MaxEnt Optimality Theory: A Guide to Analysis

August 2026

To appear, after final review and revisions, with Cambridge University Press

This distribution includes all nine chapters.

Bruce Hayes and Claire Moore-Cantwell

Department of Linguistics

UCLA

Contents

Chapter 1: Purpose and orientation

1.1 Rationale for this book

MaxEnt grammars (Smolensky 1986; Goldwater and Johnson 2003) are a formal apparatus for constraint-based linguistics. MaxEnt as a grammatical model is an outgrowth of Optimality Theory (Prince and Smolensky, 1993/2004), but works with probabilities, making it suitable for dealing with gradient phenomena. These include free variation (multiple outputs from a single input), gradient well-formedness (e.g., in syntactic judgments or phonotactics), and the matchup of native speaker judgments to statistical patterns found in their language. MaxEnt grammars can also be used in providing concrete interpretations for linguistic experiments. Increasingly, linguists having been using MaxEnt for all of these purposes.

We start by explaining the original impetus for writing this book. On various occasions, we have found ourselves conversing with linguists who have gathered interesting data that might have important implications for theory, yet are not in a position to construct a rigorous analysis of their data because they are quantitative in character. Our intent is to help such linguists overcome the barriers to using MaxEnt grammars for this purpose. MaxEnt offers both quantitatively precise analysis as well as statistical testing, which permits each constraint in the analysis to be checked for whether it is making a meaningful contribution.

We are phonologists and have primarily used phonological examples here. But, we think linguists working in other branches of the field might also find the approach here helpful to their research. In Chapter 8: we survey examples from other areas of linguistics: phonetics, metrics, morphology and syntax, in which researchers have pursued a MaxEnt approach or something like it.

1.2 The research scene in linguistics: a role for MaxEnt?

As just noted, the starting point of this work is the essentially practical one of equipping linguists with a tool that they may find useful. But there is a broader setting for this goal, based on our judgment that the whole field of linguistics is evolving in directions for which a research tool like MaxEnt is useful. More than ever before, linguists today rely on multiple sources of evidence: along with traditional consultant work and grammatical analysis, the field is now heavily engaged in corpus study, experimentation, and acquisition work; and there is a widely-shared hope that these diverse sources of evidence can be woven together to reach better-supported conclusions about the language faculty. We judge that in this enterprise, linguistic theory should be placed at the core, for it is the only mechanism that can bind together into a single picture the results coming in from all the various forms of empirical inquiry. Since the data gathered from the newer forms of research are almost always gradient, it makes sense to express the theoretical core of the research enterprise using an approach like MaxEnt, which is specifically designed for the researcher to create analyses that are both founded in mainstream theory and able to engage with gradience.

Lastly, we note that MaxEnt is hardly a neutral research tool: it has definite intellectual commitments and makes empirical predictions all by itself; in other words, it is a theory.

Necessarily, the predictions of MaxEnt are relatively abtract and quantitative in character, but they are specific enough to be empirically testable. We cover some of them in Chapter 9.

The overall organization of the book reflects the discussion just given. We start at the beginning in fully tutorial mode (Chapter 2), then add further "how-to" chapters, each covering specific analytic techiques. Starting with Chapter 8, the book turns more to advocacy, first giving a survey of MaxEnt applications, then, with Chapter 9, some general arguments that MaxEnt is mathematically a good match for how human language works.

1.3 Why is it called "MaxEnt"?

One stumbling block for this theory is its name, "Maximum Entropy," which in the context of linguistics may come across as alien or even pretentious. This is not a problem we are in a position to fix, but in brief: "Maximum Entropy" in the context of linguistics designates the mathematical apparatus that was attached to Optimality Theory by Goldwater and Johnson (2003) to render it probabilistic. In current linguistics usage, "Maximum Entropy" or "MaxEnt" are often used to refer to the whole package; the OT architecture coupled with the MaxEnt math. The intellectual connection to actual entropy (as in physics or information theory) is, in practice, rather remote and perhaps best ignored at this stage. For further discussion see §4.5.

1.4 Content

The chapters of this book are arranged in the order which we believe will be most useful to the beginning practitioner. Chapter 2 offers an introduction to MaxEnt in the form of a tutorial, starting with a brief review of classical Optimality Theory. We illustrate both theoretical issues and the practicalities of how to conduct a MaxEnt analysis in an Excel spreadsheet, using a series of toy languages based on a lenition pattern in Khalkha Mongolian. The toy languages start simple and eventually work up to a complex pattern fairly close to what is observed in real Khalkha.

Chapter 3 provides a second worked-out example, this time using data from Tagalog. This case study is an example of **lexical frequency matching**, in which probabilistic patterns observed across the words of a language are productively applied to novel words by speakers in experiments. This type of data is broadly represented in the literature, and is often analyzed with MaxEnt. We conclude the chapter with an overview of frequency matching in general. By the end of Chapter 3, readers should feel well-equipped to develop MaxEnt analyses on their own.

Chapter 4 is devoted to mathematical issues. We first cover the details of the MaxEnt math, aiming to provide an intuitive understanding of what the math is doing and why. We also show how MaxEnt expands on (subsumes) classical Optimality Theory. The chapter then covers various issues that have both mathematical and theoretical relevance, including how to interpret constraint weights, the issue of zero weights, infinite weights, indeterminate weights, and priors. The chapter also includes a discussion of how the Excel Solver and similar systems compute weights. In the final sections of this chapter, we turn to the relationship of MaxEnt to statistics, particularly multinomial logistic regression. We show that while the statistical models and MaxEnt OT share the same math, their purposes and appropriate procedures are quite different.

The name "MaxEnt" is inherited from the theory's statistical underpinnings, which we discuss, along with the history of the framework, in the final section.

Chapter 5 focuses on research methods. There are three sections: the evaluation of MaxEnt analysis with graphical and statistical methods; the construction of models (grammars) through the informed addition (or subtraction) of constraints from a possible constraint set; and lastly a survey of ways in which a researcher might obtain probabilistic data in the first place: field notes, corpora, and experiments.

Chapter 6 covers the MaxEnt analysis of phonotactics, using Turkish vowel sequencing as an example. Phonotactic analyses illustrate many of the methods covered in Chapters 2 and 3, but tableaux are set up differently, with one large candidate set including all logically possible structures of interest. A good model should allocate probability to the structures which actually occur in the language, and away from the structures that are banned. The discussion also addresses the MaxEnt approach to the classical Subset Problem in language learning.

Chapter 7 is the main chapter addressing the theoretical problem of language learning. The key idea introduced is **bias modeling**: augmenting the weight-fitting algorithm to bias certain constraints to have high or low weights. The idea is to use this method to express a hypothesis about the language faculty (UG), which could help explain differences between speakers' behavior/intuitions and the patterns available to them during learning. We use Hungarian vowel harmony as our example, since this is a system where speakers match their lexicon on some generalizations but not others. At the end of the chapter we address two further issues from the MaxEnt perspective: hidden structure and gradual learning.

In Chapter 8 we give a brief overview of further possible uses of MaxEnt, including suggested readings for readers who wish to learn more. We briefly discuss several research questions within phonology not covered elsewhere in the book, as well as loanword adaptation, lexical strata, sound symbolism, underlying representation inference, as well as work in metrics, phonetics, morphology, and syntax.

The concluding Chapter 9 discusses specific predictions that MaxEnt makes as a theory of grammar. We take up empirical evidence for using probabilistic grammars in the first place, then move on to discuss predictions that are specific to MaxEnt, comparing it to other potential models of probabilistic data: Partially Ordered Constraints, Stochastic OT, Noisy Harmonic Grammar, neural networks, and analogical models.

1.5 Our MaxEnt spreadsheets and their locations (Supplementary Materials)

We are enthusiasts for the use of spreadsheets for MaxEnt calculations, favoring them for their transparency and ease of use. We find them particularly useful for advising graduate students: an adviser can jump-start the student's analysis by providing a small, conceptual spreadsheet on which the student can build.

We have placed a variety of spreadsheets discussed in the text on the internet, from which they can be downloaded; moreover, since websites can be fragile things, we offer the following places from which the spreadsheets are currently available:

https://brucehayes.org/
https://www.clairemoorecantwell.org/
https://www.lingbuzz.net/ (search for author)

We also anticipate placing these materials on the publisher's website. In the text we refer to the spreadsheets with boldface file names. To locate a file, visit the portion of the supplementary materials for the chapter you are reading, then search for the file name.

1.6 Policy on mathematical presentation

MaxEnt involves a certain amount of math. It can be used by a linguist as a tool without understanding any of it, but is more satisfying and reliable to the extent that one does understand the math. We address a broad linguistic audience here and seek to be helpful to readers with a range of mathematical backgrounds. To this end we have specifically targeted our writing to an imaginary reader who studied math in high school, and when reminded of various theorems and formulas taught there, can more or less call them to mind. We try to be helpful by showing and justifying all the steps of the mathematical derivations. This is impossibly verbose by the standards of a technical publication, but far more explicit than is given in (for example) popular science magazines. We beg the patience of readers whose mathematical background exceeds that of our hypothetical targeted reader. The mathiest sections have been placed in what we call Excursus Boxes; these can be skimmed if necessary but we do recommend at least giving them a try.

