Bruce Hayes
UCLA

BLS Workshop, UC Berkeley
February 8, 2020

Assessing grammatical architectures through their quantitative signatures

OVERVIEW

1. Gradient phenomena in phonology

- Types of phonology where we need numbers and probabilities:
    - Generating **alternative surface forms**, at varying frequencies, from the same underlying form (much of the research literature in sociolinguistics)
    - **Frequency-matching the lexicon** when generating novel forms (Zuraw 2000 et seq.). E.g. Hungarian [Ciː C] stems take about 7% Back harmony in both the lexicon and in wug-testing; Hayes et al. 2009.
    - **Gradient well-formedness judgments**; e.g. ✓[kɪp], ?[pɔɪk], *[bzɑɹʃk], which frequency-match the patterns of the ambient language (Hayes and Wilson 2008)

2. Frameworks for analysis of gradience

- **MaxEnt grammars** (Smolensky 1986, Goldwater and Johnson 2003)
- **Noisy Harmonic Grammar** (Boersma and Pater 2016, Hayes 2016)
- **Stochastic Optimality Theory** (Boersma 1998, Boersma and Hayes 2001)
    - These often behave similarly and are all currently "in contention" as frameworks.

3. Strategy taken here

- Think a little bit abstractly about these frameworks, in a particular way:
- We want to find general predictions that will guide us in theory-evaluation.
    - These might be called **quantitative signatures**
- Here, I will do several, for each:
    - describe and explain the quantitative signature
    - cite real-world cases
    - say which frameworks possess these signatures

DESCRIPTION OF MAXENT

4. Basics

- In linguistics, MaxEnt is a version of Optimality Theory (Prince and Smolensky 1993). We have:
    - inputs
    - candidate outputs

  - constraints used to make the decision
- The theory is probabilistic, so it assigns a probability to every member of GEN (most of them essentially zero).
- With these assigned probabilities, we can assess the predictions of the analysis against quantitative data.

5. MaxEnt and common sense

- Suggestion: think of constraint violations as *evidence* that helps you *decide* which candidates should win or lose (better: have high or low probability)
- MaxEnt can be viewed as a mathematicization of how evidence is sensibly brought to bear on decisions.
- I suggest that, as such, it is a **mathematically close embodiment of common sense**.

6. The MaxEnt formula deriving probability of candidate *x* from its tableau

$$\Pr(x) = \frac{\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))}{Z}\text{, where } Z = \Sigma_j\ \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$$

- "The probability of candidate *x* is derived from the tableau information as …"
- We will cover the whole formula one step at a time.

7. Weights

- Every constraint has a nonnegative number, its **weight**, which tells you how strong it is.
  - More specifically, how much it lowers the probability of candidates that violate it.
  - In (6), this is w_i for each constraint i.
  - Weights are intuitive —we know that reasons differ in cogency.

8. MaxEnt, Step 1

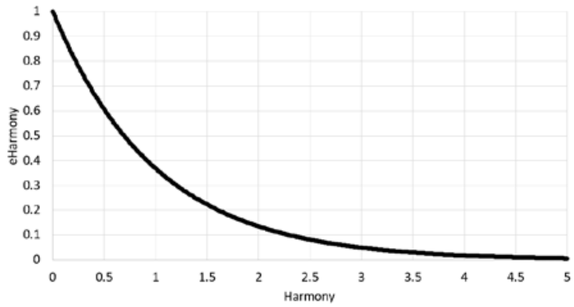
- For each tableau cell, multiply the number of violations by the weight of the constraint.
  - In (6), this is $w_i f_i(x)$ (x is candidate, f is number of violations)
  - This is intuitive, in the sense that ** is plausibly "twice the evidence" of *.

9. MaxEnt, Step 2

- Add result of Step 1 across tableau rows to get a single value.
- This is an aggregate penalty score for a candidate, called the **Harmony**.
- In formula (6), Harmony is represented by $\Sigma_i\, w_i \mathrm{f}_i\,(x)$
- Harmony is *intuitive*:
  - When we make rational decisions, we appropriately weigh *all* the evidence.
  - Classical OT is bravely counterintuitive: the decision between two rival candidates is made *solely* by the highest ranked constraint that distinguishes them.
  - The claim to be made here is: yes, brave, but empirically wrong

10. MaxEnt, Step 3

- Take every Harmony value and compute from it the corresponding **eHarmony**.[1]
- Specifically, negate the Harmony, then take *e* (about 2.7) to the result (graphed here).
    - In formula (6), eHarmony is: $\exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x))$.
    - Graphing eHarmony against Harmony:



- eHarmony performs a sort of "squishing": If Harmony gets very big, eHarmony is already close to zero and gets only slightly smaller:
- I claim that eHarmony is *intuitive*:
    - if we are at probability .5 for choosing a candidate, we welcome evidence to help decide and are seriously influenced by it (steep
    - But for a candidate already heavily penalized (e.g. .001), even a great deal of evidence may only move us to .0005.
    - Same for candidates close to one: their rivals are already penalized by a lot of Harmony and increase will only move the top candidate e.g. .999 → .9995
    - The principle: *certainty is evidentially expensive*.
    - This will matter below.

