Bruce Hayes
University of California, Los Angeles

Talk at Stanford University
September 22, 2018

Some remarks on MaxEnt grammars

Prepared for the workshop "Analyzing Typological Structure: From Categorical to Probabilistic Phonology"

1. Topics

- MaxEnt and its match with commonsense understanding
- Probability and learning in generative linguistics
- MaxEnt and restrictiveness

PART I: MAXENT AND COMMONSENSE UNDERSTANDING

2. Theme

- Everything in the maxent framework shows a close correspondence with common sense.

3. MaxEnt: background

- 19th century physics
- Smolensky (1986) — in connectionist cognitive science
- Goldwater and Johnson (2003) — as framework for constraint-based linguistics
- Ernestus and Baayen (2003), Wilson (2006) — first applications to new language data
- Quite a few papers since; see Hayes (in progress) for more bibliography

4. Maxent OT

- In its linguistics-avatar, MaxEnt is a version of Optimality Theory (Prince and Smolensky 1993).
- It inherits some virtues of OT, e.g.
  - Reduction of linguistic analysis to simple formal ingredients
  - Explains variation as the result of conflicting priorities (Anttila 1997ab, Boersma 1998)
  - Makes possible a serious effort to solve the problem of language learning, using computational modeling (Tesar and Smolensky 1993, followed by Anttila, Boersma, Magri, Pater, Jarosz, etc. etc.)

- **MaxEnt: specifics**

  - Among OT-descendents, MaxEnt is a form of Harmonic Grammar (Legendre et al. 1990, Legendre et al. 2006, Potts et al. 2010, Boersma and Pater 2016), because it uses constraint weights rather than ranking.
  - It has two parts:
    - A formula ((5) below) that inputs tableaux and generates probabilities for candidates.
    - A scheme for learning constraint weights from data, backed by proof.

5. The MaxEnt formula deriving probabilities from tableaux

$$\Pr(x) = \frac{\exp(-\sum_i w_i f_i(x))}{Z}, \text{ where } Z = \sum_j \exp(-\sum_i w_i f_i(x_j))$$

6. Explication of the formula

Compute this	*Name of what is computed*	*How it is computed*
a. $\sum_i w_i f_i(x)$	Harmony (Smolensky 1986)	Multiply x's violation counts for each constraint (designated $\mathbf{f_i(x)}$) by the weight of the constraint ($\boldsymbol{w_i}$), then add up the results across all constraints ($\boldsymbol{\sum_i}$).
b. $\exp(-\sum_i w_i f_i(x))$	eHarmony (Wilson 2014)[1]	Negate the harmony of x and then compute the function **exp( )** on the result, where exp(x) is a typographic convenience for e^x, $e \approx 2.72$. [2]
c. $\sum_j \exp(-\sum_i w_i f_i(x_j))$	Z, the "normalizing constant"	Compute the eHarmony of every candidate and sum these values.
d. $\frac{\exp(-\sum_i w_i f_i(x))}{Z}$	Probability of x	Divide the eHarmony of x by Z (and similarly for all other candidates).

7. Here is the common sense basis of MaxEnt

… stage by stage

[1] Wilson was joking in inventing this name (which also denotes a dating web site), but I feel it is quite helpful as a mnemonic.

[2] In some implementations of MaxEnt the negation step is skipped by making all weights negative in the first place. This article follows the practice of Wilson (2006), with positive weights. The difference is purely notational.

8. w_i

- This is the weight of the *i*th constraint.
- It reflects the notion that *different factors differ in their importance* in establishing a conclusion.
  - OT does this with ranking.

9. Harmony: $\Sigma_i w_i f_i(x)$ **(Step (6a))**

- Harmony is the weighted sum of constraint violations.
- It results (unavoidably) in **ganging**.

10. Ganging (Jäger and Rosenbach 2006, etc.)

- Two kinds:
  - Two constraints can gang up to overcome a third that is stronger than either one of them alone.
  - Several violations of a weaker constraint can count more than fewer violations of a stronger one.
- Compare OT, where all is decided by the highest ranking constraint that distinguishes two candidates.

11. Does ganging reflect common sense?

- It is commonsense to make decisions based on a suitable weighing of *all the available evidence*.
- OT rather bravely goes against this, relying solely on the most important relevant criterion and discarding all others.

