Calculating probabilistic phonotactics

24.964—Fall 2004
Modeling phonological learning

Class 4 (30 Sept 2004)

The context

- Surface phonotactics is one of the first things that infants show some knowledge of
- Rough chronology:
  - Six months: begin to show weak form of categorical perception for native contrasts (Kuhl and colleagues); by 11 mos, have lost ability to perceive non-(some) native contrasts (Best et al, Werker et al)
  - Nine months: can discriminate possible from impossible sound combinations (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk 1993

What are they using?

One possibility: statistical information about co-occurrence of segments

- Saffran, Aslin, and Newport (1996): 8 month old infants can

Saffran, Aslin & Newport (1996)

- Infants heard two minutes of synthesized "speech" from a language made up of 3-syllable nonsense words, in pseudo-random order (45 of each word), no pauses or intonation
  - E.g., *tupiro*, *golabu*, *bidaku*, *padoti*
- Then, were tested on "words" (*tupiro*, *golabu*) and very similar non-words (*dapiku*, *tilado*)
- Result: infants listened significantly longer to non-words (novelty preference)

Saffran, Aslin & Newport (1996)

Familiarization preference procedure

- Infant seated on parent's lap in booth, experimenter outside, peeking through small holes (so infant can't see)
- Green light directly in front of infant blinks, infant looks at it
- Then a red light on left or right blinks, and stimuli are played from a speaker on that side
- In an ideal trial, the infant's gaze goes to the red light when it starts blinking, gets fascinated by the sounds coming from that side, and keeps looking until bored

- Experimenter presses a button when infant looks away (for more than 2 secs)

- Parent and experimenter both wearing headphones, listening to masking recording (words or music); control for "Clever Hans" effect

Saffran, Aslin & Newport (1996)

How do the babies know that *dapiku* isn't a word?

- Syllables *da*, *pi*, and *ku* all appear in the familiarization set (*bidaku*, *tupiro*, *bidaku*
- But crucially, never in that order

Saffran, Aslin & Newport (1996)

But never occurring together is not enough to distinguish non-words

- *kupado* is also not a real word of this language, but could arise in the “phrase” *bidaku padoti*

Saffran, Aslin & Newport (1996)

Experiment 2:

- Same training as before, but this time tested on real words vs. "part-words" (like *kupado*)
- Still showed longer listening preference for part-words

Saffran, Aslin & Newport (1996)

How do the babies know that *kupado* isn't a word?

- Words: *tupiro, golabu, bidaku, padoti*
- Within-word sequences always occur together (*tu* always followed by *pi* in this language)
- Across-word sequences only occur that way a fraction of the time (*ku* sometimes followed by *pa*, sometimes by *tu*, sometimes *go*)
- *ku-pa* has a lower *transitional probability*

Transitional probability

Definition: transitional probability of *xy* (or $y|x$)

- The transitional probability from *x* to *y* is the probability that the next thing after an *x* is a *y*
- In other words, given an *x*, how likely is a *y*?
- $\mathrm{P}(xy) = \dfrac{\text{Freq of } xy}{\text{Freq of } xa, xb, xc, \ldots\ xz} = \dfrac{\text{Freq of } xy}{\text{Freq of } x}$

Acquisition of phonotactics

So we know that infants can keep track of statistics concerning sequences of syllables

- Incidentally, so can cotton-top tamarins (Hauser, Newport & Aslin 2001, *Cognition* 78)
- Suggests they should be able to do more than distinguish occurring from non-occurring phoneme sequences; what about likely vs unlikely sequences?

Jusczyk et al 1994

Jusczyk, Luce, and Charles-Luce (1994): tested sensitivity to high vs low frequency patterns in English

- No training: we are interested in what the infants already know prior to entering the experimental booth
- Infants are presented with lists of high-probability, and legal but low-probability non-words
  - High probability: [rɪs], [rɪn], [ʃæn], [sɛt͡ʃ], etc.
  - Low probability: [jaʊd͡ʒ], [θɔʃ], [fuv], [huʃ], etc.