11. MaxEnt, Last step

- Sum the eHarmony for every candidate for this input, call the result Z.
- In (6), this is: $\Sigma_j\ \exp(-\Sigma_i\, w_i \mathrm{f}_i\,(x_j))$.
- The *probability* of a candidate is its eHarmony divided by Z; i.e. its share in Z
- This is also intuitive: a candidate is less likely if it has strong rivals.
- Formula (6) is now explicated in full.

TWO OTHER CONSTRAINT-BASED PROBABILISTIC FRAMEWORKS

12. Noisy Harmonic Grammar (Boersma and Pater 2016)

- Compute Harmony as in MaxEnt.
- Imagine the grammar being used on a series of "evaluation times".
- At each evaluation time, let each weight be "perturbed" by a value taken from a Gaussian distribution (normal curve).
- The winner for that evaluation time is the candidate with the lowest Harmony penalty.

---

[1] Term comes from Wilson (2014), who was joking (eHarmony is a dating website), but I like the mnemonic.

- Over multiple evaluation times, we get a probability distribution, which we can check against data.

13. Stochastic Optimality Theory (Boersma 1998, Boersma and Hayes 2001)

- Instead of weights, "ranking values".
- Again, evaluation times: perturb the ranking values with a Gaussian distribution.
- Now, sort the constraints by ranking value and proceed just as in classical OT.
- Do this over many evaluation times and you will get a probability distribution over candidates.

A FIRST QUANTITATIVE SIGNATURE: THE SIGMOID CURVE

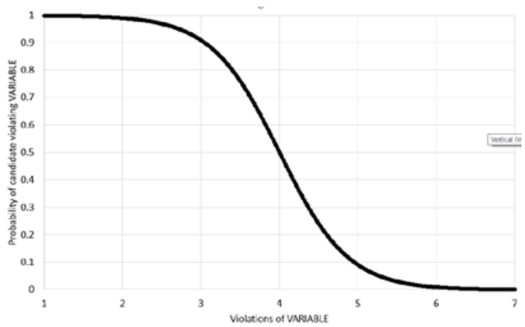
14. Step 1: How a sigmoid curve emerges in MaxEnt

- Imagine a setup with :
  - One single constraint, called ONOFF, conflicting with
  - A constraint, or set of constraints, defining a **scale**.
- Some scales:
  - A **family of assimilation triggers of varying strength —** e.g. vowels, triggering vowel harmony
  - A series of **nested levels** (varying in "cohesion") in Lexical Phonology
  - A set of **phonology-triggering affixes** that vary in their propensity-to-trigger
- Imagine a theory that takes these ingredients and computes a probability for all possible outcomes along the scale.
- Simplest case first: the scale is defined by one single constraint, VARIABLE, with multiple violation levels.

15. Concretizing a bit

- Let VARIABLE have seven values, 1-7.
- It is opposed by ONOFF.
- As throughout this talk, each input has but two viable candidates:
  - One obeys VARIABLE, violates ONOFF
  - One obeys ONOFF, violates VARIABLE
- We plot a probability function:
  - Horizontal axis: value for VARIABLE
  - Vertical axis: probability that the candidate that obeys ONOFF wins.
- We will plot for *all* values, not just the integers 1-7, since the curve emerges more clearly that way.

16. Do this in MaxEnt — you get a sigmoid

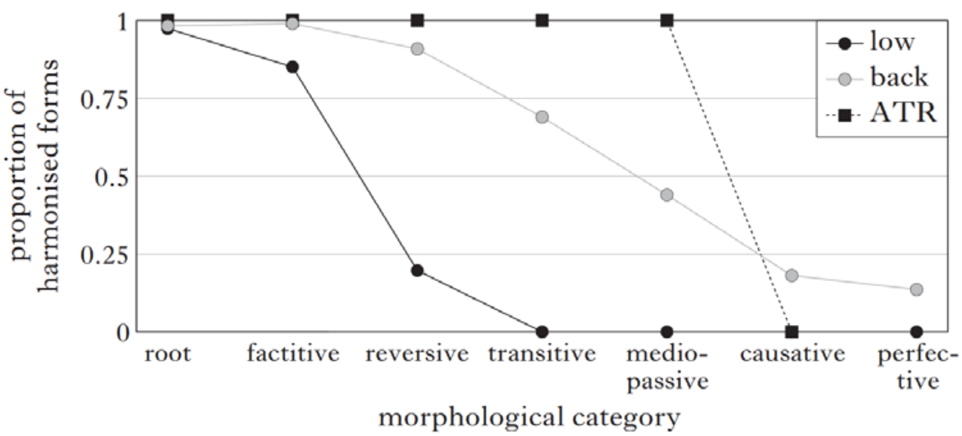


- The sigmoid asymptotes at its extremes to 1 and 0 —assuming that empirical cases exist covering enough of the horizontal axis.
- It is symmetrical about the 50% probability mark.
- For the equations that relate the shape of a maxent sigmoid to the constraint weights, see McPherson and Hayes (2016).

