unlikely rimes ([həɹp] + [jɪd͡ʒ])

Kessler & Treiman (1997)

Summary:

- Cooccurrence restrictions go beyond categorical prohibitions; many statistical tendencies can also be observed
- Most cooccurrence restrictions are on VC sequences (rimes?)
  - Some CV restrictions are also observed (at least in the lexicon)
- Speakers demonstrate, in various ways, that they are aware of these tendencies

(Note that K & T do not actually provide a full model of calculating how "good" a word sounds! They just decompose the factors, giving us some ideas about some sensible ways to proceed)

Kessler & Treiman (1997)

Other findings in a similar vein:

- Coleman and Pierrehumbert (1997)
- Frisch, Large & Pisoni (2000) Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language* 42:481-496.

Some questions that we are left with

- What is the model that produces phonotactic probabilities? (That is, how do speakers represent this knowledge?)
- Is this the same model the produces phonological outputs for words? (i.e., is it within the grammar?)
  - (Richness of the base issue here!)
- How to scale up to longer words, with more complex structure?

Bailey and Hahn (2001)

An attempt to commit to a particular model of overall wordlikeness

- Motivated by a practical need: control for well-formedness of stimuli in experiments
- Motivated also by a psychologist agenda: show that wordlikeness depends literally on similarity to words rather than more abstract knowledge about phonotactic probabilities
- Strategy: get ratings on a bunch of words, and then compare predictions of two models:
  - One that calculates scores based on phonotactic probability
  - Another that calculates scores based on similarity to existing words

Bailey and Hahn (2001)

A variety of possibilities about how phonotactic probabilities play a role

- Speakers store a rich set of statements about phonotactic probability, and consult it when evaluating novel words
- Speakers have no direct knowledge of probabilities; they poll their lexicons in evaluating novel words, and probability effects emerge based on the number of similar words that are found
- Hybrid model: both play a role somehow

Bailey and Hahn (2001)

An important problem in teasing apart these possibilities

- Existing words tend to make use of high-probability sequences (that is, after all, what makes them high probability)

- So, in many cases, calculating phonotactic probability and finding support from the lexicon will yield the same answer

- Difficult to test independent effects of two highly correlated factors

Bailey and Hahn (2001)

A sensible approach to this problem (p. 570)

- Create a set of items that vary considerably in wordlikeness (not just high and low probability, where models are especially likely to agree)

- Large amounts of variance make it unlikely that any model works terribly well

- This maximizes the chance that we can observe the virtues of each model independently (if they have virtues)

Bailey and Hahn (2001)

Model 1: phonotactic probability

- Very simple: n-gram models (transitional probability)
  - Tested both bigram and trigram models
  - (Curious claim, bottom p. 570: co-occurrence probability and transitional probability are, in practice, highly correlated. Would we expect this to be true in general?)
- p. 571: “To calculate a composite value for an entire word or nonword, we took the geometric mean of conditional segment probabilities across the whole item, giving a single average bigram probability.”
  - Frisch, Large, and Pisoni (2000): length effects (longer words get lower ratings); averaging cannot capture this
- Calculated both on pronunciation and orthography (since they showed spellings in one experiment)

Bailey and Hahn (2001)

Model 2: influence of particular existing words through neighborhood density

- Luce (1986): For a word *w*, the neighborhood of *w* is the set of all words that can be produced by changing, adding, or deleting a single segment in *w*
- Example: *cat* has neighbors *hat*, *mat*, *chat*, *cut*, *cap*, *cast*, etc.
- Neighborhood density (NNB) = number of neighbors that a word has

Bailey and Hahn (2001)

Problems with classical NNB definition:

- Naive about edits: not all modifications are equal! Some produce very similar words, while others have a substantial effect on perceived similarity
  - Some segments are more similar than others
  - Some parts of the syllable (and word) probably more important than others
  - Some changes even alter syllable count, etc.
- Cut-off of one edit is totally arbitrary
- In fact, when considering words with clusters, there are often few or no single-edit neighbors

Bailey and Hahn (2001)

Improving on the classic NNB model

- Start with a simple modification: allow 2 edits instead of one
- (Meant to be the “stupid baseline” model—but couldn’t use classic one-edit NNB model because they weren’t finding enough neighbors)

Bailey and Hahn (2001)

A more sophisticated model: the Generalized Neighborhood Model

- Inspired by the Generalized Context Model (Nosofsky 1986), an influential model of how novels items are categorized based on the influence of other, similar items
- Intuition: when deciding what to do with a new item, consult your database of existing items
  - Existing items that are more similar should get more say
  - Patterns that cover many items are also more influential

Bailey and Hahn (2001)

Nosofsky's GCM:

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

Where

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

Bailey and Hahn (2001)

Adapting the GCM for neighborhood effects

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

- 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*)

Bailey and Hahn (2001)

Calculating similarity of items

- 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

(We can return to both these techniques at the end of the semester, if you are interested in them)

Bailey and Hahn (2001)

One last modification

- Have a hypothesis that 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)

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

(Where f_j = token frequency of item *j*)

Bailey and Hahn (2001)

Testing the models

- Made up some non-word "isolates" that differed from nearest neighbors by two phoneme edits (e.g. *drolf* = *golf* $\rightarrow$ *grolf* $\rightarrow$ *drolf*, or *draw* $\rightarrow$ *drawl* (or *drawf* $\rightarrow$ *drawlf*)[1]
- Also included the intermediate "near-misses" (*grolf*, (*drawl*), *drawf*).
  - Clever: this guarantees that *all* neighborhood models will prefer the near-misses, even if future research develops a better way to calculate similarity of two words that takes syllable structure, etc., into account.
- Huge number of items: 22 isolates, 250 near misses

[1] Actually, *drawl* is a real word, which appears to be one edit away from *drolf* (making *drolf* a near-miss, not an isolate). Either *drawl* isn't in CELEX, or there's a vowel difference in British English.