1.7 Acknowledgments

Many linguists have provided, and continue to provide, helpful comments and input on the preparation of this work. In particular, we would like to thank Adam Albright, Aixiu An, Arto Anttila, Michael Becker, Canaan Breiss, Joan Bresnan, Jessica Coon, Tim Hunter, Gerhard Jäger, Gaja Jarosz, Jongho Jun, René Kager, Aaron Kaplan, Shigeto Kawahara, Stefan Keine, Giorgio Magri, Connor Mayer, Laurel Perkins, Geoff Pullum, Anthi Revithiadou**,** Robert Staubs, Donca Steriade, Megha Sundara, Anthony Woodbury, as well as the UCLA "Cans of Worms" reading group. Our apologies to anyone we may have forgotten. We would also like to thank the many graduate students over the years who have helped us hone our teaching of MaxEnt at UCLA. Special thanks to Corrina Fuller, who generously shared her Khalkha Mongolian data and expertise. We also thank Helen Barton at Cambridge University Press for all her help, as well as the six anonymous CUP reviewers who provided valuable comments on a draft version.

Chapter 2: The analysis of variation in outputs

We begin with a review of classical **Optimality Theory** (OT; Prince and Smolensky 1993/2004), then develop the exposition of MaxEnt by stages to cover, in the end, a nontrivial analysis. For pedagogical purposes we work with a sequence of imaginary languages, to be called K1 through K5. These languages incorporate successively more complex data patterns, for which we employ first classical OT, then OT with free rankings (Anttila 1997), and finally MaxEnt itself.

The K1-K5 languages form successively better approximations to a real language, namely Khalkha Mongolian. For real Khalkha, our reference source is Fuller (2024), which includes substantial corpus data and its own detailed MaxEnt analysis. Our K1-K5 are intended to build up gradually, ultimately approximating the full complexity seen in Fuller's research. In §2.7.2 we briefly mention how real Khalkha goes beyond our hypothetical languages.

2.1 Analyzing the K1 language in Classical OT

The key (indeed, only) phenomenon of our first hypothetical language, K1, falls under the category of **lenition**, whereby stops and other obstruent sounds take on weakened degrees of closure. Lenition is a common process across languages (Lavoie 2001; Kirchner 2001;); some classical instances include the lenition of Spanish /b,d,g/ to [β,ð,ɣ] (Harris 1969), or of English /t,d/ to the tap [ɾ] (Hayes 2008). In our hypothetical K1 language, we will suppose an underlying phoneme /p/ that in non-phrase-initial position is normally lenited to the homorganic glide [w], as in (1a) below. The pattern involves two complications. First, lenition is blocked whenever the /p/ is preceded by a nasal, as in (1b). Typologically, it is common for nasals (which, articulatorily, are stops) to disfavor nonstop articulations immediately after them; thus Spanish, with lenition of /b/ to [β] in many contexts, does not lenite /b/ after [m]; e.g. *la bomba* /la bomba/ → [la βomba] 'the bomb/pump'. Second, the lenition process of K1 is confined to function words (like [poɮ] 'become', or [pai] 'be'); when a /p/ appears in a content word, it never alternates; see (1c).

(1) *The patterning of /p/ lenition in K1*

a. Default realization in non-phrase-initial position: [w]

/mʊntəg **p**oɮ[1]-ən/ → [mʊntəg **w**oɮən]
great become-NPST

'will be great'

[1] The underlying form for 'become' with /p/ is justified by the appearance of [p] in other contexts, such as in /xun pəɮ-əx-ig/ → [xun pəɮəxig] (person become-FUT-ACC, *becoming a person*).

b. Retention of [p] after nasals

/ʃit-t͡ʃʰəx-sə**n** **p**ai-sən/ → [ʃitt͡ʃʰəxsə**n** **p**aisən]
decide-PERF-PST be-PST

'had decided'

c. Retention of [p] in content words

/ɔrʊɮ-ə**x** **p**ɔɮəm-t͡ʃtʰɔi / → [ɔrʊɮə**x** **p**ɔɮəmt͡ʃtʰɔi]
involve-FUT opportunity-have

'will have the opportunity to involve'

The pattern just given is the sort to which classical Optimality Theory (OT) has often been applied, and it reveals the basic ideas behind OT in action. We give here only the briefest review of classical OT. Readers who want better background in this theory are advised to consult one of the three excellent textbooks: Kager (1999), McCarthy (2002), and McCarthy (2008). Much of what MaxEnt has to say reflects its status as an outgrowth of OT, and indeed many of the controversial issues for MaxEnt are controversial for OT as well.

In OT, the problem of phonological derivation is approached as searching for a particular **candidate output** that best satisfies a ranked hierarchy of phonological constraints. This **optimal** candidate is taken to be the (unique) output of the grammar. The theory assumes a function GEN that provides a full set of candidates — in the limit, the infinite set of all possible phonological strings. The function EVAL takes in a phonological input, the candidates of GEN, along with the ranked constraint set, and outputs a winner. Candidate selection works as follows: if the highest ranked constraint $\mathbb{C}$ that treats candidates $Cand_1$ and $Cand_2$ differently is violated more times by $Cand_2$ than $Cand_1$, then $Cand_2$ cannot be the winner. Similar comparisons, applied iteratively, will normally return a unique winning candidate.

We illustrate with K1. We will make use of four constraints, as follows. First, in order to account for the lenition in the first place, we will use the constraint *p:

(2) *p

Assign a violation to every non-phrase-initial [p].

This can be taken as one member of a larger family that penalizes segments that impose high articulatory effort, such as stops; see for instance Kirchner (2001)'s constraint LAZY.

If *p were the only constraint in the grammar, then of course all underlying /p/ in non-phrase-initial positions would be converted, often wrongly, to some other sound. But in K1 the process is "held back" in two particular contexts, namely after /n/ (1b) and in content words (1c).

We analyze the resistance to alternation in content words using the well-known family of **positional faithfulness** constraints (Beckman, 1997); for constraints specifically distinguishing content from function words see Casali (1997) and Alderete (2003). Faithfulness constraints penalize candidates that change something going from input to output; and the standard taxonomy of faithfulness constraints (McCarthy and Prince, 1995) is a systematic enumeration of

all the possible (minimal) changes that could affect a phonological representation. Here, for simplicity, we assume that [p] and [w] differ solely in the feature [sonorant],[2] and our crucial positional faithfulness constraint will be called IDENT(sonorant)$_{content}$.

(3) IDENT(sonorant)$_{content}$

Output correspondents of an input [αsonorant] segment in a content word (as opposed to a function word) are also [αsonorant].

We can now begin to see the rationale for constraint ranking: *p alone will cause /p/ to be lenited everywhere except initially, but if *p is outranked by IDENT(sonorant)$_{content}$, lenition will not occur in content words, and will be confined to function words.

A comparable analysis can cover the other case of lenition blockage — after /n/ — discussed above. We use the constraint *N[+cont], from Coetzee (2000), which penalizes postnasal continuants. It penalizes the bad candidate *[nw] for underlying /np/; this is because [n] falls in the class [+nasal] and [w] is [+continuant].

(4) *N[+cont]

Nasal sounds should not be followed by a [+continuant] segment.

It is plain that *N[+cont] must dominate *p; or, to use the normal notation, *N[+cont] ≫ *p.

We now assemble the ingredients for an analysis: three ranked constraints, plus schematic representative forms that, if properly derived, lead us to expect that the relevant forms of K1 can be derived in general. We are specifically interested in inputs where (a) /p/ is non-initial, but occurs in a content word; (b) /p/ is non-initial, but follows /n/; (c) /p/ is non-initial and neither of these two conditions are met. We want a form with [w] to be the winning candidate in case (c), but not in (a) and (b).

Aside from these constraints, we will also include in our analysis the constraint IDENT(sonorant) (the general case, *not* limited to content words). This constraint is needed to ensure that phrase-initial /p/ does not also lenite. We will not model this explicitly here, but include the constraint for completeness, and because it will become important later when variation is introduced. In general, careful analysis in OT includes constraints violated by "unfaithful winners" (McCarthy, 2008). It should be clear that *p must dominate IDENT(sonorant); else lenition to [w] would never happen at all.

We can now provide candidate tables, often called **tableaux** (singular **tableau**). At the top of each tableau, we list the constraints in an order consistent with the necessary pairwise rankings. Tableau rows give candidates from GEN, which for simplicity consist just of candidates for lenition and non-lenition, and asterisks indicate violations. It is easy for the eye to scan across

[2] In the feature set of Hayes's (2008) text, they also differ in quite a few additional features: [voice], [round], [continuant], [consonantal], [approximant], [high], [back], [dorsal], and [tense]. A full analysis would include IDENT constraints for these features, as well as constraints expressing the impossibility of other possible outcomes such as [β] or [ʍ] that could repair a non-initial /p/.

such a tableau and observe how, going from left to right, the losing candidates duly drop out of the competition, leaving the winner (indicated with the traditional "pointy finger") in place at the end.[3]

(5) *Tableaux for K1*

a. Simple lenition in a function word

/mʊntəg **p**oɮ-ən/ great become-NPST	IDENT(son)$_{content}$	*N[+cont]	*p	IDENT(son)
mʊntəg **p**oɮən			*!	
☞ mʊntəg **w**oɮən				*

b. Blockage of lenition after nasals

/ʃɪt-t͡ʃʰəx-sə**n** **p**ai-sən/ decide-PERF-PST be-PST	IDENT(son)$_{content}$	*N[+cont]	*p	IDENT(son)
☞ ʃɪtt͡ʃʰəxsə**n** **p**aisən			*	
ʃɪtt͡ʃʰəxsə**n** **w**aisən		*!		*

c. Blockage of lenition in a content word

/ərʊɮ-əx **p**ɔɮəm-t͡ʃtʰɔi / involve-FUT opportunity-have	IDENT(son)$_{content}$	*N[+cont]	*p	IDENT(son)
☞ ərʊɮəx **p**ɔɮəmt͡ʃtʰɔi			*	
ərʊɮəx **w**ɔɮəmt͡ʃtʰɔi	*!			*

We note that the rankings that were actually needed were really just *p ≫ IDENT(sonorant) (so lenition will happen); *N[+cont] ≫ *p (so lenition won't happen after nasals), and IDENT(sonorant)$_{content}$ ≫ *p (so lenition won't happen in content words). The constraints *N[+cont] and IDENT(sonorant)$_{content}$ never conflict with each other (both militate *against* lenition) and so there is no basis for assigning them a ranking (the order we pick for appearance on the page is arbitrary). The inessential ranking is shown with the dotted line in the tableaux of (5).