SIGMOIDS ARE EVERYWHERE

17. In phonology

- Rate of three process of vowel harmony in Tommo So, for all seven levels of the lexical phonology (McPherson and Hayes 2016)
    - ➢ VARIABLE = AGREE$_{\text{MorphologicalLevel}n}$(vowel feature)
    - ➢ ONOFF = FAITHFULNESS(vowel feature)



- Zuraw and Hayes (2017) point out multiple sigmoids (on which more below) for Tagalog Nasal Substitution, Hungarian Vowel Harmony, and French Liaison.

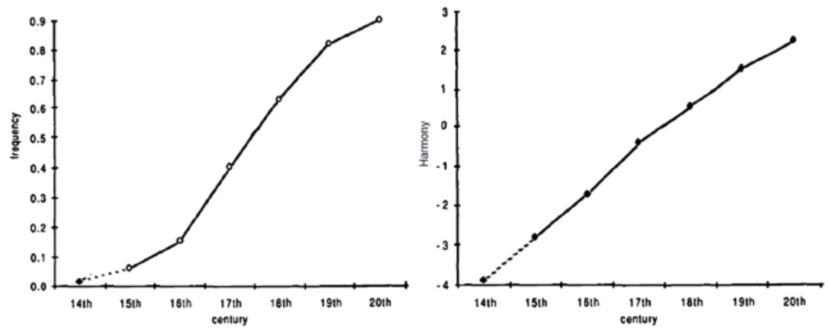
18. In speech perception

- Speech perception studies for decades have demonstrated sigmoid curves
    - ➢ Horizontal axis is a phonetic parameter, like Voice Onset Time.
    - ➢ Vertical axis probability of a percept, e.g. /p/ instead of /b/.
- Boersma (1998) suggests that speech perception works like a "backward" stochastic grammar, obtaining probabilities for input phonemes from parameters of the signal.

- Experts in speech perception are already familiar with MaxEnt models and use them, though under a different name.

19. In syntax — with a diachronic wrinkle

- Sigmoids in language change
  - ➢ The classic paper is by Kroch (1989); many followups.
  - ➢ On the left is a sigmoid for increase in use of Portuguese definite article before possessive ((*os*) *seus livros*), '(the) his books'



  - ➢ Kroch replots his data (on the right, edited) in what he calls the "logit domain" and what we will call Harmony: constant rise, 1.0 units of Harmony per century.
- Key point:
  - ➢ you can model this in maxent or NGH by supposing that the weight of constraint rises or falls at a constant rate over time
  - ➢ empirically, this produces a sigmoid in the domain of observable probabilities.[2]
- This is an oversimplification — see more below.

20. Sigmoids and quantitative signatures of the rival frameworks

- Uninteresting case: the horizontal axis is the result of a **bundle of different constraints** (like for different vowel harmony triggers).
  - ➢ Here, all of our theories (MaxEnt, Noisy Harmonic Grammar, Stochastic OT) can describe any pattern; nothing is at stake.
- The interesting case is single gradient constraint (our VARIABLE), as in Tommo So.

21. Noisy Harmonic Grammar

- Basically the same as MaxEnt, but with a complication discussed below.

22. Stochastic OT

- Cannot generate sigmoids with a single gradient constraint.
- Reason: **it is stochasticized Classical OT —**
  - ➢ Per above, Classical OT *discards evidence*

[2] This said, we still need a mechanism — presumably speakers perceive the synchronic pattern of variation *in the Harmony domain*, and mimic/exaggerate accordingly.

- ➢ Here, the evidence related to violation count (other than relative count).
- ➢ E.g. * vs. ** is not distinct in classical OT than * vs. *******.

- Full disclosure: there is a possible, still little-explored way to get sigmoids in Stochastic OT (“exploded” gradient constraints), proposed in Boersma (1998) and discussed in McPherson and Hayes (2016:fn. 21) (but not here).

