Class 2, 4/2/2018: Moore-Cantwell Paper, F-Voicing, Nuts and Bolts of MaxEnt

1. Assignments

- Read Bruce Hayes (ms.) "Assessing grammatical architectures through their quantitative signatures"
  - On web site.
- Start on your homework — medial clusters.
  - This is probably the biggest homework and is due in two weeks.
  - For now, working on finding your corpus and extracting the clusters from it.
  - I'll discuss this homework later on in this class.

2. What have we got so far?

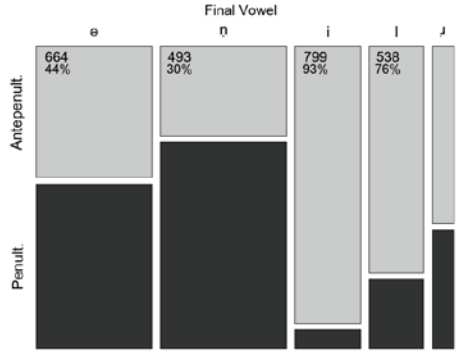
- Phonology extends its goals:
  - from its ur-homeland in providing "satisfying accounts" of patterns in a data corpus
  - … to attempts at prediction
  - … broadening of research methods
- We are seeking a good framework of constraint-based linguistics, hoping to make predictions.
  - Hence the advent of linguistic theories that resemble (better, modern) statistical modeling, for this very purpose.

A BIT ON THE MOORE-CANTWELL READING

3. Outline

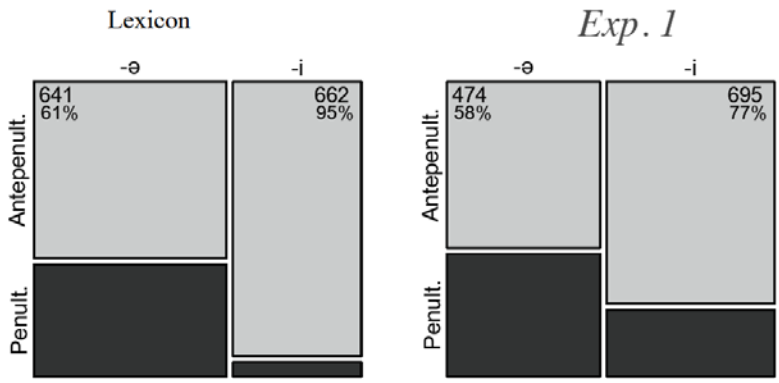
- This represents trends in contemporary phonology:
  - electronic corpus study
  - experimentation to test productivity
  - modeling that attempts to predict experimental outcomes
- We have a clear case of interacting factors:
  - the penult-weight effect
  - the final-[i] effect

4. The two effects: lexicon

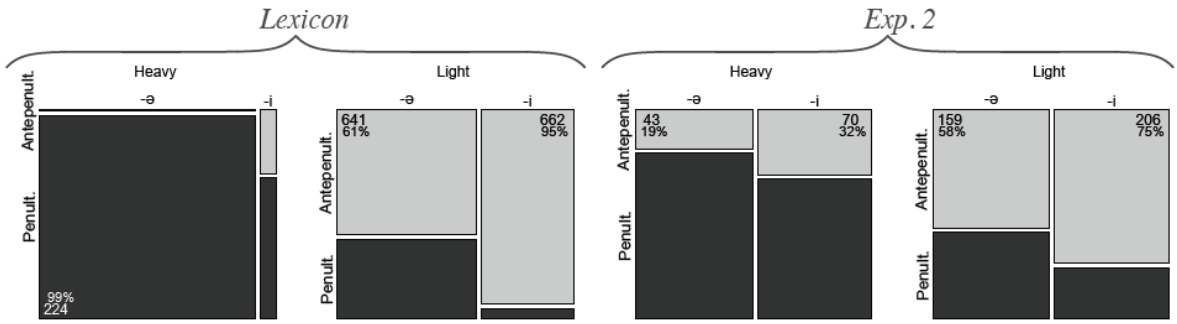


5. The two effects: blick test

- Method: play three syllables, force participants to concatenate into one word
- Light-penult words. Experimental results compared with lexicon.



