Deriving the Wug-shaped curve: A criterion for assessing formal theories of linguistic variation*

Bruce Hayes
UCLA

March 2021

Submitted to *Annual Reviews of Linguistics*

Abstract

I assess a variety of constraint-based formal frameworks that can treat gradient phenomena, such as well-formedness intuitions, outputs in free variation, and lexical frequency matching. The idea behind this assessment is that data in gradient linguistics fall into natural mathematical patterns, which I will call **quantitative signatures**. The key signatures treated here are the **sigmoid curve**, going from zero to one probability, and the "**wug shaped curve**," which combines two or more sigmoids. I argue that these signatures appear repeatedly in linguistics, adducing examples from phonology, syntax, semantics, sociolinguistics, phonetics, and language change. I suggest that the ability to generate these signatures is a trait that can help us choose between rival frameworks.

*I would like to thank [names withheld during review] for helpful input and comments on this project.

1. Introduction: probabilistic phenomena in linguistics

This article addresses linguistic phenomena in which we need to characterize **gradience**, usually using probability, in the analysis. I offer examples from several fields of linguistics: phonology, syntax, phonetics, historical change, and semantics.

I suggest there are three phenomena that ought to be addressed by linguistic theory using probability. First, we frequently need to model cases where **alternative surface forms** are generated, at varying probabilities, from the same underlying form. This is a key empirical focus of sociolinguistics (§6.2); but is also an important topic in theoretical phonology (§6, §6.1) as well as in some approaches to syntax (§6.4).

Second, speakers of a language have the ability to **frequency-match** statistical patterns in the lexicon. For phonology, this result emerged in experimental work by Zuraw (2000) and Ernestus and Baayen (2003); and has been generally confirmed in studies since the. For instance, when Hungarian speakers undertake a nonce-probe task testing their intuitions about vowel harmony, their responses statistically match the harmony pattern of the Hungarian lexicon (Hayes and Zuraw 2017). In syntax, speakers statistically track the selectional properties of verbs (*suspect* prefers S over NP; *remember* prefers NP over S), and they use this information during sentence perception (Jurafsky 2003, Linzen et al. 2016).

Third, **native speaker judgments** are characteristically gradient and can be modeled probabilistically. These include phonological well-formedness judgments (Scholes 1965) and grammaticality judgments in syntax (see Lau et al. 2017 for a recent overview and proposal).

To treat these three kinds of gradience within generative grammar, we need frameworks that can generate outputs on a probability scale. As a reference point, against which alternatives will later be compared, I will cover **Maximum Entropy Harmonic Grammar** (Goldwater and Johnson 2003, Wilson 2006, Jäger 2007, Hayes and Wilson 2008) — for short, “MaxEnt” — which is a probabilistic version of Optimality Theory (Prince and Smolensky 1993/2004). The apparatus in MaxEnt that assigns probabilities is identical to the well-known statistical procedure of **logistic regression** (Jurafsky and Martin 2020); and I will alternatingly use “MaxEnt” and “logistic regression” below to refer to the same mathematics, depending on context.

Here I will evaluate MaxEnt against alternative formal conceptions of probabilistic linguistics. The strategy adopted is to examine characteristic quantitative behaviors of the frameworks. Specifically, we use simple math to locate quantitative patterns characteristically generated under each theory, patterns identifiable visually when we plot them on a graph. I will call such patterns **quantitative signatures**. My work follows up on earlier studies of this kind (Jesney 2007, Zuraw and Hayes 2017, Hayes 2017, Smith and Pater 2020). I seek to extend this work by offering a way to visualize the signatures that I believe is informative; and to apply this method to all areas of the grammar.

The discussion covers two related signatures. For each, I will describe the pattern, cite real-world cases, and demonstrate mathematically which frameworks possess these signatures; this in turn is taken to reflect on the empirical adequacy of these frameworks.

In pursuing this inquiry I ended up examining about 20 different cases in various fields. This brief paper cannot accommodate them all; yet I feel that rigor compels me to report them, to avoid cherry-picking. To this end, I have created a web site, the *Gallery of Wug-Shaped Curves*, (linguistics.ucla.edu/people/hayes/GalleryOfWugShapedCurves/). For each case, the site includes an illustrative graph illustrating the relevant quantitative signature, as well as the spreadsheet calculations that generated. There are also a few of brief essays covering points that would not fit into this article.

2. MaxEnt

I begin with an exposition of MaxEnt. This will be more than an overview, because developing a close *intuitive* understanding of MaxEnt helps with the task of assessing quantitative signatures, hence theory-comparison.

2.1 As a species of Optimality Theory

In linguistics, MaxEnt is a version of Optimality Theory (OT; Prince and Smolensky 1993/2004). In OT, one analyzes a language system using a set of **inputs**, sets of candidate **outputs** for each input, and a set of **constraints** used to choose from among candidates. The theory derives outputs not with a serial derivation, but by defining in advance the set of all possible outputs (GEN), and employing a metric (EVAL) that selects the best one. The task of finding the optimum is dealt with separately; as in Eisner (1997), Riggle (2004).

The metric used for candidate selection is thus: constraints are strictly ranked, part of the language-specific grammar. Between any pair of candidates for a given input, the decision is made by the highest-ranking constraint that prefers (assigns fewer violations to) one of them. Similar decisions made across the whole candidate set determine a unique overall winner, which is the output of the grammar.

In probabilistic versions of OT, selection of a unique winner is replaced by assignment of probability to every member of GEN. In some systems, such as MaxEnt, every candidate receives a positive probability, but typically the vast majority of probabilities are so vanishingly low as to be the equivalent of zero. In variable phenomena, often more than one candidate receives non-negligible probability, serving as an account of gradient free variation or preference. The probabilities are precise numbers and can be tested against quantitative data from corpora or experiments.

3. The MaxEnt math and its intuitive rationale

MaxEnt replaces the strict-winner selection system of classical OT with the mathematics of logistic regression, with constraint violations taking the role of predictors. The purpose of this section is to take apart this math, step by step, and show the each step is intuitive and sensible. This will help later as we examine how the math behaves in language examples. More generally, I hope to portray MaxEnt as a *mathematicized embodiment of common sense.*

The key will be is to think of MaxEnt as a decision procedure. The constraint violations are, in essence, *evidence* bearing on which candidates should be assigned high or low probability.

We start by looking at the whole formula, given in (1).[1]

(1) *The MaxEnt formula*

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

The formula calculates **Pr(*x*)**, the probability of candidate *x* for some input. The information in the formula includes everything needed to calculate this probability, including the set of output candidates, the constraints, a violation count for each constraint/candidate pair — and one other item, the weights, which form part of the grammar. We will now reconstruct the formula in stages, starting from its smallest parts.

3.1 Constraint weights

In a MaxEnt grammar, every constraint bears 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.[2] In (1), this is $\boldsymbol{w_i}$ for each constraint *i*. Weights in MaxEnt take on the role played by constraint ranking in classical OT. Assigning weights to constraints is intuitive, because reasons differ in cogency.

3.2 Multiple violations

In (1), we twice see the expression $\boldsymbol{w_i f_i(x)}$, where *x* is the candidate being evaluated, $f_i(x)$ is the number of violations that candidate incurs for the *i*th constraint, and w_i is the weight of the *i*th constraint. Thus, weights are multiplied by violation counts. This is intuitive in the sense that two violations are plausibly "twice the evidence" of one. As we will see below, other approaches adopt a different view.

3.3 Harmony

Once we have carried out the multiplication step for each combination of candidate and constraint, we calculate a sum, going across all of the constraints, for one single candidate. This sums acts as an aggregate penalty score for the candidate, and it is often called the **Harmony** of a candidate (Smolensky 1986).[3] In (1), Harmony is represented by $\boldsymbol{\Sigma_i\, w_i \mathrm{f}_i\,(x)}$, where Σ_i represents summation across constraints.

---

[1] The formula appears in, e.g., Goldwater and Johnson (2003, ex. (1)); or the logistic regression chapter of Jurafsky and Martin (2019).

[2] Actually, if we reverse the sign of a constraint it will *increase* the probability of a candidate that "violates" it; this possibility is controversial in linguistics but quite standard in other applications. We will not take a stand here on whether reversed-sign weights are to be tolerated.

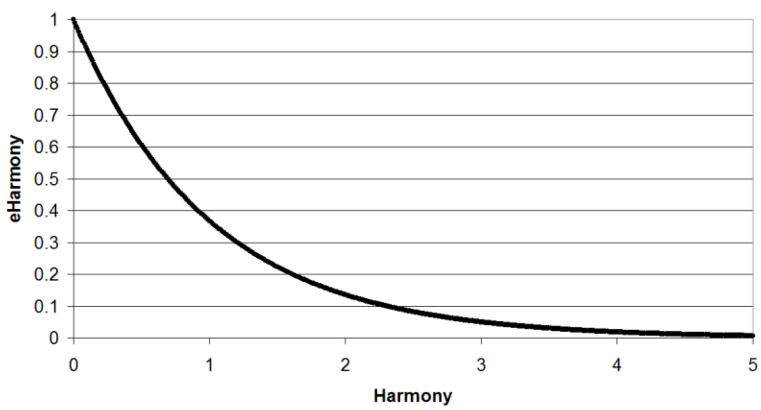
[3] Caveat: in some presentations either the constraint weights or the violations are expressed with opposite sign, so Harmony itself is negative and counts as a reward. This difference is purely notational.