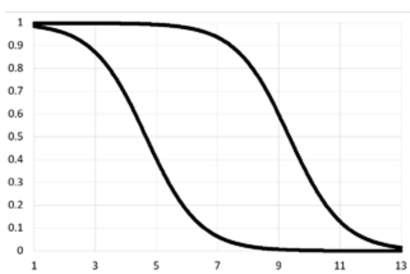
A MORE COMPLEX CASE: THE WUG-SHAPED CURVE

23. Scenario

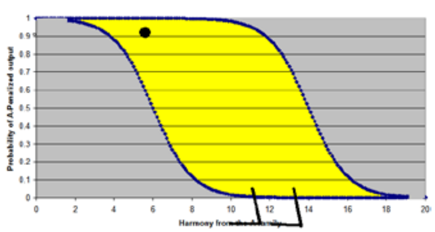
- Let us augment the primal case of (16); i.e., *one constant constraint* like ONOFF vs. *one variable constraint* (like VARIABLE) *or family*.
- Now, **double the input set**, adding a new batch of inputs identical to the first *except* that they violate PERTURBER — a constraint defined on an independent dimension.
- Example of a perturber (to be covered more below): in Hungarian, stems take front harmony more often if they end in a sibilant; hence *BACK AFTER SIBILANT.

24. Effect of perturbers in Maxent

- You can perhaps already guess: they create a **second sigmoid**, shifted over relative to the original sigmoid by a particular amount, namely the weight of PERTURBER.



- To some, these sigmoids evoke the adorable Emblematic Animal of Linguistics, and so have been called the **wug-shaped curve**.[3]



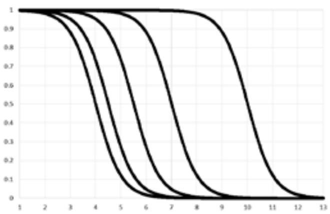
- Skinny wugs, fat wugs: the wug-shaped curve is fatter when the weight of PERTURBER is bigger.

25. Multiple perturbers?

- Nothing is stopping us, and indeed there are empirical cases (below).

[3] Thanks to Dustin Bowers for noticing this and coining the name.

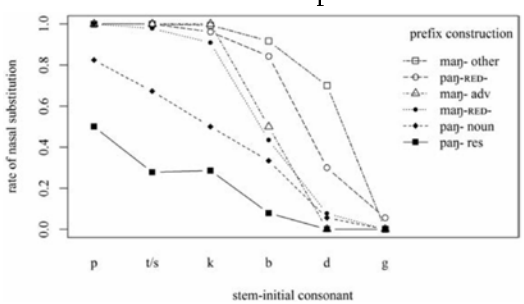
- PERTURBER1, PERTURBER2, PERTURBER3, etc. each define a separate sigmoid, according to their weights.
- The result, if you can bear this level of cuteness, might be called the **Stripey Wug**:



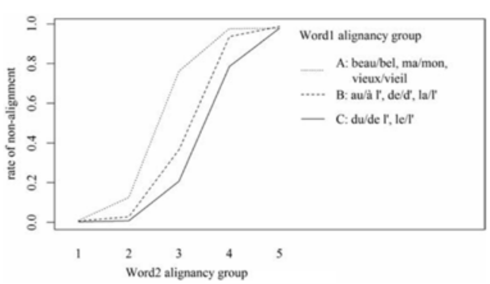
STRIPEY WUGS IN REAL LIFE

26. Zuraw and Hayes (2017)

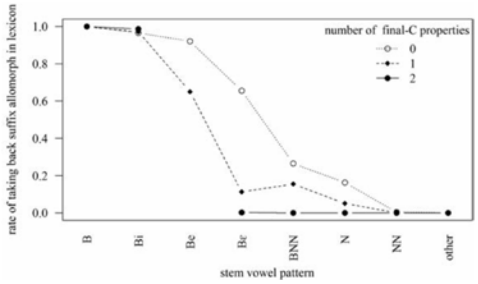
- They put forth three cases.
- Tagalog Nasal Substitution (e.g. /ŋ+p/ → [m], /ŋ+t/ → [n], etc.)
  - ➢ base constraint set: features of stem-initial consonants
  - ➢ perturber constraint set: each prefix has own propensity to induce mutation, formalized with its own perturber constraint.



- French Liaison
  - ➢ base constraint set: lexical degree of *h-aspiré-ité* (tendency to behave as if beginning with a silent consonant)
  - ➢ perturber constraint set: lexical propensity to respect h-aspiré preference of the following word



- Hungarian Vowel Harmony
  - ➢ Base constraint set: phonological environments with differing harmony probabilities
  - ➢ Perturber constraint set: stem-final consonant environments (Hayes et al. 2009)



27. Wug-shaped curves and stripey wugs in speech perception

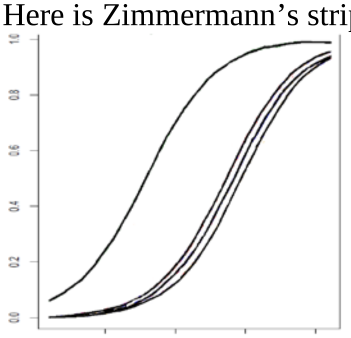
- Ubiquitous; the standard way to assess the strength of some perturbing effect.