Both light and heavy penults:



- This is a classical result, I believe:
  - The two effects combine smoothly in a cross-classifying pattern, modeled by Moore-Cantwell using MaxEnt.
  - The experiment shows *weaker* effects than the lexicon — which is something we must explain ☞ **how?**

6. Rest of lecture

- General background on predicting things in phonology.
- Doing /f/ Voicing in English
- More technical material on MaxEnt
- More background on the homework assignment

THE F-VOICING EXERCISE

7. Baby example, to get up and running with MaxEnt

- North American English tapping (*butter* = [ˈbʌɾɚ])
- Applies only in "ambisyllabic" position (intervocalic pre-atonic) (compare *attend*)
- Probably more obligatory to tap with /d/ than /t/: *writing* vs. *riding*[1]
- Impossible (more or less) to tap when the ambisyllabic environment is not met: *write*, *ride*

- Bring up /f/ Voicing spreadsheet, where this appears as a baby example.

[1] There is empirical work on corpora, which I ran out of time to check. Try e.g. Byrd, Dani. 1993. 54,000 American stops. *UCLA Working Papers in Phonetics* 83, 97-116.

8. Brief review of the /f/ Voicing data

- BH's personal set of voicers (☞ are you the same?)

Obligatory	Optional
calf elf half knife leaf life loaf scarf self sheaf shelf thief wife wolf	*behalf dwarf epitaph hoof roof* [2] *wharf*

- BH's personal set of nonvoicers

autograph	cuff	huff	phonograph	scuff	trough
bailiff	dandruff	jeff	photograph	serf	turf
beef	duff	kerchief	plaintiff	sheriff	unicef
belief	f	laugh	pontiff	skiff	waif
biff	fief	lithograph	poof	sniff	whiff
bluff	fife	lymph	prof.	snuff	woof
brief	fluff	massif	proof	spoof	
buff	gaff	mastiff	puff	staff	
caliph	gaffe	mimeograph	ralph	stiff	
carafe	giraffe	mischief	rebuff	strife	
chaff	goof	molotov	reef	stuff	
chef	graph	monograph	ref	surf	
chief	grief	motif	riff	tariff	
clef	gulf	muff	rough	telegraph	
cliff	handkerchief	nymph	ruff	tiff	
cough	hieroglyph	paragraph	safe	tough	

9. ☞ Think for awhile about what might predict voicing

10. Spreadsheet work

- Open spreadsheet, look at data, consider factors, implement the MaxEnt analysis.

11. Looking at the output of the grammar

- **Self-check** the analysis against the existing lexicon, as a sort of sanity check.
- Do a **probability sort** within categories and plot.
- Are the forms predicted to be impossible, impossible?
- What are the most likely [f] plurals to be pronounced *innovatively* with [v]?
- What of Berko's *heaf* form, where we already have a modest real probability value?
- What distinctions are made among the existing forms?

[2] I really couldn't say *rooves* myself but I accept it from other people as non-bizarre.

DEEPEST BACKGROUND ON MAXENT: PROBABILITY THEORY

12. Probability

- We mostly know probability as likelihood of random events — which is sort of valid, e.g. "probably Bruce will say *dwarves* next time he utters the word".
- But there is an influential alternative conception: ***probability as quantification of degree of (rational) belief***.
  - I.e. the belief that Bruce holds that *dwarves* is the true and correct way to pluralize *dwarf*.
- This conception was worked out by extraordinary minds in the 20th century and serves as the basis of a major intellectual trend often labeled **Bayesianism**.
- By "Cox's theorem" (Cox 1946), probability theory and its axioms emerge as the *only possible formalization* of inductive reasoning compatible with our comment-sense notions.
- With this mathematization-of-common-sense, we can use the same fundamental principles by which we reason, applying them to much harder problems with the support of the mathematics of probability

13. Readings on probability for the curious

- Probability as only basis for inductive logic:
  - Jaynes, Edwin (2003) *Probability Theory: The Logic of Science*, Cambridge.
  - I kept a copy of the now-unavailable PDF, which I am happy to share.
- The maxent framework we will examine is likewise defended as "inevitable" in
  - Skilling, John (1989) Classic Maximum Entropy. In J. Skilling, ed., *Maximum Entropy and Bayesian Methods: Cambridge, England 1988.* pp. 45-52. Dordrecht: Kluwer.