Harmony is an intuitive idea because when we make rational decisions we find it appropriate to weigh *all* of the evidence. In this respect, classical OT is bravely counterintuitive, because the choise between two candidates is made solely by the highest ranked constraint that distinguishes them, ignoring the testimony of all lower-ranked constraints (Prince and Smolensky 1993:§5.2.3.2). The view taken here is that Prince and Smolensky's move to discard evidence was brave, but emerges in the end as empirically wrong.[4]

3.4 eHarmony

Once Harmony values are computed for each candidate, they are converted to what Wilson (2014) has called **eHarmony**.[5] This is done by negating Harmony, then taking *e* (about 2.72) to the result. Thus, in formula (1), the term for eHarmony is: $\mathbf{exp(-\Sigma_i\, \mathit{w_i} f_i\, (\mathit{x}))}$, where exp($x$) is an abbreviation for e^x. The eHarmony function is plotted in (2) below.

(2) *eHarmony plotted against Harmony*



The eHarmony function *rescales* the evidence: if Harmony is increased from an already-large value, then the eHarmony, being already close to zero, and gets only slightly smaller; whereas if Harmony is not very big in the first place then small differences in Harmony result in large differences of eHarmony.

I suggest that this rescaling reflects intuitively sensible decision making. For suppose we are trying to predict output probability for a candidate for which we know, as a rough guess, that the probability is going to be about .5. In such a case, we are quite uncertain, and additional information to inform our choice is welcome and taken seriously. If on the other hand, if a candidate is heavily penalized by information we already have (e.g. probability .001), than even a great deal of evidence may change probability by only a small amount; say, .0005. And for

[4] Classical Optimality Theory finds an echo, perhaps, in the research of Gigerenzer (Gigerenzer and Gaissmaier 2011) which emphasizes cases in which people make choices using single salient facts rather than by weighing all the evidence. Giegerenzer's examples involve real-world decisions (like sending a patient to intensive care) where prediction is difficult and data scarce. The "grammatical decisions" described in this article, however, are different: they have a whole childhood's worth of data behind them. It will emerge below that people do indeed blend a rich variety of evidence when they make unconscious linguistic choices. So whatever the merits of Gigerenzer's ideas in general, I doubt they should be applied to grammar.

[5] Wilson was joking (eHarmony is a dating website), but the mnemonic seems useful.

most people, I suspect, to become *absolutely* certain requires a vast, perhaps infinite, amount of evidence.

This difference in evidentiary scaling is what eHarmony accomplishes. Thus, in graph (2), a 50/50 candidate would lie on the steeply sloped part of the curve, so that small differences of Harmony results in large differences of eHarmony; and this will eventually result in large differences of probability. A candidate whose approximate probability is known to be near zero is heavily penalized in Harmony, and lies far to the right on (2). Here, the slope is shallow, so additional doses of Harmony have only a small effect. The same goes for candidates whose probability is close to one: their rivals are already heavily penalized, and increasing their Harmony penalty will only move the top candidate upward by a small amount.

A slogan that expresses these patterns is: *certainty is evidentially expensive*: to move probability around when it is already close to zero or one requires large infusions of evidence. Converting Harmony into eHarmony is the way MaxEnt implements this principle.

3.5 Computing probability

For the last two steps of the MaxEnt derivation, we sum eHarmony for all the candidates assigned to a given input, calling this sum Z. In formula (1), Z is expressed as: $\boldsymbol{\Sigma_j \exp(-\Sigma_i w_i \mathrm{f}_i(x_j))}$, where j is the index intended to denote candidates. The *probability* of a candidate is obtained by dividing its eHarmony by Z; i.e. we calculate its share in Z. This division appears in the complete formula for P(x) in (1).

The addition-then-division procedure is intuitive, since it says that a candidate is less likely if it has strong rivals. Further, we see now that the probability of any candidate is *proportional* to its eHarmony; hence the discussion in the preceding section, showing how exponentiation makes certainty evidentially expensive, carries through to the final probability relations.

Summing up, the MaxEnt computation is claimed here to be intuitive at every stage:

(3) *MaxEnt and common sense*

a. Constraints differ in their evidential force (§3.1).
b. Multiple violations of the same constraint make a candidate less probable (§3.2).
c. All evidence is considered, none thrown out (§3.3).
d. Evidence has a smaller effect as we approach certainty (§3.4).
e. Candidates are less probable when they compete with powerful rivals (§3.5).

To the extent that these five properties reflect sensible principles for arriving at conclusions from evidence, MaxEnt (or any framework that has these properties) can be said to have an *a priori* claim on our attention.[6]

---

[6] Obviously, there is much more to say about MaxEnt/logistic regression from the technical point of view. For logistic regression as a statistical inference technique, with applicable methods of significance testing, see the textbooks by Johnson (2008) and Baayen (2008). On logistic regression in computer science, with the standard method of calculating the best weights to fit the data (and the proof of its convergence), see Jurafsky and Martin

4. First quantitative signature: the sigmoid curve

With this background we can turn to the main topic: quantitative signatures, their derivation under different theories, and their distribution in the real world.

We focus on simple cases in which for each input, there are just two viable output candidates. In OT, including MaxEnt, this means that all other conceivable candidates are ruled out by powerful constraints. This is normal in OT, and I will not bother with formulating the necessary constraints below.

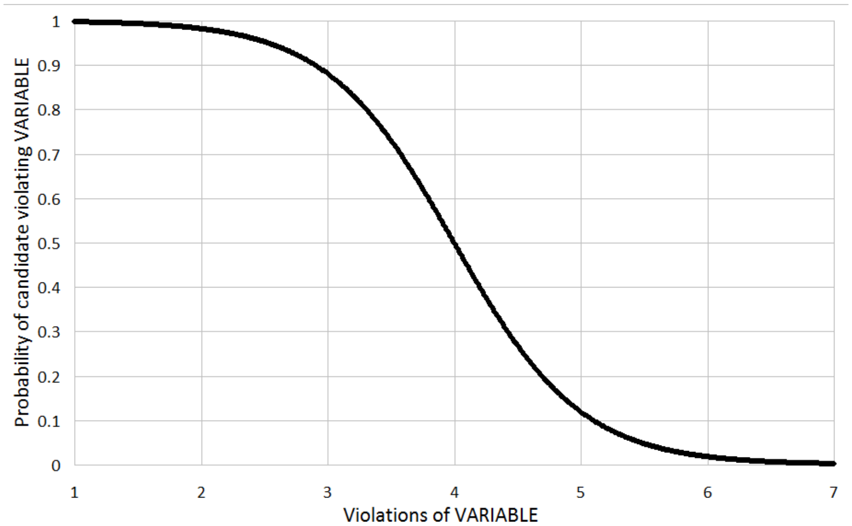
The two viable candidates compete on the basis of less-powerful constraints. Suppose that one of these constraints may be violated either *once* or *not at all*; call it ONOFF. Let the other be a constraint, or a set of constraints, defining a **scale**. Scales are familiar in constraint-based linguistics (Prince and Smolensky 1993/2004 §5.1; de Lacy 2004); and linguists have developed analyses in which the scale is formalized either with a single, multiply-violable constraint, or with families of related constraints.

Let us deal first with the simplest case, where the scale involves multiple violations of a single constraint. We call this constraint VARIABLE and assign it violation levels ranging (for concreteness) from 1 to 7. The candidate competition has two viable candidates per input, of which one obeys VARIABLE and violates ONOFF, and the other obeys ONOFF and violates VARIABLE some specified number of times, depending on the choice of input. Under this setup, we can derive the output probabilities using (1), then plot a function: the horizontal axis gives the number of violations of VARIABLE across inputs, and the vertical axis gives the probability that the candidate violating VARIABLE wins. For clarity, I will plot this function for *all* values on the horizontal axis, not just the 1-7 that would occur for particular input forms.

The curve that MaxEnt derives under these conditions is a **sigmoid** (S-shaped) function, illustrated in (4).

(2019). For MaxEnt specifically applied as a method of analysis in generative grammar, see Goldwater and Johnson (2003), Jäger (2007), and Hayes and Wilson (2008).

(4) *A sigmoid curve generated in MaxEnt*



Here are crucial properties of the MaxEnt sigmoid, often called the “logistic” function.[7] (a) It is symmetrical, and the symmetry point falls where probability crosses 50% — in (4), this is at 4 violations of VARIABLE.[8] (b) It asymptotes on either end at 1 and 0. (c) It is steepest at the symmetry point, and becomes more level as one proceeds in the positive or negative direction. (d) The uphill/downhill orientation depends on whether the constraint weight of VARIABLE is positive or negative; and its steepness is greater when the weight of VARIABLE is larger. (e) The relative right/left position of the curve is determined by the weight of ONOFF. For more details see McPherson and Hayes (2016).

These properties must be kept in mind when we later assess whether an empirically-observed curve is properly to be considered as a sigmoid. Markedly asymmetrical curves, or curves that asymptote at a value other than one or zero, or curves that at some point reverse their slope, would not qualify. On the other hand, the language under study might not provide a full range of values for how much VARIABLE is violated, so in the empirical domain we will often find truncated sigmoids.

4.1 *Illustration: a sigmoid from a phonetic experiment*

In illustrating the math, it is helpful to start with a case in which the horizontal axis of the sigmoid is uncontroversial, being a physical quantity rather than an analytic construct. Such cases arise frequently in phonetics, in the context of speech perception experiments. Suppose, for instance, that we plot on the horizontal axis a phonetic parameter like stop closure duration, as varied in synthesized experimental stimuli. On the vertical axis we plot the probability that an experimental participant will experience a certain percept, such as the sound [b] as opposed to

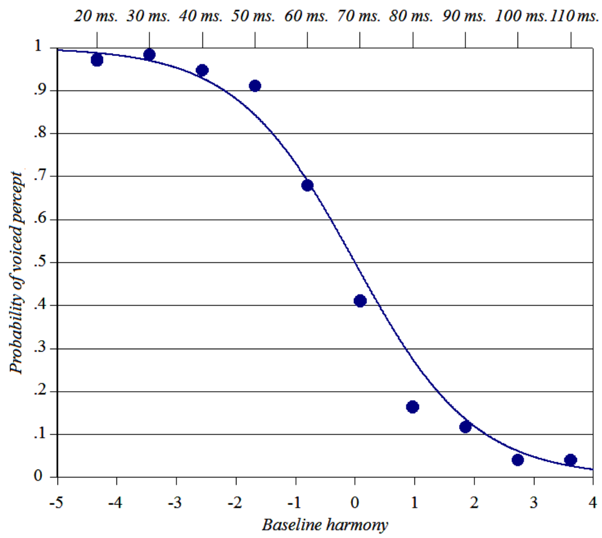
---

[7] The logistic function was named in 1845 by its discoverer, the Belgian mathematician Pierre-François Verhulst. For history of the function and of logistic regression, see Cramer (2002).