28. Wug-shaped curves and stripey wugs in historical change[4]

- We've already covered Kroch's "constant rate" hypothesis.
- As he notes, the deeper and more meaningful aspect of the hypothesis is its application when the same basic change occurs *in a set of different contexts*.
- Kroch's theory says that the change is *constant rate when measured as Harmony*.
- So the data, plotted as probability, forms a stripey wug.

29. Richard Zimmermann's (2017) stripey wug

- English has gradually changed by shifting possessive *have* from Aux toward main verb.
- Zimmermann explored this in four contexts:
  - Negation (*I haven't any*, *I don't have any*.)
  - Inversion (*Have you a penny? Do you have a penny?*)
  - Ellipsis (*You have a flair*, *you really have/do*.)
  - Adverbs (*He has already the approval of the nation*/ … *already has*
- Each may plausibly be assumed to be affiliated with additional constraints acting as Perturbers.
- The diachronically-shifting constraint governs whether possessive *have* can be used as an Aux.
- Here is Zimmermann's stripey wug in stripped-down form:



- From left to right, the sigmoids are for adverbs, negation, inversion, ellipsis

[4] Thanks to Tony Kroch and Beatrice Santorini for help with this section; they obviously not responsible for misuse or misunderstanding of their suggestions on my part.

30. Excursus: What are the prospects for synchronic MaxEnt syntax?

- Some very nice work has already been done: Velldal and Oepen (2005), Bresnan et al. (2007), Bresnan and Hay (2008), Irvine and Dredze (2017)
- The experimental program of Featherston (2005, 2019) makes a sensational claim:
  - ➢ We can *measure* Harmony directly.
  - ➢ We just need to use Harmony-based syntax,[5] and gather the judgments using Magnitude Estimation (Bard et al. 2006).
  - ➢ I.e. each syntactic violation substract a characteristic, consistent amount on the scale, consistent with Harmonic Grammar.
  - ➢ See work of Keller (2000, 2006) for similar results.

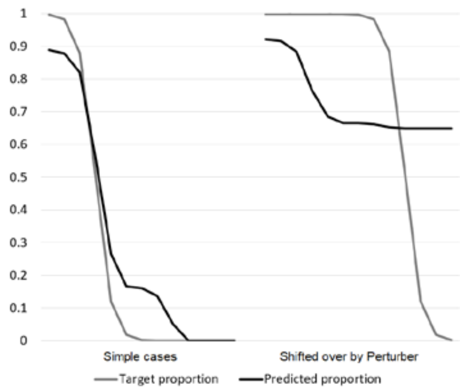
31. What about Stochastic OT: Can it derive wug-shaped curves?

- In the general case, Stochastic OT *does not provide analyses* that match wug-shaped curve data.
- For example, here is a wug-shaped curve done in MaxEnt, with 13 discrete data points (done for convenience, to ease the Stochastic OT comparison).
  - ➢ Instead of wug-format, I used two separate curves on two graphs.

  - ➢ Fit is almost perfect, so that the black "predicted" curves actually cover up the gray "to be modeled" curves.

[5] … under another name; Featherston calls it the Decathlon Model.

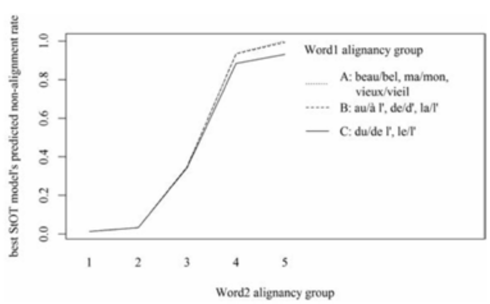
- Now, the same data modeled in Stochastic OT.
  - ➢ The curves emerge as attenuated and ill-fitted.



- Why? Intuitively, PERTURBER cannot be in two places at once; it can only struggle to model the separate sigmoids.
- This can be traced to OT's assumption that decisions are made only by the highest-ranking constraint that cares — the key assumption called into question by Zuraw and Hayes's paper ("Intersecting constraint families: An argument for Harmonic Grammar")

32. Stochastic OT and empirical instance of stripey wugs

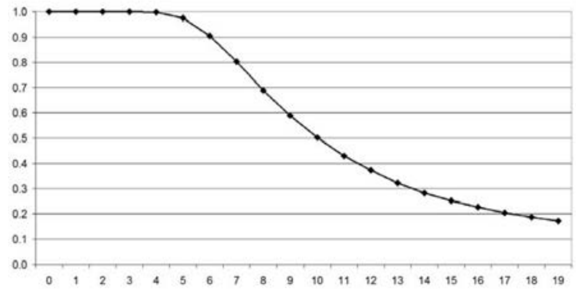
- Unsurprisingly, Stochastic OT proves to be a poor tool for analyzing the effects of intersecting constraint families. [6] E.g., here is the outcome for French (compare (26) above with MaxEnt):



33. What about Noisy Harmonic Grammar?

- There is at least one cloud on the horizon: the sigmoids it generates, in its classical version, are *asymmetrical*:

[6] The full problem for Stochastic OT is even worse than (32) implies: a Divergence Theorem proven by Giorgio Magri and reported in Zuraw and Hayes (2017:529-530) designates a broad range of cases in which Stochastic OT cannot generate anything like a wug-shaped curve.