SOME TECHIE BACKGROUND ON MAXENT

14. A source I found very helpful

- Jurafsky, Dan and James Martin (forthcoming) *Speech and Language Processing*, 3rd ed.
- Download it while it's still free! https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- The maxent chapter is Chapter 5.

15. The sound foundation

- Maxent normally uses the **maximum likelihood** method to set the weights.
- Due to a wonderful theorem reviewed in Jurafsky and Martin, an adequate set of weights can be found essentially with certainty.
  - Caution: there may be more than one, and in huge cases the search might take too long.
- This is because the search space is "convex"; on which more later.

- For algorithms that explore the search space, see this fn.[3]

16. The nomenclature

- Jurafsky and Martin call it **logistic regression**, and this may indeed be a better long-term name.
- I sense that MaxEnt was the jargon of a group of computer scientists who fell in love with it; but logistic regression may be a little more grown up …

17. Software for practical analytic use

- I find the solver very convenient.
- Others like the "Maxent Grammar Tool" of Wilson and George: https://linguistics.ucla.edu/people/hayes/MaxentGrammarTool/
- The most high-powered software I know was put together by Tim Hunter:
  - https://github.com/timhunter/loglin
- You can also try logistic regression in R, which would also let you use mixed-effect models…

18. How to cite MaxEnt

- My personal practice is to cite:
  - Smolensky, Paul (1986) Information processing in dynamical systems: Foundations of Harmony Theory. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models, ed. by James L. McClelland, David E. Rumelhart and the PDP Research Group, 390-431. Cambridge, MA: MIT Press. All the math is here — just no language!
  - Goldwater, Sharon & Mark Johnson. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), Proceedings of the Stockholm Workshop on Variation within Optimality Theory, 111–120. Stockholm: Stockholm University. Brought the framework back to notice, reanalyzing examples from Boersma and Hayes (2001.

EVALUATING MAXENT (AND OTHER THEORIES) FOR INTUITIVENESS

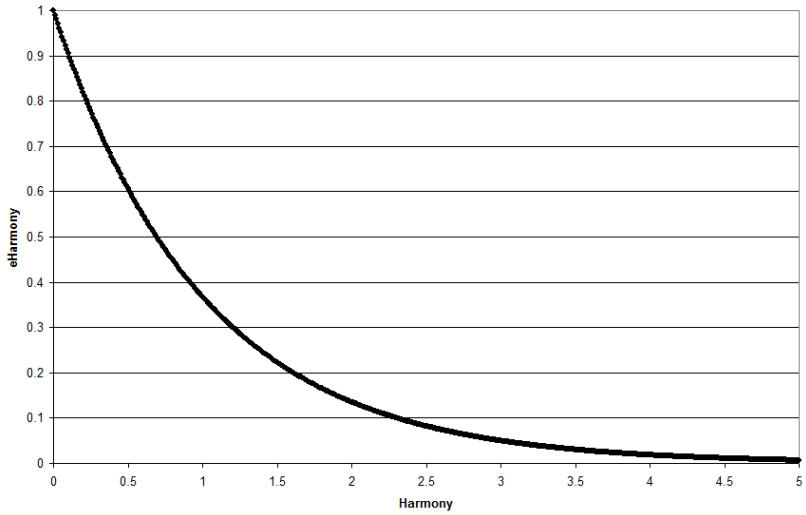
19. Key idea

- Think of constraints as *evidence* in favor of or against particular candidates.
- Then we can compare the MaxEnt decision procedure to our own internalized sense of rational decision making.

[3] The big papers appear to be: Berger, Adam; Stephen Della Pietra; and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22. 39-71. and Della Pietra, Stephen, Vincent J Della Pietra & John D Lafferty. 1997. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19. 380–393.

20. The maxent procedure thus assessed

- Every constraint has a weight, a non-negative number.
  - This is intuitive; reasons differ in cogency.
- Every candidate is given a **harmony score**.
  - = weighted sum of its violations
  - i.e. pairwise multiply weights and violation counts, and sum up
  - This is intuitive: weigh **all** the evidence. ☞ what theory does not?
- Every harmony score is converted to an eHarmony[4] score, by negating it and taking *e* to that power. *e* is about 2.718.