[8] The symmetry point is determined by the constraint weights: it falls at the value VARIABLE/ONOFF. In making the graph in (4), I used the weights VARIABLE = 8, ONOFF = 2, which is why the symmetry point is 4.

[p]. Kluender et al. (1988) report such an experiment, and their data indeed emerged as an approximate sigmoid. A subset of the data is replotted in (5); the narrow line behind the data points represents the predictions of the MaxEnt model to be developed immediately below.

(5) *Sigmoid curve relating closure duration to voicing percept, adapted from Kluender et al. (1988)*



I assume the reader's agreement that the sigmoid curve superimposed on the data in (5) is a decent fit, and that small deviations may be attributed to sample or measurement error; the same holds for the remaining graphs in this article.[9] We turn, then, to the reanalysis of the data in MaxEnt terms.

For present purposes it will be useful to adopt a stance proposed by Boersma (1998): that speech perception be regarded as a form of grammar. Boersma sets up a constraint-based, probabilistic theory in which the grammar inputs the acoustic signal and outputs a probability distribution for the set of possible phonemes (or words, etc.) that are inferred from the signal. The particular framework he uses to do this (not MaxEnt) is discussed below in §7.2.1.

Pursing Boersma's imperative in MaxEnt terms, we can arrange our grammar as a simple target-and-penalty system. The grammar inputs closure duration values and selects between the percepts [b] and [p]. As before, we exclude all other percepts by fiat; in a full grammar, they would violate highly-weighted constraints, resulting in essentially zero probability.

Let the constraint VARIABLE penalize the percept of [b] to the extent that closure duration deviates from the extreme value of 20 ms. (which we adopt as the idealized target for [b]). VARIABLE assesses a penalty for every millisecond by which a [b] candidate exceeds this target. We also include a baseline ONOFF constraint, which simply penalizes all [p] candidates. VARIABLE and ONOFF conflict, and the computed [b]-probability will depend on the state of this conflict for a particular number of milliseconds of closure duration in the stimulus.

[9] Of course, as we move from exploration (the goal here) to demonstration (the long-term goal), it becomes essential to assess model fit quantitatively. For the standard techniques, see Johnson (2008) and Baayen (2008).

Using a spreadsheet, it is easy to find the weights that produce the most accurate model for the Kluender et al. data.[10] These turn out to be 0.088 VARIABLE and 4.34 for ONOFF. Using these weights, we can then use the MaxEnt formula (1) to calculate the probability of the voiceless candidate for all values; these are plotted as the narrow line in (5). A detailed review of how these calculations are done has been included in Appendix A, in the Gallery.

The units of the lower horizontal axis in (5), labeled "Baseline harmony," require comment. By simple math, applied to the MaxEnt formula, it turns out that in a system with just two viable candidates, we can recapitulate all the information need to calculate their probability with a single number, the *difference* of Harmony between the two rival candidates. For how this works, see Appendix B in the Gallery. The use of differences is helpful because we can encapsulate the relevant analytical information as a single value on the *x* axis.

Summing up so far: MaxEnt applied to the simple VARIABLE + ONOFF constraint system yields a sigmoid as its quantitative signature, and this signature is well attested in speech perception.

5. Second quantitative signature: the wug-shaped curve

Our next scenario works as follows. As before, we have an ONOFF constraint and a VARIABLE constraint, but this time we double the input set, adding a new batch of inputs identical to the first except that they violate a constraint we will call the PERTURBER: a constraint defined on an independent dimension.

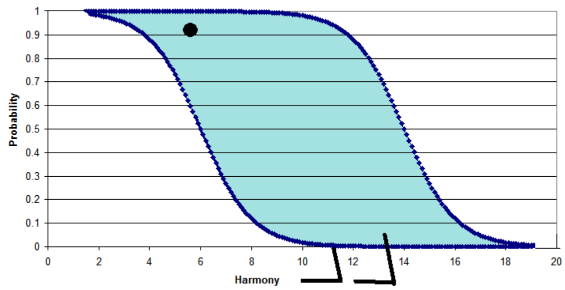
Let us first establish the MaxEnt predictions. The subpopulation of candidates that violate PERTURBER will have their Harmony values increased or decreased, depending on whether PERTURBER is "allied" with ONOFF or with VARIABLE. Other than that, these candidates will behave just like their counterparts that do not violate PERTURBER. Hence, if in a graph similar to (4), we plot the two populations of candidates separately, we will get a second sigmoid, *shifted over from the first* by an amount corresponding to the weight of the PERTURBER. This is illustrated in (6).

As Dustin Bowers suggested to me, it is not hard to imagine in this double-sigmoid shape the perky creature who in recent years has been adopted as the emblematic animal of linguistics; hence we can call it the **wug-shaped curve**, honoring its inventor, (Berko 1958). I have artistically embellished (6) to emphasize the resemblance.

---

[10] In brief, one locates the constraint weights that maximize the product of the predicted probabilities of all data points; i.e., one maximizes likelihood (Goldwater and Johnson 2003, (2)). In Excel, the Solver utility does well for this purpose on modest-size data sets. For weight setting in general, see Hayes and Wilson (2008:385-389), and for the particular calculations done throughout this paper, see the spreadsheets posted in the Gallery.

(6) *The wug-shaped curve*



The weight of the Perturber can be read off the wug's torso: it is the difference of Harmony values at the level where the two sigmoids cross the 50% probability mark. In (6), the this is 14 − 6 = 8. Thus, high and low Perturber values are graphed as fat and skinny wugs.

5.1 Multiple perturbers

Nothing stops us from using more than one Perturber; and when this happens, we will have multiple, parallel sigmoids, spaced as the weights dictate. It is tempting to think of this as a "stripey wug," but I will use "wug-shaped curve" for these cases as well. Most of the curves described below will have multiple Perturbers.

5.2 Wug-shaped curves in the real world

My own involvement with wug-shaped curves arose from participation as second author on Zuraw and Hayes (2017), which may be viewed as an effort to adduce real-life cases of wug-shaped curves and from them make some of the arguments about frameworks that I give in §7. I will first review a particular case we studied; Later, the discussion will expand to other phonologists' work and then to other disciplines of linguistics.

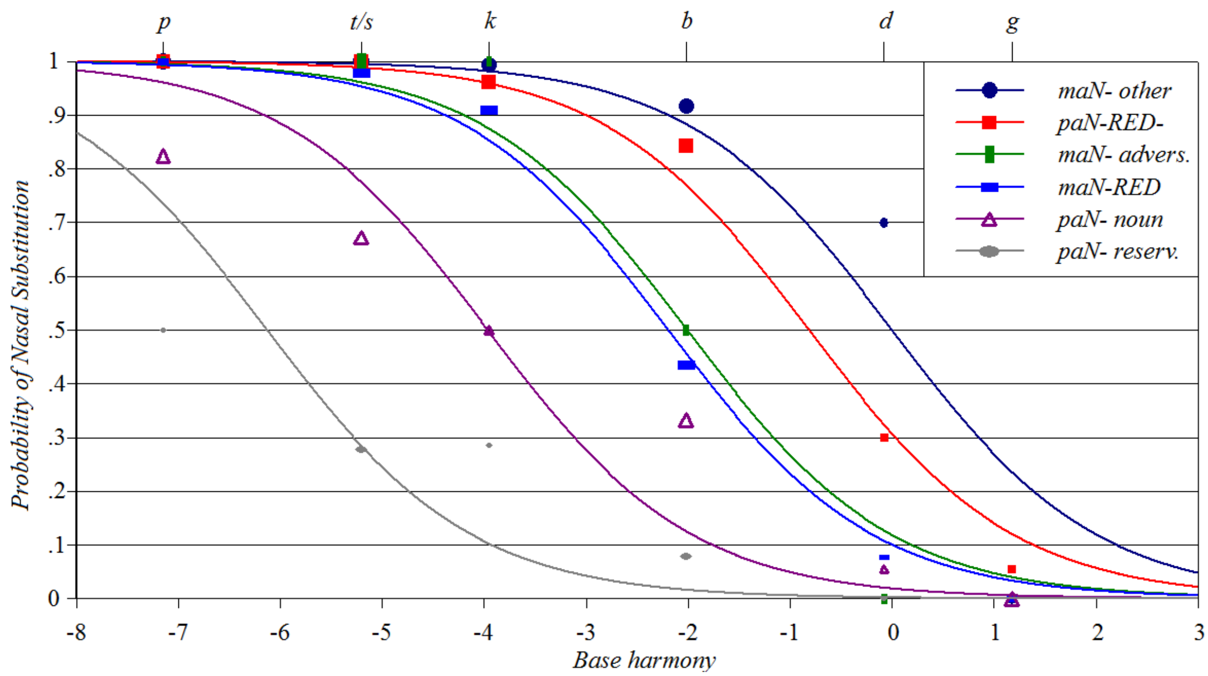
5.2.1 Tagalog Nasal Substitution

Zuraw (2000, 2010), working on Tagalog, was the first to observe wug-shaped patterns and treat them with probabilistic frameworks. The Tagalog consonant [ŋ], when suffix-final, often *merges* with a following consonant, creating an output that blends the nasality and voicing of [ŋ] with the place of articulation of the stem consonant; thus /ŋ+p/ → [m], /ŋ+t/ → [n], etc. The process is lexically optional, applying on a word-by-word basis, and the wug-shaped pattern of application emerged when Zuraw calculated application rates from a language-wide corpus, supported by a nonce-probe study.

