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

Bruce Hayes
UCLA

September 2020

Draft of a shorter paper to be 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 [ various people kept anonymous for review purposes ] and [ ________ your name here ] for helpful input and comments on this project.
1. Introduction: probabilistic phenomena in linguistics

The empirical domain covered in this article is linguistic phenomena in which we need to generate gradience, usually in the form of probabilities, in the analysis. I will employ examples from a variety of fields of linguistics: phonology, syntax, phonetics, historical change, and semantics. Phonology is the starting point for me, since it is my own research area, but I have been intrigued to find evidence that other areas of linguistics work similarly.

I think there are basically three phenomena that ought to be addressed by linguistic theory in terms of probability. First, we frequently need to model cases where alternative surface forms are generated, at varying frequencies, from the same basic or underlying form. In phonology, this is ubiquitous. For instance, in my own variety of English the word Centinela (a street in Los Angeles) is underlyingly something like /ˌsɛntəˈnɛlə/ and can be realized as any of [ˌsɛntsəˈnɛlə], [ˌsɛŋtəˈnɛlə], or [ˌsɛtəˈnɛlə], according to regular phonological processes. The derivation of multiple surface forms from single underlying forms (and the social factors that influence the outcome) is studied intensively in sociolinguistics, discussed in §5.2. We will also see (§5.3) an approach to syntax that derives multiple variants from a single base form.

Second, speakers of a language have the ability to frequency-match statistical patterns in the lexicon. For phonology, this result emerges from early work by Zuraw (2000), Ernestus and Baayen (2003), and others, and has been generally confirmed in experimental work since then. Thus, for instance, when speakers of Hungarian undertake a nonce-probe task testing their intuitions about vowel harmony in their language, their responses statistically match the vowel harmony pattern of the Hungarian lexicon; see §4.2.1 and §6.5 below. In syntax, experiments show that speakers frequency-match the selectional properties of verbs (e.g. suspect prefers S over NP; remember prefers NP over S), and they use this information to facilitate their perception and parsing of sentences (e.g. Jurafsky 2003, Linzen et al. 2016).

Lastly, native speaker judgments are characteristically gradient. These include, for instance, phonological well-formedness judgments: controlled studies suggest that speakers judge some nonce forms as perfect, others as possible-but-imperfect (with all intermediate values), and completely bad; see Scholes (1965) and much later work. For syntax, the literature on this point is very substantial; see Lau et al (2017) for a recent overview and proposal. In phonetics, native speaker judgments are often made on the basis of phonemic classification: what contrasting speech sound (phoneme) does a listener hear when presented with a gradely varying stimulus?

2. Theory

To treat these three kinds of gradience within generative grammar, we need formal frameworks that can generate outputs on a scale, which is most often a probability scale. As a central 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) that uses the same mathematics as the well-known statistical procedure of logistic regression (Jurafsky and Martin 2020). The connection between
MaxEnt as a linguistic theory and logistic regression as a branch of statistics is interesting and will be discussed below in §6.4. As we will see, many scholars have adopted logistic regression as a practical way of understanding complex data, but the agenda here will ultimately concerns as a possible element of formal linguistic theory.

MaxEnt vies with a number of alternative formal conceptions of probabilistic linguistics, which we will address here. The evaluative strategy to be taken here is to think abstractly about these frameworks in a (straightforward) mathematical way. The math is used to locate quantitative patterns characteristically generated under a theory, patterns readily identified visually when we plot the data 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 hope to usefully extend this line of work by offering a particular way of visualizing the signatures that I believe is helpful. I also offer coverage not just of phonology, as in the work just cited, but of multiple fields of linguistics, which converge in a way I have found intriguing.

The discussion covers two related signatures. For each, I will describe the pattern, cite some real-world cases, and demonstrate mathematically which frameworks possess these signatures; this in turn may be taken to reflect on the empirical adequacy of these frameworks. All of my cases will be taken earlier work; I hope to add value here with a consistent exploration of the quantitative-signature idea. I should add that doing this work has pushed me well beyond my usual scholarly limits, and the examples from outside phonology should be taken as more tentative than those within.

I begin with an exposition and review of MaxEnt. This will be more than a quick overview, and this is for a reason: I believe that understanding the functioning of MaxEnt as closely as possible at an intuitive level will help in the task of assessing quantitative signatures and theory-comparison.

2.1 MaxEnt 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 by proposing a set of inputs, sets of candidate outputs for each input, and a set of constraints used to make the choice(s) from among candidates. As in much contemporary computer science, the theory derives its outputs not by following a step-by-step derivational procedure, but by defining the set of all possible outputs (GEN), and defining a metric that selects the best one. The task of finding it is left to theory-external, if interesting, means.¹

The metric of relative goodness that OT uses for candidate selection works like this. The constraints are all strictly ranked, as 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 will determine a unique overall winning candidate, and this is the output of the grammar.

¹ For OT see e.g. Eisner (1997), Riggle (2004).
In probabilistic versions of OT, selection of a unique winner is replaced by some mechanism that assigns a probability to every member of GEN. In some systems, such as MaxEnt, every member of GEN receives a positive probability, but typically the vast majority of candidates get such a minutely low probability as to be in all essentials at zero. In the study of variable phenomena, often more than one candidate receives non-negligible probability, and this serves as an account of gradient free variation or preference, as in §1 above. The probabilities are precise numbers and can be tested against quantitative data from corpora or experiments.

Given its ancestry, MaxEnt OT inherits virtues and flaws of Optimality Theory itself. My personal judgment is that OT was, indeed, a breakthrough: the earlier rule based systems of generative linguistics were complicated, often missed generalizations, and had no way of relating the specifics of language-particular analyses to general principles (in particular, principles of Markedness). Within specific empirical areas, OT has led scholars to develop language-independent constraint sets that shed light on typology as well as providing the basis for integrating language-particular analysis with typological generalizations. For detailed defense of these views, see for instance the original presentation of OT by Prince and Smolensky, or the excellent textbook account in McCarthy (2002)

OT also led, I think, to a broadening of the scope of linguistic theory, particularly in phonology. Notably, constraint-based linguistics, with its simple ingredients, has proven highly suited to the creation or adaption of learning algorithms, making possible substantial, empirically tested proposals on how language might be learned; see Jarosz (2019) for a recent overview. (Indeed, these algorithms have indeed have been deployed throughout the discussion below to find the best-fit analysis for any combination of theory, data, and constraint set.) Further, the constraints OT posits have proven to have applications going beyond that of just mimicking the effect of the older rules: constraints can form the basis of theories of speech errors (Goldrick and Daland 2009), of loanword adaptation (Kang 2003), of sound symbolism (Kawahara, in press), and of phonological effects in sentence formation (Breiss and Hayes 2020).

2.2 The MaxEnt math and its intuitive rationale

Let us return now specifically to MaxEnt. As noted above, it is one way of making OT probabilistic. MaxEnt couples together OT (a sound and effective linguistic framework) with the mathematics of logistic regression (a sound general method of probabilistic reasoning) to achieve a theoretical approach that is both empirically comprehensive and probabilistically effective — or so I would judge.

Math is just math and it does what it does however we may think or feel. However, in the present context I think it is helpful to take apart the math, step by step — for we will find that every step turns out to be intuitive and sensible. Further, this understanding will help later as we examine how the math behaves in actual language examples. Thus, the section that follows is my attempt to present MaxEnt as a mathematicized embodiment of common sense.

The key here will be is to think of MaxEnt as a decision procedure. The constraint violations of the candidates are, in essence, evidence bearing on which candidates ought to win or lose; or more generally, receive high or low probability.
It is best to start by looking at the whole formula, given in (1). Our purpose will be to reconstruct the formula from its parts, pointing out intuitiveness at each stage.

(1) The MaxEnt formula

\[ \Pr(x) = \frac{\exp\left(-\sum_i w_i f_i(x) \right)}{Z} \], where \( Z = \sum_j \exp(-\sum_i w_i f_i(x_j)) \)

This is the formula for calculating \( \Pr(x) \), the probability of candidate \( x \) for some input. The information on the right side of the equals sign embodies (as we will see) everything that is need to calculate this probability for a particular input, including a set of output candidates, a set of constraints, a violation count for each constraint/candidate pair — and one other item, the weights, to be discussed in the next section.

2.2.1 Constraint weights

In a MaxEnt grammar, every constraint is affiliated with a nonnegative number, its weight, which tells you how strong it is; or more specifically, how much it lowers the probability of candidates that violate it. In formula (1), this is \( w_i \) for each constraint \( i \). From the viewpoint of commonsense reasoning, weights are intuitive — we know that reasons differ in cogency. Weights in MaxEnt take on the role played by constraint ranking in classical Optimality Theory.

2.2.2 Multiple violations

In (1), we twice see the expression \( w_i f_i(x) \), 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 in §6.3.1, other theories of probabilistic constraint-based linguistics adopt a different view.

2.2.3 Computation of Harmony

Once we have carried out the multiplication step for each combination of candidate and constraint, we can 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). In formula (1), Harmony is represented by \( \sum_i w_i f_i(x) \), where \( \sum \) is the operator for summing across constraints.

---

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

3 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.