Jusczyk et al 1994

Result: infants attend significantly longer to lists of high-probability items

- Desired interpretation: they know these items are more English-like, and listen more closely
- Another possibility, though: maybe the high probability items are just *inherently* more interesting to babies (for reasons unknown to us), and babies are fascinated by this intrinsic property, not by their relation to English

Jusczyk et al 1994

Experiment 2:

- Same as exp. 1, but with 6-month olds
  - 6-month olds are like 9-month olds in that they are human babies. But they are different, in that they don't know as much about English.
- Result: no preference for high-probability items
- Interpretation: the result with 9-month olds shows that they have learned something about English

Jusczyk et al 1994

Experiment 3:

- Same as exp. 1, but with different lists of items, controlled for vowel quality
  - Exp 1: [rɪn], [bæp] vs. [jaʊd͡ʒ], [ʃɔb]
  - Exp 3: [sʌʃ], [kəɹm] vs. [Đʌd͡ʒ], [ʃəɹg]

- Basically same result (slightly weaker)

High vs. low probability words

On the whole, the difference between Jusczyk et al's high and low probability stimuli looks pretty good

[rɪs], [rɪn], [ʃæn], [sɛt͡ʃ] vs. [jaʊd͡ʒ], [θɔʃ], [fuv], [huʃ]

There are, however, a few items that make you wonder ([dɪʒ] = high, but [ʃaʊd] = low?); especially in Exp. 3 stims

High vs. low probability words

Two tasks:

1. Check up on Jusczyk et al's claims of high and low probability
2. Simulate the type of knowledge that the babies might be using

High vs. low probability words

Jusczyk & al 1994, p. 633

> "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 [refs]. The average summed phoneme probability was .1926 for the high-probability pattern list and .0543 for the low-probability pattern list."

High vs. low probability words

Positional phoneme frequency:

> “A high-probability pattern consisted of segments with high phoneme positional probabilities. For example, in the high-probability pattern /ɹɪs/, the consonant /ɹ/ is relatively frequent in initial position, the vowel /ɪ/ is relatively frequent in the medial position, and the consonant /s/ is relatively frequent in the final position.”

High vs. low probability words

Biphone frequency

> "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-to-final 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 /ɪ/ to /s/"

High vs. low probability words

Looking first at positional probabilities

- “how often a given segment occurs in a position with a word”

What are some problems that arise in turning this description into a procedure?

High vs. low probability words

Some vagaries in J & al's description:

- What is a "position"? ("initial, medial, and final"??? works final for CVC nonce words, but how do you count from the wordlist of real words?)
- How do you compare existence vs. non-existence in a position?
  - Example: Coda nasals in a language like Japanese; if we just compare the set of coda consonants, /N/ has 100% probability (making nasal-closed syls very probable). If we also include open syls, then closed syls are less probable

High vs. low probability words

Some vagaries in J & al's description:

- How are words aligned so they can be compared? Do examples like /tɹɪst/ contribute to the goodness of /ɹɪs/, by providing examples of /ɹ/ onsets and /s/ codas?
- What precisely does "log-frequency weighted values" mean?

High vs. low probability words

What J & al actually did:

> Vitevitch, M.S. & Luce, P.A. (submitted). A web-based interface to calculate phonotactic probability for words and nonwords in English.Âă Behavior Research Methods, Instruments, & Computers.

(More complete description of “operational procedure” which has been used by now in many papers)

Positional probability

The definition of positions

1	2	3	4	5	...

Positional probability

Aligning word by “position”

@	b	I	l	I	t	I		
k	@U	I	g	z	I	s	t	s
f	I	g	z					
k	i:	O	s	k	s			
p	aI	n	I	N				
S	r	i:	k					
v	E	n	@	m	@	s		
p	l	V	g					
p	&	tS						
m	I	N	g	l,	d			
l	E	t	d	aU	n			

(Is this what you were expecting? Why or why not?)