In the rendition of Zuraw and Hayes (2017), the Baseline constraints form a family, each of which forbids NC clusters with various features (place, voicing). This family, all of whose members receive different weights in the best-fit analysis, distinguishes six categories: {p, t/s, k, b, d, g}. These categories can be identified by the labels just above the graph in (7). Zuraw also showed that each [ŋ]-final prefix has its own propensity to induce mutation; this is formalized using the constraint weights for a family of prefix-specific Perturber constraints. The horizontal axis in (7) plots the baseline Harmony resulting from the consonant-specific constraints, and the

Perturbers are represented by giving each its own sigmoid. Point sizes reflect the number of cases from which the probability is calculated. The plot is essentially the same as in Zuraw and Hayes (2017), except the horizontal axis is scaled to reflect Baseline harmony.

(7) *The wug-shaped curve in Tagalog Nasal Substitution (after Zuraw and Hayes 2017:Fig. 10)*



The visual fit of the wug-shaped curve to the data strikes me as reasonably good; for quantitative testing of model fit, see Zuraw and Hayes (2017:§2.7).

An important aspect of the wug-shaped curve is that the magnitude of the effect of the Perturbers depends on where we are located on the Baseline scale: it is maximal in medial position and diminishes gradually toward the peripheries; see the vertical spacing of the dots in (7). This pattern, pointed in Zuraw and Hayes (2017) and Smith and Pater 2020)[11] is, as can be shown, a consequence of the MaxEnt formula. From the perspective of §3.4, this pattern is intuitive: the evidence from a Perturber buys you a lot in the middle, where you are uncertain, but will buys little at the peripheries, where you are already close to certain.

The remaining two cases discussed in Zuraw and Hayes 2017 (French Liaison and Hungarian vowel harmony), when replotted using the format described here, again yield wug-shaped curves; these plots and the calculations supporting them may be viewed in the Gallery.

[11] Smith and Pater's clear discussion is focused on the interpretation of an experiment they carried out on French vowel-zero alternations. Their observation becomes visible if one replots their data in the format employed here; it is a clear wug-shaped curve. I have included such a plot in the online Gallery.

6. Prospecting the linguistics literature for wug-shaped curves

Encouraged by the results from work in which I myself participated, I undertook for purposes of this paper a sort of intellectual hiking trip, browsing through classic works of probabilistic linguistics and replotting their data with arrangements of Baseline and Perturbers.

My criteria for choosing cases were as follows. First, the probability of candidates had to approach one at one end, or zero at the other, or ideally both. Otherwise we only see vaguely parallel lines that are uninformative. Second, examples had to be abundant enough so that each data point would represent multiple observations, preventing random fluctuations from obscuring the result. When partitioning the constraints into Baseline and Perturber sets, I favored a Baseline set that would yield a broad probability range. I also favored partitions that gave the Perturber set (where possible, both sets) a unified, intuitively distinct rationale.

Everywhere I went, I found wug-shaped curves. I will demonstrate this in six areas of linguistics: more phonology, sociolinguistics, syntax, phonetics, historical linguistics, and semantics.

6.1 Other work in phonology

The cases replotted in the online Gallery come from the following studies: Anttila's (1997) pioneering demonstration of constraint-based modeling of variable outputs, with data from Finnish genitive plurals; Ernestus and Baayen's (2003) modeling (including MaxEnt) of the ability of Dutch speakers to project the underlying forms of finally-devoiced consonants on the basis of the phonological properties of stems; Ryan's (2019) study of stress placement in Hupa; and Smith and Pater's (2020) study of vowel-zero alternations in French. All four cases yielded patterns reasonably interpreted as wug-shaped curves.

6.2 Sociolinguistics

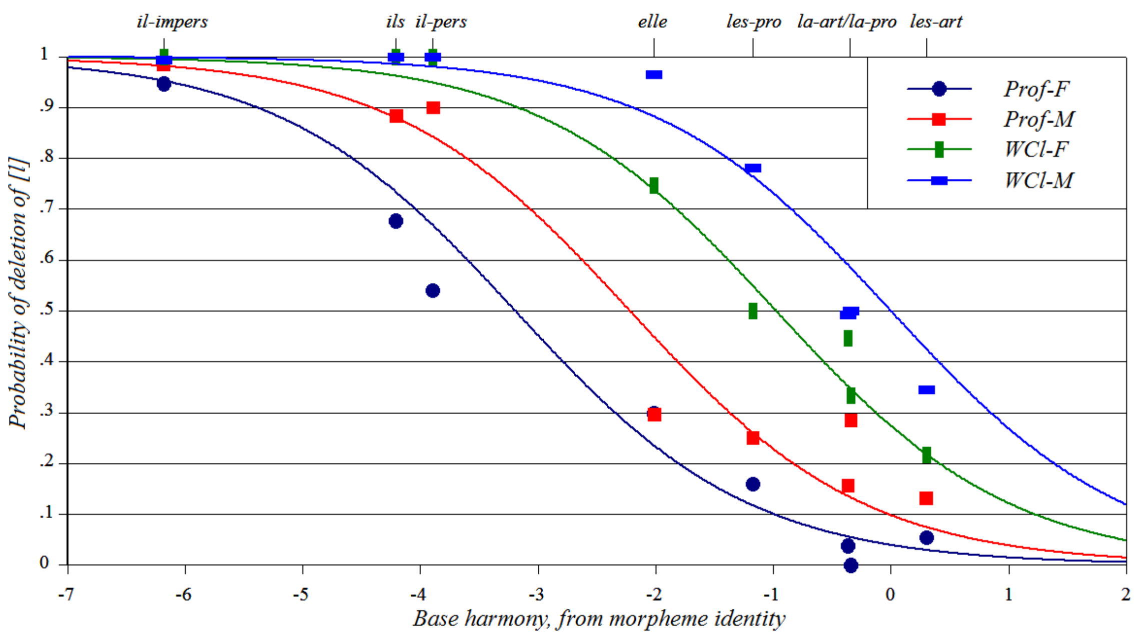
The essential ideas of this paper — MaxEnt analysis, Perturbers, and wug-shaped curves — were all innovated in work by sociolinguists circa 1970. Labov's (1969) study of Black English copula deletion established the systematicity of linguistic variation. It also demonstrated the existence of Perturbers and their ability to affect output probabilities across the Baseline range. MaxEnt was introduced, under the label of logistic regression, by researchers centered on David Sankoff (Cedergren and Sankoff 1974, Rousseau and Sankoff 1978, Sankoff and Labov 1979); and the wug-shaped curve was discovered by Bailey (1973:106); though he did not treat it with the new forms of quantitative analysis. Since then, quantitative modeling of variable phonology in sociolinguistics has continue apace both empirically and theoretically (§7.3); for general background, see Chambers and Schilling (2013) and Mendoza-Denton et al. (2003).

In my survey of sociolinguistic wug-shaped curves, I recalculated and plotted wug-shaped curves for several classic studies: Labov's (1969) (covering both contraction and deletion), Wolfram (1969) on Cluster Simplification in Detroit Black English, and three studies from Cedergren and Sankoff (1974): *que*-dropping in Québec French, [r] spirantization in Panamanian Spanish, and (with Labov's data) [r]-Dropping in New York City English. All of these may be found in the Gallery.

The illustration I give here replots the data for which Bailey first pointed out a wug-shaped curve.[12] He took his main example from research by Gillian Sankoff (ms., 1972/1978) on Québec French; this involved optional deletion of [l] in function words. In my MaxEnt reconstruction, the Baseline constraints are (a) a general Markedness constraint disfavoring the realization of [l]; (b) lexically-specific MAX constraints, militating against [l]-loss in particular function words. [13] The Perturbers are, in superficial terms, further MAX constraints based on the sex and socioeconomic status (professional/working class) of the speaker. I personally doubt that such factors actually appear in the grammars of individual speakers; it seems more reasonable to suppose that speakers set the weight of MAX(l) differently in various social contexts, in ways that respond to sex and social class.[14] Thus in the present case, sex and social class are treated as proxies for the varying weight of MAX(l).

The wug-shaped curve I obtained is in (8). The vertical compression of the data points Bailey observed, particularly at the "beak," is evident.

(8) *The wug-shaped curve in Quebec French [l]-deletion*



[12] Bailey's method of plotting the curve, using contour lines, strikes me as infelicitous, but his verbal description of cross-classifying Baseline and Perturber effects does capture the key point: "the statistics are more bunched in the bottom and top percentages and more spread out toward the middle percentages" (p. 106).

[13] MAX, penalizing deletion, is a key constraint family in the standard theory of phonological Markedness constraints, proposed by McCarthy and Prince (1995). The tendency of function words to have morpheme-specific behavior has been long known, see Kaisse (1985).

[14] The response of phonology to social context is a vast research area, and the essays in Part III of Chambers et al. (2013) offer a useful guide. As far as varying MaxEnt weights by speaking context, I know of little work, but I think the proposal of Coetzee and Kawahara (2017) would be helpful as a means of taking on this topic.