Our analysis thus far has been carried out using a relatively trivial GEN, with only two candidates for each input. A more valid analysis of K1, or any language, would select the right candidate out of an infinite GEN, and would require many more constraints and rankings. For example, if we added to (5a) the candidate [mʊntəg oɮən], in which the offending [p] has been deleted rather than just lenited to [w], it could be ruled out by ranking a MAX constraint (McCarthy and Prince 1995) over IDENT(sonorant). The more trivial candidate [mʊntəg **ə**poɮən], with an epenthetic schwa that does not help with the violation of *p, would be ruled out with a

[3] We follow standard practice in other aspects of our tableaux: an exclamation point marks the particular asterisk (sometimes not first in its cell) that kills off a candidate. Gray shading indicates cells which are irrelevant to evaluation, because the winning candidate has already been selected somewhere to the left.

DEP constraint, as would even more bizarre candidates, like [ʊmʊntəg poɮən] or [mʊntəg poɮən**t**]. In actual analytical practice scholars tend to present a GEN that includes anything that seems reasonable and bears on the problem at hand, but they must be on guard that they haven't missed something. Combinatorics, and software, can be helpful here (Karttunen, 2006; Tesar, 2004). We discuss some issues of GEN specific to MaxEnt in §6.1 below.

OT has been extremely influential in phonological theory, adopted by many who had previously made use of the rule-based framework of *SPE* (Chomsky and Halle 1968). Before we go on, we mention two reasons why this is so.

First, OT accomplishes what many feel to be a scientifically sensible goal, namely stripping down the analysis of phonological systems to their minimal ingredients. Here, these are the Markedness principles that militate against particular surface forms and the Faithfulness constraints that atomize the set of ways in which surface forms can differ from their underlying forms. In contrast, rule-based theory deploys, in essence, *bundles* of Markedness-cum-Faithfulness: a rule like A → B / C___D combines a tacit "phonological problem" (a.k.a. Markedness constraint) that can be expressed as *CAD, together with the licensing of the change from A to B to solve this problem. (This point was first made in Prince and Smolensky, 1993:4). That such generality is worth achieving becomes evident when we note that the same Markedness constraint gives rise to different Faithfulness-violating remedies, both across languages (Kiparsky 1973) and across contexts within the same language (Kisseberth 1970). In other word, giving a distinct status to Markedness constraints solved the "conspiracy problem" that had baffled theorists since the 1970s.

Beyond this, a Markedness constraint sometimes can have the effect of *blocking* a change in some languages but *causing* it in another. Thus *N[+cont] blocks lenition in K1 and Spanish; but actively causes fricatives to become stops in languages with "postnasal hardening" (Rosenthall 1989).

A further benefit from OT's separation of Markedness problems from Faithfulness remedies is that we can make language-specific phonological analysis more responsible to typology: when things go well in OT (as they often do), we find that a very detailed language-specific analysis can be reduced to a language-specific ranking of principles each of which has broad cross-linguistic support. In the present case of our K1 analysis, we find that the three principles we appeal to would (with suitable library research) all would be found to have extensive precedents beyond K1 (or more precisely, the real-life Khalkha on which K1 is based.) Stops lenite non-initially in a great number of languages Gurevich (2011), nasals frequently impede lenition, and content words often resist alternation more than function words do. Thus our analysis of K1 can be argued to have been assembled from "wholesome," typologically-sensible ingredients. At the very least, OT forces the analyst to declare which constraints in the analysis have broad cross-linguistic support and which ones are parochial; as we will see in Chapter 7, this distinction can have detectible consequences.

2.2 K2 and free variation

Let us move to our next hypothetical language, K2. In K2, pretty much everything is the same, except that the process of /p/ lenition is optional. While it is possible for a native speaker

of K2 to say [mʊntəg **w**oɮən] for item (1a), it is also possible to say [mʊntəg **p**oɮən]; and similarly for all similar items. Since both are grammatical in the language, but no meaning difference arises, the phenomenon is commonly referred to as **free variation**.

Free variation is perhaps more common in language than some linguists (trained on problem sets) may be aware. The field of **variationist linguistics**[4] (Labov, 1969, 1973; Wolfram, 1993; Guy, 1993; Walker, 2010) has long studied free variation processes and achieved careful documentation of them in a variety of languages. Our own impression is that even in work with individual language consultants, free variation will often manifest itself provided the researcher is alert to the possibility — it often pays to ask the consultant "is also possible to say it this way?". Typically, carefully analyzed free variation is not a sea of chaos, but is phonologically structured (Cedergren and Sankoff 1974; Sankoff 1988), in ways that we will suggest are well suited to being addressed with MaxEnt grammars.

For the simplest cases of free variation, there is an immediately obvious possible approach, which occurred to theorists (notably Kiparsky, 1993, 2006ab; Anttila, 1995, 1997) almost as soon as OT was invented: instead of imposing a full ranking of the constraints in the grammar, we let certain constraint pairs be ranked freely. In the case of K2, we would posit free ranking for the constraints *p and IDENT(sonorant), as shown in the pair of tableaux in (6) below. Note that both tableaux retain the former free ranking of IDENT(sonorant)$_{\text{content}}$ and *N[+cont]— this is a "trivial" free ranking, in that it has no effect on the outcome. The essential free ranking is that of *p and IDENT(sonorant), since when these are ranked differently, different candidates win.

(6) *A tableau pair for a case in which two constraints are freely ranked*

/mʊntəg **p**oɮ-ən/ great become-NPST	IDENT(son)$_{\text{content}}$	*N[+cont]	*p	IDENT(son)
mʊntəg **p**oɮən			*!	
☞ mʊntəg **w**oɮən				*

PLUS

/mʊntəg **p**oɮ-ən/ great become-NPST	IDENT(son)$_{\text{content}}$	*N[+cont]	IDENT(son)	*p
☞ mʊntəg **p**oɮən				*
mʊntəg **w**oɮən			*!	

This use of multiple tableaux is clumsy, and normally we just write one, with a dotted vertical line separating the freely ranked constraints, just as it does in (6) for IDENT(sonorant)$_{\text{content}}$ and *N[+cont].

[4] Also called **sociolinguistics**, especially in earlier work, though that term is now used more broadly. See Walker (2010) ch.1 for discussion.

(7) *A single tableau with free ranking that predicts multiple outputs*

/mʊntəg **p**oɮ-ən/ great become-NPST	IDENT(son)$_{content}$	*N[+cont]	*p	IDENT(son)
☞ mʊntəg **p**oɮən			*	
☞ mʊntəg **w**oɮən				*

Here in (7), we have just one pair of freely ranked constraints affecting the outcome, but the theory permits in principle any number of pairs (or triplets, quadruplets, etc).

2.3 The interpretation of free ranking as probability

The grammar illustrated in (7) is a rough-and ready account of free variation: it simply tells us that two outcomes are possible, accepting both [mʊntəg **p**oɮən] and [mʊntəg **w**oɮən]. Such a grammar would not make any predictions about which outcome was more common, and under what circumstances. However, in the work of Kiparsky (1993, 2006ab) and Anttila (1997), grammars like the one illustrated in (7) are used to predict precise probability distributions over candidate outputs.

This body of work proposes and refines an important and still-advocated theory of variation in OT. It has various names (Kiparsky, 1993; Anttila, 1997; Riggle, 2010), but we will call it the **Theory of Partially Ordered Constraints**, abbreviated "**POC**", following Coetzee and Pater (2011). To explain how the theory works, we use the example of tableau (7). Going across the constraints in this tableau from left to right, we let each group of constraints delimited by solid lines be called a *constraint stratum*. Thus for (7) the constraint strata are as in (8).

(8) *Constraint strata for tableau* (7)

a. *First stratum*

IDENT(sonorant)$_{content}$, *N[+cont]

b. *Second stratum*

*p, IDENT(sonorant)

Within each stratum (no matter how many constraints are in it), ranking under POC is assumed to be free. Further, it is assumed that any constraint in a higher stratum must outrank all constraints of a lower stratum. In order to predict probabilities over outputs, we first need to count the rankings compatible with these strata. Here, each stratum with two constraints in it gives rise to two possible rankings of those constraints; and because we have two strata each with two freely ranked constraints in them, the total number of rankings is $2 \times 2 = 4$. To be specific, the four rankings are as given in (9):

(9) *Rankings compatible with* (8)

a. IDENT(sonorant)$_{content}$ ≫ *N[+cont] ≫ *p ≫ IDENT(sonorant)
b. IDENT(sonorant)$_{content}$ ≫ *N[+cont] ≫ IDENT(sonorant) ≫ *p
c. *N[+cont] ≫ IDENT(sonorant)$_{content}$ ≫ *p ≫ IDENT(sonorant)
d. *N[+cont] ≫ IDENT(sonorant)$_{content}$ ≫ IDENT(sonorant) ≫ *p

In the general case, a stratum with *n* constraints gives rise to *n!* (*n* factorial, or n × (n − 1) × (n − 2) ...) possible rankings. We calculate n_i! for each freely ranked stratum *i*, then multiply out all of these terms to find the ranking count.