4 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 concept, because when we make rational decisions we find it appropriate to weigh all of the evidence. In this respect, classical Optimality Theory is bravely counterintuitive: in classical OT, the decision between two rival candidates is made solely by the highest ranked constraint that distinguishes them, discarding the testimony of all lower-ranked constraints (for the full original discussion, see Prince and Smolensky 1993:§5.2.3.2). The view taken here is that Prince and Smolensky’s move to discard evidence was indeed brave, but in the end emerges as empirically wrong. This is also true, I believe, with regard to the similar issue mentioned in the preceding subsection, namely ignoring all differences of violation count beyond just relative differences.5

2.2.4 Computation of eHarmony

Once Harmony values are computed for each candidate, they are converted to what Wilson (2014) has called eHarmony.6 This is done by negating Harmony, then taking $e$ (about 2.72) to the result. Thus, in formula (1), the term for eHarmony is: $\exp(-\sum w_i f_i(x))$, where $\exp(x)$ is an abbreviation for $e^x$. The eHarmony function is plotted in (2) below.

(2) eHarmony plotted against Harmony

As can be seen, the eHarmony function induces a rescaling of 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.

Once again, I suggest that this transformation corresponds to 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 likely

5 Classical Optimality Theory finds an echo, perhaps, in the research of G. Gigerenzer (Gigerenzer and Gaissmaier 2011) which emphasizes cases in which people make choices using single salient facts rather than by weighing all the evidence. Gigerenzer’s examples involve real-world decisions (like picking stocks or sending a patient to intensive care) where prediction is difficult and data scarce. The “grammatical decisions” described in this article, however, are very 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 judge they should not be applied to grammar.

6 Wilson was joking (eHarmony is a dating website), but the mnemonic seems useful.
to be ambivalent, and any additional information to inform us more about our choice is welcome and taken seriously. If on the other hand, a candidate is already heavily penalized by the information we have (e.g. probability .001), than even a great deal of evidence may only move us only a small distance, to (say) .0005. And for most people, I suspect, to become absolutely certain requires a vast, perhaps infinite, amount of evidence.

This difference in evidentiary scaling is what the eHarmony mapping accomplishes. Thus, in graph (2), a 50/50 candidate lies (as will shortly be shown), on the steeply sloped part of the curve, so that small differences of Harmony results in relatively large differences of eHarmony; and this will turn out to 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 penalized by a large quantity of Harmony, and increasing this penalty will only move the top candidate’s probability upward by a small amount.

A slogan, perhaps intuitive, that expresses these patterns is: certainty is evidentially expensive: to move probability around when it is already very close to zero or one requires large infusions of evidence. The mapping of Harmony into eHarmony is the way in which MaxEnt implements this principle.

2.2.5 Computation of probability

For the last two steps of the MaxEnt derivation, we sum the eHarmony for all of the candidates assigned to a given input, and call this sum Z. In formula (1), this is expressed as: \( \Sigma_j \exp(-\Sigma_i w_{ij}(x_j)) \), where \( j \) is the index intended to denote candidates. Then, the probability of a candidate is its eHarmony divided by Z; i.e. its share in Z; once we have carried out this division, we get the full formula for \( P(x) \) as it appears in (1).

I offer two comments on the addition-then-division procedure. First, it is intuitive, since it says that a candidate is less likely if it has strong rivals. Second, 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, which are merely scaled by the factor Z, retaining their relative magnitudes.

Summing up, the MaxEnt computation is claimed here to be intuitive at every stage, as summarized in (3).

(3) Summary: MaxEnt and common sense

a. Constraints differ in their evidential force (§2.2.1).

b. Multiple violations of the same constraint are predicted to make a candidate less probable (§2.2.2).

c. All evidence from the constraints is duly considered in proportion to their weights; and no evidence is thrown out (§2.2.3).

d. Evidence is scaled to make it have less effect as we approach certainty (§2.2.4).

e. Candidates become less probable when they compete with powerful rivals (§2.2.5).
To the extent that the reader agrees that these five properties reflect sensible principles of arriving at conclusions from evidence, the MaxEnt framework (or any framework that has these properties) can be said to have a certain a priori claim on our attention.

Obviously, there is much more to say about MaxEnt/logistic regression from the technical point of view, and in this context I will only suggest some readings. On 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 as an area of 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 (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).

With this background we can turn to the main topic: quantitative signatures, their derivation under different theories, and their distribution in the real world. As noted earlier, we will first explore the signatures under MaxEnt, then later (§6) under alternative theories.

### 3. First quantitative signature: the sigmoid curve

The cases we will focus on here are in one sense very simple: we suppose that for an input there are just two viable output candidates. In Optimality Theory and its descendents, this means that all other conceivable candidates are ruled out by very powerful constraints. This is quite normal in OT, and I will not bother with formulating the necessary constraints below. The two viable candidates compete with other, on the basis of the less-powerful constraints that they violate.

Let us suppose that one of these constraints is one that is violated either once or not at all; let us call it ONOFF. Let the other be a constraint, or a set of constraints, defining a *scale*. Scales are a familiar item in constraint-based linguistics (see e.g. Prince and Smolensky (1993/2004 §5.1) and de Lacy 2004); and linguists have has repeatedly developed analyses in which the basic scale-defining ingredients are formalized either with single constraints (multiply violable), or else families of related constraints. The three cases in (4) below include both types.

(4) *Some gradient constraints or constraint families in linguistics*

a. **Phonology**  
The vowels of a language vary in how likely they are to trigger vowel harmony; one constraint for each vowel (Hayes and Londe 2006, and §6.5 below).

b. **Syntax**  
In English genitive constructions, *N of NP* is favored over *NP’s N* to the extent that *NP* is long (Szmelesanyi et al. 2017, and §5.3). One gradient constraint, with violations corresponding to word count.

c. **Semantics**  
The quantifiers of a language can be arranged on a continuum for their tendency to take broad scope (AnderBois et al. 2012 and §5.6); one constraint per quantifier.
Let us deal with the simplest case first: the scale at hand is defined by one single constraint, to be called \textit{VARIABLE}, with multiple violation levels, ranging here (for concreteness) from 1 to 7. Assume as above a candidate competition with just two viable candidates per input, and that one of the two obeys \textit{VARIABLE} and violates \textit{ONOFF}, while the other obeys \textit{ONOFF} and violates \textit{VARIABLE} some specified number of times, which will depend on the particular input chosen. Under this setup, we can plot a probability function, where on the horizontal axis we have the number of violations of \textit{VARIABLE} across inputs, and on the vertical axis the probability that the candidate that violates \textit{VARIABLE} wins. For clarity, I will be plotting the function for \textit{all} values on the horizontal axis, not just the integers 1-7 that would occur for particular input forms.

The curve that MaxEnt derives under these conditions is a \textit{sigmoid} (S-shaped) function. In (5) I give an example.

(5) A sigmoid curve generated in MaxEnt

\begin{center}
\begin{figure}
\begin{center}
\includegraphics[width=0.8\textwidth]{sigmoid_curve.png}
\end{center}
\end{figure}
\end{center}

Here are the crucial properties of the MaxEnt sigmoid, which is often called the “logistic” function.\footnote{The logistic function was named in 1845 by its discoverer, the Belgian mathematician Pierre-François Verhulst. No one knows why he chose this name. For helpful discussion of the history of the function and of logistic regression, see Cramer (2002).} (a) It is fully symmetrical, and the symmetry point falls where probability crosses the 50\% line — in (5), this is at 4 violations of \textit{VARIABLE}.\footnote{Further, this point is determined by the constraint weights: it falls at the value \textit{VARIABLE}/\textit{ONOFF}. In making the graph in (5), I used the weights \textit{VARIABLE} = 8, \textit{ONOFF} = 2, so the symmetry point is at 4.} (b) It asymptotes on either end at 1 and 0. (c) It is steepest at the symmetry point, and becomes ever more level as one proceeds in the positive or negative direction. (d) When derived in a MaxEnt grammar, the uphill/downhill orientation of the logistic depends on whether the constraint weight of \textit{VARIABLE} is positive or negative; and its steepness is greater when the weight of \textit{VARIABLE} is larger. (e) Again assuming MaxEnt, the relative right/left position of the curve is determined by the weight of \textit{ONOFF}. 

\begin{itemize}
\item \textit{VARIABLE}/\textit{ONOFF}
\end{itemize}
These and other mathematical properties of the MaxEnt sigmoid are derived in reader-friendly form in McPherson and Hayes (2016; Supplemental Materials). We will keep them in mind here because we will be assessing in qualitative terms of whether an empirically-observed curve is properly to be considered as a sigmoid. Thus, markedly asymmetrical curves, or curves that asymptote at a value other than one or zero, or curves that are at some point convex or concave (shifting the sign of the 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 should also expect sometimes to see truncated sigmoids, and these will indeed show up later on.

3.1 Sigmoid curves are ubiquitous in language data

I suggest that those who collect and plot quantitative linguistic data — in all areas of grammar — frequently encounter sigmoid curves. The bulk of the cases I will cover here actually involve multiple sigmoids appearing on the same axis; and we will start focusing on these in §4 below. But as a warm-up exercise, let us start with a representative singleton sigmoid. It will be helpful to begin with a case in which the horizontal axis of the sigmoid is totally uncontroversial, being a physical quantity rather than an analytic construct, as in later practice. Phonetics provides many good cases.

3.1.1 A sigmoid from a phonetic experiment

Within phonetics, speech perception studies have for decades revealed the presence of sigmoid curves. A typical case would arise if we plot on the horizontal axis a phonetic parameter such as 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 [p]. Kluender et al. (1988) report a well-executed experiment of this kind, in which their data emerged as an approximate sigmoid. A subset of their data (more to follow) is replotted in (6).