Bailey and Hahn (2001)

Experiment 1: written questionnaire

- “Does minth sound like a typical word of English?”
- Told to focus on (imagined) sound, not spelling
- Participants circled 1 (very non-typical) through 9 (very typical)

Use of written materials is a grave error here! Are participants reading *minth* as [mɪnθ] or [maɪnθ]? Items were supposed to be chosen to be orthographically unambiguous, but there are various problems of this sort, including not only ambiguities of native spelling rules (*prolf*, *swuft*, *sandge*), but also items with possible “hyperforeignistic” pronunciations (*zin*, *sulp*, *sesk*, etc.)

Bailey and Hahn (2001)

Results

- Participants sort of agreed with one another (not fantastically high agreement rate, but significant)
- Significant effect of phonotactics: all models correlate significantly, but phonotactic trigrams show the biggest effect
- Neighborhoods: GNM does best, but NNB also plays a distinct role (not an ideal result)
- Putting them together: multiple regression shows that both neighborhoods and phonotactics have significant independent effects
- Lexical frequency: removing this from the GNM model hurts it a little

Bailey and Hahn (2001)

Experiment 2

- This time, recorded in a “frame” (sort of—see p. 580)
- Participants heard the words and rated them

Bailey and Hahn (2001)

Similar results

- Participants agreed with one another to the same extent (and sort of agreed with exp. 1 participants)
- Significant effect of phonotactics (this time bigrams did best; little or no effect of orthographic ngrams)
- Neighborhoods: again, GNM does best, but NNB also plays a distinct role
- Putting them together: multiple regression shows that both neighborhoods and phonotactics have significant independent effects
- Lexical frequency: again, removing this from the GNM model hurts it a little (31% to 30%)

Bailey and Hahn (2001)

Another overall result:

- Even these kitchen sink models, taken all together, explain rather little of the variance (around 30% total)
- Bailey and Hahn: “there’s room for improvement”

Bailey and Hahn (2001)

Some attempts to modify the models slightly: phonotactics

- Calculate ngrams over just monosyllables: much worse
- Weight bigram counts by token frequency (as in Jusczyk et al): makes no difference at all (does this make sense?)
- Derive scores by multiplying logs of transitional probabilities, rather tahn simply averaging (Coleman and Pierrehumbert 1997): this helps
- Incorporate individual phoneme probabilities: also helps
- Onset-Rime probabilities instead of bigrams: this does not help

B&H conclusion: many possibilities left to explore, but not clear that of these techniques promises to improve things

Bailey and Hahn (2001)

Some attempts to modify the models slightly: neighborhoods

- Tried two different ways of calculating similarity of words
  - Instead of counting edits, consider how many precedence relations are preserved (*stick* = /stɪk/, /stɪ/, /tɪk/, /stk/, /sɪk/, /st/, /tɪ/, etc.)
  - Weight mismatches in onsets differently from codas
- Result: weighting onsets a little more heavily helps (surprisingly); excluding vowels altogether actually helped (?!?!?!?!) (B&H point out that their vowel similarity values were probably not that great)

Bailey and Hahn (2001)

The picture that emerges

- People know words, and this plays a role in deciding how wordlike a novel item is
- Yet it appears that they also know phonotactic probabilities (in some form), as distinct, abstracted knowledge

Bailey and Hahn (2001)

Even so…

- Putting together a “best shot” model still only explains 38% of the variance in participant responses
- p. 585: “This result confirms the conclusion that an entirely adequate account of wordlikeness has not yet been found”

Just a few of many mysteries

Some of my favorites: (feel free to take these up for final projects)

- Why are sC_1VC_1 words so bad?
- Similarly, why is [smaɪ] so bad? (sm + tense vowel in open monosyllables)
- Why is [bʊt] so bad? [bʊ] does occur in *book*, and [ʊk] is fairly common. It also has lots of neighbors (*book*, *bat*, *but*, *bet*, . . . , *put*)

Does these follow somehow from the statistics of English? Do we need to assume that such irrational constraints are innate, or could they be learned somehow?)

Summary of our look at statistical models

- These are easy to implement because they just involve counting.
- But they are hard to explore because there are so many possibilities
- Any even halfway sensible model will get the easy cases; getting all the intermediate gray stuff is much harder
- Psychologists like these models because they are general-purpose (after a fashion)
- They are a useful baseline

Some general questions as we move on

How do these models relate to what we think about "grammar"?

- Should an OT grammar try to capture all of these effects? (If so, how could it be incorporated?)
- Where would you look for evidence for a distinct grammar, separate from these gradient probabilistic effects?
- If there are two systems determining well-formedness, how do they interact? (and why do we have them?)

For next time

We are moving on to OT. Please read:

- Tesar and Smolensky (1996) Learnability in OT (short version)
- Prince and Tesar (2004) Learning phonotactic distributions