Now that we can count rankings, we can say that the **probability** assigned to any candidate under POC is simply the number of rankings for which that candidate is the winner, divided by the total number of rankings. The actual computation of this probability when the constraint set is intricate may be difficult, but the underlying concept seems straightforward. In the case of (7), there are four possible rankings (2! times 2!), of which two derive [mʊntəg **p**oɮən] and two derive [mʊntəg **w**oɮən].[5] Hence the prediction is that the lenited form [mʊntəg **w**oɮən] will surface exactly half the time, a probability of 0.5. Note that no ranking produces lenition in content words, or after nasals, so surface forms like *[ʃittʃ͡ʰəxsə**n** **w**aisən] or *[ərʊɮəx **w**ɔɮəmʧ͡ʰɔi] are predicted to have a probability of zero.

What we have just established is sufficiently important that it is worth describing in more general terms: unlike classical OT, a grammar under POC is a **probabilistic grammar**. Instead of just selecting one or more candidates from GEN as grammatical outputs for the given input, it assigns a probability value (often zero) to each member of GEN. It is probabilistic grammars with which we will concern ourselves in this book, as we seek to find analyses for K2 and other more complex cases.

POC is, arguably, the minimal extension of classical OT that can be used to create probabilistic grammars. POC grammars have been adopted for the analysis of a variety of systems with free variation (e.g. Pater 2000; Zuraw 2002). For the reasons we think POC is inadequate as a theory of gradient grammar, see §9.5.4 below; here, we move on directly to MaxEnt.

2.4 K3: modeling arbitrary probabilities

When researchers turn to quantitative examination of large datasets (texts, corpora, dictionaries, or datasets collected from speakers), it is common to find that variants are not equally likely as they were in K2; for discussion, see §9.1.

[5] Recall that the two constraints of the first stratum, IDENT(son)$_{content}$ and N[+cont], do not conflict, so that their ranking does not matter.

To explore this type of situation, we introduce our third hypothetical language, K3, which will involve us with quantitative data. To give a sense of realism, we look ahead to the discussion (§2.7.2) of real Khalkha, as studied by Fuller (2024). Fuller obtained her data by carefully transcribing the phonetic realization of every instance of phonemic /p/ in TEDx talks (https://tedxulaanbaatar.com/), totaling about 53 minutes of speech, given by four native speakers of Khalkha in Ulaanbaatar, the capital city of Mongolia. From these data, Fuller collected 1334 instances of /p/, sufficient to shed light on the quantitative patterning of /p/ lenition in relatively formal Khalkha speech. Importantly, each individual talker in her data exhibited similar variation. Our synthetic data for K3 follow the same trends that Fuller observed, but have been constructed to be a maximally clear pedagogical example.

In K3, when a function word is preceded by a non-nasal sound, as in /mʊntəg **p**oɮ-ən/ above, our constructed data include 300 instances of [p] and 110 instances of [w]. Examples are given in (10). Crucially, surface [p] and surface [w] are not equally likely (as the tableau in (7) above, with partially-ranked constraints, would predict); in fact, the probability of weakened [w] is 110/(110+300) = 0.27. The remaining data are similar to K1 and K2; that is, lenition never occurs in content words or after nasals.

(10) *Synthetic data for K3*

UR type	*Example*	*SR*	*count*
... [−nasal] [$_{\text{function}}$ p...	/mʊntəg **p**oɮ-ən/ great become-NPST	p	300 (73%)
		w	110 (27%)
... [+nasal] [p...	/ʃit-t͡ʃʰəx-sə**n** pai-sən/ decide-PERF-PST be-PST	p	212 (100%)
		w	0
... [$_{\text{content}}$ p...	/ɔrʊɮ-əx **p**ɔɮəm-t͡ʃʰɔi / involve-FUT opportunity-have	p	720 (100%)
		w	0

In order to model the data of K3, we seek a probabilistic grammar which can predict the exact lenition rates that we see in the data: 0% in content words and after nasals, and 27% in function words. This will consist of a MaxEnt grammar, so we first need to take a fairly substantial detour to explain what a MaxEnt grammar is.

2.5 MaxEnt: an introduction

Maximum Entropy Grammar is a variant of Optimality Theory, designed for the purpose of formulating probabilistic grammars. The key difference is that instead of using constraint ranking, MaxEnt assigns a numerical value to each constraint, its **weight**, which intuitively reflects its strength. We reemphasize here that in most respects MaxEnt is very much like OT (§4.2): it evaluates candidates in parallel, uses a potentially infinite GEN, and atomizes phonological analyses into simple, independent constraints; thus solving the conspiracy problem and linking language-specific analyses to phonological typology.

2.5.1 Preview: Classical Harmonic Grammar

To present MaxEnt clearly, we will begin with a simpler theory which will serve as a kind of way-station: **Classical Harmonic Grammar**, a theory originating with Legendre et al. (1990). For comparison with classical OT, see Pater (2009). In Classical Harmonic Grammar, the constraint weights and violations are used to calculate a **Harmony** ($\mathcal{H}$) score for each candidate, which, despite its name, acts as a kind of penalty. The output of the grammar in this theory is the least-penalized member of GEN for a given input. Note even though Classical Harmonic Grammar uses numbers, it is *not* a probabilistic theory: other than in rare cases of tied harmony, it produces one single winner for each candidate competition, just as Optimality Theory does.

A Harmony score is calculated as follows:

(11) *Calculating the Harmony of a candidate*

a. For each constraint, multiply the number of violations the candidate incurs by the constraint weight.
b. Sum the result obtained in step (a) across all constraints.

The same computation is expressed as a formula in (12):

(12) *Formula: Harmony for candidate* x

$$\mathcal{H}(x) = \sum_{i \in Constraints} w_i \times v_i(x)$$

where $v_i(x)$ is the number of violations incurred by candidate x for constraint i.

We illustrate with an example. Since Classical Harmonic Grammar allows only one single winner per input, we return here to the non-varying data of K1 ((1) above). The crucial tableaux are shown in (13), which are Harmonic Grammar versions of our OT tableaux in (5). For these tableaux, we have hand-selected the weights for each constraint.

(13) *Classical Harmonic Grammar tableaux for K1*

	IDENT(son)$_{cont.}$	*N[+cont]	*p	IDENT(son)	
weights:	3	3	2	1	Harmony
a. /mʊntə**g** **p**oɮ-ən/ great become-NPST					
mʊntə**g** **p**oɮən			1		2
☞ mʊntə**g** **w**oɮən				1	1
b. /ɔrʊɮ-əx **p**ɔɮəm-t͡ʃʰɔi / involve-FUT opportunity-have					
☞ ɔrʊɮəx **p**ɔɮəmt͡ʃʰɔi			1		2
ɔrʊɮəx **w**ɔɮəmt͡ʃʰɔi	1			1	4
c. /ʃit-t͡ʃʰəx-sə**n** **p**ai-sən/ decide-PERF-PST be-PST					
☞ ʃitt͡ʃʰəxsə**n** **p**aisən			1		2
ʃitt͡ʃʰəxsə**n** **w**aisən		1		1	4

We start with an example of Harmony calculation. Candidate [mʊntəg **p**oɮən] violates *p once. The weight of *p is 2, hence the Harmony contribution from this constraint for this candidate is 2. Since [mʊntəg **p**oɮən] *only* violates *p, its total Harmony score is therefore 2. Candidate [ʃitt͡ʃʰəxsə**n** **w**aisən] incurs one violation of *N[+cont], which has a weight of 3, and one violation of IDENT(sonorant), with a weight of 1. So its Harmony is 1 × 3 + 1 × 1 = 4.

Once we have calculated Harmony, it is straightforward to determine the unique winner for each input, namely the candidate with the lowest Harmony penalty. This candidate is marked in (13) with a pointy finger.

In (13), each constraint is only violated once. To illustrate the multiplication of violation counts by weights, we consider another example from Fuller (2024), which contains multiple p's.

(14) *Multiple lenition sites in a single input*

Odoo	*khümüüs*	*oroi*	***untdag***	***polson***	***pai-na***
			/ʊntʰ-təg	pɔɮ-sən	pai-n/
now	people	late	sleep-HAB	become-PST	be-NPST

'Nowadays, people sleep late.'

The Harmony of neither, just one, or both /p/'s leniting is shown in (15):

(15) *Calculating Harmony in Classical Harmonic Grammar with multiple constraint violations*

	IDENT(son)$_{cont.}$	*N[+cont]	*p	IDENT(son)	
weights:	3	3	2	1	$\mathcal{H}$
/ʊnt^h-təg **p**ɔɮ-sən **p**ai-n/					
ʊnt^{h}təg **p**ɔɮsən **p**ain			2		4
☞ ʊnt^{h}təg **w**ɔɮsən **p**ain			1	1	3
ʊnt^{h}təg **p**ɔɮsən **w**ain		1	1	1	6
ʊnt^{h}təg **w**ɔɮsən **w**ain		1		2	5

As can be seen, for candidate [ʊnt^{h}təg **p**ɔɮsən **p**ain], the constraint *p contributes 4 units of Harmony, the product of two violations and its weight of 2. The winner here is actually [ʊnt^{h}təg **w**ɔɮsən **p**ain], with lenition on the first, but not the second, /p/. This one violates *p only once, and IDENT(sonorant), which has a lower weight, once. This makes its Harmony just 3, lower than the 4 of the fully faithful candidate. Either candidate where the second p is lenited (the last two) also violates N[+cont], with its higher weight of 3.

2.5.2 Shifting to MaxEnt

The discussion of Classical Harmonic Grammar just given provided the mechanisms for converting weights and violation patterns into Harmony scores. For MaxEnt itself, there is one more step, namely to convert Harmony scores into probabilities. MaxEnt is a probabilistic version of Classical Harmonic Grammar, just as POC is a probabilistic version of OT. The math we will cover here will have two essential properties. First, it will assign lower probability to candidates that have higher Harmony scores. Second, the probability assigned to the full set of candidates for a given input will sum to one; this is what it means to be a probability distribution.