(6) Sigmoid curve relating closure duration to voicing percept, adapted from Kluender et al. (1988)
I assume the reader’s agreement that the curve in (6) is a reasonable approximation to a sigmoid, and that small deviations from this ideal may be attributed to sample or measurement error; the same holds for the remaining graphs in this article. We turn, then, to the task of analysis and modeling.

For purposes of this article, it will be useful to adopt an interesting stance proposed and developed by Boersma (1998): that speech perception should 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 probabilistic framework he uses to do this (not MaxEnt) is discussed below in §6.3.1.

Pursuing Boersma’s imperative in MaxEnt terms, we can arrange our grammar as a simple target-and-penalty system. The grammar will input closure duration values and select between the percepts [b] and [p]. As before we exclude all other percepts by fiat; in a real grammar, they would be ruled out by constraints with very high weights, giving them 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 ideal target value for voiced [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 with each other, and the value computed by MaxEnt will depend on the particular state of this conflict for a particular number of milliseconds of closure duration in the stimulus.

Using a spreadsheet, it is easy to find the weights that produce the most accurate model for the Kluender et al. data. The best-fit weights for VARIABLE and ONOFF turn out to 0.088 and 4.34, respectively. From this, we can calculate the probability of the voiceless candidate as 60 msec. using the MaxEnt formula (1). Table (7) carries this out step by step.

(7) Calculating the MaxEnt probability of one data point from Kluender et al. (1988) (probability of [b] for closure duration 60)

| 60 − 20 = 40 | Violations of VARIABLE for the input 60 msec., candidate [b]: deviation from target at 20. |
| 1 | Violations of ONOFF for candidate [p] |
| \(H_b = 0.088 \times 40 = 3.53\) | Multiply violations by weights to obtain Harmony. |
| \(H_p = 4.34 \times 1 = 4.34\) |

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 some discussion, see §6 below, and for presentation of standard techniques see the references given above at the end of §2.2.5.

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 very 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 online Gallery.
Take e to the minus Harmony to obtain eHarmony.

Add eHarmony across candidates to obtain Z.

Divide eHarmony for [b] by Z, obtaining its predicted probability.

The value obtained, .69, is felicitously close to the observed value of .68; the predictions for other data points are decent but not always so close. Such deviations from prediction are expected, given that the number of participants in the experiment was only 16 (sampling error) and errors arising from the experimental setup are also essentially inevitable.

Performing similar calculations on all the data, I obtained the results plotted in (8).

(8) The Kluender et al. (1988) data replotted, superposed with model predictions

Along with the data, I have plotted the sigmoid that represents the model predictions, not just for the observed cases but smoothly, for any value in the relevant range; this helps make it clear what the underlying theory is saying, even at values where it has not been checked against a real-world observation.

The units of the lower horizontal axis in (8), 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 just a single number, which is the difference of Harmony between the two rival candidates. For how this works, see Appendix. The practice is helpful, since it means we can encapsulate all of the relevant analytical information as a single value on the x axis, as I did in figure (8). To give a concrete example, the fifth data point from the left is the input with closure duration 60 msec., whose predicted value for percept [b] was calculated above in (7) as .69. There, we saw that $H_b = 3.53$ and $H_p = 4.34$, so $H_b - H_p = -0.80$. The fifth dot from the left in (8), representing the actual observation, is plotted at ($-0.80, .68$), a tiny bit below the predicted value at .69.
We will return to MaxEnt in phonetics later, but one point specific to phonetics should be made here: the MaxEnt math, under its logistic regression name, is a standard tool for phoneticians who do experiments with the (now-classical, still-active) paradigm exemplified by the Kluender et al. experiment. In modeling their data, phoneticians characteristically use the math of MaxEnt to provide a mathematical characterization of these sigmoid curves. A helpful guide to this work is provided in Morrison (2007).

4. A more complex quantitative signature: the wug-shaped curve

We return to our main goal, the evaluation of theories through their quantitative signatures. As before, we use MaxEnt as our basic mathematical language. The scenario to be considered now 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. In real life this happens all the time, and offers interesting analytic angles.

Let us first establish the predictions that MaxEnt makes. The subpopulation of candidates that violate PERTURBER will have their harmony values increased or decreased, depending on whether PERTURBER is “allied” in effects 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 (5) above, 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 outcome is illustrated schematically in (9).

(9) The double sigmoid resulting from the presence of a PERTURBER constraint

As Dustin Bowers once suggested to me, it is not hard to imagine in this shape the perky creature who in recent years has been widely adopted as the emblematic animal of linguistics; since it is brief and vivid I adopt Bowers’s suggestion of wug-shaped curve (honoring its inventor, Berko 1958), and giving my artistic rendition of it below.
(10) *The wug-shaped curve*

The degree to which the curve will resemble its Berkovian original will depend on how the weights are set and the curve is plotted. Most saliently, high vs. low values of PERTURBER will correspond to fat and skinny wugs, respectively.

4.1 *Multiple perturbers*

Nothing is stopping us from having more than one Perturber; and indeed there exist many empirical cases, discussed below. Assuming distinct weights, each Perturber will define a separate sigmoid, which will be spaced as the weights dictate. It is tempting to think of this as a stripey wug, or indeed as a flock of wugs marching in parallel,11 but I will simply use the term “wug-shaped curve” for these cases as well.

(11) *The wug-shaped curve with multiple sigmoids*

4.2 *Wug-shaped curves in the real world*

My own initial empirical involvement in wug-shaped curves arose from my 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 §6. I will first go through the three cases we studied, since the analytic work was done in MaxEnt and no adaptation at all is needed to the present context. Later, the discussion will expand to other people’s work and other disciplines of linguistics.

A caveat for all of these examples is that for reasons of space I can only give the sketchiest outline of the data pattern, or of the VARIABLE, ONOFF, and PERTURBER constraints employed. For full detail the reader should consult the original articles, as well as an online supplement I

---

11 Thanks to an anonymous talk audience member for this appealing image.
have created for this article, the Gallery of Wug-Shaped Curves, posted at https://linguistics.ucla.edu/people/hayes/GalleryOfWugShapedCurves/. This site includes the spreadsheets I used to recalculate the analyses and plot the data.

4.2.1 Hungarian Vowel Harmony

In Hungarian, most suffixes have two variants, one with a phonetically back vowel and the other with a front vowel (e.g. [-ńk]/[-ńk] for the Dative). The choice of variant depends on the phonological properties of the stem to which the suffix is attached. The most important factor in determining the choice of suffix variant is the character of the last vowel, or last few vowels, of the stem. Since typically the suffix comes to agree in phonetic backness with a stem vowel, this is a case of vowel harmony. Hungarian vowel harmony has been carefully studied for decades; analyses using MaxEnt appear in Hayes and Londe (2006), Hayes et al. (2009), and Zuraw and Hayes (2017).

To understand Hungarian vowel harmony we will adopt the standard taxonomy of Hungarian vowels, namely Back (B), Front Rounded (F), and Neutral (N). To give just a subset of the harmony principles, suffixes must always be front after a vowel of class F, always back after B; and the individual vowels of the N class ([i, iː, eː, ɛ]) give rise to a variety of probabilistic environments. In the references just cited, a suitable family of constraints is put forth to predict vowel harmony from these vocalic factors. These constraints have internal justification from both typological and phonetic study (Kaun 2004). Together, they will serve as our source of baseline harmony, the equivalent of VARIABLE in (8) above, but with the work dispersed among multiple constraints.

Our Perturber constraints arise from a surprising discovery made by Hayes et al. (2009) (using corpus and experimental data): Hungarian vowel harmony is probabilistically influenced by consonants. For example, if a stem ends in two consonants, it is more likely to take front suffixes than otherwise. The same is true for stems ending in a bilabial stop, or a sibilant, or a coronal sonorant; see Hayes et al. (2009:386) for details. In order to obtain enough data points to justify the data plots, I will follow Zuraw and Hayes (2017:533) in reducing the scheme of Perturber violations to a three-way system, namely total violations of 0, 1, or 2 Perturber constraints.

To sum up, the system involves two constraint families, a Baseline family that expresses traditional wisdom about the effect of vowels on harmony, and a Perturber family (three values) that adds in the effects of consonants. In (12) below I replot the pattern of responses from Hayes et al.’s (2009) nonce-probe experiment, superimposing it on the predictions of the best-fit MaxEnt grammar.
(12) The wug-shaped curve in Hungarian vowel harmony

On the horizontal axis the data labels represent the various sequences of stem vowels that are distinguished by the vowel-governed constraints, using the B/N/F taxonomy. It can be seen that the Perturbers duly separate out the three levels of consonant influence, and that the three sigmoids are essentially parallel.

A particular aspect of this wug shape, which we will see throughout our set of examples, is that the magnitude of the effect of the Perturbers depends on where we are located on the scale of Baseline harmony: it is maximal in medial position and diminishes gradually to essentially nil at the peripheries. This pattern, already pointed in Zuraw and Hayes (2017) and Smith and Pater (2020)¹² is, as can easily be shown, a direct consequence of the MaxEnt formula. From the perspective of §2.2.4, this finding is a sensible and intuitive one: since certainty is evidentially expensive, the evidence from a Perturber will buy you a lot in the middle, where you are uncertain, but will buy you very little at the peripheries, where you are already close to certain.