6.2.1 How MaxEnt is used in classical sociolinguistics

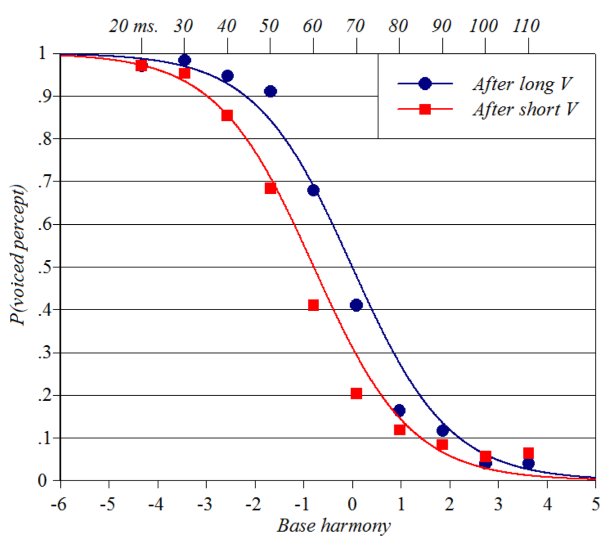
The use of logistic regression in sociolinguistics depended on (to describe it anachronistically) an interesting hybrid, which blended the rule-based phonology of *SPE* (Chomsky and Halle 1968) with constraint-based MaxEnt phonology. Its key mechanism, the **variable rule**, was much like an *SPE* phonological rule, but bolted on to it was a small MaxEnt grammar, which had just two output candidates, *Apply Rule* and *Don't Apply Rule*. The constraints of this grammar — often identifiable with constraints later to be used in OT phonology — were used to make this choice. On any given speaking occasion, the little MaxEnt grammar would render its probabilistic verdict, and the *SPE* rule coupled to it would apply accordingly.

Empirically, this remains a viable model, but I feel it is unnecessarily complex. As OT shows, the constraints themselves, just slightly amplified, can cover the whole grammar without the use of rules, creating a theory that is simpler and more responsive to typology. Observe further that variable rules offer the analyst greater freedom, since the weights of the Perturber constraints are set separately for each rule, rather than being a property of the grammar as a whole. It would be interesting to see if the older theory might be empirically defended by adducing cases where this freedom is necessary.

6.3 Phonetics

We return to sigmoid (5), from Kluender et al. (1988), discussed in §4.1. For simplicity, (5) plots only one of the two data series from this paper. The authors' actual research interest focused on a Perturber, namely the length of the vowel preceding the [b]/[p]. Their hypothesis was that, since vowels are normally longer before voiced stops, the presence of a longer vowel would bias perception in favor of [b]. That this hypothesis panned out is shown by (9) below.

(9) *Wug-shaped curve for closure duration to voicing percept under two conditions; adapted from Kluender et al.(1988)*



The MaxEnt grammar that underlies (9) is like the one for (5), except that it includes a Perturber, *VOICED PERCEPT AFTER SHORT VOWEL; this penalizes the [b] candidate when in this context. When I fit the full Kluender data, including both long and short vowels before [b] and [p], this constraint received a weight of 0.84, and the result was a clear if skinny wug.

Plots like (9) frequently appear in the work of phoneticians, who use MaxEnt (under the logistic regression rubric) to quantify the influence of the Perturber. Following the MaxEnt math, it is straightforward to rescale Perturber harmony as actual milliseconds. In the present case it emerges that the ms. value for *VOICED PERCEPT AFTER SHORT VOWEL is about 9.5 msec, in rough agreement with what Kluender et al. found using a different method. The repeated successful use of MaxEnt by phoneticians in modeling this sort of data (Morrison 2007) has probably created the largest population of wug-shaped curves in the linguistics literature.

6.4 Syntax

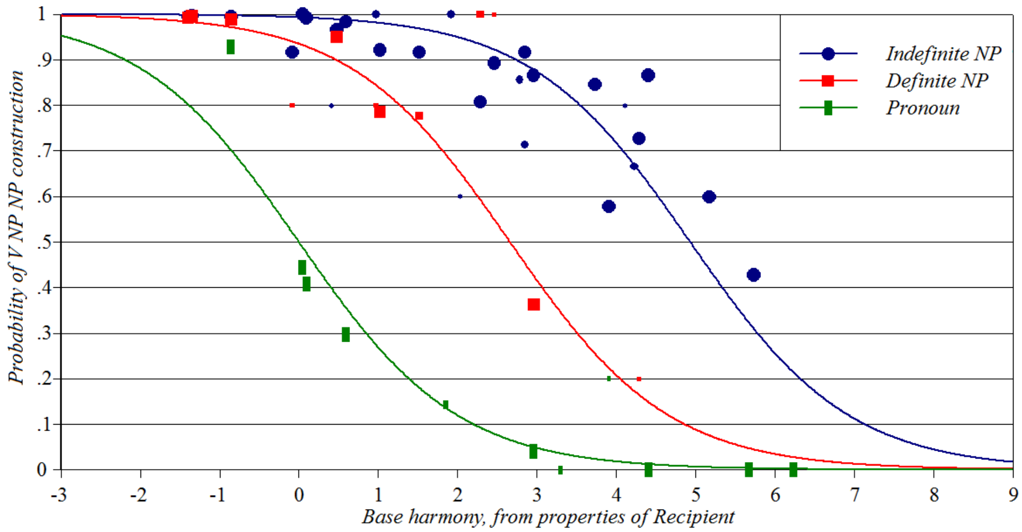
Many studies in syntax have engaged with gradience of the types described in §1, using MaxEnt or similar models; see, for example, Velldal and Oepen (2005), Bresnan et al. (2007), Bresnan and Hay (2008), and Irvine and Dredze (2017). There are also experimental studies that, while not incorporating an explicit probabilistic component, nevertheless support the concept of Harmony, in supposing that violations of distinct syntactic principles assess particular, additive decrements of well-formedness; Featherston (2005, 2019), Keller (2000, 2006).

The research addressed here, by Bresnan and colleagues, focuses on a syntactic *microdomain*: instances in which the same communicative intent can be expressed with two different syntactic encodings. An example is the two ways that English offers to express the arguments of a verb of giving: NP NP (*Mary gave John a book*) and NP PP (*Mary gave a book to John*). In such cases, it has proven possible to identify probabilistic factors that favor one or the other outcome. In analyzing such cases, Bresnan et al. have used MaxEnt and similar tools. Their studies show that choices like *NP NP* vs. *NP PP* are, as it were, *semipredictable*, provided one uses an appropriate statistical model. The refined distinctions predicted by their constraint weights are supported empirically in that they show up as clear if modest distinctions between dialects, such as New Zealand and American English. These distinctions emerge in experimentation (Bresnan and Ford 2010) and corpus work.

A recent article in this tradition, Szmrecsanyi et al. (2017), uncovers dialect-specific patterns for four varieties of English (US, UK Canada, New Zealand) for two syntactic choices; the dative one just mentioned as well as the genitive choice for, e.g., *the king's palace* vs. *the palace of the king*. In my replottings, I abstracted away from these differences and merged the data from all four dialects.

For the datives, we can take as Baseline constraints the following: (1) those which depend on Szmrecsanyi et al.'s taxonomy of verb semantics, distinguishing "transfer," "communication," and "abstract"; (2) those dependent on properties of the *recipient* NP, such as animacy, definiteness, and pronounhood; (3) a constraint based on relative length (in words), which prefers placing longer phrases second. This array of constraints produces a rich baseline with multiple values (so, for details the reader should consult the original paper and the Gallery). For Perturbers, I selected the constraints and data series that single out three categories of the *theme* NP (that which is given): indefinite full NP, definite full NP, and pronoun. The wug-shaped curve that emerged under this re-plotting is shown in (10).

(10) *The wug-shaped curve in English dative constructions, after Szmrecsanyi et al. (2017)*



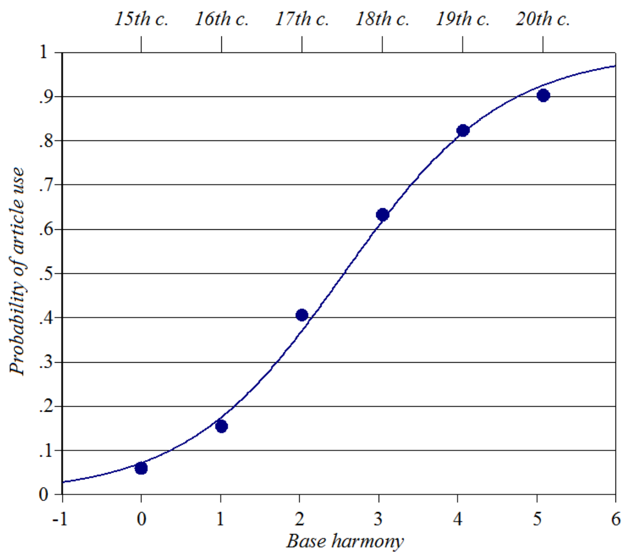
The genitive data in Szmrecsanyi et al. (2017) are of special interest because the numbers suffice to inspect the effect of one single gradient constraint, which favors the *N of NP* construction when the possessor NP is long, as measured in words. The replotting of these data, in the Gallery, demonstrates (at least in a limited range) the form of wug-shaped curve that arises (like (5)) from a system in which a single constraint with multiple violations forms the Baseline.

6.5 Historical linguistics[15]

Kroch (1989) inspected old texts across time, tracking the relative frequencies of syntactic variants as a language gradually changes. One of Kroch's data series, from Oliveira e Silva (1982), documents a centuries-long syntactic change in Portuguese, whereby NP that include a possessor would also include a definite article; thus over time, the former *seus livros* 'his books' was gradually replaced by *os seus livros* '(the) his books'. In (11) is a chart, adapted from Kroch, which shows the frequency with the definite-article variant is employed.

[15] For general background on probabilistic modeling of language change, including more detailed discussion of the material treated here, see Zuraw (2003).