12. Is ganging empirically motivated?

- One clue is that classical OT analysis repeatedly has had to resort to **constraint conjunction** (Smolensky 1995), which overrides the "highest ranked constraint decides" character of the theory.
- Zuraw and Hayes's (2017) study:
  - They take on "intersecting constraint families" — seeking cases where there is maximum *opportunity* for constraints to gang.
  - They *do* gang — across the board, it would seem.
  - As a result, frameworks that have only "incidental" ganging under special circumstances (Anttila 1997, Boersma 1998, Magri 2012) cannot handle the Zuraw/Hayes data.

13. Exponentiation: $\exp(-\Sigma_i w_i f_i(x))$ (Step (6b))

- Harmony is exponentiated, forming eHarmony, on its way to being converted to probability. Why so?
- I believe this too matches common sense: *much evidence is required to approach certainty*.
- Example:
  - to shift the predicted probability of a candidate downward from **50%** to **49.001%**, one needs to assign it only **0.040** units of additional harmony
  - to shift the predicted probability of a candidate downward from **1%** to **0.001%** requires **6.92** units.

14. Implicit comparison with alternatives: the role of Z (Steps (6c,d))

- To obtain probability, we divide the eHarmony of a candidate by Z, the sum of the eHarmonies of all the candidates.
- Common sense principle: a candidate should get a lower probability if it competes with strong alternatives.

15. Summing up

- *Every* element of (5) ($w_i f_i(x)$, $-\Sigma_i$, exp(), $\frac{1}{Z}$) shows qualitative correspondence with common sense.
  - Differing weights (credibility) for differing sources of evidence.
  - *All* the evidence is weighed.
  - Requirement of more evidence to approach certainty.
  - Assignment of lower probability in the presence of strong alternatives.

PART II: PROBABILITY THEORY AND GENERATIVE LINGUISTICS

16. My recent reading activity

- Probability theory
  - This topic really bored me a lot when I was a college freshman!
  - I hated the dice, the urns, the colored balls …
- Perhaps I was learning the wrong stuff? Now probability theory seems meaningful, possibly even deep …
- I am especially taken with Edwin Jaynes's widely-cited book *Probability Theory: The Logic of Science* (2003), which I have found difficult but also stimulating and inspirational.

17. What is the subject matter of probability theory?

- There are different views, both dating back to Laplace (early 19th cen.)
  - The "study of random events"
  - The study of *the relationship of rational belief to evidence*.

- The latter is the view advocated by Jaynes.
- For him, the 0 to 1 scale of probability theory is a *scale measuring justifiable belief in the face of limited evidence*.

18. Jaynes's vision

- We start with very simple, completely uncontroversial instances of common sense.[3]
- We use them (plus lots of calculus) to *prove* the postulates of probability theory.[4]
  - This is **Cox's Theorem** (Cox 1946, deepened by Jaynes 1957, 2003).
- The postulates of probability theory are then used to prove a vast edifice of theorems.[5]
- We then use the theorems to carry out inductive reasoning on a vast scale — **assisted induction**, which is:
  - computationally-implemented
  - extracts far more meaning from complex data than we could ever do with our unassisted common sense.
  - … and completely rigorous and trustable, because backed by proof.
- Jaynes died in 1998 but *his dream is being realized right now all over the world*.
  - See McGrayne (2011), an excellent popularizing book covering this development.

19. What about generative linguistics?

- Lots of linguistics, myself included, endorse the Chomskyan view that explaining the quasi-miracle of language acquisition is the key task in theoretical linguistics.
- Hence the development of computational models of language learning based on OT and its progeny (see (3) above) seems to me a very encouraging development.
- For all but the most hard-line innatists, language acquisition is a clear case of how we
  - *justify a belief* ("my grammar has this…")
  - … *in the face of limited evidence* (*n* short years of childhood)
- In other words, language acquisition, our key theoretical issue, may be a problem best addressed by probability theory.

---

[3] One example (Jaynes p. 30) is that as the plausibility of A goes up, the plausibility of *not* A goes down; there are others. See Smith and Erickson (1989) for a perhaps clearer presentation of these than in Jaynes.

[4] These are the Product Rule, $P(AB|C) = P(A|BC)P(B|C)$; and the Sum Rule, $P(A|B) + P(\bar{A}|B) = 1$.