4.2.2 French liaison and elision

This example concerns two-word constructions in French; call them Word1 + Word2. A variety of words when occurring in Word1 position have two surface forms, one deployed (approximately) before words beginning with the consonant, the other before words beginning with a vowel; thus the feminine definite article: [la kwiːʒɛt] ‘the zucchini’, but [lɔbɛʒɛ̃] ‘the eggplant’. The pattern is complicated by cases of Word2 that, for historical reasons, start phonetically with a vowel but nevertheless take the pre-consonant version of Word1: [la ʁɛ] ‘the axe’. Although classical French generative phonology treated such cases as an either/or matter,

¹² 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.
corpus and other evidence shows that the propensity of Word2 types to take the vowel-appropriate versions of Word1 is gradient, forming a scale. This scale, though continuous, is approximated by Zuraw and Hayes by forming the set of words into a hierarchy of five groups, and the constraints that implement this scale determine Baseline harmony in the data plot given here.

The Perturber constraint set likewise reflects gradient lexical propensity, namely the preference of a Word1 for its prevocalic allomorph; again invoking groups of similar words such as \{au, de, ma\}. In (13), I have replotted the Zuraw/Hayes data along with the sigmoid curves that model them in MaxEnt.

(13) The wug-shaped curve in French Liaison

Once again, we see that the vertical spacing of data points that share the same base harmony is closer when the base harmony is closer to its extreme values, a key aspect of this quantitative signature.

4.2.3 Tagalog Nasal Substitution

The phonology involved, worked out in detail in Zuraw (2000, 2010), is as follows. The sound \[\text{ŋ}\], when it appears at the end of a suffix, often merges with a following consonant to create a blended consonant, which preserves the nasality and voicing of the prefix \[\text{ŋ}\], along with the place of articulation of the stem consonant; thus \(/\text{ŋ}+\text{p/} \rightarrow \text{[m]}, /\text{ŋ}+\text{t/} \rightarrow \text{[n]}, \text{etc. The process is lexically optional, applying on a word-by-word basis, and the intricate pattern of application emerged when Zuraw calculated the application rates from a language-wide corpus. Further, when Zuraw did a nonce-probe study, Tagalog speakers in their responses roughly approximated the pattern, suggesting they are productively internalize it.}

In the MaxEnt analysis, the Baseline constraints form a family, each of which forbid NC clusters with various features (place, voicing) for the C position. This family, all of whose members receive different weights in the best-fit analysis, distinguishes among six categories:
\{p, t/s, k, b, d, g\}. Points corresponding to these categories can be identified by consulting the labels just above the graph in (14) below.

As for Perturbers, it was shown in Zuraw’s earlier corpus research that each [ŋ]-final prefix in Tagalog has own distinct propensity to induce mutation; this is formalized using the constraint weights for a family of prefix-specific Perturber constraints. On the horizontal axis of (14) below is plotted the baseline Harmony resulting from the consonant-specific constraints, each militating against retaining the unchanged NC of various kinds, 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 more data, the more trustable that data point, so it is plotted larger.

(14) 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). This wug-shaped curve, with a broad range of base harmony and six distinct sigmoids, is the most substantial I know of.

5. 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 the classic works of probabilistic linguistics and replotting their data (helpfully provided by the authors) with arrangements of Baseline harmony 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 I made choices for partitioning the constraints into Baseline and Perturber sets, I favored a Baseline family that would yield a broad probability range. I also favored partitions that gave the Perturber set (and where possible, both sets) a unified, intuitively distinct rationale.

Everywhere I went, I found wug-shaped curves. I will proceed through six areas of linguistics: phonology (done by other people), sociolinguistics, syntax, phonetics, historical linguistics, and semantics.

5.1 Earlier work in phonology

I restudied two seminal works of probabilistic phonology to see if their data might reveal the wug-shaped curve pattern.

5.1.1 Ernestus and Baayen (2003) on Undoing Dutch Final Devoicing

Ernestus and Baayen’s work was, along with Zuraw’s, the first demonstration of a point that by now is well-replicated: that speakers of a language tacitly know in some detail the statistical phonological patterns present in their lexicon, and roughly duplicate this pattern when they take a nonce-probe test. Ernestus and Baayen were also the first to use a MaxEnt model to model such patterns. Specifically, they used corpus and experimental data to demonstrate that Dutch speakers can use lexical knowledge to predict the underlying value for [voice] for consonants at the end of a word, a context where in Dutch this value is not actually phonologically realized, being neutralized to voiceless. Thus, hearing a nonce stem ending with voiceless [x], Dutch speakers can guess that it is very likely to appear in a suffixed form with voiced [y], since this is the predominant pattern in the lexicon of Dutch for stems of this class. Dutch speakers can also match the likelihood that phonetic [p] will alternate or not with [b], and similarly for the pairs [f] ~ [v], [t] ~ [d], and [s] ~ [z]. Further, speakers are also tacitly aware of environmental effects: that of a preceding sonorant or obstruent consonant, and that of the length of the closest preceding vowel.

As Baseline constraints I adopted from Ernestus and Baayen the ones regulating voicing by consonant pair, as just given; including as well the constraints representing the effect of a preceding sonorant or obstruent; the top axis labels in (15) abbreviate these configurations. For Perturbers, I adopted the constraints whereby particular types of vowels encourage voicing according to their length: long vowels encourage it the most, high tense [i, u, y] (which are known to have an ambiguous status in Dutch phonology) are intermediate, and short vowels encourage voicing the least. The resulting data plot, somewhat noisy but still identifiably wug-shaped, is given in (15).
A similar curve, not shown here, emerges if one selects the consonant-specific constraints as the Perturbers and the others as Baseline.

5.1.2 Anttila (1997) on Finnish genitive plurals

This paper is notable as one of the first attempts to employ constraint-based grammars to treat quantitative corpus data. This work proved highly influential, with repeated efforts to restudy and reinterpret Anttila’s striking findings. The empirical pattern is based on the Finnish genitive plural suffix, which surfaces as either [-iden] or (approximately) [-jen], depending on prosodic properties of the stem to which it is attached. Subsequent analysts have tried to diagnose the pattern by trimming it down to its essence.\(^{13}\) Goldwater and Johnson (2003) found that several of the basic stem shapes involved share the relative number of violations in the constraint set and can thus be pooled (they receive the same prediction), leaving just eight patterns to deal with. I confirmed this, and also found that it was possible to trim the constraint set quite a bit, keeping just four constraints from the original work while still deriving an appropriate fit. In particular, the data require a probabilistic ban on consecutive light syllables, a very strong ban on stressless heavies, and a gradient series banning stress on light syllables that applies, with greater successive stringency, when the vowel is nonlow, or high (for grounding of such constraints in phonetics and in phonological typology, see Gordon 2007).\(^{14}\)

---

\(^{13}\) N.B.: Such simplifications are diagnostic only, since many of Anttila’s constraints must be retained for the role they play elsewhere in Finnish phonology.

\(^{14}\) Only two constraints are needed for the three-way vowel height continuum. As is often true in MaxEnt, one can simply remove the weakest of a family of related constraints with no loss of accuracy; a behavior that will be familiar to researchers who do regression modeling.
On replotting, I found that the Antillean pattern is a sparsely-populated wug-shaped curve, with a sigmoid for each height value. In addition, below the wug is a horizontal line, reflecting the outright ban on [-den] when its use would create a stressless heavy syllable.  

(16) Anttila’s (1997) genitive plural pattern as a sparse wug

As noted earlier, the closer spacing of the three vowel heights in the tail of the wug is a characteristic pattern of all wug-shaped curves.

5.2 Sociolinguistics

Sociolinguists were, to my knowledge, the first theorists to propose and deploy sophisticated probabilistic models in linguistics; and the search for improved probabilistic models continues among sociolinguists today (§6.4). MaxEnt first appeared in sociolinguistic work in the late 1970s. Here, I will focus on some of this early work, in which the conception of Perturbers originated and came to be understood in MaxEnt terms.

5.2.1 Labov (1969)

This influential article covered the syntactico-phonological processes that lead to the contraction and deletion of various forms of the verb be in vernacular American Black English. Labov observed several Perturber effects in his data, and noted how they exert an influence across the board in a systematic way. Labov’s findings were soon analyzed in a probabilistic model by Cedergren and Sankoff (1974), who employed a multiplicative system (§6.2) that was an immediate predecessor within their line of research to MaxEnt. The effectiveness of this kind of modeling quickly became apparent, especially after the kinks were worked out with the model, replacing the multiplicative system with MaxEnt (a.k.a. logistic regression, Rousseau and

---

15 This being MaxEnt, the bottom curve would indeed rise up in sigmoid fashion like the others, but only at extremely low values for Baseline harmony that don’t actually occur in the data.

16 This brief section on sociolinguistics dips its toes into an ocean-sized research literature. For useful orientation to the latter, see Chambers and Schilling’s (2013) encyclopedia and Mendoza-Denton et al. (2003).
Sankoff 1978), and MaxEnt thus became the basis for a large literature. The use of MaxEnt by sociolinguists was greatly facilitated, even in a time of primitive computers, by Rousseau and Sankoff’s widely distributed “VARBRUL” program. The quantitative modeling of variable phonology in sociolinguistics has continued to evolve since; see §6.4 below.

Figure (17) is based on my recalculation of Labov’s (1969) Black English findings, with the data imported from Cedergren and Sankoff’s paper. The constraints I chose as Baseline collectively reflect the factors that occur to the left side of the targeted copula, and involve a standard phonological Markedness constraint (avoidance of complex syllable codas, arising in Boot is → Boot’s) as well as a preference for (what I suspect to be) lexically-listed portmanteaux forms based on pronouns, such as he’s. The Perturbers reflect the differing toleration for contraction of four different right-side syntactic environments. The presence of a wug-shaped curve—in this case just the tail—should be plain; once again, the effect of the Perturbers is weaker at the periphery.