(11) *Use of the definite article in Portuguese possessed noun phrases, after Kroch (1989)*

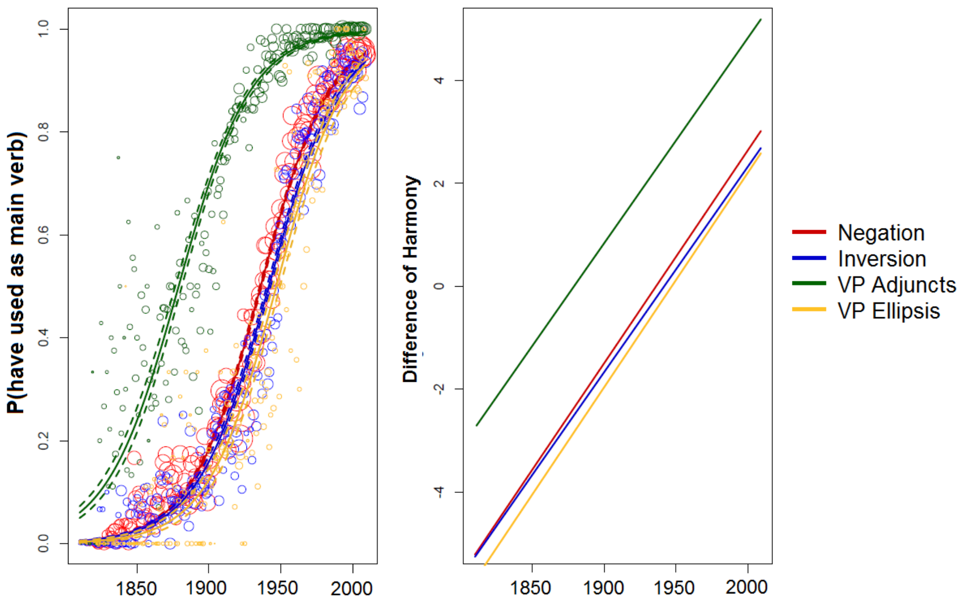


My replotting of the data (which, as Kroch noted, form a sigmoid) is based on my recreation of the MaxEnt calculations Kroch carried out in the 1980s. The lower horizontal scale gives the Harmony-difference values ($H_{article} - H_{noarticle}$). That these values are evenly spaced in time illustrates one of Kroch's main points: if MaxEnt/logistic regression is used to model the data, we get a clean generalization, namely that *in the Harmony domain*, the propensity to use the novel construction increases at a constant rate. For the Portuguese case, the rate of increase turns out to be 1.02 Harmony units per century.

Kroch's "constant rate" hypothesis becomes even more interesting when we include, as Kroch did, some Perturbers. While the Variable constraint weight increases over time, the Perturbers are assumed to be stable. This gives rise to a diachronic wug-shaped curve, with identical sigmoids spaced apart following the weights of the Perturbers. This is what Kroch's and subsequent work has found: variation in rates of change across contexts that look nonsensical when measured as probability look coherent when measured as Harmony.

A study on these lines, Zimmermann (2017), traces the evolution of English *have* from an auxiliary to a main verb. This change is manifested in four contexts: *negation* ("I (haven't/don't have) any"); *inversion* ("(Have you/Do you have) a penny?"), *ellipsis* ("You have a flair; you really (have/do)") and *adverb placement* ("He (has already/already has) the approval of the nation."). Each context is assumed to be affiliated with constraints acting as Perturbers. The VARIABLE is the diachronically-shifting constraint governing whether *have* functions as an Aux or main verb. Zimmermann, tracing each phenomenon across two centuries, obtains the wug-shaped curve in (12). The left panel depicts the four sigmoids that were found, with error bars and circles indicating the size of the text from which each datapoint derives (Zimmermann 2017:107).

(12) *A wug-shaped curve in syntactic change, adapted from Zimmermann (2017:107)*



The right panel implements a practice developed by Kroch: the four sigmoids are plotted not as observed proportions, but as the Harmony difference from the MaxEnt model. These lines are straight and parallel, illustrating clearly what is meant by the "constant-rate hypothesis."

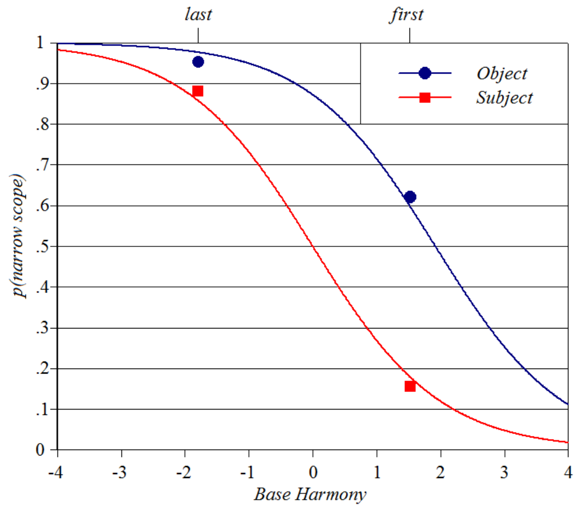
Kroch's 1989 study spawned much similar work; Blythe and Croft (2012:279-280) list dozens of language changes involving sigmoid curves. There has also been intriguing work addressing *why* language change should typically show a constant rate; for discussion see Appendix C in the Gallery.

6.6 *Semantics/Pragmatics*

Quantifier scope ambiguities occur in sentences like *A student saw every professor*. Corpus study (e.g. AnderBois et al. 2012) suggests that an appropriate system for predicting quantifier scope is likely to be a probabilistic one: judgments reflect a blend of conflicting factors, and my sense is that they are gradient, not categorical.

I offer here a tiny curve based on a subset of AnderBois et al.'s data. The Baseline reflects linear order (leftward favors broad scope) and the Perturber reflects grammatical relations (subjects prefer broad scope). The fit seems good (unimpressively so, given that they are few data points; but see a rival model in §7.1). The graph does display the narrowing of the influence of the Perturber at the periphery that MaxEnt predicts.

(13) *A wug-shaped curve in quantifier scope, adapted from AnderBois et al. (2012:107)*



6.7 Sound symbolism

For an intriguing Pokémon study, yielding a MaxEnt wug-shaped curve, see Kawahara (in press).

7. What formal models can generate Wug-shaped curves?

We turn to the other goal of this paper, framework assessment. This involves critiquing models that demonstrably fail to generate wug-shaped curves, and asking about models whose behavior is yet undiagnosed.

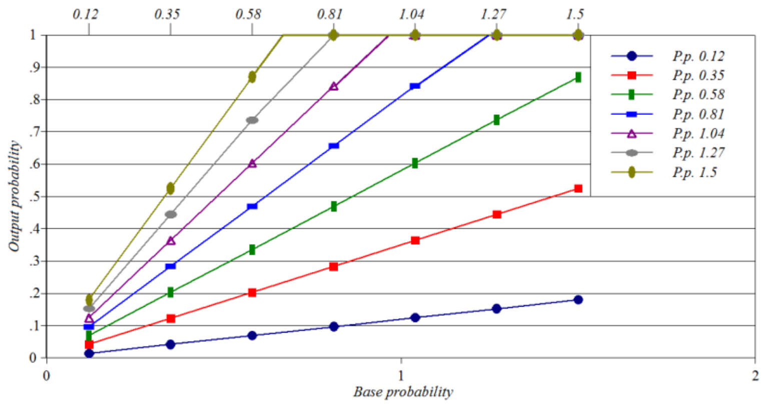
In inspecting the results of various frameworks applied to the same data, I have found a consistent pattern: often a defective framework gets lucky in fitting a particular batch of data. To evaluate a framework properly, we need to examine its performance in a variety of situations.

7.1 Some simple alternatives to MaxEnt

MaxEnt is not the only way to map from constraint violations and weights to probability, nor even the simplest. Some alternatives were considered in the early days of quantitative sociolinguistics, before the field shifted toward MaxEnt (Sankoff and Labov 1979).

In a **Multiplication-cum-Cutoff** model, every constraint violation has the effect of multiplying candidate probability by the weight of the constraint. Constraints are allowed to have values greater than one, so they can increase as well as reduce probability. Since probabilities cannot go above one, this model prevents impossible values by imposing a ceiling of one by fiat. In (14) I give a schematic quantitative signature of this approach, assuming seven Baseline probabilities and seven Perturber probabilities (P.p.).

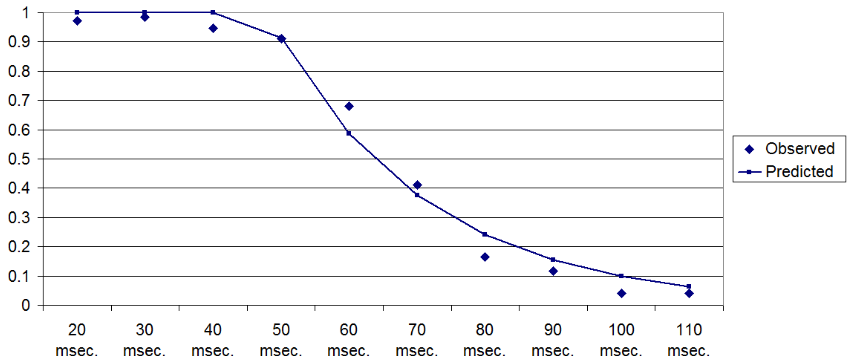
(14) *Quantitative signature of the Multiplication-cum-cutoff model*



As can be seen, the prediction is that probabilities for particular Perturbers will converge in one direction, diverge in the other, up to the point where the cutoff prevents further divergence. I have never encountered data patterns like this and would be curious to know if they exist.

A second quantitative signature of the Multiplication-cum-Cutoff is obtained when it is applied to cases where a single constraint is violated a variable number of times, as in (9). What we find is a curious shape, rather like the MaxEnt sigmoid on the left, but a declining exponential on the right. The curve for the Kluender et al. data (post-long-vowel series) looks like (15).