In (16), we give the MaxEnt transformation from Harmony to probability in the form of a recipe.

(16) *Calculating probability from Harmony values in MaxEnt*

a. Calculate the **Harmony** ($\mathcal{H}$) of each candidate, as in (12).

b. For each Harmony value, first make it negative, and then exponentiate it using Euler's number (*e*).[6] We will call this value **eH**:

$$eH = e^{-\mathcal{H}}$$

c. For each input, sum up the eH over all the candidates for that input. Call this **Z**.

$$Z = \sum_{i \in candidates} eH_i$$

[6] Euler's number *e*, useful in mathematics for many reasons, is about 2.718. MaxEnt would work perfectly well with some other base, such as 10, but we are following standard practice here.

d. For each candidate, divide its eH by Z. This is its predicted **probability**.

$$P(i) = \frac{eH_i}{Z}$$

To illustrate these mechanisms, we return to K3, the language in which [mʊntəg **p**oɮən] occurs 73% of the time and [mʊntəg **w**oɮən] 27%. We seek to formulate a MaxEnt grammar that matches this percentage, and also derives the invariant outcome for the remaining two inputs.

Formally, a MaxEnt grammar consists of a set of constraints, each paired with a weight. Our proposed grammar is given in (17).

(17) *A MaxEnt grammar for K3*

Constraint	*Weight*
*p	3
IDENT(son)	4
$\text{IDENT(son)}_{\text{content}}$	20
*N[+cont]	20

The weights in (17) were chosen by hand. While it is possible to find a good fit to the data by choosing weights by hand for a very small dataset like this one, realistic analyses generally require the weights to be fit by machine, a topic taken up in §2.6.

For now we will simply accept the weights of (17) and proceed with the computation. All the needed values appear in the tableaux in (18) below, which may also be consulted in spreadsheet form in the supplementary materials for Chapter 2 as **K3.xlsx**. The weights of (17) are shown in (18) underneath their respective constraint names; and the constraints assign violations in the same way as in previous tableaux.

(18) *Tableaux: MaxEnt analysis of K3*

/mʊntə**g** **p**oɮ-ən/ great become-NPST			IDENT(son)$_{content}$	*N[+cont]	*p	IDENT (son)				
		weights:	20	20	3	4				
	Freq.	**prob.**					$\mathcal{H}$	eH	Z	p
mʊntə**g** **p**oɮən	300	0.73			1		3	0.050	0.068	0.73
mʊntə**g** **w**oɮən	110	0.27				1	4	0.018	0.068	0.27
/ɔrʊɮ-əx **p**əɮəm-t͡ʃʰɔi / involve-FUT opportunity-have										
ɔrʊɮəx **p**əɮəmt͡ʃʰɔi	720	1			1		3	0.049	0.050	≈1
ɔrʊɮəx **w**əɮəmt͡ʃʰɔi	0	0	1			1	24	3.77×10^{-11}	0.050	≈0
/ʃit-t͡ʃʰəx-sə**n** **p**ai-sən/ decide-PERF-PST be-PST										
ʃitt͡ʃʰəxsə**n** **p**aisən	212	1			1		3	0.049	0.050	≈1
ʃitt͡ʃʰəxsə**n** **w**aisən	0	0		1		1	24	3.77×10^{-11}	0.050	≈0

To illustrate the calculations, we first derive step by step the probability of the first candidate in the first tableau, [mʊntəg **p**oɮən], which competes with [mʊntəg **w**oɮən] as the output for /mʊntəg **p**oɮ-ən/. First, its Harmony is calculated in the same way as in Classical Harmonic Grammar; it comes out as 3 × 1 = 3. To convert this Harmony value into a probability, we follow the procedures in (16). eH, following (16b), comes out to $e^{-3} = 0.050$. Next, (16c) tells us to calculate Z by summing eH values for both candidates. Here, we calculate eH for the competitor candidate [mʊntəg **w**oɮən], which comes out to 0.018, then add the two eH values to obtain 0.067. Lastly, we calculate the probability of [mʊntəg **p**oɮən] according to (16d), by dividing its eH by Z. This is 0.050/0.068 = 0.73. The last three columns of (18) illustrate the computations for all candidates.

As you can see, we have chosen our weights such that the MaxEnt-predicted probabilities closely match the frequencies of our toy data set, given as raw numbers in the second column of (18) and as proportions in the third. The generation of probabilities like 0.73 is beyond the capacity of Classical OT or Classical Harmonic Grammar, which are nonstochastic theories. Using POC (§2.3) would also be insufficient, since only a probability distribution of 50/50 can be generated.[7] MaxEnt, in contrast, can generate any probability distribution in such cases with a suitable choice of weights.

[7] For a refinement of this point see §9.5.4.

At this point we have illustrated the core of MaxEnt, the math that converts Harmony ($\mathcal{H}$) to probability. To wrap things up, we reexpress the MaxEnt recipe of (16) as a single mathematical formula, given in both terse and verbose form; the latter spells out the Harmony calculation.

(19) *Maxent formula*

a. Terse

$$P(x) = \frac{e^{-\mathcal{H}(x)}}{Z} \text{ where } Z = \sum_i e^{-\mathcal{H}(cand_i)}$$

Key: x is a candidate for some input
$P(x)$ is the probability the grammar assigns to x
e^t is e taken to the tth power
$\mathcal{H}(x)$ is the harmony of candidate x
$cand_i$ is the ith candidate for the given input
Σ_i denotes the sum over all candidates

b. Verbose

$$p(x) = \frac{e^{-\sum_{i \in Con} w_i \times v_i(x)}}{\sum_{y \in candidates} e^{-\sum_{i \in Con} w_i \times v_i(y)}}$$

w_i is weight of the ith constraint
$v_i(y)$ is the number of times candidate y violates the ith constraint
Σ_i denotes the sum over all constraints

Before going on, we offer a caution to the reader: in the literature one will find variants of (19) that are essentially the same calculation. In particular, if we put negative signs in front of every violation, as in Coetzee and Pater (2011), then Harmony will come out negative, and we can omit the negation step in (16b). The same will hold true if we negate the constraint weights. Our own practice (positive weights to express constraint strength, violations counted in the ordinary way) derives from the original paper of Goldwater and Johnson (2003). In addition, the very word "Harmony" sometimes is used to designate something else; for instance Magri (2025) uses "Harmony" to designate what we call eH. These variants will generally be intelligible in context.

2.6 K4: Perturbers as "hidden" phonology in variation

We now move to our next hypothetical language, K4. This toy language is intended to represent a very common situation in patterns of variation. Not only is the probability of the two possible outcomes unequal, but additionally it is *affected by context*; that is, there is at least one independent factor — we will call it a **perturber** — that changes the output probabilities in a systematic way (Labov 1969, Bayley 2002). In a MaxEnt grammar, this additional factor will usually be treated using what we will call a **perturber constraint**. Candidates that violate the perturber constraint will have a probability distribution different from what is seen among non-perturber-violating candidates.

Our K4 language is set up to illustrate the effect of a perturber. As in K3, lenition does not occur in content words or after nasals. However, in other environments the lenition rate is affected probabilistically by the preceding context: specifically, lenition is more likely after a vowel than after a consonant. For example, in an input like /mʊntə**g p**oɮ-ən/, with a preceding consonant, lenition occurs about 17% of the time. But in /tʊtʰ**ʊ p**ai-ga/, with a preceding vowel, lenition occurs considerably more often: 83% of the time. Hence, a preceding vowel is a perturber that increases the probability of lenition. The pattern is summarized in (20).

(20) *Synthetic data for K4*

UR type	Example	SR	*count*
... C [$_{\text{function}}$ p ...	/mʊntəg **p**oɮ-ən/ great become-NPST	p	*290 (83%)*
		w	*60 (17%)*
... V [$_{\text{function}}$ p ...	/tʊtʰʊ **p**ai-ga/ missing be.PROGPRES	p	*10 (17%)*
		w	*50 (83%)*
... [$_{\text{content}}$ p ...	/ɔrʊɮ-əx **p**ɔɮəm-t͡ʃʰɔi / involve-FUT opportunity-have	p	*720 (100%)*
		w	*0*
...[+nasal] [p ...	/ʃit-t͡ʃʰəx-sə**n** pai-sən/ decide-PERF-PST be-PST	p	*212 (100%)*
		w	*0*

Our K4 language illustrates a very common situation when researchers examine probabilistic phonological patterns (whether this be lexica, corpora, or experiments): While at first it may appear that all varying forms pattern together (as was true in K3), further inspection reveals non-uniformity: after a vowel, we have predominant lenition (83%), whereas after a consonant lenition is far less usual (17%). If we did not examine the specific rates of lenition in these two contexts, but pooled them together (as we likely would have done if we were only interested in modeling the mere fact of variation), we would have failed to notice the effects of the perturber environment.

To formalize the perturber effect in K4, we adopt a commonplace lenition constraint as our perturber constraint, defined in (21):

(21) POSTVOCALICLENITE

Assign a violation for every sequence of the form [+syllabic][−sonorant].

The idea here is that, as in in many languages, K4 disprefers postvocalic stops, and replaces them with a more sonorous consonant, in this case [w].

For the moment, we stick with the constraint weights we used for K3, but add in POSTVOCALICLENITE, with a weight of 2, chosen by hand. The tableaux in (22) illustrate the predictions of this new grammar. The new input /tʊtʰ**ʊ p**ai-ga/ is now included, for which the POSTVOCALICLENITE constraint affects the predicted probability distribution.