(17) The wug-shaped curve in Labov’s (1969) Black English contraction data

5.2.2 Bailey (1973)

Bailey (1973:106), working only slightly after Labov, was the first linguist, as far as I know, to identify a wug-shaped curve. He had no way of mathematically formalizing it, though this was to come soon. Bailey took his main example from research by Gillian Sankoff (ms., 1972/1978) on Québec French. The phenomenon Sankoff studied was optional deletion of the phoneme [l] in function words. I have taken the Baseline constraints to be (a) a general Markedness constraint disfavoring the realization of [l]; (b) lexically-specific constraints of the

---

17 His 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 qualitative point: “the statistics are more bunched in the bottom and top percentages and more spread out toward the middle percentages” (p. 106). Bailey also identified and plotted single sigmoids (p. 77).
MAX family,\(^{18}\) militating against [l]-loss in particular function words (the tendency of function words to have morpheme-specific behavior has been long known, see Kaisse 1985). 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.\(^{19}\) Thus in the present case, sex and social class are treated as proxies for the varying weight of MAX(l). The Bailey/Sankoff curve is given with my MaxEnt replotting in (18).

(18) *The wug-shaped curve in Quebec French [l]-deletion, after Sankoff and Bailey*

\[\text{(18)}\]

5.2.3 *A remark on how MaxEnt is used in classical sociolinguistics*

The use of logistic regression in the classical Labovian/Sankovian framework depended on (to describe it anachronistically) an interesting hybrid, which blended classical rule-based phonology (from *SPE*; Chomsky and Halle 1968) and 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 the MaxEnt 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.

\(^{18}\) MAX, penalizing deletion, is a key constraint family in the standard theory of phonological Markedness constraints, proposed by McCarthy and Prince (1995).

\(^{19}\) The response of phonology to social context is a vast research area, and the essays in Part III of Chambers et al. (2004) 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.
Empirically, this remains a viable model, but I tend to feel it is unnecessarily complex. The constraints themselves, just slightly amplified, can cover the whole grammar without the use of rules, creating a theory that is simpler and also more responsive to typology. Indeed, the whole debate of rule-based vs. OT phonology has its sociolinguistic version, with variable rules stacked against probabilistic versions of OT; and many of the arguments favoring OT (see §2 above) carry over to this domain. I note finally that the theory of variable rules offers the analyst more freedom, since the weights of the Perturber constraints are set separately for every 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 defensible by adducing cases where this freedom is genuinely necessary.

5.3 Syntax

Many studies in syntax have engaged with gradience of the three types described in (4), 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 research programs that, while not incorporating an explicit probabilistic component, nevertheless engage with the key theoretical concept of Harmony; specifically in supposing that violations of distinct syntactic principles assess a particular, consistent, and additive decrement of well-formedness, as assessed experimentally; this is a research result of both Featherston (2005, 2019) and Keller (2000, 2006).

One way to deal with the great complexity of syntax is to focus on a microdomain, i.e. 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 the probabilistic factors that favor one or the other outcome. This has been for some time a research focus of Bresnan and her colleagues, who use MaxEnt and similar statistical tools in their work. These studies show that choices such as the NP NP/ NP PP one just given are, as it were, semipredictable, provided one uses an appropriate statistical model to make the prediction. Moreover, the refined distinctions predicted by the constraint weights can be assigned a compelling reality, in that they show up as clear if modest distinctions between closely related dialects, such as New Zealand and American English; and these distinctions emerge in experimentation (Bresnan and Ford 2010) as well as corpus work.

A recent article in this tradition, Szmrecsanyi et al. (2017), uncovers dialect-specific patterns for four dialects of English (U.S., U.K, Canada, New Zealand) for two syntactic choices; the dative one just mentioned as well as the genitive choice between, e.g., the king’s palace and the palace of the king. In the graphs below, I will abstract away from these differences and merge the (web-posted) 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 which depend 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 on the data points, the reader will need to consult the original paper as well as the spreadsheet in the online Gallery cited above). For the Perturbers, I selected
the constraints and data series that single out the three common categories of the theme NP (that which is given): indefinite full NP, definite full NP, and pronoun. The curves that emerge under this re-plotting are shown in (19).

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

In the same article, the genitive data are of particular 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 word count; this is part of a long-known principle of syntax that longer constituents tend to be placed rightward. In (19) below, we take this constraint, something to the effect of *NP’s N if NP LONG, as the Baseline constraint, and all other constraints from Szmrecsanyi et al.’s analysis as Perturbers. In the chart, the data series Group 1 through Group 4 are arbitrary, as they simply group together similar regions of Perturber harmony to permit the plotted points to reflect similar numbers of data; again see the online Gallery for details.
(20) The effect of a single Variable constraint (possessor length) in expressing English genitive constructions, after Szmrecsanyi et al. (2017)

The key point is that, although the gradient length constraint is of insufficient weight to induce a full-size sigmoid at any one level of Perturber harmony, it does seem that we are seeing “sigmoid snippets,” each reflecting the gradient constraint as it shows its effects against a particular Perturber harmony level. The same data yield a figure (not shown) similar to the dative figure (19) when we replot the data in the opposite way, with the length treated as the Perturber.

5.4 Phonetics

Let us return to the Kluender et al. (1988) example from §3.1.1 above. There, for expository simplicity, I plotted only one of the two data series from this paper. In fact, the authors’ research interest lay elsewhere, namely in the influence of a Perturber, the length of the vowel preceding the [b]/[p]. Their sensible 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 chart (21) below.
The MaxEnt grammar that underlies (21) is like the one for (6), except that it includes a Perturber I will call *VOICED PERCEPT AFTER SHORT VOWEL; it penalizes the [b] candidate for 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 combined data emerged as a clear if skinny wug.

In the context of speech perception work, my “reanalysis” of the Kluender data is no such thing, because logistic regression is a standard tool for interpreting experiments of this kind. In particular, it is one way to quantify the influence of the Perturber: with straightforward application of the MaxEnt math we can rescale Perturber harmony as milliseconds it emerges that the strength of vowel about 9.5 msec, in rough agreement with what Kluender et al. found (8.4 msec.) using a somewhat different method.

The repeated successful use of MaxEnt by phoneticians in modeling their perceptual data has probably created the largest population of wug-shaped curves in the linguistics literature. If, as Boersma suggests, speech perception is the product of a grammar, then it is plausible to suggest that it is a MaxEnt grammar.

5.4.1 Phonetics II: Production

In principle, it would be nice to cover the opposite side of theoretical phonetics, namely the theory of how phonological categories are rendered physically in speech production. Here, the pioneering work was that of Liberman and Pierrehumbert (1984) and Pierrehumbert and Beckman (1988), who used rule-based systems; this tradition has more recently been revived and continued by scholars whose phonetic grammars have made use of constraints. Harmonic
Thus, we can ask if wug-shaped curves can be found at the production end of phonetics. It would be nice if they could, but there is a strong impeding complication: given the state of (apparently random) variation that is found in any phonetic measurement experiment of sufficient scope, it would seem appropriate for any phonetic model at a detailed level to predict probability distributions of quantitative outputs on some physical scale (such as fundamental frequency, vowel formants, or duration). The simple tracking of a two-choice output frequency, in response to a range of Harmony values, as we can do for so many areas of linguistics, is not and available choice for modeling.

This is not to say that quantitative signatures are not present in the field of production phonetics. Notably, Flemming (2001) derived the famous “locus effect” (Sussman et al. 1993), found everywhere in phonetic data, from a simple postulate (parabolic violation profiles) about phonetic constraints in a harmonic grammar. Lefkowitz (2018), employing hemiparabolic violation profiles in a MaxEnt duration grammar, mathematically derives a number of quantitative signatures that are supported in his experimental data. Thus, I think the future for seeking quantitative signatures of probabilistic theories of phonetics is bright, though these will be different signatures and most of the research has not been done yet.

5.5 Historical linguistics

Kroch (1989) employed the method of inspecting old texts across time in order to track the relative frequencies of syntactic variants as a language gradually changes. For example, 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 (22) below is a chart, adapted from Kroch, which shows the frequency with the definite-article variant is employed.

---

20 Further afield, much of the vast literature of Articulatory Phonology (Browman and Goldstein 1986 et seq.) also falls into this domain. However, my own (diffident) belief is that a mishosen set of initial guiding assumptions (the choice of mass-spring systems as the computational basis) has impeded the search for authentic quantitative signatures in this theoretical approach; see Windmann et al. (2015) for phoneticians who likewise advocate expanding the range of possible mathematical working assumptions for theoretical phonetics.

21 For general background on probabilistic modeling of language change, including more detailed discussion of the material treated here, see Zuraw (2003).
My own replotting of the data, which as Kroch noted, clearly form a sigmoid, is based on my spreadsheet replication of the MaxEnt calculations Kroch carried out in the 1980s. The lower horizontal scale of (22) gives the harmony-difference values ($H_{\text{article}} - H_{\text{noarticle}}$). 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 very 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 about 1.02 Harmony units per century.