- McPherson and Hayes (2016) show this can be pernicious; it yields slightly inferior fits to the Tommo So data.
- Yet, there are many *different versions* of NHG (Hayes 2016), and some of them generate perfectly good sigmoids (and wug-shaped curves, and stripey wugs, Zuraw and Hayes 2016).

WHERE ARE WE IN THE CHOICE OF FRAMEWORKS?

34. Stochastic OT strikes me as being in trouble

- It generates
  - ➢ Sigmoids only by fiat (hence not when one constraint embodies a scale)
  - ➢ Wug-shaped curves and stripey wugs only under special, lucky, conditions (see Zuraw and Hayes (2017) for discussion)

35. Maxent is doing fine by the data given here

- … but is under attack on other grounds, specifically overgeneration (Magri and Anttila 2019)
- Linguists will differ on the strength of overgeneration arguments; which, empirically, are the argument from silence (how well have the world's languages been checked?)

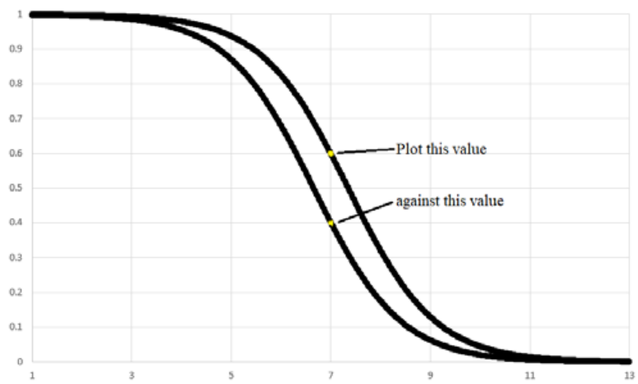
36. Noisy Harmonic Grammar is also in the running

- … particularly if we use a variant (Hayes 2016) that doesn't suffer from the asymmetrical-sigmoid issue.
- However, NHG cannot replace MaxEnt as a model of well-formedness (see ✓[kɪp]/?[pɔɪk]/*[bzɑɹʃk] in (1) above), at least if we use the probabilities-to-GEN strategy of Hayes and Wilson (2008).

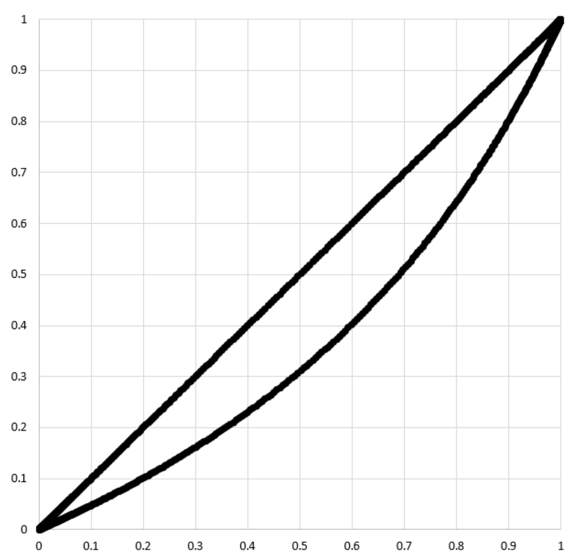
EPILOGUE: THE BANANA-SHAPED CURVE

37. This is really the same math as the wug-shaped curve, but visualized differently

- Take a wug-shaped curve and replot it, as in the following example.
  - ➢ For clarity, we start with a skinny wug:



- The lower curve represents probabilities as affected by PERTURBER constraint.
- We select all "vertical pairs" as shown (they share baseline value), and replot as scattergram, obtaining a **probability-against probability** curve.
  - I.e. comparable pairs, differing in whether PERTURBER is violated.
- In the replotting, I include the diagonal line (*y* = *x*), so we are comparing the two patterns with each other, one with aviolation of PERTURBER, one without.
- Here is the result, a **banana-shaped curve**:



- Diagonal line: non-perturbed cases, given as comparison.
- Sagging line: perturbed cases
- It could equally well have been an upward rather than downward bulge, depending on which candidates are penalized by PERTURBER.

38. Intuitive implication of the banana-shaped curve

- See (10) above, on evidential expensiveness of certainty or near-certainty
- PERTURBER has its main effects in the *medial region*, where baseline harmony level isn't already forcing the probability close to zero or one.