(15) *Fitting the Multiplication-cum-Cutoff model to the data of Kluender et al. (post-long vowel series)*



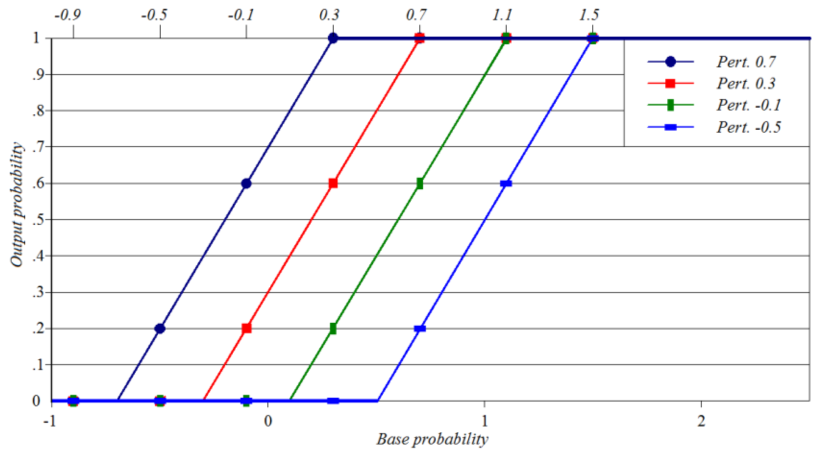
The data misfit is admittedly modest, but the typological implication seems wrong — to my knowledge, nowhere in the speech perception literature is it claimed that the curves from identification experiments are systematically asymmetrical in this way; and the same would hold for the literature on historical-change sigmoids.[16]

In an **Addition-cum-Cutoff** model, probability is linearly related to the violations of VARIABLE, with each PERTURBER adding to, or subtracting from, the base probability by a constant. To avoid impossible probabilities, we impose cutoffs at 0 and 1. This model was put

[16] For further discussion of the problems with this model, see Sankoff and Labov (1979).

forth as a straw man by Cedergren and Sankoff (1974). Its quantitative signature is the "Z-shaped curve," with parallel lines going diagonally between the cutoffs and sharp angles at the transition.

(16) *Quantitative signature of the Addition-cum-Cutoff model*



Where MaxEnt smoothes off the ends, gradually leveling the slope, Addition-cum-Cutoff crashes into the limits at zero and one. In actual model-fitting, this difference often produces only trivial differences in accuracy, because the noise present in almost any data means that it is hard to prove that the ends of the sigmoid really are smooth rather than angular.[17] However, even the very simple data of (13) are modeled poorly in Addition-cum-Cutoff, as the two lines "want" to have different slopes.[18]

The models just covered can be addressed more broadly, in terms of the ways that a constraint-based framework could express a rational inductive procedure. §3.4 showed that MaxEnt is varies how strongly evidence (here, constraint violations) bears on probability: in the middle of the probability range, violations are influential; at either periphery, less so; and this embodies the sensible principle that certainty should be evidentially expensive. Neither Multiplication-cum-Cutoff nor Addition-cum-Cutoff does this. Multiplication-cum-Cutoff says that the value of evidence is strongly asymmetrical with respect to the defining scale. Addition-cum-Cutoff says evidence is equally informative throughout the zone between cutoffs, then suddenly becomes 100% uninformative.

[17] The closest thing I have seen to a data pattern that looks like (16) is the Pokémon data, as originally created by company employees, given in Kawahara (in press). However, the fit of the Addition-cum-Cutoff in this case is only marginally better than the fit of MaxEnt model.

[18] The quantitative signature of the Addition-cum-Cutoff model would also poorly model an important finding in speech perception: small *differences* in the physical signal become progressively more informative to the hearer as one approaches the category boundary. This is a natural consequence of the MaxEnt sigmoid, whose derivative, depicting sensitivity, is a mountain-like shape. It would not be expected under Addition-cum-Cutoff, whose derivative is an all-or-nothing block, predicting that small differences should be uniformly effective in the local zone, useless outside it. The curves obtained by McMurray et al. (2008), for instance, strongly support MaxEnt under this interpretation.

7.2 Frameworks originating in Optimality Theory

Two further approaches I will discuss have an ancestry in Optimality Theory; both are like MaxEnt in attempting to render OT probabilistic.

7.2.1 Stochastic Optimality Theory

In Stochastic OT (Boersma 1998), the key idea is that the content of the grammar is itself probabilistic: constraints comes with a number ("ranking value") that expresses how highly ranked they are in general, but each time the grammar is deployed ("evaluation time"), these ranking values are adjusted by a small random noise factor. The adjusted values are then used to sort the constraints, and at this point the choice of winning candidate follows classical OT. Repeated application of this procedure will yield a probability distribution through sampling.

Applied in the cases discussed here, Stochastic OT exhibits two failings. First, it cannot treat cases of variation from single VARIABLE constraints that vary in their violation count. This is because the classical-OT decision procedure is indifferent to the "margin of victory", caring only about relative differences. As we have seen, such indifference is problematic for dealing with speech perception (§4.1, §6.3), word-count effects in syntax (§6.4) or syllable counts in sound symbolism (§6.7).[19]

Second, in Stochastic OT a Perturber can only perturb "within its own zone"; that is, when its own ranking value is within shouting distance of the constraints that it interacts with. But in real life, the effect of a Perturber is *across the board*: it interacts with constraints that are mutually very far apart on the ranking scale. This point is covered in detail in Zuraw and Hayes (2017; §2.6). The present paper includes a similar conspicuous failure, in the Québec French case of §6.2; see Gallery for the failed graph.

Both failings are rooted in traits of the classical OT on which Stochastic OT is based: contrary to principles (3b,c) above, it ignores relevant data, either in the form of violation counts, or constraints ranked too low to have influence.

7.2.2 Noisy Harmonic Grammar

The primary reference is Boersma and Pater (2016). Like MaxEnt, this is a species of Harmonic Grammar, and the procedure for assigning probabilities to candidates starts in the same way, with the computation of Harmony for each candidate. At this point Noisy Harmonic Grammar becomes like Stochastic OT, in that we suppose a series of evaluation times at which the grammar gets altered by random shifts, chosen from a Gaussian distribution. The framework comes in several varieties (Hayes 2017), which differ in which part of the calculation gets tweaked: we can alter constraint weights, violations, tableau cells (violations times weight; as in Goldrick and Daland 2009) or the Harmony scores of candidates.

[19] A proposal made by Boersma (1998: §6, §8.4) actually *can* derive sigmoids from variable constraints in Stochastic OT. The idea is to replace single gradient constraints with *bundles* of constraints, each having equivalent effect but a slightly different target value. The complexity of implementing this approach has perhaps been a factor in its still being underexplored.

These types differ in their quantitative signatures. Adding noise to constraint weights is, in my opinion, probably not a good idea, for its quantitative signature for variable violation count cases looks much like a declining exponential, which we already saw, and criticized, for the Multiplication-cum-Cutoff model.[20] The reason this happens, and an example of the resulting inferior fit, is given in McPherson and Hayes (2016:156). The situation is actually somewhat worse than for the Multiplication-cum-Cutoff, since the curve asymptotes to the right at a positive value, never approaching zero.

The opposite end of the continuum of types for Noisy Harmonic Grammar[21] is randomly altering the Harmony scores of candidates. As Flemming (2017) demonstrates, this theory is extremely close to MaxEnt, and the sigmoid curves it generates as a quantitative signature are virtually indistinguishable from the MaxEnt sigmoid.[22] It is hard to imagine language data being sufficiently accurate to distinguish the two shapes. Unsurprisingly, this version of Noisy Harmonic Grammar has all of the common-sense properties put forth above in (3) as traits of MaxEnt: it gives different strengths to different constraints, avoids discarding evidence (either from violation counts or from weaker constraints), it makes certainty evidentially expensive, and it gives lower credence to an option that competes with strong alternatives.[23]

7.3 Other models

Among the probabilistic descendants of Optimality Theory, MaxEnt differs in origin from Stochastic OT and Noisy Harmonic Grammar in that it was not home-grown; it imported its math from existing work in statistics. However, from the viewpoint of statistics itself, MaxEnt is a retrograde, representing the avant garde of the 1970’s (Cramer 2002). In more recent decades, it has become normal for experimental and corpus work that uses logistic regression to employ the *mixed-effects* version of the model (e.g., Baayen 2008, Johnson 2011), which controls for the idiosyncrasies of individual items or participants.[24] There are other models more elaborate than MaxEnt, e.g. neural network models (e.g. Goldberg 2017), or random forest models (Tagliamonte and Baayen 2012). Some of the authors whose empirical work is surveyed here have made use of these more sophisticated statistical approaches; e.g. Zimmermann (2017) employs mixed-effects regression and Szmrecsanyi et al. (2017) employ random forests. Within

[20] A graph of this function is given in the Gallery.

[21] I will skip the violation-noise and tableau-cell noise variants, which are intermediate in their behavior.

[22] A graph of the MaxEnt sigmoid superposed with the Noisy Harmonic Grammar sigmoid is given in the Gallery.

[23] Noisy Harmonic Grammar has been put to use, and thus scrutinized, far less than MaxEnt. It is of particular interest in that one version of it, constraint-noise NHG, has the property, shared with OT, of assigning zero probability to “harmonically bounded” candidates; this may inoculate it against the possibility of generating typologically aberrant predictions, as suggested by Anttila and Magri (2018) and Magri et al. (2018). Unfortunately, the most restrictive version of the theory in this respect is the very same one that has the most problematic quantitative signature; so it is not clear that NHG is able to resolve all our worries.

[24] Indeed, the balancing of lexical vs. general preferences is a current live issue in phonological theory, and we are seeing efforts to set this balance appropriately (Moore-Cantwell and Pater 2016), including by using mixed-effects regression (Zymet 2018).

sociolinguistics, the methods of statistics employed have also evolved greatly since the introduction of simple logistic regression in the 1970s (Johnson 2009).