Kroch’s “constant rate” hypothesis becomes even more interesting when we include, as Kroch did, some Perturbers. The basic assumption, in present terms, is that the Variable constraint is steadily increasing its weight over time, but the Perturbers are diachronically stable. Such conditions will give rise to a diachronic wug-shaped curve, with identical sigmoids spaced apart by an amount corresponding to the differences in the weights of the Perturbers. This is what Kroch and his co-workers have repeatedly found in their historical studies. The result is a pleasing one: the variation in rates of change across contexts may look nonsensical when measured directly as probability, but they are coherent and orderly when measured in their natural linguistic units of Harmony.

A recent study illustrating these ideas is Zimmermann (2017), who traces the evolution of have in English from an auxiliary to a main verb. This change shows up in four different contexts, listed in (23).

(23) Four contexts in which have can be either an aux or a main verb

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>As Aux</th>
<th>As Main Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>I haven’t any</td>
<td>I don’t have any.</td>
</tr>
<tr>
<td>Inversion</td>
<td>Have you a penny?</td>
<td>Do you have a penny?</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>You have a flair; you really have.</td>
<td>…do.</td>
</tr>
<tr>
<td>Adverbs</td>
<td>He has already the approval of the nation</td>
<td>… already has</td>
</tr>
</tbody>
</table>
Each context may plausibly be assumed to be affiliated with additional constraints acting as Perturbers. The diachronically-shifting constraint governing whether possessive *have* acts as an Aux or main verb serves as the VARIABLE constraint. Zimmermann, tracing each phenomenon across two centuries, obtains the wug-shaped curve in (24). The left panel depicts the sigmoid curves found in the study in schematic form; for full details see Zimmermann (2017:107).

(24) 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* obtained in the MaxEnt model. These lines are straight, and parallel. The parallelism of the lines is not an automatic consequence of the analytic method, but an empirical finding; since the four sigmoids were fitted separately. This depiction of the data mostly clearly illustrates what is meant by “constant-rate hypothesis” in this research tradition.

My literature search suggests that the approach established in Kroch’s (1989) study has been very influential, and it would be impossible to survey here all the studies it has spawned. Blythe and Croft (2012:279=280) offer a table containing dozens of cases of language change that involve sigmoid curves, along with citations of ample supporting literature. There has also been intriguing work addressing the hard question of why language change should typically show a constant rate (as measured with Harmony); see in particular Blythe and Croft (2012) and Stadler et al. (2016). A perhaps overly-concrete interpretation of this work is as follows: language changes that progress steadily are the work of adolescents, who are in the process of fixing their grammars into adult form. These crucial individuals take as their role model speakers somewhat older than themselves, but they exaggerate to some degree the ways in which the slightly-older
speakers themselves differ from full-grown adults. Since all generations behave roughly alike in this respect, the rate of change tends to be constant.\footnote{It is tempting to suggest that exaggeration takes place on a quantitative scale that employs the natural units of grammar, hypothesized here to be Harmony. Stadler et al. employ instead the Multiplicative-cum-Cutoff model described in §6.2 below, noting, however, that it doesn’t seem to matter much what particular imitation mechanism they employ in their model.}

5.5.1 The research just reviewed and syntactic theory

It is plain that the research tradition that Kroch founded shares some of its ideas with the work on synchronic syntax by Bresnan and colleagues discussed in §5.3. It seems reasonable to me to place the same theoretical gloss on both research traditions: both are accessing a fundamentally correct theoretical conception, in which the probability patterns emerge from simple interaction of conflicting principles. These principles operate abstractly in the Harmony domain but are observable in the probability domain. Bresnan and her coworkers measure the VARIABLE, ONOFF, and various Perturber constraints interacting in real time within the mind/brain of a single speaker as she frames her thoughts in sentences. For Kroch and colleagues, the facts under study reflect the steady rise or fall of VARIABLE against backdrop of stable Perturbers, as adolescents imitate and extend the diachronic drift of VARIABLE. However, the mechanism of both approaches is the same, if something like MaxEnt is part of the language faculty.

5.6 Semantics/Pragmatics

A key area here is quantifier scope, i.e. the wide/narrow ambiguity in sentences like A student saw every professor. Experimental work (e.g. AnderBois et al. 2012) indicates that the appropriate kind of system for predicting quantifier scope judgments is likely to be a probabilistic one. Native speaker judgments result from a blend of conflicting factors, and the judgments are gradient rather crisply categorical.

I have only the tiniest curve to offer here, namely a subset of AnderBois et al.’s data, covering only the two well-known factors of linear order (leftward favors broad scope) and grammatical relations (subject position favors broad scope). The fit seems good (unimpressively so, given that they are few data points; but see a rival model below in §6.2). The graph does display the narrowing of the influence of the subject/object distinction at the periphery that MaxEnt predicts.
(25) A wug-shaped curve in quantifier scope, adapted from AnderBois et al. (2012:107)

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

With our survey concluded, we turn to the other goal of this paper, namely framework assessment. This requires addressing both the models that demonstrably cannot generate wug-shaped curves, and asking about models whose behavior is yet undiagnosed.

6.1 The need for consistent performance

In inspecting the results of various frameworks applied to the same data, I find a consistent pattern: often a defective framework gets lucky, working quite well to fit a particular batch of data. Thus, to evaluate a framework, we need to examine its performance in a variety of situations.

6.2 Some simple alternatives to MaxEnt/logistic regression

MaxEnt is hardly the only way to map from constraint violations and weights to probability, nor even the simplest; and it is useful to examine a couple of alternatives. These alternatives were considered seriously in the early days of quantitative sociolinguistics, before the field shifted toward MaxEnt as a better solution; see Sankoff and Labov (1979) for discussion.

Consider first a **Multiplication-cum-Cutoff** model, in which every constraint violation has the effect of multiplying candidate probability by a particular value, namely the constraint’s weight. Here, we will allow constraints to have values greater than one, so that they can increase as well as reduce probability. Since probabilities cannot go above one, this model prevents logically impossible probability values by a ceiling of one imposed by fiat. In (26) I give a schematic for the quantitative signature of this approach, assuming a series of seven baseline probabilities and seven Perturber probabilities.
(26) The basic quantitative signature of the Multiplication-cum-cutoff model

As can be seen, the basic prediction is the probabilities for particular Perturbers will converge in one direction, diverge in the other (up to the point where the cutoff at 1 prevents further divergence). I have not seen data patterns like this in my explorations and would be curious to know if they ever occur.

I have tried out the Multiplication-cum-Cutoff model on a variety of cases and, per what was said above, it often works surprisingly well — fitting the parameters appropriately with data having a limited range can conceal the model’s fundamental “preferences” as revealed in (26). However, when applied to cases when a single constraint is violated a variable number of times, as in (8) above, Multiplication-cum-Cutoff is likely to perform poorly. The quantitative signature for such cases is a curious shape: rather like the MaxEnt sigmoid on the left, but a declining exponential on the right; call it the Asymmetrical Sigmoid. The best-fit Asymmetrical Sigmoid for the Kluender data (post-long-vowel series) is shown in (27) below.

(27) Fitting the Multiplication-cum-Cutoff model to the data of Kluender et al. (post-long vowel series)
same would hold for the literature on historical-change sigmoids. For further discussion of the
problems with a model that is inherently asymmetrical, see Sankoff and Labov (1979).

We can consider likewise an **Addition-cum-Cutoff** model, in which probability is linearly
related to the violations of VARIABLE, with each PERTURBER adding to, or subtracting from, the
base probability by a constant. We again impose by fiat a floor at 0 and a ceiling at 1; this model
was put forth as a straw man by Cedergren and Sankoff (1974). The empirical signature of the
Addition-cum-Cutoff model is the “Z-shaped curve,” with parallel lines going diagonally
between the cutoffs, with sharp angles at the transition.

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

![Image of Z-shaped curve](image)

The main qualitative difference is that MaxEnt smooths off the ends, gradually leveling the
slope, whereas Addition-cum-Cutoff crashes directly 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 often hard to prove that the ends of the sigmoid
really are smooth (MaxEnt) rather than angular (Addition-cum-Cutoff). However, it is certainly
the widespread research experience of both phoneticians and historical linguists that empirically,
we see gradual sigmoids, not abrupt cutoffs. The very simple data of figure (25) are modeled
poorly in Addition-cum-Cutoff, as the two lines “want” to have different slopes.

It is worth addressing these failings in deeper terms; that is, in terms of the ways that a
constraint-based framework expresses a rational procedure for making inferences from the facts.
We saw in §2.2.4 above that MaxEnt is specifically designed to vary how strongly facts (here,

---

*23* The closest thing I have seen to a data pattern that looks like (28) 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.

*24* The quantitative signature of the Addition-cum-Cutoff model would poorly model an important finding in
speech perception. It appears that 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 graceful mountain-like shape. It would not be expected under the Addition-
cum-Cutoff function, 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.
constraint violations) affect probability: in the middle of the probability range, violations are influential; at either periphery, less so; and this embodies the sensible principle that certainty (in either direction) should be evidentially expensive. Neither the Multiplication-cum-Cutoff nor the Addition-cum-Cutoff models do 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 all of a sudden becomes 100% unininformative. Neither of these principles corresponds to sensible inductive reasoning.

6.3 Frameworks originating in Optimality Theory

Two further approaches I will discuss have an ancestry in Optimality Theory. In its classical form, OT is a nonprobabilistic framework that cannot be used to treat gradient data. However, not long after OT was invented in 1993, scholars undertook the task of extending OT in probabilistic versions, concretely by having it assign a probability distribution over GEN. The idea of accomplishing this task by grafting existing ideas from the field of statistics onto OT emerged only later, after linguists had first worked on homegrown approaches.