39. The banana-shaped curve in real life: Moore-Cantwell’s and Kush’s (2019) English stress study

- This was a “blick” test assessing people’s intuitions about stress placement in English.
- Should a CVCVCV nonce word receive *penultimate* or *antepenultimate* stress?
- Experimental method (from Guion et al. 2003): blend together three nonsense syllables.

- A startling result the authors got:
  - The subjects disagree with each other *enormously* in whether they should prefer antepenultimate or penultimate stress in general.
- Nevertheless, they show there is considerable order in their data!

40. The Perturber: Moore-Cantwell’s (2016) “Final [i]” stress constraint

- Trisyllabic words ending in [i] should have antepenultimate stress.
  - Cf. words like ˈ*burgundy*, ˈ*cavalry*, ˈ*dynasty*, ˈ*galaxy*, ˈ*majesty*, which on other grounds (Chomsky and Halle 1968 *et seq.*) “ought to” have penultimate stress.
  - Compare schwa-final words like *a*ˈ*genda*, *a*ˈ*lumna*, *bo*ˈ*nanza*, *ca*ˈ*nasta*, *la*ˈ*sagna*
- How to formalize this? Nontrivial, but Moore-Cantwell has done it; see her work (2016) for a full account.
- For present purposes, we can use the deadpan constraint “Have antepenultimate stress if [i]-final.”

41. But what is the *baseline* in Moore-Cantwell and Kush’s experiment?

- I.e., why are the participants so amazingly variegated in their baseline preference for antepenultimate stress??
- Here is a conjecture.

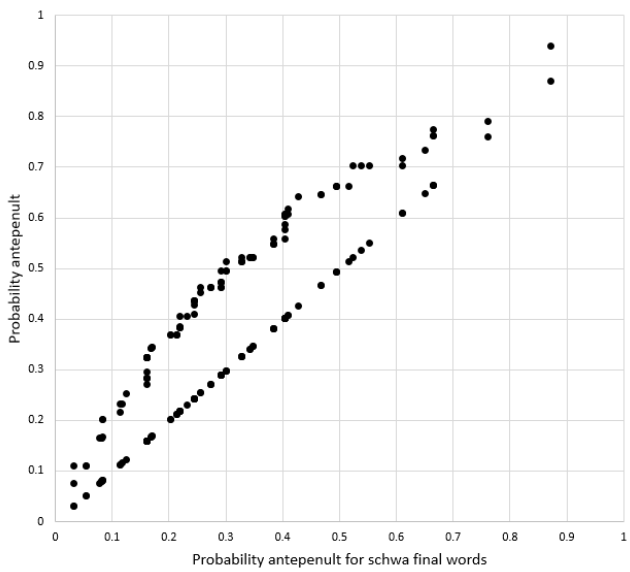
42. Vocabulary strata in English stress

- English stress has major effects of **vocabulary stratum** (cf. *SPE*, Ito and Mester 1995).
- Words perceived as **[+foreign]** (i.e. truly “exotic”, not just Latinate) tend to obey the following, perhaps Spanish-derived, stress pattern:
  - Stress the penult if the final is CV;
  - Else stress the final (i.e. if CVC).

- Foreign words often obey this rule in the speech of Anglophones, even when the result *both* violates the native-English stress norm *and* produces the wrong answer in the source language! (Janda et al. 1994)
  - Final CVC: Hebrew **Shi*ˈ*mon* **Pe*ˈ*res*, *Menachem* **Be*ˈ*gin*[7], **Ra*ˈ*bin*; Yiddish **Man*ˈ*del*, Aklan **Ak*ˈ*lan*, Spanish **Cha*ˈ*vez* (all final in many people's English, penultimate in source)
  - Final CVCV: Spanish *Sepul*ˈ*veda*, Japanese *O*ˈ*saka*, Italian *Cristo*ˈ*fori*, Hungarian *pa*ˈ*prika* (penultimate in many people's English, antepenultimate in source)
- Moore-Cantwell and Kush left it up to the participants whether to regard the experimental words as foreign or native, and this perhaps was the source of the massive variation.
  - We might learn more by manipulating frame sentences.[8]

43. The banana-shaped curve in Moore-Cantwell and Kush (2019)

- The diagonal set of points simply encodes the cases where one or more participants assigned the probability (on either axis) of antepenultimate stress to the [ə]-final stimuli.
- Aligned above this point: the mean value (same subset of participants) assigned to the [i] final stimuli.



- The bulge is upward, since in this case Perturber constraint *favors* antepenultimate stress.
- For rigor, you can check the maxent math: take *logs* of probabilities, then see if the regression equation $y = x + b$ fits ok; r = .981.

[7] The public got this right in the end; Janda et al. report the earlier stages of hyperforeignization.

[8] "We sang the fine old English folk song [ˈmæʃəbi/məˈʃæbi]"; "Hyman served up delightful steaming plates of [ˈdɛləsə/dəˈlɛsə]".

