Linguistics 219 Phonological Theory III
B. Hayes Spring 2019

Class 1, 1/9/17: Goals; Maxent I

1. Go over syllabus

2. Assignments for this week

- First reading assignment:
  - (2006) Bruce Hayes and Zsuzsa Londe. "Stochastic phonological knowledge: the case of Hungarian vowel harmony". *Phonology* 23:59-104. On course web site.
- Want more background on the maxent covered today? Read sections 2 and 3 of
  - (2008) Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39, 379-440. On course web site.
- Medial cluster exercise to be handed out on Wednesday.
- Start thinking about what you want to do your term paper on.

3. Goals today

- Goals of phonology: rationalization vs. prediction
- Frameworks as ways to predict, and how Junior is in the same position as me.
- Plunge into a specific case (/f/ Voicing) with a specific framework.

COMMENTS ON GOALS AND THE HISTORY OF PHONOLOGY

4. A historical shift?

- Historically, the field has been corpus-oriented and bent on **rationalization** of the data:
  - "this gives a **satisfying account** for X". [1]
- This sometimes seems troublesome to me (due to objectivity of "satisfyingness") but I think it is essential as a reconnoitering of the empirical territory.
- I think we are moving toward a new era in which the name of the game is to **make correct predictions**; see below.

5. Data rationalization: case study of English stress

- *SPE* (Chomsky and Halle 1968) was a classical work of this area.
- It sought a satisfying account of English stress, and also of the intricate alternations of vowel quality found in the learned vocabulary stratum of this language.
- In seeking this, it achieved a degree of analytical detail seldom observed since.

[1] "satisfying account" & "phonology" makes an interesting search on Google.

- It also was "recreational" to an extent seldom observed, positing highly radical analytical moves on behalf of the tiny numbers of actual forms.

6. How SPE did it

- All regularities were encoded by rule.
- Irregularities were handled by assigning abstract representations, e.g. /giræffɛ/ to get final stress in *giraffe*.

7. A tiny bit of *SPE*: one concrete research question

- In an English word of the form [X V C_0 V C C $\begin{bmatrix} V \\ -\text{stress} \end{bmatrix}$ C]$_{\text{word,}}$ will stress be penultimate or antepenultimate?[2]
- This is an old research question from *SPE*, where they sought to predict English stress.
- I checked with my primitive search software.
- By far the norm is penultimate stress, which is what the *SPE* rules predict.
- But let's look at the rarer words that have antepenultimate stress. (Source: my personal edited version of CMU, own search software).

invertebrate	IH2 N V ER1 T AH0 B R AH0 T
cerebral	S EH1 R AH0 B R AH0 L
vertebral	V ER1 T AH0 B R AH0 L
ambassadress	AE2 M B AE1 S AH0 D R AH0 S
ambergris	AE1 M B ER0 G R IH0 S
integral	IH1 N T AH0 G R AH0 L
ludicrous	L UW1 D AH0 K R AH0 S
inadequate	IH0 N AE1 D AH0 K W AH0 T
adequate	AE1 D AH0 K W AH0 T
harlequin	HH AA1 R L AH0 K W AH0 N
discipline	D IH1 S AH0 P L AH0 N
talisman	T AE1 L IH0 S M AH0 N
armistice	AA1 R M AH0 S T AH0 S
pedestal	P EH1 D AH0 S T AH0 L
idolatrous	AY2 D AA1 L AH0 T R AH0 S

- A few words I will spare you: they have inflectional or consonant-initial suffixes, which are known to be ignored for participation in stress:

Wellington	W EH1 L IH0 NG T AH0 N
Parkinson	P AA1 R K IH0 N S AH0 N
singleton	S IH1 NG G AH0 L T AH0 N
Christendom	K R IH1 S AH0 N D AH0 M

[2] Pre-antepenultimate is vanishingly rare.

☞ **Exercise**: find an appropriate characterization — a "satisfying account" of the "allows antepenultimate" clusters.

8. Some cleanup points for the English example

- Here are the clusters, with word counts, that take penultimate stress:

K SH	94	L D	3
S T	81	R D	3
K T	70	R F	3
N T	54	R K	3
N SH	42	Z M	3
N CH	32	D R	2
P SH	25	M N	2
N D	19	DH M	1
L Y	16	F R	1
N S	16	G N	1
P T	15	G W	1
N Y	10	L G	1
R T	9	L JH	1
S CH	8	L M	1
L SH	7	L T	1
R SH	7	M F	1
M B	6	P L	1
R M	6	P R	1
K N	5	P S	1
M SH	5	R B	1
NG G	5	R JH	1
T R	5	R N	1
L S	4	R P	1
M P	4	R S	1
B L	3	S K	1
B R	3	TH L	1
B Y	3	V R	1
K S	3		

- The [st] clusters are not what they seem: virtually all of them precede *-ic,* a pre-stressing suffix (*majestic*).
- Ditto for [tr] in *geometric, geriatric,* others
- *Establish*, with [bl], likewise has a pre-stressing suffix.
- *Cathedral*, with [dr], has a long vowel, which is itself stress-attracting.
- What would you say about *injustice*, with penultimate stress?