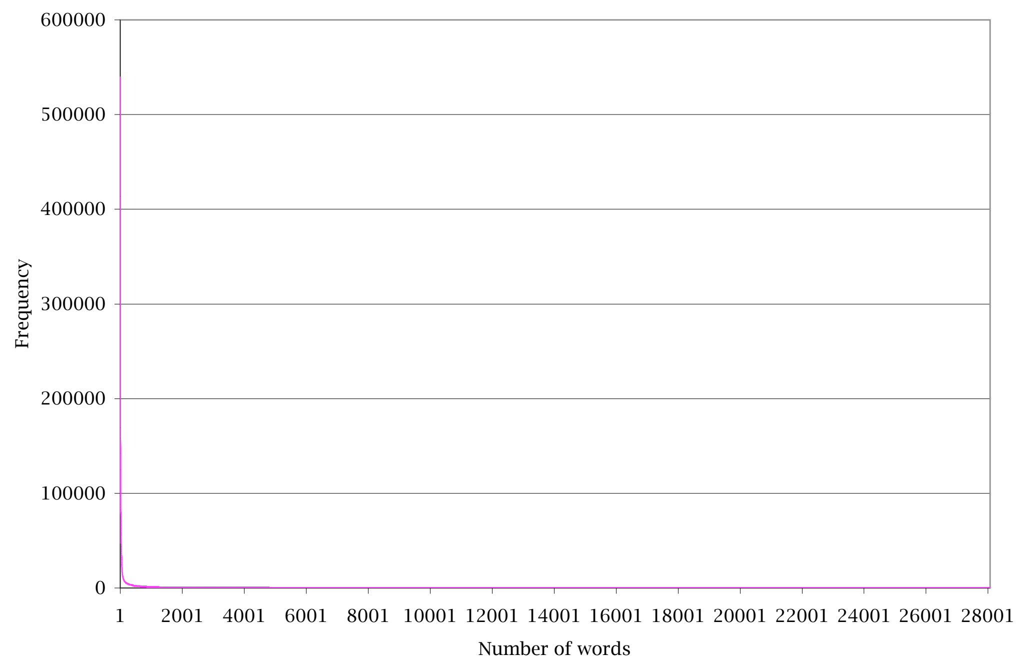
Positional probability

Counting: Vitevitch & Luce, p. 6

- "*Positional segment frequency* was calculated by searching the electronic version of Webster's (1964) Pocket Dictionary for all of the words in it (regardless of word length) that contained a given segment in a given position. The log (base 10) values of the frequencies with which those words occurred in English (based on the counts in Kucera & Francis, 1967) were summed together, and then divided by the total log (base 10) frequency of all the words in the dictionary that have a segment in that position to provide an estimate of probability. Log values of the Kucera & Francis (1967) word frequency counts were used because log values better reflect the distribution of frequency of occurrence and better correlate with performance than raw frequency values [refs]. Thus, the estimates of position specific frequencies are token- rather than type-based estimates of probability."