Unlike MaxEnt/classical logistic regression, these statistical approaches do not (to my knowledge) have a simple analytic solution for when they would generate wug-shaped curves, and it would be worth exploring their behavior further. I anticipate that they would probably give rise to wug-shaped curves frequently; for a simple reason: they too, are interpretable as mathematical embodiments of common-sense inductive reasoning, and no system that implements these principles faithfully would be likely to deviate all that far in the curves it generates. This point is pursued in AppendixD in the Gallery, where I suggest adherence to the principles of (3) virtually guarantees a theory that generates something approximating MaxEnt sigmoids and wug-shaped curves. Appendix E discusses another statistical issue for which no space is available here, namely the use interaction terms, known in OT as conjoined constraints.

8. Conclusions

8.1 Is the pattern found here meaningful, and if so, how?

To put a brave slant on the content of this paper: I raise the possibility that there exist general quantitative principles, along the lines of MaxEnt, that establish the normal patterning of variation in human languages, and that this is what leads to the repeated appearance of wug-shaped curves when we plot data from the various fields of linguistics. Of course, it is unlikely that with further scrutiny, *all* observed patterns of variation will line up as prettily as the ones seen here, and indeed I have found a few cases (all posted in the Gallery) where the presence of a wug-shaped curve is less compelling than described here. However, there is a fact that encourages me in thinking that a broader inquiry would confirm the basic pattern, namely that *logistic regression has proven popular wherever it has been adopted in linguistics*. This suggests that, on the whole, it has made possible accurate modeling of variation data; which, given the math discussed above, means that if we partition the constraints into Baseline and Perturber families, we will probably find more wug-shaped curves. That this is a non-trivial finding emerged from §7.1 and §7.2, which explored alternative approaches: MaxEnt and related theories generate wug-shaped curves, others don't.

8.2 Is MaxEnt part of the language faculty?

When we ponder whether some principle pervasive in language is part of the "language faculty," there are two senses in which this is meant. One is, "an innate principle specific to human language"; the others "an innate cognitive capacity possessed by humans, employed in language." I suggest that if MaxEnt is part of the language faculty, it is probably in the latter, broader sense. Here are three related points.

1. Under the label of logistic regression, MaxEnt is a standard model of mathematical psychology, with numerous applications throughout cognitive science. The early stages of its spread are documented by Cramer (2002). The earliest proposal of MaxEnt as a cognitive model I am aware of is Smolensky (1986) — but Prof. Smolensky in 1986 was not yet a linguist, and his proposal was intended as a model of cognition in general.

2. Profs. AnderBois and Brasoveanu, whose work on quantifier scope is discussed above, point out to me that scope judgments likely involve more than just linguistic principles; they invoke real-world inferences and other forms of thought. It would be feasible mathematically — though perhaps unsettling to our current habits of thinking — to use MaxEnt to fold extralinguistic principles of reasoning together with the grammatical system, forming a broader system for computing scope judgments. I suspect such a system would likewise exhibit wug-shaped curves.

3. One might also question whether the specific mathematics of MaxEnt is innate. I suspect what is innate is some powerful capacity to make rational inductive choices at the unconscious level, following the principles of (3). MaxEnt serves as a good, provisional, mathematical approximation of this capacity.

References

AnderBois, Scott, Adrian Brasoveanu, and Robert Henderson (2012) The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, vol. 16, no. 1, pp. 15-28.

Anttila, Arto (1997) 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.

Anttila, Arto, and Giorgio Magri (2018) Does MaxEnt overgenerate? Implicational universals in maximum entropy grammar. In *Proceedings of the Annual Meetings on Phonology*, vol. 5.

Baayen, R. Harald (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bailey, Charles-James N. (1973) *Variation and linguistic theory*. Washington: Center for Applied Linguistics.

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

Blythe, Richard A., and William Croft (2012) S-curves and the mechanisms of propagation in language change. *Language* 88:269-304.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy, eds. (2003) *Probabilistic Linguistics*. Cambridge: MA: MIT Press.

Boersma, Paul (1998. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. Ph.D. 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 John McCarthy and Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press

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 Jennifer Hay (2008) Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English." *Lingua* 118, no. 2 (2008): 245-259.

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

Cedergren, Henrietta J., and David Sankoff (1974) Variable rules: Performance as a statistical reflection of competence. *Language* 50:333-355.

Chambers, Jack K., Peter Trudgill, and Natalie Schilling-Estes, eds. (2013) *The handbook of language variation and change*. Oxford, UK: Wiley-Blackwell, 2013.

Chomsky, Noam and Morris Halle (1968) *The Sound Pattern of English.* New York: Harper and Row.

Coetzee, Andries W., and Shigeto Kawahara (2013) Frequency biases in phonological variation. *Natural Language & Linguistic Theory* 31:47-89.

Cramer, Jan S. (2002) The origins of logistic regression. *Tinbergen Institute Discussion Papers* No 02-119/4, Tinbergen Institute.

de Lacy, Paul (2004) Markedness conflation in Optimality Theory. *Phonology* 21:145–199.

Eisner, Jason (1997). Efficient generation in primitive Optimality Theory, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Morgan Kaufmann. pp. 313-320.

Ernestus, Mirjam and R. Harald Baayen (2003) Predicting the unpredictable: interpreting neutralized segments in Dutch. *Language* 79:5–38.

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.

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011) Heuristic decision making. *Annual Review of Psychology* 62:451–82.

Goldberg, Yoav (2017) *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan and Claypool.

Goldrick, Matthew and Robert Daland (2009) Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology* 26:147–185.

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

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

Irvine, Ann and Mark Dredze (2017) Harmonic Grammar, Optimality Theory, and syntax learnability: An empirical exploration of Czech word order. arXiv preprint arXiv:1702.05793.

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

Jäger, Gerhard (2007) Maximum entropy models and stochastic Optimality Theory. *Architectures, rules, and preferences. Variations on themes by Joan W. Bresnan*, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467-479. Stanford: CSLI Publications.

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.

Jesney, Karen (2007) The locus of variation in weighted constraint grammars. Paper given at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford, CA.

Johnson, Keith (2011) *Quantitative methods in linguistics*. John Wiley & Sons.

Johnson, Daniel Ezra (2009) Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3:359–383.

Jurafsky, Dan (2003) Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod et al. (2003), pp. 39-96.

Jurafsky, Dan and James H. Martin (2020) *Speech and Language Processing* (3rd ed. draft), web.stanford.edu/~jurafsky/slp3/

Kaisse, Ellen M. (1985) *Connected speech: The interaction of syntax and phonology*. San Diego: Academic Press.

Kawahara, Shigeto (in press) A wug-shaped curve in sound symbolism: The case of Japanese Pokémon names. To appear in *Phonology*.

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.

Kluender, Keith R., Randy L. Diehl, and Beverly A. Wright (1988) Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics* 16:153-169

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

Labov, William (1969) Contraction, deletion, and inherent variability of the English copula. *Language* 45:715-762.

Lau, Jey Han, Alexander Clark, and Shalom Lappin (2017) Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science* 41: 1202-1241.

Linzen, Tal and T. Florian Jaeger (2016) Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40:1382-1411.

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*

McCarthy, John and Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Suzanne Urbanczyk and Laura W. Dickey (eds.) *University of Massachusetts occasional papers in linguistics 18: Papers in Optimality Theory*. 249–384.

McMurray, Bob, Richard N. Aslin, Michael K. Tanenhaus, Michael J. Spivey, and Dana Subik (2008) Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance* 34:1609-1631.

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

Mendoza-Denton, Norma, Jennifer Hay, and Stephanie Jannedy (2003) Probabilistic sociolinguistics: Beyond the variable rule. In Bod et al. (2003), pp. 97-139.

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 Joe Pater (2016) Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15: 53-66.

Morrison, Geoffrey S. (2007). Logistic regression modelling for first- and second- language perception data. In M. J. Solé, Pilar Prieto, Joan Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology*, pp. 219–236. Amsterdam: John Benjamins.

Oliveira e Silva, Giselle (1982) Estudo da regularidade na variacao dos possessivos no portugues do Rio de Janeiro. Ph.D. dissertation, Universidade Federal do Rio de Janeiro.

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]

Riggle, Jason. (2004) *Generation, recognition, and learning in finite state Optimality Theory*. Ph.D. dissertation, UCLA.

Rousseau, Pascale and David Sankoff (1978) Advances in variable rule methodology. In David Sankoff, ed., *Linguistic variation: Models and methods*, pp.57-69.

Ryan, Kevin (2019) *Prosodic Weight: Categories and Continua*. Oxford: Oxford University Press.

Sankoff, David, and William Labov (1979) On the uses of variable rules. *Language in Society* 8: 189-222.

Sankoff, Gillian S. (ms.) A quantitative paradigm for studying communicative competence. Paper given at the Conference on the Ethnography of Speaking, Austin, Texas.

Scholes, Robert (1965) *Phonotactic Grammaticality.* The Hague: Mouton.

Smith, Brian W. and Joe Pater (2020) French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5(1)24.

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.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte, and Simon Todd (2017) Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2: 86.1-27.

Tagliamonte, Sali A. and Baayen, R. Harald (2012). Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 34, 135-178.

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.

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.

Wolfram, Walt, and Ralph W. Fasold (1974) The study of social dialects in American English. Englewood Cliffs, NJ: Prentice Hall.

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 (2000) Patterned exceptions in phonology. Ph.D. dissertation, UCLA.

Zuraw, Kie (2003) Probability in language change. In Bod et al. (2003), pp. 139-176.

Zuraw, Kie (2010) A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language & Linguistic Theory*, 28: 417-472.

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

Zymet, Jesse (2018) *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. Ph.D. dissertation, UCLA.