- Exceptions that lack "excuses" of this kind end up few:
  - *intestine*, *Nicaraguan*, *asbestos*, a few others

9. We can put this all together in classical OT

NONFIN forces feet to not cover final syllables.
*CODA keeps onsets maximal
FOOTBIN-MAX (of the moraic trochee type) avoids unstressed heavy penults
FOOTBIN-MIN avoids stressed light penults.
ALIGN-FOOT-RIGHT enforces the stress window.
-IC IS FOOTED might give us the prestressing property of *-ic*.

Do: *adequate, connective*, *Mississippi*

10. The point at hand

- Syllabification provides what seems a "satisfying account" of the patterning of penultimate/antepenultimate stress in English.
- It establishes a (loose) connection with word-initial clusters and stress patterning.

11. Footnote about *SPE*

- This actually is the most conspicuous *failure* of the work!
- A whole series of last-minute footnotes were added, apologizing for ugly rules which were necessitated by the lack of a theory of the syllable.
- The remedy was quickly made in the early 1970's, by phonologists such as James McCawley, Daniel Kahn, and Lisa Selkirk.

12. A focus of this sort of work

- It tends to get **focused on a corpus** — this was especially true of the old literature on English stress.
- Once everything is derived, we are sort of done, but this is more:
  - What is *not* out there?
  - What is out there, but quite abnormal?

13. Complaints about English stress in SPE

- There is no clear sense of what is well-formed, or semi-well-formed.
- Abstract geminates let us derive words like

  antenna
  abscissa
  Nutella

- Silent /x/ lets us derive otherwise puzzling velar nasals in words like *dinghy*

/dɪnxi/	UR
ŋ	Nasal Assimilation
∅	x Deletion
dɪŋi	Surface representation

- I think these devices eventually lead us to lose a clear conception of phonological legality; e.g. with underlying

  /mafɛxxɛ/

  we derive *[maˈfɛ], which is outright impossible in English.

- I.e. assuming a perfectly regular phonology and abstract UR's to handle the exceptions eventually loses us our grip on the essential concept of phonological normalness, which tends to be more surfacey.
- It would have been better to adopt a theory that lets us approach normalness more directly.

A MORE RECENT APPROACH TO PHONOLOGICAL THEORY: PREDICTION

14. Some examples of predictions that might be made by a phonological analysis

- "When Hungarian speaker attempts to say the dative of [bortog] (nonce form), she will say [bortog-nɔk]"
  - see readings
- "English speaker will forthrightly reject [vzɛp] as sounding un-English (or whatever)"
  - see Scholes (1965), *Phonotactic Grammaticality*
- "Speaker will be ambivalent about saying [haːdeːl-nɛk] Y or [haːdeːl-nɔk] for dative of [haːdeːl]"
  - see readings
- "Speaker will be ambivalent about the well-formedness of [vlɛp]".
  - again example from Scholes

15. Even the SPE stress analysis makes *some* predictions

- To my knowledge, there is no way in *SPE* to derive either of these:

  [ˈpɑdəktəl] 'podectal' (Liberman and Prince 1977)
  [ˈpæmələnə] 'Pamelana'

  and I would judge that they sound distinctly un-English.

ENGLISH F-VOICING AS A CHALLENGE FOR PREDICTION

16. The phenomenon

- [f] is replaced by [v] in final position in plurals of about 20 English nouns.

17. The historical origin of this phenomenon

- This is discussed in my textbook, *Introductory Phonology* (2008:231).
- /f/-Voicing in plurals is a historical survival of an old allophone:
  - Old English: no /v/ phoneme
  - [v] as intervocalic allophone of /f/.
  - The plural suffix had a schwa vowel at the time.
- Historical (not synchronic) derivation:

sɛlf	sɛlf-əz	'self-sg./pl.'
—	v	Intervocalic voicing of fricatives (cf. *baths¸houses*)
—	∅	Loss of schwa in inflectional endings
[sɛlf]	[sɛlvz]	outcome

- A few relics elsewhere in the system, like *breath ~ breathe*
- There has been leveling since then, and novel forms have usually not undergone the voicing.