[5] Jaynes p. 51: "Essentially all of conventional probability theory as currently taught, plus many important results that are often thought to lie beyond the domain of probability theory, can be derived from [this] foundation."

20. In practical terms …

- Generative linguistics stands out among the sciences in insisting on inventing all of its own tools. Should this continue?
- … or might there be great stuff sitting out there waiting for us to use it?

21. Some importation of probability theory into linguistics already taking place

- Bayesian reasoning (work of Josh Tenenbaum, Sharon Goldwater, others)
- MaxEnt:
  - After Boersma promulgated his Stochastic OT in the late 1990's, several computer scientists independently read it and felt impelled to suggest that we try out MaxEnt instead.[6]
  - They pointed out MaxEnt's good properties (solid mathematical foundations, principled basis, modeling accuracy, learning algorithm backed by proof)
  - Maxent started to be used in phonological research; Ernestus and Baayen (2003) and Wilson (2006)

22. Does MaxEnt embody common sense?

- See Jaynes (2003:ch. 11), where he carries out for MaxEnt the same strategy he adopts for probability theory as a whole.
- Key idea: given some facts, and a theory to analyze them, *make no commitments other than those justified by your facts.*
  - The MaxEnt analysis is the "most honest" (Jaynes) in light of limited information.
  - When the math is done, the analysis settles down to a status that does maximal justice to the learning data.
- Indeed I am often stunned by the ability of MaxEnt software to match data with extreme precision.

PART III: MAXENT AND RESTRICTIVENESS

23. This is a topic of great current interest

- … and can be attacked with mathematical sophistication; Anttila and Magri (2018).

24. Is restrictiveness overrated as a criterion for theory-evaluation in linguistics ?

- I used to be whole-heartedly committed to this criterion.
  - I spent much of my career (1978-1993) trying to invent the most restrictive theory I could for metrical stress (Hayes 1995).
- Since then, I've become much more skeptical about how informative restrictiveness is as a criterion for evaluating theories.

[6] I have located four such papers: Eisner (2000), Johnson (2002), Manning (2003), and Goldwater and Johnson (2003).

25. What are the scientific reasons for pursuing restrictiveness in linguistic theorizing?

- Predicting what is out there in the world (typology).
- Explaining how language is learnable.

26. I. Explaining what is (not) out there in the world

- It has long been pointed out that the formal theory of language structure is not the only game in town for explaining gaps in the typological pattern.
- Diachrony can play a role, too: things are missing from the world's languages because they have **no possible diachronic origin**.
  - See Myers (2002), Blevins (2004), Moreton (2008)
  - Kiparsky (2008) gives a nuanced account in which the role of UG in constraining diachrony plays a role.
- In any event, UG explanations of typology are not cost-free; our successors will have to explain how the principles of UG arose in natural selection.

27. II. Restrictiveness and explaining language acquisition

- We should ponder: *why* does a restrictive theory help explain language acquisition?
- An explicit theory of learning involves:
  - a hypothesis space
  - a means of exploring this hypothesis space as guided by the input data
- So: "*restrictiveness is good because it shrinks the hypothesis space that the learner needs to explore*".

28. This is a weak argument

- My source in saying this: Tesar and Smolensky (2000)
- Sometimes, *larger* hypothesis spaces, such as the factorial typologies of OT, may be more easily searched than *smaller* ones, such as parametric theories.
  - This is because searchability itself is related to the structure of the UG theory, as Tesar and Smolensky show for OT.

29. A not-obviously-wrong research path for linguistics

- Consider theories of UG that are *not* especially restrictive but highly amenable to search.
- Perhaps work in progress will find the right combination of restrictiveness and effective searching to explain acquisition.

30. The feebleness of our data with respect to restrictiveness arguments

- Our knowledge of the detailed phonology of the world's languages is seriously limited.
- Given that phonological theory is conducted using a small data sample, it is no surprise that counterexamples to restrictiveness claims frequently appear.