(22) *Tableaux for K4 (hand-fit weights)*

a. Simple leniting environment (no perturber)			IDENT(son)$_{content}$	*N[+cont]	*p	IDENT (son)	POSTVOCLENITE				
/mʊntə**g** **p**oɮ-ən/ great become-NPST	**Freq.**	**prob.**	20	20	3	4	2	$\mathcal{H}$	eH	Z	p
mʊntə**g** **p**oɮən	290	0.829			1			3	0.050	0.068	0.73
mʊntə**g** **w**oɮən	60	0.171				1		4	0.018	0.068	0.27
b. Leniting environment with postvocalic perturber											
/tʊt^{h}**ʊ** **p**ai-ga/ missing be.PROGPRES											
tʊt^{h}**ʊ** **p**ai-ga	10	0.167			1		1[8]	5	0.007	0.025	0.27
tʊt^{h}**ʊ** **w**ai-ga	50	0.833				1		4	0.018	0.025	0.73
c. Blockage in stems											
/ɔrʊɮ-əx **p**ɔɮəm-t͡ʃ^{h}ɔi / involve-FUT opportunity-have											
ɔrʊɮəx **p**ɔɮəmt͡ʃ^{h}ɔi	720	1			1			3	0.050	0.050	≈1
ɔrʊɮəx **w**ɔɮəmt͡ʃ^{h}ɔi	0	0	1			1		24	3.77* 10^{-11}	0.050	≈0
d. Blockage after nasal											
/ʃit-t͡ʃ^{h}əx-sə**n** **p**ai-sən/ decide-PERF-PST be-PST											
ʃitt͡ʃ^{h}əxsə**n** **p**aisən	212	1			1			3	0.050	0.050	≈1
ʃitt͡ʃ^{h}əxsə**n** **w**aisən	0	0		1		1		24	3.77* 10^{-11}	0.050	≈0

Our predicted probabilities are reasonably close to the observed ones. (We will see in §2.6.2 that fitting the weights by algorithm obtains a closer fit.) The key point for now is that we have successfully incorporated the effect of POSTVOCALICLENITE as a perturber constraint: it lowers the output probability of candidates that violate it (here 0.27 for [tʊt^{h}**ʊ** **p**ai-ga] vs. 0.73 for [mʊntə**g** **p**oɮən]), but does not rule out violating candidates entirely.

2.6.1 More on perturber constraints

It is common for constraints that act as a perturber in one language to appear as a fully determinative factor in another language. For example, Spanish is said to be a language in which post-vocalic lenition (of voiced stops) is entirely obligatory: *navidad* /nabidad/ 'Christmas' *must* be [naβiðað], and cannot realize any of the obstruents as a stop. Although our use of postvocalic lenition as a perturber in Khalkha is invented, we cover in §9.1.3 several real-language pairs of

[8] Note that since POSTVOCALICLENITE is assessed for all segment pairs, there should actually be actually many violations present throughout the tableaux. We record only the violations that distinguish candidates; the outcomes remain the same.

phenomena where a factor that fully determines the outcome in an obligatory process in one language or languages acts as a perturber in another language. To give just one example here, in Iraqi Arabic a ban on triple consonant clusters (*CCC) is categorical: when such sequences arise across morpheme boundaries, they are obligatorily repaired (Broselow, 1980). But in English, the same constraint acts as a perturber; it merely increases the frequency with which CCC clusters are repaired in surface forms. For example, the final /t/ of *tanked* [tæŋkt] is optionally deleted more often than the final /t/ of *tact* [tækt] (Labov, 1989). In the context of a constraint-based theory like MaxEnt, this pretheoretical observation has a clear theoretical interpretation, namely, that variable and categorical phenomena use the same constraint set. Obligatoriness vs. optionality is simply a consequence of the particular weights that have been assigned.

The pervasive appearance of perturbers in variation is perhaps the best argument for the use of MaxEnt or similar frameworks as an analytical tool: without creating a probabilistic analysis of the data, one is liable to miss important parts of the language system. In Classical Optimality Theory, a nonstochastic theory, there is no way to treat perturber phenomena — it was simply not designed for this purpose. MaxEnt gives us a way to integrate perturber phenomena into language-specific analysis and, indeed, into the study of typology.

We also suggest that MaxEnt might also be of assistance to the working linguist in trying to make sense of new data: it encourages the analyst to seek out gradient generalizations and test them through explicit formal analysis.

2.6.2 Fitting the weights

Establishing the weights is, of course, part of the MaxEnt analysis, analogous to finding a workable ranking in Optimality Theory. However, weighting is different from ranking in that it is almost impossible to do accurately by hand. We have found it a useful classroom exercise in teaching beginning MaxEnt to ask our students to do hand-weighting; we find that it is possible but generally difficult, and usually leads to a less accurate grammar than we can obtain by using a machine-implemented algorithm. Following standard usage, we use the term **fitting** to denote the process of finding the best available weights for a given set of data and constraints.

MaxEnt weights can be fit by algorithm using a variety of different software types, which we cover briefly in §2.11. For the MaxEnt beginner, we highly recommend the use of Excel for implementing the MaxEnt math, and the Excel Solver add-on for fitting weights. We recommend this approach mainly for transparency. The cells of the spreadsheet display in full the math that calculates harmony and converts it to probability. Furthermore, the Excel format makes it easy to implement changes to the analysis and understand their consequences. More experienced MaxEnt practitioners may also find the Excel format useful, as it makes it easy to introduce novel elements to the structure of the theory. We will illustrate many examples of this throughout the later chapters of the book. Lastly, Excel sheets can be placed into online Supplementary Materials of a published paper, so that the calculations become freely inspectable, and modifiable, by anyone with access to Excel.

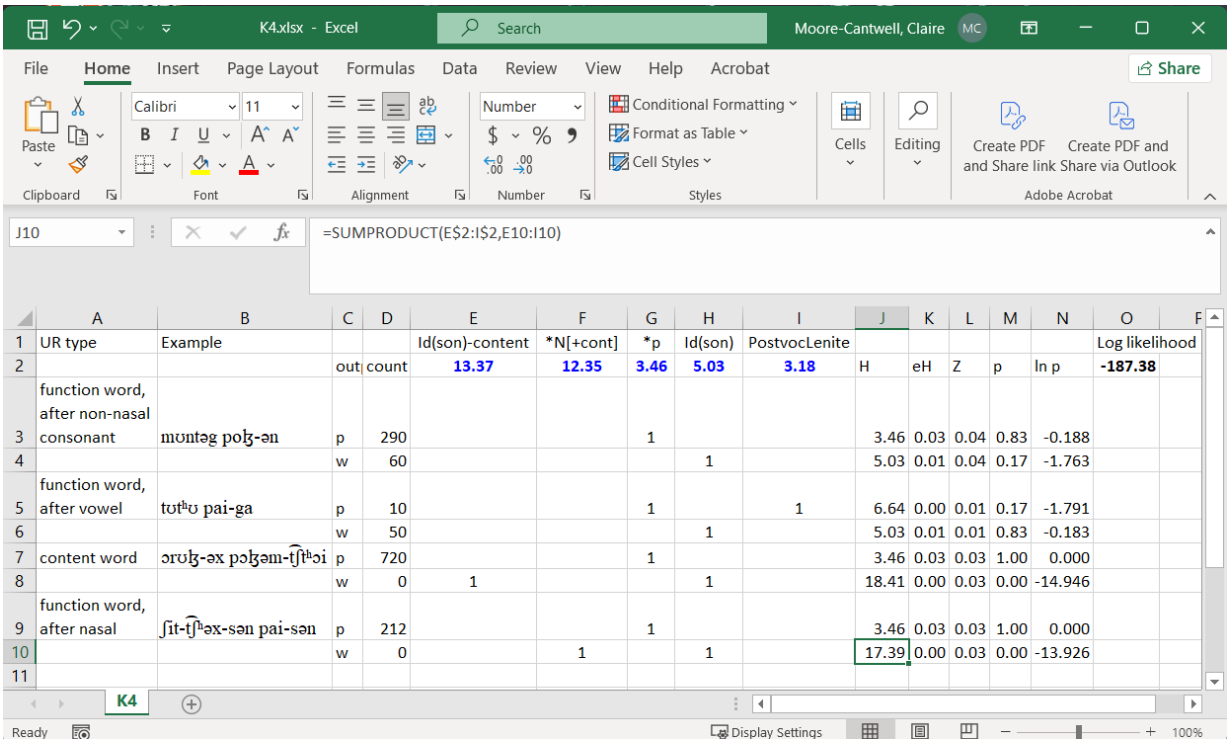
What follows is a "how-to" for implementing a MaxEnt analysis in Excel. We first show how to implement the MaxEnt formula (19) in the format of an Excel spreadsheet (§2.6.3). With this in place, we show how to calculate the weights (§2.6.4).

2.6.3 Implementing the MaxEnt formula on a spreadsheet

The first step is to implement the MaxEnt formula, given above in (19), using the resources of Excel. For this purpose we return to our K4 example. The spreadsheets discussed here may be obtained in the Supplementary Materials for this book, available at the web locations listed in §1.5; the particular spreadsheet we will be working with can be found among the Chapter 2 files as **K4.xlsx**. This spreadsheet is given in print below, and you may find it helpful not to download it but rather to type it out yourself as a study aid.

Figure 1 is a screen clip of the actual Excel interface. It is opaque, as one sees the "underlying" values of the cells only when you hover the mouse over them. The content of cell J10 has been selected and thus shows up in the formula bar (large white window).