18. Data

- I found on my personal English database a full set of the nouns ending in /f/.
- I sorted them for whether they undergo /f/-Voicing in the plural.
- Different speakers are different (e.g. some people tolerate *gulves* for *gulf*).

19. My own set of voicers

Obligatory	Optional
calf	*behalf*
elf	*dwarf*
half	*epitaph*
knife	*hoof*
leaf	*roof* [3]
life	*wharf*
loaf	
scarf	
self	
sheaf	
shelf	

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

thief
wife
wolf

- Note that absolute-core character of the obligatory voicers.
  - Key words of life
  - Folkloric words (*elf, sheaf, wolf*)
- Note *dwarf* as an *extension* of the alternation; its historical plural was *dwarrow* and *dwarves* has newly entered into competition with *dwarfs*.

20. Excursus: splits in usage?

- *Dwarfs* is appropriate for Disney, *dwarves* for Tolkien.
- *Shelfs* is marginally ok and individuates the shelves; *shelves* feels more a like a unified group of shelves.
- *Loaves* works well as a measure word: *three loaves of bread*. *Loafs* is marginally ok but would not be appropriate as a measure word.

21. My own set of non-voicers

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	

PREDICTION I: HOW WOULD I RESPOND IN A WUG TEST?

22. The first wug test tested this!

- Berko, Jean (1958). The child's learning of English morphology. *Word* 14: 150-177.

16. Plural. One insect, then two. "This is a heaf /hiyf/. Now there is another one. There are two of them. There are two ——."

- Responses for 12 adults[4]: 5 [hivz], 7 [hifs]
- Responses for 89 children (half pre-schoolers, half first graders):
  - 9 [hif] zero change is very common in wug-testing very young children
  - 4 [hifəz] treating [f] as a sibilant?
  - 3 [hivz] precocious!
  - 73 [hifs] favoring the "regular" or uniform-paradigm outcome
- I personally am ambivalent and would be happy with either [hivz] or [hifs].

HOW DO WE PREDICT THINGS?

23. Options

- Here some theorist might want to go with some form of analogy.
  - We will discuss this later if time.
- For many cases (and we'll include this one), we want to produce a grammar.
  - For this case, we'll have to countenance some rather parochial constraints!

24. Starting assumptions about the speaker

- He knows words that "go both ways".
- He (rationally?) expects that novel words will behave rather like the known words.
- Having undergone phonological acquisition in childhood, he has a grammar that tracks the properties of words that are relevant to /f/ Voicing.
- He may also bring some **UG biases** to the problem — see the Becker/Nevins paper, later on, which claims precisely this for /f/ Voicing.

25. Paean to constraint-based grammars

- Intellectually, it seems a good idea to break down hard problems into simple ingredients — like (many) OT constraints.
- As so often in OT, we can use a ranking (or weighting) to get intricate patterns to emerge from simple ingredients.
- … and setting the rankings can sometimes be done by algorithm more accurately than people can do it.

[4] I suspect: wandering up and down the halls of a university; several consultants were described as having graduate degrees.

26. Optionality and ambivalence

- All of the classical OT literature assumes one winner for each input.
- But we are already facing six forms where the native speaker states that both are ok.
- So let's try various frameworks that permit nuances to be taken into account. Just one of them today.

27. 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. belief that Bruce has 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

28. 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.

INTRODUCTION TO THE MAXENT FRAMEWORK

29. Desiderata for a theory (met, I believe, by maxent)

A. The candidate set is assigned **probability values**.
  - Maxent: often, these are so close to 0, or 1, as to recapitulate OT.
  - Every OT grammar has a maxent translation.[5]

B. The **total probability** of the candidate set is **1**.
  - This is just part of the theory of probability.

[5] Caution: you need to limit violation counts to some finite number; see Prince (19xxx) "anything goes".

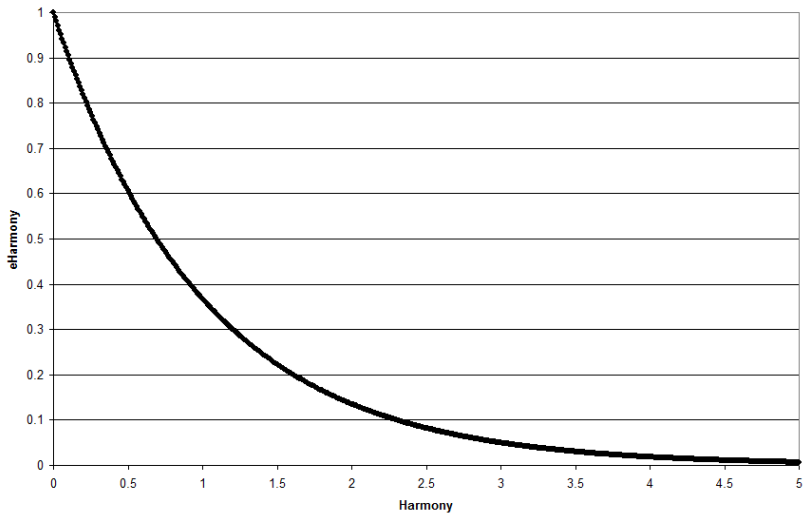
C. Constraints vary in how much they influence the probability of a candidate.
- They are "strong" or "weak".

