Linguistics 219 Phonological Theory III
B. Hayes Spring 2020

Class 1, 3/20/20: Orientation; Maxent I

1. Go over syllabus

2. Assignments for this week

- First reading assignment: Claire Moore-Cantwell (ms.) "Weight and final vowels in the English stress system". On CCLE Web site.
- First homework exercise to be handed out on Thursday.
- Start thinking about what you want to do your term paper on.

3. Topics for today

- Goals of phonology: rationalization vs. prediction, and what this implies
- Frameworks as ways to predict.
- Work out/play with a specific case (/f/ Voicing) in MaxEnt.

SHIFTS IN PHONOLOGY: RATIONALIZATION VS . PREDICTION

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".
  - This sometimes seems troublesome to me (due to objectivity of "satisfyingness") but I think it is still important as a reconnoitering of the empirical territory.
- I think we are probably moving toward a new era in which the name of the game is to **make correct predictions**; see below.

5. Data rationalization: English stress as a case study

- *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; hundreds of difficult-to-read pages!
- *SPE* — and the work it immediately inspired —was also "**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 either with rule-exception features, or more often by positing abstract underlying representations.
- Hence Chomsky and Halle said — perhaps somewhat cheaply? — that "English stress is predictable."

7. Example of the data-scrutiny approach of *SPE*: one (small) research question

- In an English word of the form [X V C_0 V **C C** $\begin{bmatrix} \text{V} \\ -\text{stress} \end{bmatrix}$ C]$_{\text{word}}$, will stress be penultimate or antepenultimate?[1] Base the answer on what **C C** is.
- *SPE* found its examples by combing through the *Pronouncing Dictionary of American English* by Kenyon and Knott (1944), a beautiful and compendious work.
- Nowadays we would search digitally, and I checked with my primitive search software. (https://linguistics.ucla.edu/people/hayes/EnglishPhonologySearch/.)
- 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.

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

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

[1] Pre-antepenultimate is vanishingly rare. Please ignore verbs, which work differently.

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
*BADONSET avoids outlandish onsets (to be formalized …)
FOOTBIN-MAX (of the moraic trochee type) avoids unstressed heavy penults
FOOTBIN-MIN avoids stressed light penults.
ALIGN-FOOT-RIGHT enforces the stress window.

Do: Cressida, discipline, Alinda, pangolin

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*

- What we just did is actually the most conspicuous *failure* of this work! They had no syllables.
- 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. In a way I love this work, and not just from nostalgia!

- The theory is really, really simple, can be done on yellow pads, and encourages us to reflect.
- Also: however diverse research methods in linguistics get (fieldwork, typology, experiment, computational simulation), **theory and analysis** can tie these threads together as a common research program.
- However, we can also anticipate that theory and analysis will evolve a lot as they serve in this role.

13. Corpus-rationalization and absent forms

- The SPE-style work, as noted, tended 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 there is more:
  - What is *not* out there?
  - What is out there, but quite abnormal?
  - This was a virtue of OT, the first theory that forced phonologists to think about this question (because of GEN).

- The really dangerous blend was based on
  - A zest for really abstract solutions *combined with*
  - Neglect of the need to not-derive what is not-there.

14. An example of the defective strategy

- *SPE* posited abstract geminates, which made stress "predictable" in words like these:

  antenna
  abscissa
  Nutella

  [☞ Specify informally the rules and ordering.]

- Also, final /ɛ/ lets you treat final syllables as if they were penultimate.

  picturesque, eclipse, bequest, romance

15. ☞ now tell me how *SPE* gives these words "predictable" stress

giraffe, acquiesce, Ronelle, lagniappe, Yvette (others?)

16. The underlying velar fricative of English

- Silent /x/ lets us derive otherwise puzzling velar nasals in words like *dinghy*
  - Note that [ŋ] is otherwise only final or before a velar consonant.

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

☞ Other such words?

17. Overgeneration?

[try combining the two devices and see what you get. Do not turn page.]