  - Note that Harmony is a badness score (penalty), but eHarmony is a goodness score (virtue)
  - The conversion of Harmony to eHarmony is intuitive, because once Harmony is very big, only very small gains in eHarmony are made when Harmony increases (and, just below, probability will be derived from eHarmony).
- Take all the eHarmony scores and add them up. By tradition this number is called **Z**.
- The probability assigned to a candidate is the share of its eHarmony in Z; in other words, divide the eHarmony value by the Z value.
  - So, probabilities will sum to 1.
  - This is intuitive, since a choice is less appealing when there are strong alternatives.

FINDING THE BEST WEIGHTS

21. Sources on this

- User friendly intro for this: Hayes and Wilson (*LI* 2008)
- The real math with proofs etc: Jurafsky and Martin, cited above

[4] Caution: I really like this term, invented by Colin Wilson. But it is tricky to use in writing, since it is a joke (eHarmony is a dating site on the internet).

22. The usual method for finding the best weights

- Computer science tells us that usually the best way to search for something is to take a preliminary step: **define what you want numerically**.
  - e.g. "set the weights so that this formula is maximized"
- A number commonly used is called the **likelihood** (e.g., the likelihood of these data, assuming these constraint weights)
- So, "find the weights whose likelihood is the maximum possible".
- Likelihood is known as an **objective function**, i.e. a function that measures goodness of solutions and the maximization[5] of which forms our objective in weight-setting.

- Phonologists are emerging from a somewhat benighted period, in which I participated heavily, in which you lay out a search procedure and hope that it works …

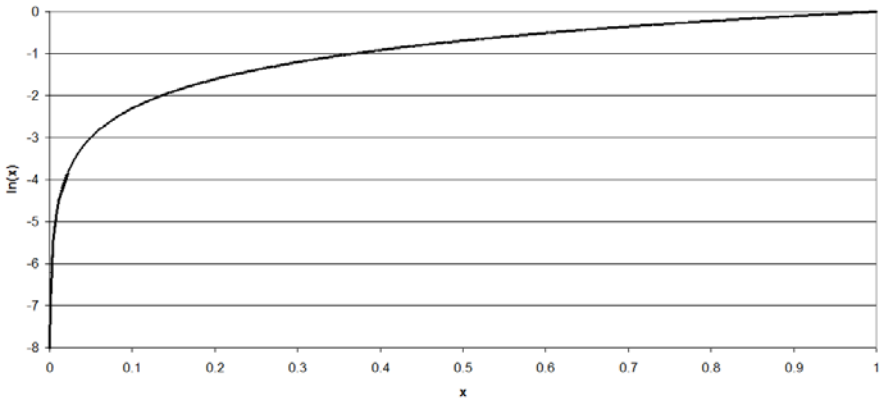
23. Computing likelihood

- The grammar assigns every observed datum a probability.
- Since probabilities multiply, we can just multiply across all the data to assign a probability to the complete dataset.
- Intuition behind the use of likelihood:
  - Probability always sums to one.
  - Divert that probability (the "probability mass") as much as you can toward the forms that actually exist.
  - This will divert it away from the forms that don't exist, yay.

24. Shifting to log likelihood

- For any realistic problem, likelihood values are extremely low, so to keep Excel from crashing we instead take the natural logarithm of the likelihood ("log likelihood").
  - This preserves the relative goodness of different solutions (monotonic) but keeps the numbers manageably small.

[5] Sometimes minimization.



25. The actual computation

- Summing the logs of numbers has the same effect as multiplying the numbers.
- So: in the spreadsheet, you multiply frequencies of candidates by the log of their probabilities, then sum up to get the log likelihood — the magic spreadsheet cell.

26. Searching for the best weights

- There are many algorithms, invented by computer scientists, that can swiftly and accurately find the maxent weights that maximize log-likelihood.
  - We don't care much about them, I suspect — our work was already done when we implemented maxent and its likelihood function.

27. A big caveat about maximum likelihood

- It is vulnerable to theories that permit **endless proliferation of hypotheses**
- The ultimate limit is the analysis that has constraints like
  - "The plural of *leaf* is *leaves*"
  - "The plural of *gulf* is *gulfs*"
  - etc.
  - ☞ What is the likelihood of the data under this account?
- The term for evil analyses of this kind is **overfitting**.
- The culture of phonology tends to protect from this, since we already are big fans of highly-general constraints.
  - And so, I will suggest, are native speakers …
- To the extent we use really parochial constraints, we must be cautious of the maximum-likelihood solution (more on this later).