31. Wonderful creatures appear on our doorstep, surprising us

- **Kager-Hamilton paradox**: "Reduplication processes never truncate the base to match the template" (McCarthy & Prince 1997, crediting Kager and Hamilton).
  - Schematically: [badupi] → [baba].
  - But Caballero (2006) found a real-life case in Guarijío: e.g. /RED_{CV}+muhiba/ → [mumu] 'to start throwing'
- **Myopia** (Wilson 2006 et seq.)
  - Harmony processes putatively cannot "block themselves" if they can see a problem lying far away in the input string.
  - A clear example of myopia in Tutrugbu has been presented in McCollum and Essegbey (2018). [7]
- **Majority rules** (Lombardi 1999)
  - Phonological processes are held not to be allowed to count violations and adopt the preference of the majority.
  - An evident counterexample from Warlpiri vowel harmony is given by Bowler (2017).
- **"Phonology is subregular"** (Heinz 2011)
  - In fact, the right conclusion is probably that phonology is not even regular.
  - In many languages reduplication is not a morphological process but a *phonotactic principle* (see Blust 2004 on Austronesian) — hence the phonotactic system must know about the trans-regular principle of copying.
  - E.g., in Balinese (Barber 1977, 1979) the CC cluster of a CVCCVC root is usually homorganic NC, but may take on a great number of additional possibilities *only if* the two CVC sequences are copies of each other. [taptap] is ok, but *[meptik] is bad. *[tap] does not exist, and similarly for all other cases.

32. So what is the right stance with regard to restrictiveness?

- Basically, to keep doing what we're doing, but maybe not so naively.
- I'm really intrigued by predictions like Kager-Hamilton, Myopia, No Majority Rules, Subregularity.
  - Even if they proved to be wrong, they were worth checking.
- But restrictiveness arguments in linguistic theory should be regarded with great skepticism.
  - They rely on an *absence* of cases in a small, understudied data world.

33. Is there anything we can do to improve the rigor of restrictiveness arguments?

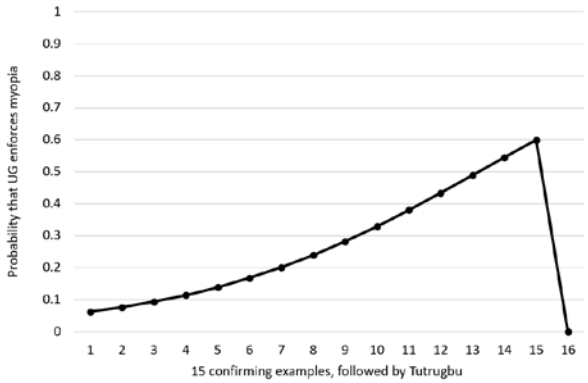
- Perhaps the answer lies in probability theory.
- I suggest to anyone proposing a restrictiveness argument that they compute a (very rough) estimate of the probability that the restrictness hypothesis is correct, based on the data we have.

[7] Similar if slightly less clear cases have been presented by Walker (2010) and Stanton (2017).

- Perhaps Bayes' Theorem would help.

34. Estimating the probability that UG enforces myopia: a very rough attempt

- A priori estimate for the probability that natural selection has created humans whose UG enforces myopia. *BH: .05?*
- A survey of *n* languages whose phonology gave them the choice of enforcing or not enforcing myopia. *BH: n = 15?*
- An estimate of the probability that myopia is expected for diachronic reasons (harmony emerging from coarticulation). *BH: .8?*
- From this, we could at least obtain a reasonable guess about whether the known typological data support the restrictive hypothesis (that UG enforces myopia).
- For the above assumptions, *p*(Myopia-UG) rises to **0.6** as we encounter more confirming cases, then drops to zero when we encounter Tutrugbu.



- This is obviously primitive, but I think following this path might increase the rigor of restrictiveness arguments.

References

Anttila, Arto. 1997a. Variation in Finnish phonology and morphology. Doctoral dissertation, Stanford University, Stanford, Calif.

Anttila, Arto. 1997b. Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. Frans Hinskens, Roeland van Hout, and Leo Wetzels. Amsterdam: John Benjamins. Rutgers Optimality Archive ROA-63, http://ruccs.rutgers.edu/roa.html.

Anttila, Arto and Giorgio Magri (2018) Does MaxEnt overgenerate? Implicational universals in Maximum Entropy Grammar. *Proceedings of AMP 2017.*

Barber, C. C. (1977) *A grammar of the Balinese language.* Aberdeen: Aberdeen University Library.