6.3.1 Stochastic Optimality Theory

As noted above in §2.1, classical OT uses a procedure for finding winning candidates that privileges higher-ranking constraints absolutely. In so doing, the procedure discards information, since if the first n ranked constraints suffice to find a winner, the testimony of the remaining constraints is ignored, no matter how much they may favor some other candidate. Classical OT also discards information about relative violation, since a set of 7 violations stacked against 1 is treated no differently from 2 vs. 1.

A way of rendering classical OT probabilistic, Stochastic OT, was invented by Boersma (1998), who later collaborated with me (Boersma and Hayes 2001) in applying the model to some classical cases of phonological analysis. The idea is that the grammar itself is probabilistic: constraints come with a number (“ranking value”) that expresses how highly ranked they are in general, but each time the grammar is deployed (“evaluation time”), the 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 the selection procedure of classical OT. Repeated application of this procedure will yield a probability distribution, through sampling.

Like other theories that I now believe to be wrong, Stochastic OT gets lucky (or so I claim) rather often when applied to particular data. For instance, Ernestus and Baayen (2003) tried out Stochastic OT on their Dutch data, and it modestly outperformed MaxEnt. However, Stochastic has proven to have two serious flaws.

First, it cannot treat continuas; i.e. VARIABLE constraints with varying number of values. The reason is has already been given: the decision procedure of Optimality Theory is indifferent to the “margin of victory”: a candidate that violates constraint seven times loses to a candidate that violates it just once, no more than a candidate that violates a constraint two times. But these differences evidently matter; examples include all of the sigmoids of speech perception (§3.1.1, §5.4), word-count effects in syntax (§5.3) and the effects of sound symbolism discovered by
Kawahara (in press), who find a smoothing increasing function for how word length increases listeners’ impression of size in fictional Pokémon creatures.\textsuperscript{25}

Second, in Stochastic OT a Perturber can only perturb “within its own zone”; that is, when its own inherent ranking value is within shouting distance of the constraints that it interacts with. But in real life, the effect of a Perturber is \textit{across the board}: it interacts with constraints that are mutually very far apart on the ranking scale. This point is covered in detail, with failed empirical examples, in Zuraw and Hayes (2017). We can add another failed case here: Stochastic OT offers a very poor fit to Sankoff-Bailey Québec French data discussed above in §5.2.2. Figure (29) below is the same as Fig. (18), except with the Stochastic OT predictions superimposed instead of MaxEnt.

\textbf{(29) The Sankoff/Bailey Québec French pattern with Stochastic OT predictions}

![Graph: Sankoff/Bailey Québec French pattern with Stochastic OT predictions]

The ill-fitting flat regions are cases in which a Perturber simply lacks the range to interact with other constraints that have greatly different ranking values.

Tracing the problem back further, we can ask \textit{why} two constraints must be close on their ranking values in Stochastic OT to provide a probabilistic outcome. The reason is that when they are far apart, strict domination applies, and the Perturber drops out of sight. The latter fact is a consequence of the standard OT decision procedure, which lets the highest ranking constraint that distinguishes two candidates be the decider, shutting out influence of dominated constraints. In other words, at the deepest level, the failure of Stochastic OT to generate wug-shaped curves is the consequence of the fundamental decision made in standard OT, noted in §2.2.3 above, to discard a great deal of potentially-relevant information in making choices.

\footnote{A proposal made by Boersma (1998: §6, §8.4) actually \textit{can} derive sigmoids from variable constraints in Stochastic OT. The idea is to replace single gradient constraints with \textit{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.}
6.3.2 Noisy Harmonic Grammar

The primary reference for this framework 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 (per §2.2.1-§2.2.3 above) of the Harmony of each candidate. At this point Noisy Harmonic Grammar becomes like Stochastic OT, in that we again suppose a series of evaluation times at which the grammar gets altered by set of a random shifts chosen from a Gaussian distribution. The framework comes in several varieties (Hayes 2016), which differ in that part of the calculation gets randomly shifted: we can alter the constraint weights, the violations, the tableau cells (violations times weight; as in Goldrick and Daland 2009) or indeed the Harmony scores of candidates.

These different types differ in their quantitative signatures. Adding noise to the constraint weights is, in my opinion, probably not a good idea, for its quantitative signature as applied to cases like our Klünder et al. example (§3.1.1, §5.4), with series of varying violation counts, will produce a quantitative signature much like a declining exponential, which we saw, and criticized, for the Multiplication-cum-Cutoff model. The reason this happens, and an empirical example of the resulting inferior fit, is given in McPherson and Hayes (2016:156).

(30) An asymmetrical sigmoid generated under classical Noisy Harmonic Grammar

![Graph of an asymmetrical sigmoid](image)

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 is adding the noise to the Harmony scores of candidates. As Flemming (2017) demonstrates mathematically, this theory is extremely close to MaxEnt, and the sigmoid curves it generates as a quantitative signature are very similar to the MaxEnt sigmoid. I illustrate this in (31), which gives a MaxEnt sigmoid (weights: VARIABLE 2, ON/OFF 8) almost duplicated by a candidate-noise Noisy Harmonic Grammar sigmoid, weights 2.5 VARIABLE 2.5, ON/OFF 10.

---

26 I will skip the violation-noise and tableau-cell noise variants, which are intermediate in their behavior.
(31) A MaxEnt sigmoid nearly replicated in candidate-noise Noisy Harmonic Grammar

It is hard to imagine language data, prone as they are to random noise, 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 when strong alternatives are available.

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.

6.4 Other models

Arrived at from the viewpoint of linguistics, MaxEnt is an evolved version of Optimality Theory; it differs in origin from Stochastic OT and Noisy Harmonic Grammar in that it was not home-grown, but rather was a case of importing ideas from statistics and computer science. Addressed from the viewpoint of statistics, however, MaxEnt is a bit retrograde, representing the avant garde of the field from around the 1970’s (Cramer 2002). In more recent decades, it has been fairly standard for experimental and corpus work that uses logistic regression to employ the mixed-effects version of the model, which provide better control for the idiosyncrasies of individual items or participants.27 The texts by Baayen (2008) and Johnson (2011) provide training in the use of this method. There are other models likewise 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 sociolinguistics, the

27 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).
methods of statistics employed have also evolved greatly since the introduction of simple logistic regression in the 1970s; see for instance Johnson (2009).

It is almost beyond question that more advanced modern statistical models permit better model fit and more rigorous testing of claims. I also would be surprised that models that are as successful at these could fail to generate wug-shaped curves, which would arise, I suspect, in any inductive system that obeys the principles given earlier under (3). However, I emphasize that the usefulness of these methods for the scientist (where they have been primarily used) is not the question addressed here. The research agenda I am suggesting is adaptation of statistical methods as part of competence models, particularly to serve as the decision procedure for candidates in (evolved forms of) Optimality Theory. Determining what procedure is actually used by humans is, in principle, an empirical question.28

In support of the idea that effectiveness for the scientist might not be the same as appropriateness for a competence model, I note that there is some evidence that human learners sometimes refuse what is clearly the best model, opting for something simpler. A case in point is given in the next section.

6.5 Interaction terms and conjoined constraints

It is standard for scholars using MaxEnt/logistic regression for purposes of statistical testing to check interaction terms. For example, given two choices A vs. B, C vs. D, it can turn out to be the case that the choice between C and D comes out different when A is true than when B is. Testing for the interaction of the A/B and C/D factors can inform us how likely it is this scenario is in effect.

Interaction terms are also widely proposed in Optimality Theory, under the title of “conjoined constraints” (Smolensky 1995); i.e. constraints that are composed of two preexisting constraints A and B, and are violated only when both A and B are violated. Deployed in an OT model in which the selection procedure is MaxEnt, these are very close to being conjoined constraints.29

A moment’s thought will reveal that free and arbitrary use of local conjunction will wipe out the possibility of making strict predictions about quantitative signatures. For any of the data points in any of the cases covered in this paper, we could add a conjoined constraint that covers exactly that data point, shifting the prediction made by the model to any value we please; hence at some level the framework ends up saying nothing.

28 Beyond this, there are models that could not be adapted as a decision procedure for OT because they rely on analogy rather than constraints, e.g. Pierrehumbert (2001), Skousen et al. (2002), and Daelemans and van den Bosch (2005). These too almost certainly generate wug-shaped curves, though it might be harder to establish an analytical basis for how they do so. It is worth noting that the neural network and analogical models slightly outperformed MaxEnt in Ernestus and Baayen’s (2003) testing against Dutch data (§5.1.1).

29 However, the OT version of conjunction is usually local conjunction: *A & B is violated only when the violations of A and B are in the same location in the string, in some formally defined sense.
Given that interaction terms and conjoined constraints sometimes are proposed in linguistics (see Shih 2017 for a well-argued recent case), we need to address their implications for quantitative signatures. There are two possible threads of inquiry.