SUMMING UP

44. Theme

- We contrive to use a little bit of math fill the gap between abstract principles and empirical work.
- The math gives us general quantitative signatures — visible to the eye when plotted.
- We can use the presence of signature-like data as a way to evaluate the theories.

45. Who is winning?

- Especially if we consider the Zuraw/Hayes cases: the winner is either form of stochastic Harmonic Grammar (MaxEnt, NHG).
    - They match the wug, stripey wug, and diagonal banana signatures.
- Stochastic OT (historically, a great way to lead phonology into the domain of quantitative modeling) seems not to be holding up under this kind of scrutiny.

46. Further work

- Sorting these issues out …
- I am astonished by Featherston's claim that Harmony can be directly measured, and would love to see this checked in the domain of phonology.
- I would also love to see work extending the Krochian diachronic-syntax tradition — itself essentially founded in MaxEnt — to synchronic syntax. The way is open:
    - existing MaxEnt syntax work by Bresnan et al. (2007) and others
    - the Featherstonian research paradigm, if valid, as a way of finding good data

References

Bard, Ellen Gurman, Dan Robertson and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. Language 72:32-68.

Boersma, Paul and Joe Pater (2016). Convergence properties of a gradual learning algorithm for Harmonic Grammar. In J. McCarthy and J. Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press

Boersma, Paul. 1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. Doctoral dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, ed. by G. Boume, I. Krämer, and J. Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86.168–213.

Featherston, Sam. 2005. The decathlon model of empirical syntax. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, ed. by Stephan Kepser and Marga Reis, pp. 187–208.

Featherston, Sam. 2019. The Decathlon Model. *Current Approaches to Syntax: A Comparative Handbook*, ed. by Andras Kertesz, Edith Moravcsik, and Csilla Rakosi, pp. 155–186.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. *Proceedings of the workshop on variation within optimality theory, Stockholm University*, 2003.

Susan G. Guion, J.J. Clark, Tetsuo Harada, and Ratree P. Wayland. (2003) Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech*, 46(4):403-427.

Hayes, Bruce (2017) Varieties of Noisy Harmonic Grammar. In Karen Jesney, Charlie O'Hara, Caitlin Smith and Rachel Walker (eds.), Proceedings of AMP 2016.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.

Hayes, Bruce, Kie Zuraw, Peter Siptár and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822-863.

Itô, Junko & Armin Mester. 1995. Japanese phonology. In *The Handbook of Phonological Theory*, ed. by John Goldsmith, 817–838. Oxford: Blackwell.

Janda, Richard D., Brian D. Joseph, and Neil G. Jacobs (1994) Systematic hyperforeignisms as maximally external evidence for linguistic rules. In Susan D. Lima, Roberta Corrigan and Gregory Iverson, eds., *The Reality of Linguistic Rules*. Amsterdam: John Benjamins.

Jurafsky, Dan and James H. Martin (2019) *Speech and Language Processing* (3rd ed. draft), web.stanford.edu/~jurafsky/slp3/

Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Ph.D. dissertation, University of Edinburgh.

Keller, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. *Gradience in grammar: Generative perspectives*, ed. by Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky, and Ralf Vogel, 270-287.

Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change* **1**. 199–244.

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41, no. 5: 1202-1241.

Magri, Giorgio, Scott Borgeson, and Arto Anttila (2019) Equiprobable mappings in weighted constraint grammars. *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*

McPherson, Laura and Bruce Hayes (2016) Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology* 33: 125–167

Moore-Cantwell, Claire. 2016. The representation of probabilistic phonological patterns: neurological, behavioral, and computational evidence from the English stress system. Ph.D. dissertation, University of Massachusetts, Amherst.

Moore-Cantwell, Claire and Dave Kush (2019) Cognitive load impairs access to the phonological grammar. Lecture given at the UCLA/USC Joint Seminar in Phonology, Fall meeting.

Prince, Alan & Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell]

Smolensky, Paul (1986) Information processing in dynamical systems: Foundations of Harmony Theory. In James L. McClelland, David E. Rumelhart and the PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models.* Cambridge: MIT Press. 390-431.

Velldal, Erik & Oepen, Stephan. 2005. Maximum entropy models for realization ranking. Proceedings of the 10th Machine Translation Summit, ed. by Jun-ichi Tsujii. Asia-Pacific Association for Machine Translation.

White, James (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*, 93(1), 1–36.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30.945–982.

Wilson, Colin (2014) Tutorial on Maximum Entropy models. Lecture given at the Annual Meeting on Phonology, Massachusetts Institute of Technology, Cambridge, MA, September 19.

Zimmermann, Richard (2017) Formal and quantitative approaches to the study of syntactic change: Three case studies from the history of English. Ph.D. dissertation, University of Geneva.

Zuraw, Kie and Bruce Hayes (2017) (2017) Intersecting constraint families: an argument for Harmonic Grammar. *Language* 93:497-548.