Figure 1: The Excel interface



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	UR type	Example			Id(son)-content	*N[+cont]	*p	Id(son)	PostvocLenite						Log likelihood
2			out	count	13.37	12.35	3.46	5.03	3.18	H	eH	Z	p	ln p	-187.38
3	function word, after non-nasal consonant	mʊntəg poɣ-ən	p	290			1			3.46	0.03	0.04	0.83	-0.188	
4			w	60				1		5.03	0.01	0.04	0.17	-1.763	
5	function word, after vowel	tʊtʰʊ pai-ga	p	10			1		1	6.64	0.00	0.01	0.17	-1.791	
6			w	50				1		5.03	0.01	0.01	0.83	-0.183	
7	content word	ɔrʊɣ-əx pɔɣəm-t͡ʃʰɔi	p	720			1			3.46	0.03	0.03	1.00	0.000	
8			w	0	1			1		18.41	0.00	0.03	0.00	-14.946	
9	function word, after nasal	ʃit-t͡ʃʰəx-sən pai-sən	p	212			1			3.46	0.03	0.03	1.00	0.000	
10			w	0		1		1		17.39	0.00	0.03	0.00	-13.926	
11															

The table in (23) shows the underlying structure of Figure 1, in particular the formulae that compute the displayed values.[9] We begin at column B, which contains the example inputs of each type (they appear shortened in (23) relative to the actual spreadsheet). Column C contains the candidate outputs, and column D contains the counts of each candidate output 'observed' in our toy data. Columns E-I contain constraint names in row 1 and weights in row 2. Constraints in columns E-G are not shown, but in the full version of the spreadsheet are as in (22).

[9] For specific Excel formulae, we find the "Help" menu within Excel itself can be useful. Many training resources for Excel exist online.

(23) *Calculating MaxEnt probabilities in Excel*

	Excel column names									
row #	B	C	D	...	H	I	J	K	L	M
1	Example			...	Id(son)	Postvoc Lenite	Harmony	eH	Z	p
2		Output	count		4.00	2.00				
3	g poʞ-	P	290				=SUMPRODUCT (E$2:I$2, E3:I3)	=EXP(-J3)	=SUM(K3:K4)	=K3/L3
4		W	60		1		=SUMPRODUCT (E$2:I$2, E4:I4)	=EXP(-J4)		=K4/L3
5	ʊ pai-	P	10			1	=SUMPRODUCT (E$2:I$2, E5:I5)	=EXP(-J5)	=SUM(K5:K6)	=K5/L5
6		W	50		1		=SUMPRODUCT (E$2:I$2, E6:I6)	=EXP(-J6)		=K6/L5
…	…	…	…	…	…	…	…	…	…	…

The next few columns implement the MaxEnt math, closely following the procedure of (16). Column J calculates the Harmony of each candidate, rendering in Excel the formula from (12), $\sum_{i \in Constraints} w_i \times v_i$(candidate). It uses the Excel function `SUMPRODUCT()`, which performs two steps at once: it multiplies together the individual values for constraint violation counts and weights (columns E-I), then sums the result across all constraints. The expression "`E$2:I$2, E3:I3`" in cell `J3` follows standard Excel conventions: `E$2:I$2` uses the '$' symbol to mean "always the cells in row 2", hence always the row of weights; but "`E3:I3`", indicates "whatever appears in cells E-I of this row" (currently, row 3). By using this notation, it is possible to write the function once, in the top cell (here, `J3`) then copy it down to fill the rest of the column. Excel will automatically update any contextual references, so that the constraint weights will be multiplied by whatever violation counts are present in a given row.

Column K computes the eH value of each candidate *x* (this is $e^{-\mathcal{H}}$, from (16b)) by first negating, then exponentiating, the result of Column J. As can be seen, the Excel notation for e^x is `EXP(x)`. As before, this value can be written once and copied down the column.

Column L calculates the normalizing factor Z for each input (16c), using the `SUM()` function to total the eH values for all (both) candidates.

Column M carries out the final step, from (16), calculating the probability assigned to each candidate by dividing the eH from column K by the Z value from the appropriate cell in column L.

2.6.4 Setting the weights by algorithm

Now that we have expressed the MaxEnt calculations in a spreadsheet, we can turn to our original problem of using software to find the most accurate constraint weights.

To begin, we first have to quantify what we mean by "accurate." We will employ a widely-used metric of accuracy: the **log likelihood** assigned by the grammar to the data (Jurafsky and Martin, 2026). Log likelihood is a measure of how likely our observed data are given our current set of weights. It is defined as the log of the probability of the observed data if the model is assumed to be correct. A higher log likelihood value implies greater accuracy. The search for optimal constraints is now defined as finding the weights that maximize log likelihood. We delay detailed discussion of the rationale for log likelihood to §4.4 below; here we simply give the steps that are needed to calculate it in Excel.

The procedure for log likelihood calculation in Excel is illustrated in (24), which is extracted from **K4.xlsx**; we show the additional columns which are used to calculate log likelihood. Column M, already seen, gives the predicted probability for each candidate. Column N converts this probability to log probability, using the log function (`LN()`) in Excel. Since many of the data types occur multiple times in the K4 data (see column D), we take account of each datum by multiplying these log probability values by the frequencies, using the convenient `SUMPRODUCT()` function, which we used earlier to calculate harmonies in column J. The result is given in cell `O2`.

(24) *Calculating log likelihood in Excel*

	Excel column names									
row #	B	C	D	...	H	I	...	M	N	O
1	`Example`			...	`Id(son)`	`Postvoc Lenite`	...	`p`	`log p`	`Log likelihood`
2		`output`	`count`		`4.00`	`2.00`				`=SUMPRODUCT( D3:D10,N3:N10)`
3	`g poʁən`	`P`	`290`					`=K3/L3`	`=LN(M3)`	
4		`W`	`60`		`1`			`=K4/L3`	`=LN(M4)`	
5	`ʊ pai-`	`P`	`10`			`1`		`=K5/L5`	`=LN(M5)`	
6		`W`	`50`		`1`			`=K6/L5`	`=LN(M6)`	
...	...	...	...	...	...	...	...	...		

Note that some candidates are never observed, so their frequency is zero; because of the multiplication, these candidates do not contribute to log likelihood.

When we view **K4.xlsx** in the normal mode of Excel, we see the values computed by the formulae of (24):

(25) *Same table, with displayed values*

	Excel column names									
row #	B	C	D	...	H	I	...	M	N	O
1	Example			...	Id(son)	PostvocLenite	...	p	log p	Log likelihood
2		output	count		4.00	2.00				-198.44
3	g poʁən	p	290					0.73	-0.313	
4		w	60		1			0.27	-1.313	
5	ʊ pai-	p	10			1		0.27	-1.313	
6		w	50		1			0.73	-0.313	
...	...	...	...	...	...	...	...	...		

As you can see, the log likelihood, −198.44, is a negative number, since the log of a probability (something less than 1) is always negative. Our goal is to maximize this value, specifically to find a negative value that is as close as possible to zero.

Now that we have a cell that calculates log likelihood, we can use it as our accuracy metric, guiding the search for the best weights. This can be carried out in Excel using the "Solver" utility, which comes with the program but must be activated.[10] Unfortunately, although Solver programs exist for Google Sheets and Libre Office, at the time of this writing these versions do not contain the algorithm necessary for fitting a MaxEnt model.

To find the best weights, we recommend carefully following these steps:

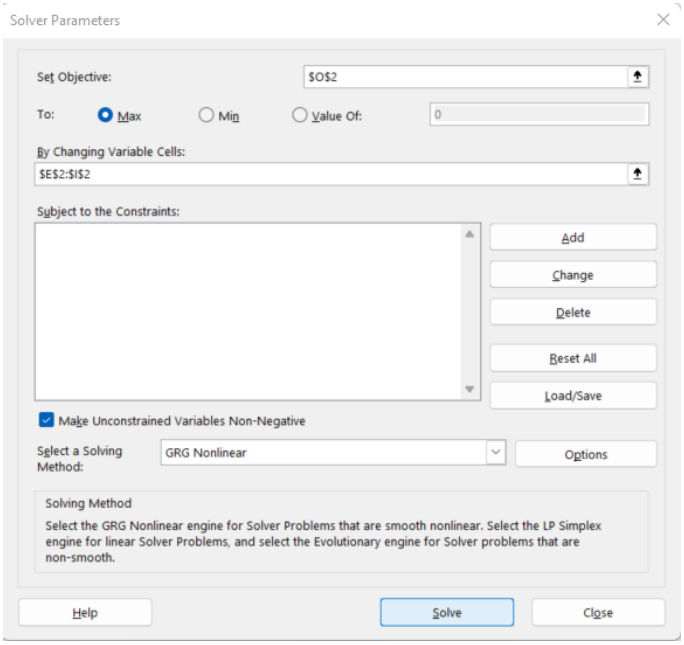
1) **Initialize:** Set all constraint weights to 0 (in **K4.xlsx**, this would be cells `E2:I2`).
2) **Select Log likelihood as objective:** Open the Solver interface, and select the cell where you have calculated log likelihood (here, cell `O2`) for the "Set Objective" field. Since you are trying to maximize log likelihood, you should select the "Max" button.
3) **Select weights:** Select the cells which contain your weights (cells `E2:I2`) in the "By changing variable cells" field.
4) **Keep weights non-negative:** In order to avoid negative weights, make sure that the checkbox next to "Make Unconstrained Variables Non-Negative" is checked.
5) **Select solving method:** Where it says "Select a solving method", choose "GRG Nonlinear" (usually the default).[11]

The settings of all these conditions for the Solver are shown in the screen clip of Figure 2.

[10] In the version of Excel we are currently using the procedure is: File > Options > Add-ins. In the Manage drop-down menu, select "Excel Add-ins" and click Go. Then, check the box next to "Solver Add-in" and click OK.