PREPARING FOR THE HOMEWORK: MEDIAL CLUSTER ANALYSIS

28. An encounter I had

- My appointment with a thoughtful grad student at Cornell.

- Says,
  - "Everybody thinks the Syllable Contact Law is relevant to my language's phonology, but I'm not so sure."
  - "It seems to me that other, independently motivated constraints will do the work we attribute to the Law, which is then perhaps not needed."
- I think we need some way of testing such claims, in principle more precisely than "satisfying account".
- Perhaps creating a complete model, assigning a probability to every logically possible medial cluster, might help — does Syllable Contact Law help, at a statistically significant level?

29. The "Markedness Only" approach to phonotactics

- To my knowledge this was invented by Hayes and Wilson (2008), though the idea is pretty obvious.
  - Bruce Hayes and Colin Wilson. (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379-440.
- Assign a probability to every form in GEN.
- Or, perhaps, every form in GEN less than 20 phonemes long…
- This can only use Markedness constraints — so things like Positional Faithfulness cannot be used.

30. How phonotactics is done in classical OT (Prince and Smolensky 1993)

- Rich Base: everything can be an input
- Grammar as filter: some inputs get changed to something else.
- The full set of "something elses" and survivors form the set of legal forms.
- I worry about the ability of this system to capture marginal cases: ?[pɔɪk] is mildly aberrant to me, but I have no inclination to repair it (e.g. to [paɪk]).

31. The goal at hand (homework)

- Suppose the Markedness Only theory of phonotactics is correct.
- Doing whole languages is a huge job (see Hayes/Wilson, and their software, which uses finite state machines to cover vast sets of strings).
- But medial consonant clusters: V<u>CC</u>V as manageable: GEN is only the square of the number of consonants.
- So: obtain a full, explicit, gradient analysis of some language's phonotactics, using maxent.

32. Note on vowel sequencing as an option

- If you want to do vowels instead, that would be fine with me.
- E.g.,
  - Does a language have tacit vowel harmony?

  - Does a known vowel harmony language obey its own rules within stems?
- I suggest you start with just disyllables, since initial vowels often are very special for vowel harmony systems.

33. Step 1: obtain an electronic lexicon from the Internet

- You want phonemic listings (IPA not essential).
- Hopefully not too huge a consonant inventory
- Perhaps useful not to have too many VCCCV.
- I have found I sometime have to steal the data one letter at a time.

34. Sample solution

- I did **Warlpiri** (Australia, a focus of colleague Margit Bowler, outstanding linguists have worked on it for decades; good online resources and book references).
- I also made extensive use of the excellent 1980 MIT dissertation by David Nash, which covers the phonology and particularly the medial clusters.

35. Grabbing the dictionary

- Download the whole online dictionary one initial letter at a time.
- Discard all but the entries:
  - In Word, replace every space with a tab.
  - Paste into Excel, and keep only the first column.
  - Sort that column and discard crud.

36. Forming a list of medial clusters with counts

- Harvest the medial clusters:
  - Paste first column of spreadsheet into Word, then
  - Replace the long vowel digraphs *aa*, *ii*, *uu* with single symbols.
  - Replace every vowel with *tab vowel tab*
  - Paste result back into Excel and intervocalic consonants and clusters are all in the same column! (no vowel initial words or hiatus)
- Reduce the medial clusters, original a list of tokens, to single counted types
  - I use my Typizer, toy software I can share.
  - In Excel a pivot table will do it.
- Discard the singletons (VCV)
- Starting with a list of the consonant phonemes, make a list in Excel of all logical combinations.
- Plug in the frequencies for the attested and zeros elsewhere.
- Now you are ready to analyse!

37. Maxent analysis of clusters on a spreadsheet

- Add a lot of columns with feature values needed for both C1 and C2.
- Then use the formula = IF(AND(….), 1, 0) to assign constraint violations.
  - … can be references to feature values, or to segment identity.
- The rest is just plain phonology: use your brain/guile to find really good constraints, and watch the log likelihood go up.
- A scattergram of observed/predicted can lead to increased analytical excitement.
- It is useful to include a column that detects the biggest overgeneration error (higher predicted probability than observed probability).