D. **The closer we want to get to certainty, the more evidence we need**.
- If your a priori belief in choice A/B is 0.5/0.5, then evidence E will likely sway us easily toward 55/45 or 45/55.
- But if you already assign 0.99 probability, then evidence E will likely only sway us up to (say) 0.999.

E. We **combine evidence from multiple sources** in forming our judgments.
- So the right theory probably embodies some sort of addition.

30. The maxent procedure

- Every constraint has a weight, a non-negative number.
  - This satisfies (0C).
- Every candidate is given a **harmony score**.
  - = weighted sum of its violations
  - i.e. pairwise multiply weights and violation counts, and sum up
  - This satisfies (0E).
- Every harmony score is converted to an eHarmony[6] 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 satisfies (0D), 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**.[7]

[6] Caution: I really like this term, invented by Colin Wilson. But it cannot be used in writing, since it is a joke (eHarmony is a dating site on the internet).
[7] which I would guess stands for German *Zahl* 'total'.

- 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, satisfying (0B).

31. Maxent on spreadsheets: a small example with North American English Tapping

- Phonological suppositions about Tapping:
  - Tapping in English is by far the norm when the context (intervocalic, pre-atonic) is met.
  - /d/ taps a little bit more readily than /t/.
  - Tapping is outright impossible when the crucial environment is met.
- A quickie set of constraints:
  - DON'T NOT TAP (= *[+syl] $\begin{bmatrix} -\text{son} \\ -\text{cont} \\ +\text{ant} \\ -\text{cor} \end{bmatrix} \begin{bmatrix} +\text{syl} \\ -\text{stress} \end{bmatrix}$]
  - IDENT(sonorant)
  - IDENT(voice)
- We proceed to a spreadsheet and implement the maxent calculations for eight candidates: the tapped and untapped versions of *put, putting, ride,* and *riding*.
  - Lurking, important assumption: all other candidates will be penalized by very highly weighted constraints. This has to work or we are sunk.

32. The beautiful 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[8] of which forms our objective in weight-setting.

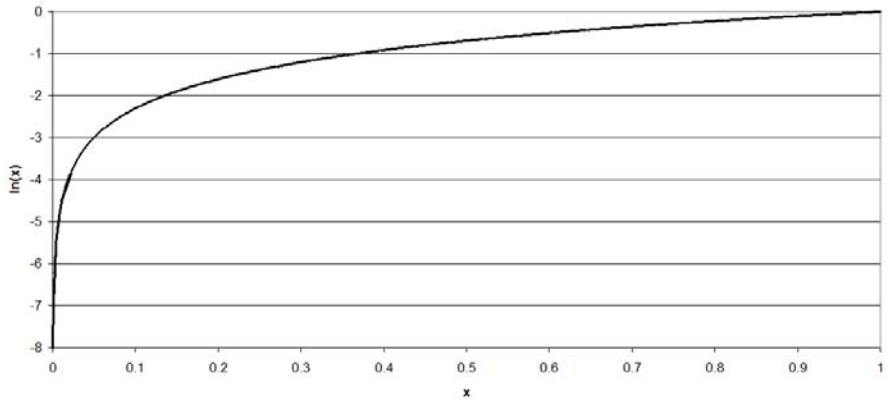
33. Computing likelihood

- 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.

[8] Sometimes minimization.

34. 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.



35. 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.

36. 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.
- Excel has a few of these algorithms, in its plug-in, free Solver.
  - I use the default settings.

37. Return to our Tapping example, letting Solver set the weights

- We need only add new columns that define the objective function, then run Solver.

38. Return to our /f/ Voicing example

- We seek a model that includes constraints embodying various factors:
  - Why should [v] be favored in general?
  - Why should [f] be favored in general?
  - What circumstances totally rule out [v]?
  - What circumstances make [v] especially likely?

- We can now easily implement this with our data file, obtaining predictions about every word — existing, or wug words like *heaf*.
  - These attempt to be a model of the native speakers *tacit degree of belief* that a noun ending in [f] should take a [v] plural.

39. Looking at the output of the grammar

- 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?