/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 — per Kisseberth (1970) — tends to be more surfacey.
- It would have been better to adopt a theory that lets us approach normalness more directly.
- This leads us, further on, to constraint-based theories of phonotactics, which are perhaps more controllable …

A MORE RECENT APPROACH TO PHONOLOGICAL THEORY: PREDICTION

18. 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]" (Hayes and Londe 2006 *Phonology*, Hayes et al. 2009, *Lg.*)
- "English speaker will forthrightly reject [vzɛp] as sounding un-English (or whatever)"
  - see Scholes's pioneering study (1965), *Phonotactic Grammaticality*
- "A Hungarian speaker will be ambivalent about saying [haːdeːl-nɛk] Y or [haːdeːl-nɔk] for dative of [haːdeːl]" (refs. as before)
- "Speaker will be ambivalent about the well-formedness of [vlɛp]".
  - again Scholes

19. Source of the predictions

- Speakers, during phonological acquisition, apprehend the phonological pattern of the lexicon, including frequency information ("Law of Frequency Matching").
- So a formal model trained on the lexicon can make predictions about their behavior under elicitation/testing.

20. 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, *LI* 1977)
  [ˈpæmələnə] 'Pamelana'

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

WHAT ARE GOOD TOOLS FOR PREDICTING THINGS?

21. Caution

- I'm really quite out of my depth here and am just trying to think through the problem superficially.

22. Mathematical laws

- This is how physics, queen of the sciences, predicts things.
- The theory involves essentially perfect understanding (I think…) and to predict we just calculate.
- Even here things are not always smooth; the analyst makes approximations and calculates with them.

23. Complex systems (e.g. epidemiology)

- We find **multiple factors** that influence an outcome, find how strong they are, how they interface.
- For phonology, we are in a curious position: we model messy lexical data, hoping to mimic how the native speaker models messy data. The theory predicts by mimicking acquisition.

24. A consequence

- Modern **statistics** is the use of sophisticated mathematical methods, founded in probability theory, to make predictions using data.
- … and modern phonological (linguistic) theories increasingly look like standard statistical models
- E.g:
  - OT, constraints, ranking … shifting toward
  - statistical models, factors, weighting
- An extreme case is Zymet (2018, UCLA diss.), who sketches a constraint-based model that is based on mixed-effects logistic regression, with word as an random effect.

25. An example of a prediction that would never have been made in the SPE era

- Moore-Cantwell's constraint to the effect that words of at least three syllables ending in /i/ have antepenultimate stress (readings).
  - This sometimes even overrides the heavy-penult constraint: *galaxy*, *Picardy*
- Nobody knows why it should be true
- It wug-tests as valid.
- It **cross-classifies** with the normal constraints on syllable count and weight; perturbing the numbers in the direction of antepenultimate.
- It looks, for all the world, like an *effect*.

ENGLISH F-VOICING AS A CASE STUDY

26. The phenomenon

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

27. 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*)
—	sɛlvz	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.

28. 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*).

29. My own set of voicers

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

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

shelf
thief
wife
wolf

- Note that absolute-core character of the obligatory voicers.
  - Key words of life
  - Folkloric words (*elf, sheaf, wolf*)
  - This is frequency in historically-archaic alternations, and it a good way to spot them.
- Note *dwarf* as an *extension* of the alternation; its historical plural was *dwarrow* and *dwarves* has newly entered into competition with *dwarfs*.

30. 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	

31. Excursus: splits in usage?

- Do you share these perhaps-subtle intuitions?
  - *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. *hooray, I've now bought three ___; I put the books on the ___.*
  - *Loaves* works well as a measure word: *three loaves of bread*. *Loafs* is marginally ok but would not be appropriate as a measure word. *I think we should produce ___.*
- If valid, what is the right story for them?

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

32. The very 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[3]: 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] These clever children will enroll in prestigious graduate programs later on.
  - 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?

33. Options

- Here some theorists might want to go with some form of **analogy**.
  - We will discuss this later if time (Google "TiMBL" or "Analogical modeling of language for two contemporary schools).
- 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!