First, I judge that often, constraints that we might think of as conjoined are really unitary: A is preferred to B in the presence of C because there is a substantive reason why the presence of C makes B harder. To give an example, in Japanese and other languages, there is a tendency to avoid voiced obstruent geminates (double consonants). Is this because the goal is to avoid two bad things at once? Many languages do indeed avoid voiced obstruents, and many languages avoid geminates, so the interaction-term approach has some appeal. Yet closer scrutiny of the phonetics of voicing difficulty, as in Westbury and Keating (1986), gives a substantive reason, based on airflows and vocal tract properties, to think that voiced obstruent geminates pose a special extreme of difficulty in maintaining voicing, which has been documented experimentally by Kawahara (2006). In other words, we would have good phonetic justification for setting up the “conjoined” constraint [+voice, +obstruent, +long] even in the absence of any information that [+voice, +obstruent] and [+long] are themselves avoided to a lesser extent. [+voice, +obstruent, +long] is plausible as a simplex constraint with richer internal specification, and this fits the facts of phonetics more directly than an appeal to constraint conjunction would.30

A second item to consider is the possibility that language learners dislike conjunction and tacitly try to avoid it. A possible example occurs in Hungarian vowel harmony, discussed in §4.2.1. We abstract away from the modest consonant effects discussed there and consider only the vowels. In words ending in a Back vowel followed by a Neutral vowel, harmony varies probabilistically, and there is a natural hierarchy of trigger strength depending on the height of the Neutral vowel (the “Height Effect”): low [ɛ] triggers front harmony more strongly than [ɛː], which is stronger than [i]. There is also a Count Effect: stems ending in Back Neutral1 Neutral2 take front harmony more often than comparable Back Neutral stems with the same final Neutral vowel. Thus there is a six way pattern, involving the possibilities Bi, Beː, Be, BNi, BNeː, BNe.

Surveys of the Hungarian lexicon find an ugly picture in terms of what a Hungarian-learning child would encounter in assessing the pattern of her language: BNeː is an outlier, showing far higher front harmony than it ought to given the Count and Height effects. But when Hungarians take a nonce probe study, the ugliness disappears: they smooth out the anomaly (Hayes and Londe 2006:73) and give responses that are well modeled by a MaxEnt grammar that embodies just the Count and Height effects, as in (32):

---

30 For a much earlier insistence on substantive support when positing interaction terms, see Sankoff and Labov (1979:205).
This result was later replicated in a more thorough wug-test study, Hayes et al. (2009).

The original lexical data is very accurately modeled if we allow ourselves a conjoined constraint/interaction term that specifically covers BNe words. Moreover, this interaction term embodies a highly significant effect ($p = .0003$, by the Likelihood Ratio Test). But speakers of Hungarian seem to have arrived at a grammar that uses unitary principles, favoring this over accuracy. Though they mimic lexical frequencies in general, they reject the opportunity for greater accuracy afforded by constraint conjunction.

To summarize this section: liberal use of interaction terms/constraint conjunction produces a theory that has no quantitative signature and thus (from one point of view) is very uninteresting. But careful attention to cases where putative conjoined constraints are justified on external grounds as singletons, and insistence that putative cases of interaction be backed by evidence from psycholinguistic testing, may result in a more interesting picture.

These reflections on conjoined constraints are applicable as well, in principle, to the use of more powerful statistical models, as mentioned in §6.4. Such models do often provide better fit to data than MaxEnt, but to be convinced by them as competence models we need to be sure that the improved fit is something that human language learners also achieve.

7. Conclusions

7.1 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 where the wug-shaped curve is not strongly
evidenced.\textsuperscript{31} 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 when we partition the constraints into Baseline and Perturber families, we will probably find more wug-shaped curves. So I think it is not out of the question to imagine that something like MaxEnt is part of the language faculty.\textsuperscript{32}

The success of logistic regression in probabilistic linguistics is sufficient to lead me to voice a methodological caution: there is a possibility that the dulling effect of experience could lead us to a trivializing interpretation of the perserviveness of wug-shaped curves — “of course the data is like that; we all know this.” The argument against such trivialization is that the right quantitative signatures emerge only if you use the right mathematical approach: MaxEnt and related theories generate wug-shaped curves; others don’t (§6.2, §6.3).

7.2 Do findings drawn from the field of statistics fall within the province of linguistic theory?

During the earlier decades of my participation in the field of linguistics, I would have found it unimaginable that statistics could form part of linguistic theory. This was partly the consequence of the character of statistics long ago (e.g., as it was taught to me as an undergraduate). I suspect I was not alone in thinking of it as leaden, dull, and tedious — a necessary and burdensome duty for experimentalists, but not relevant to theorizing.

But statistics kept evolving; in the intervening decades it seems to have become a far more lively, truth-seeking research activity.\textsuperscript{33} Nowadays, we are better at using math to extract valid conclusions from noisy data than we used to be; indeed, the real world, even baseball, attests to this. And extracting valid conclusions from noisy data is precisely what young human language learners must do. If there exist rational and effective mathematical modes of inductive reasoning, then it is not unreasonable to suppose that natural selection has equipped human beings with some analogue of these principles.

\textsuperscript{31} Let me confess them here. (a) Cedergren and Sankoff’s (1974) Panamanian Spanish example, in addition to lacking data points at the probability extremes (hence uninformative with regard to wug-shapedness) also has a lot of noise, particularly among the [+lateral] cases. (b) The Black English copula deletion data in Cedergren and Sankoff (1974) is not very orderly (some data lines actually cross), and reflects perhaps an incorrect analytical choice, i.e. modeling a Deletion probability conditioned on Contraction. (c) The data set from Detroit Black English cluster simplification in Wolfram and Fasold (1974:132) shows somewhat closer spacing of the perturbed data points at a wug-medial position than a neighboring wug-peripheral position; (d) Ryan (2019:91), plotting a wug-shaped curve from Hupa stress, notes a similar small medial anomaly. In none of these cases, however, does the MaxEnt model create a really bad fit, as in (29); the problem seems to be unexplained outliers, rather than a wholesale missed pattern. For calculations and graphs for these cases, please visit the online Gallery, cited above.

\textsuperscript{32} Or better: part of human cognition, adapted to use in the language faculty. One of the earliest formulations of MaxEnt, Smolensky (1986), was meant as a contribution to cognitive science in general; and it is hardly unusual in contemporary research to try to explain universals of language with principles that are not specifically linguistic.

\textsuperscript{33} A popularization I have enjoyed is McGrayne (2011), which conveys some of the liveliness of modern statistical inquiry as well as a sense of drama about how the field came to evolve into its present form.
Beyond this, it appears that some of the principles of statistics are simply true in the sense that the theorems of Euclidean geometry are true. The core principles of probability theory (the Sum Rule and the Product Rule, and their corollary Bayes’ Theorem) have been proven formally on the basis of simple and intuitive axioms (Jaynes 2003, amplifying Cox 1946); and from this starting point a vast set of theorems have been proven. MaxEnt itself been advocated on mathematical grounds. It is proven that the method is indeed entropy-maximizing: per Johnson and Riezler (2002:244) MaxEnt models “contain the minimum additional information over and above the information contained in the training data”; i.e. maximally avoid commitments unjustified by the data, and thus again have an a priori command on our attention.\textsuperscript{34}

In light of these developments, I advocate importation: our linguistic theories will benefit from having a strong statistical foundation. The current linguistic theories that most easily engage with this foundation are, I think, constraint-based ones like Optimality Theory, where linguistic knowledge is deliberately “atomized” to the point that statistical principles can readily engage with it. Such theories leave plenty for the linguists to do, for instance in understanding the families of constraints and their origin or in understanding the form of linguistic representations. But we should be happy to import our forms of probabilistic reasoning if that is what works best.

Approaching the same issue from the other side, I feel there is also a need to incorporate linguistic theory into statistical data analysis. Some of the research I have read for purposes of writing this article strikes me as unusually agnostic with regard to theory: it is quite easy in working practice to adopt purely-empirical classifications of the facts, plug them into a good statistical model, and obtain accurate results. But as a generative grammarian I hope to see future work use constraints that are themselves the result of extensive theoretical development and typological testing. Such scaling up of the inquiry will, I think, ultimately make the research results more explanatory and more convincing.

**Appendix: Recoding Harmony into a single value for a two candidate system**

Assume the Kluender et al. example from §3.1.1, with two viable candidates [b] and [p]. We can calculate the probability of [b] like this:

\[
\text{Pr}(b) = \frac{\exp(-\sum_i w_i f_i(b))}{Z},
\]

where
\[
Z = \sum_j \exp(-\sum_i w_i f_i(x_j))
\]

\[= \frac{\exp(H_b)}{Z}, \quad \text{where } Z = \sum_j \exp(-\sum_i H_j)
\]

\[= \frac{\exp(H_b)}{\exp(H_b) + \exp(H_p)}
\]

MaxEnt formula, from (1)

Use a single symbol for Harmony instead of spelling out the calculation.

Specify \(Z\); there are just two viable candidates [b] and [p]; and all others are very close to zero.

\textsuperscript{34} For an attempt to prove the principles of MaxEnt as optimal from a set of primitive/intuitive axioms, see Skilling (1988), and for skepticism Uffink (1997).
\[
\frac{1}{1 + \exp(H_p)\exp(H_b)} = \frac{1}{1 + \exp(H_p - H_b)}
\]

Divide top and bottom by \(\exp(H_b)\).

Division is same as subtracting exponents.

Given the constraints we adopted in §3.1.1, the harmony difference \(H_p - H_b\) will be a simple function of the closure duration of any particular input (i.e., value of closure duration in msec.); and, as the derivation shows, this number suffices for computing \(P(b)\). Thus, by putting \(H_p - H_b\) on the \(x\) axis, we can obtain a simple two-dimensional figure that gives the analytic basis of the sigmoid.

References


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


Smolensky, Paul (1995). On the internal structure of the constraint component Con of UG. Talk given at UCLA.