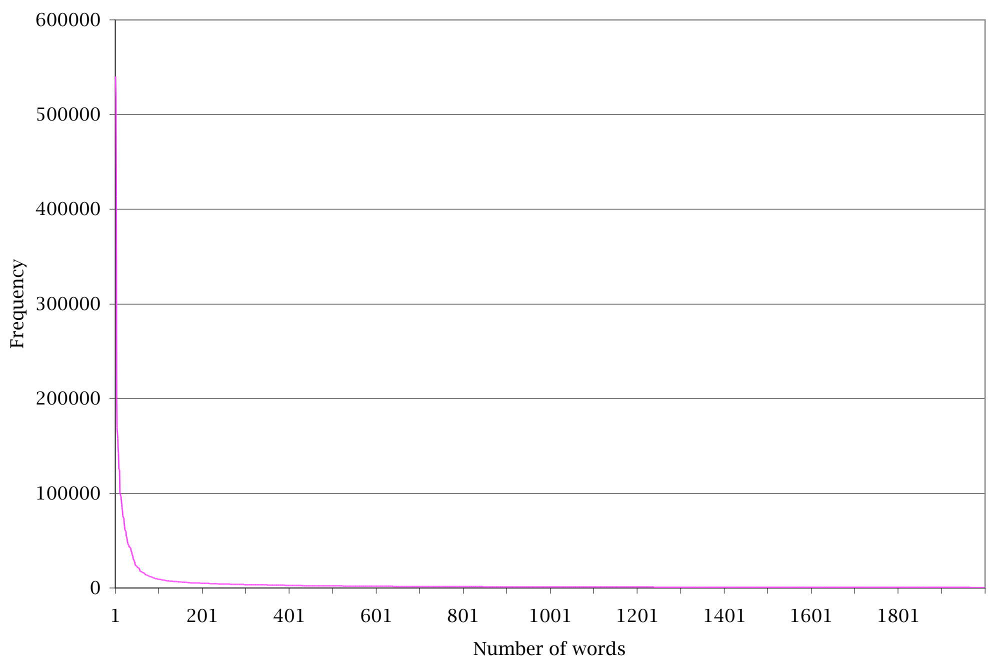
Positional probability

Distribution of words by raw token frequency



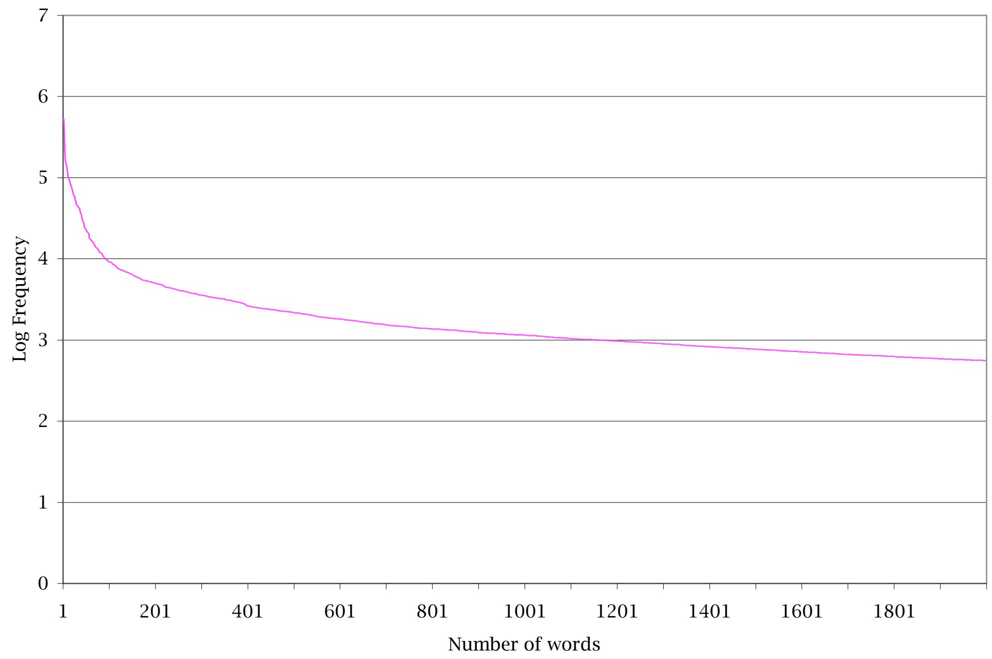
Positional probability

Distribution of words by raw token frequency



Positional probability

Distribution of words by log token frequency



Positional probability

Probability of a phoneme in a position = Sum of log10 frequencies of all existing words that contain that phoneme in that position, divided by sum of log 10 frequencies of all words that contain *anything* in that position

Positional probability

Probability of a word = sum of all of its positional probabilities

- Why is this OK for comparing CVC stimuli, but not OK as a general model of well-formedness?

Biphone probability

A common model of sequencing constraints: *transitional probabilities*

- The transitional probability from x to y is the probability that the next thing after an x is a y
- In other words, given an x, how likely is a y?
- $\mathrm{P}(xy) = \dfrac{\text{Freq of } xy}{\text{Freq of } xa, xb, xc, \ldots\ xz} = \dfrac{\text{Freq of } xy}{\text{Freq of } x}$
- *n-grams* (bigrams = 2, trigrams = 3, etc.)

Probability of $xyz = \mathrm{P}(xy) \times \mathrm{P}(yz)$

Biphone probability

Here, too, this is not what Jusczyk & al did; instead, they calculated *probabilities of two-phoneme sequences*, not the transitional probabilities from one phoneme to the next (V & L p. 7)

@	b	I	l	I	t	I		
k	@U	I	g	z	I	s	t	s
f	I	g	z					
k	i:	O	s	k	s			
p	aI	n	I	N				
S	r	i:	k					
v	E	n	@	m	@	s		
p	l	V	g					
p	&	tS						
m	I	N	g	l,	d			
l	E	t	d	aU	n			

High vs. low probability words

Putting this together into a procedure: `VitevitchLuce.pl`

What would a smarter program do?

Some problems you may have encountered

- A more sensible use of positions?
- How to handle words with different numbers of items in the relevant position?

My own attempt to do this part of the task

Perl script: `PositionalProbability.pl`

- Fails to divide up complex onsets/codas
- Only works to derive predictions for CVC items

Other questions that seem relevant

- Is it even right to be weighting counts based on token frequency? What about probabilities based purely on type frequency?
- Should all parts of the word count equally in determining its score?
- Are the same counts relevant for all parts of the syllable? (See Kessler & Treiman 1997)
- Should we be comparing monosyllables with polysyllabic words? (What are some ways in which monosyllables are not actually composed of strings of possible monosyllables?)

Assignment for next week

- Keep tweaking your program; in its final state, it should produce files that list the scores for each word, so it is relatively easy to give it a set of nonce words, and browse through the list of predicted probabilities

- It should also calculate probabilities in two different ways (one based on positional probability, and one based on biphones, in some fashion)

Assignment for next week

- Finally, I would like you to pick one of the questions raised here, and see how it affects the results, e.g.
  - ◦ Transitional probability vs cooccurrence probability
  - ◦ Different alignments
  - * Different weighting of different parts of the syllable
  - ◦ Use of type vs token frequency
  - ◦ Training on monosyllabic vs all words
- Specifically, discuss how a different way of modeling phonotactic probability affects the match to the data found in AlbrightHayes.txt

Readings

(Discussants, anyone?)

- Kessler and Treiman (1997) Syllable Structure and the Distribution of Phonemes in English Syllables

- Bailey and Hahn (2001) Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language* 44, 568-591.