[I will leave this old handout material in place but prefer to plunge into the maxent analysis now.]

34. Starting assumptions about the speaker

- She knows words that "go both ways".
- She (rationally?) expects that novel words will behave rather like the known words.
- Having undergone phonological acquisition in childhood, she 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 — we will discuss work by Nevins and colleagues bring this to bear on /f/ Voicing.

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

35. 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, MaxEnt.

36. The /f/-voicing exercise

- Open spreadsheet, look at data, consider factors, implement the MaxEnt analyis, see how well it works.

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

- North American English tapping (*butter*)
- Applies only in "ambisyllabic" position (intervocalic pre-atonic) (compare *attend*)
- Probably more obligatory to tap with /d/ than /t/: *writing* vs. *riding*[4]
- 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.

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?

[4] There is empirical work on corpora, e.g. by Dani Byrd, which I ran out of time to check.

MAXENT FROM FIRST PRINCIPLES

40. 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

41. 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.

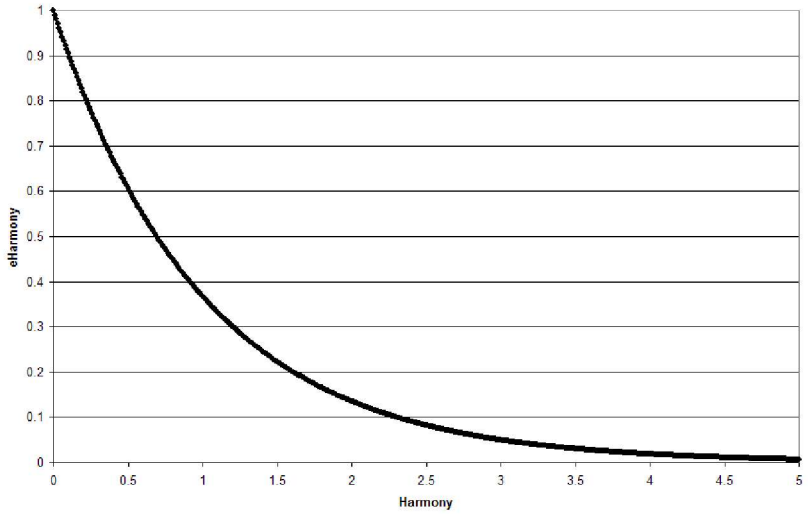
42. Evaluating MaxEnt (and other theories) for intuitiveness

- 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.

43. The maxent procedure

- 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[5] 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.

44. 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[6] of which forms our objective in weight-setting.

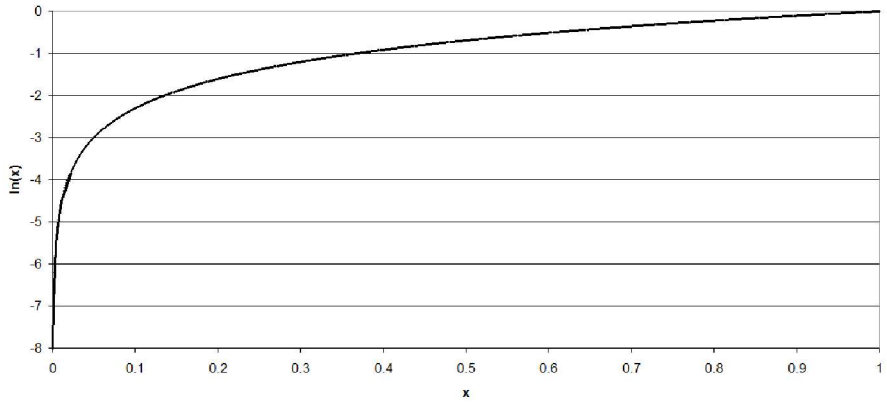
[5] Caution: I really like this term, invented by Colin Wilson. But it is tricky to use in writing, since it is a joke (eHarmony is a dating site on the internet).
[6] Sometimes minimization.

45. 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.

46. 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.



47. 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.

48. 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.