Barber, C. C. (1979) *Dictionary of Balinese*-English. Aberdeen: no publisher given.

Blevins, Juliette (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blust, Robert (2009) *The Austronesian Languages*. Canberra: Pacific Linguistics.

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

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

Bowler, Margit (2013) Majority rules effects in Warlpiri vowel harmony. Ms. UCLA.

Caballero, Gabriela (2006) "Templatic backcopying" in Guarijio abbreviated reduplication. *Morphology* 16: 273–289.

Cox, R. T. (1946) Probability, frequency, and reasonable expectation. *American Journal of Physics* 14:1–13.

Eisner, Jason (2000) Review of Kager: "Optimality Theory". *Computational Linguistics* 26:286–290.

Ernestus, Miriam and Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* **79**. 5-38.

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.

Hayes, Bruce (1995). *Metrical stress theory: principles and case studies*. Chicago: University of Chicago Press.

Hayes, Bruce (in progress) Maxent grammars: rationale and spreadsheet tutorial. Ms., UCU

Heinz, Jeff (2011) Computational Phonology – Part I: Foundations. *Language and Linguistics Compass* 5/4:140–152.

Jäger, Gerhard, and Anette Rosenbach. 2006. The winner takes it all—almost. Linguistics 44:937–971.

Jaynes, Edwin T. (1957), 'How does the brain do plausible reasoning?', *Stanford University Microwave Laboratory Report* 421. Reprinted in Erickson, G. J. & Smith, C. R. (1988), eds., *Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1, Foundations; Vol. 2, Applications*, Kluwer Academic Publishers, Dordrecht, Holland.

Jaynes, Edwin T. (2003) *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Johnson, Mark (2002) Optimality-theoretic Lexical Functional Grammar. In Paula Merlo and Susan Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. 59–74. John Benjamins, Amsterdam, The Netherlands.

Kiparsky, Paul (2008) Universals constrain change, change results in typological generalizations. In Jeff Good, (ed.) *Linguistic universals and language change*, OUP 2008.

Laplace, P. S. (1819), *Essai Philosophique sur les Probabilités*, Courcier Imprimeur, Paris.

Legendre, Geraldine, Yoshiro Miyata & Paul Smolensky. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006. The Optimality Theory - Harmonic Grammar connection. In Smolensky and Legendre 2006, 339–402.

Lombardi, Linda (1999). Positional faithfulness and voicing assimilation in optimality theory. *Natural Language and Linguistic Theory* 17:267-302.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29:213–269.

McCarthy, John and Alan S. Prince (1997) Faithfulness and Identity in Prosodic Morphology. ROA 216.

McGrayne, Sharon (2011) *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven: Yale University Press.

McCollum, Adam G. and James Essegbey (2018) Unbounded harmony is not always myopic: Evidence from Tutrugbu. In Wm. G. Bennett, Lindsay Hracs, and Dennis Ryan Storoshenko (ed.) *Proceedings of the 35th West Coast Conference on Formal Linguistics*, pp. 251-258.

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

Moreton, Elliott. (2008). Analytic bias and phonological typology. *Phonology* 25:83-127.

Myers, Scott (2002) Gaps in factorial typology: The case of voicing in consonant clusters. ROA 509.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker (2010). Harmonic Grammar with linear programming: From linear systems to linguistic typology. *Phonology* 27, 77-117.

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]

Smith, C. Ray and Gary Erickson (1989) From rationality and consistency to Bayesian probability. In J. Skilling, ed., *Maximum entropy and Bayesian methods*. Dordrecht: Kluwer, pp. 29-44.

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.

Stanton, Juliet. Gurindji nasal cluster effects as trigger deletion. Ms., NYU.

Tesar, Bruce, and Paul Smolensky. 1993. The learnability of Optimality Theory: An algorithm and some basic complexity results. Ms. Department of Computer Science and Institute of Cognitive Science, University of Colorado at Boulder. Rutgers Optimality Archive ROA-2, http://ruccs.rutgers.edu/roa.html.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.

Walker, Rachel (2010). Nonmyopic harmony and the nature of derivations. *Linguistic Inquiry* 41:169-179.

Wilson, Colin (2006) Unbounded spreading is myopic. Paper presented at the workshop on Current Perspectives on Phonology. Indiana University, Bloomington, June 23, 2006.

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

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