[11] GRG Nonlinear in general terms can be studied in Lasdon et al. (1974), though the specific implementation in the Solver is proprietary. GRG Nonlinear is just one of many algorithms that can be used to effectively compute MaxEnt weights; the choice of algorithm is in most cases a purely practical matter and will not affect the analytic results obtained; see §4.4.2.

Figure 2: Launching the Solver on the K4 problem



With these conditions set, one can click on the Solve button, and at least for schematic problems of the size described here, weights are found quite quickly. The weights for the K4 analysis are as follows:

(26) *Constraint weights for the K4 analysis*

IDENT(son)$_{content}$	13.37
*N[+cont]	12.35
*p	3.46
IDENT(son)	5.03
AGREE-round	3.18

These weights provide a close fit to the observed probabilities in K4, as (27) shows.

(27) *Observed vs. predicted probabilities for the K4 analysis*

Input (see Fig. 1*, column A)*	*candidate*	*observed probability*	*predicted probability*
function word, after non-nasal consonant	p	0.829	0.828
	w	0.171	0.172
function word, after vowel	p	0.167	0.167
	w	0.833	0.833
content word	p	1.000	1.000
	w	0.000	0.000
function word, after nasal	p	1.000[12]	1.000
	w	0.000	0.000

We now have a complete MaxEnt analysis of K4, which fits the data quite well. All steps in the creation of the analysis have been described explicitly, with the one exception of the GRG Nonlinear algorithm, internal to the Solver, that locates the sets of weights under which log likelihood has been optimized; this alone has been left as a black box.

If you have consulted this book with an eye to analyzing your own data, this might be an appropriate point to begin. You can use **K4.xlsx** as a template, adding or subtracting cells to match your own data.

2.7 K5: Multiple perturbers

Having introduced the basics of MaxEnt grammar via the toy languages K3 and K4, we now turn to a final toy language, which is closely based on the real Khalkha data analyzed by Fuller (2024). This more complex example will illustrate a more intricate type of data pattern that shows up widely in MaxEnt analysis.

First, a trivial change: we will treat the post-nasal environment not as an absolute blocker of lenition, but merely a downward perturber; thus forms like /ʃitt͡ʃʰəxsə**n p**aisən/, from K4, do occasionally undergo lenition, though substantially less often than forms without a nasal.

More fundamentally, in K5 we adopt a crucial observation made by Fuller (2024), who observed different rates of lenition in different function words. In *pi*, the first person pronoun, she observed almost no lenition (just 7 lenited forms out of 191). On the other hand for the particle *pol*, she observed 112 cases of lenition out of 163. She observed different rates in two other function words as well, one of which is homophonous with the particle *pol*. In this case we adopt for K5 some values actually taken from Fuller.

[12] Note that these predicted probabilities are not exactly 1 and 0, but rather very close to 1 and zero; close enough that Excel displays only 1 or 0.

(28) *Data from Fuller (2024), Figure 10*

Function word:		*Instances with p*	*Lenited instances*	*Lenition rate*
pol	particle used for interrogatives, conditionals, and focus	51	112	68.7%
pol-	copula, "to become"	55	57	50.9%
pai	copula, "to be"	142	99	41.1%
pi	1st person pronoun	184	7	3.7%

Fuller found that these preferences of each function word coexist alongside the grammatical effects of blockers (simplified as the nasals in our toy languages; see §2.7.2) and of the lenition-preferring environment (simplified as post-vowel). Each function word, though it has its own baseline probability of leniting, lenites more in the lenition-preferring environment, and less in the blocking environment.

In the present case, morpheme identity can be treated as another type of perturber, in this case lexical rather than phonological. If, hypothetically, we think of the lenition rate for *pol* as a sort of default, then using the other morphemes *pol-, pai*, or *pi* in effect perturbs the lenition rate downward.

To give the full data set, we cross-classify the effects of phonological environment already seen with the effect of morpheme identity, as in (29). The counts are based on Fuller's actual corpus data, but simplified somewhat.

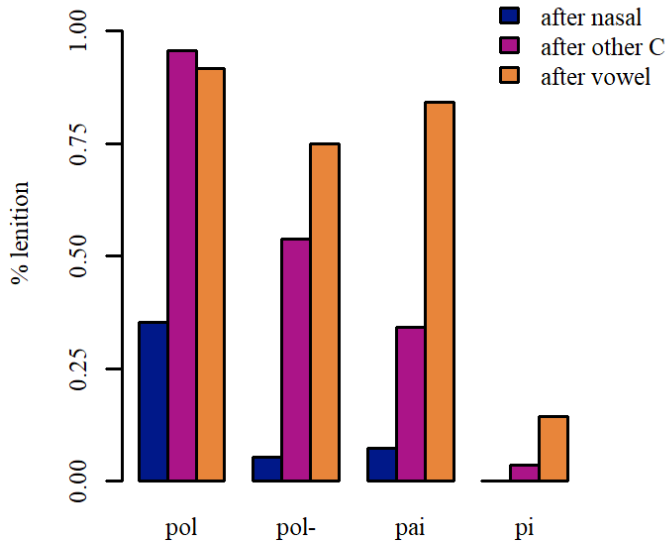
(29) *The K5 data pattern (**K5.xlsx**)*

Function word	*UR type*	*Example*	*SR*	*Observed*
pol	... C [$_{\text{function}}$ p ...	[natə**t p**ɔɮ]	p	3
			w	65 (95.6%)
	... V [$_{\text{function}}$ p ...	[ɔrəɮt͡sʰ**ɔ p**ɔɮ]	p	2
			w	22 (91.7%)
	... [+nasal] [p ...	[ju**m p**ɔɮ]	p	46
			w	25 (35.2%)
pol-	... C [$_{\text{function}}$ p ...	[t͡ʃiɮ**t p**ɔɮsən]	p	31
			w	36 (53.7%)
	... V [$_{\text{function}}$ p ...	[xit͡sʰ**u p**ɔɮt͡ʃʰəxsən]	p	6
			w	18 (75.0%)
	... [+nasal] [p ...	[xu**n p**ɔɮəxig]	p	18
			w	1 (5.3%)

pai	... C [function p ...	[pɔttə**g paisən**]	p	58
			w	30 (34.1%)
	... V [function p ...	[tʊt^{h}**ʊ** paiɢa]	p	10
			w	53 (84.1%)
	... [+nasal] [p ...	[ʃi͡tʃəxsə**n paisən**]	p	75
			w	6 (7.4%)
pi	... C [function p ...	[pajərla**t pi**]	p	131
			w	5 (3.6%)
	... V [function p ...	[ətɔ **pi**]	p	12
			w	2 (14.3%)
	... [+nasal] [p ...	[tɛgsə**n pi**]	p	41
			w	0 (0%)

Note that while every function word has its own preference for lenition or not, every function word is still affected by the phonological environment. That is, for all function words, the proportion of lenition after vowels is greater than elsewhere. Likewise, after nasals, the proportion of lenition is always lower. This can be seen clearly in Figure 3, which shows the proportion of lenition in each phonological environment for each function word. The orange bars indicate the proportion of lenition after vowels, and are highest for every function word except *pol* where rates of lenition in the elsewhere context and after vowels are near ceiling. Likewise, the dark blue bars, which indicate proportion lenition after nasals, are lowest for all function words.

Figure 3: Lenition rates in K5: intersecting effects of function word and phonological environment



What we see here in K5 is another, more complex, case of interacting pressures in the grammar. It is not enough to label some function words as exceptions. It would be tempting, for instance, to label *pi* as exceptionally non-leniting. But even *pi* lenites the most (percentage-wise)

after vowels, and the least after nasals. Both individual lexical preferences and grammatical pressures jointly affect the likelihood of lenition.

To model this in MaxEnt, Fuller (2024) proposes lexically-indexed copies (Pater 2010) of the IDENT(sonorant) constraint as follows:

(30) *Lexically-indexed IDENT constraints*

a.	IDENT(son)$_{pol}$	Output correspondents of an input [αsonorant] segment in the lexical item *pol* are also [αsonorant].
b.	IDENT(son)$_{pai}$	Output correspondents of an input [αsonorant] segment in the lexical item *pai* are also [αsonorant].
c.	IDENT(son)$_{pol\text{-}}$	Output correspondents of an input [αsonorant] segment in the lexical item *pol-* are also [αsonorant].
d.	IDENT(son)$_{pi}$	Output correspondents of an input [αsonorant] segment in the lexical item *pi* are also [αsonorant].

These lexically-indexed constraints act just like their general counterpart IDENT(sonorant), but only assign violations within the lexeme to which they are indexed.

Fuller uses one lexically-specific version of IDENT(sonorant) for each function word in her data, each with a different weight. Because they are all versions of IDENT(sonorant), they will all conflict with *p. Their weight relative to *p reflects that function word's overall propensity for lenition.

The tableaux in (31) shows the lexically-specific constraint analysis for K5. The four IDENT(sonorant) constraints assign violations identically, but apply to different inputs, only assigning violations when their indexed lexical item is present in the input. (For convenience we leave out general IDENT(sonorant).) The lexically-indexed IDENT(sonorant) constraints interact with the general phonological constraints from our K4 analysis: *p, *N[+cont], and POSTVOCALICLENITE.

(31) *Tableaux for K5*

word	Context	cand.	obs.	*N[+cont] 2.71	*p 2.23	PostVocLenite 1.63	Ident(son)$_{pol}$ 0.00	Ident(son)$_{pai}$ 2.59	Ident(son)$_{pol-}$ 2.22	Ident(son)$_{pi}$ 5.55	$\mathcal{H}$	p	Obs. p
pol	elsewhere	p	3		1						2.23	.10	.04
		w	65				1				0.0	.90	.96
	after V	p	2		1	1					3.86	.02	.08
		w	22				1				0.0	.98	.92
	after nasal	p	46		1						2.23	.62	.65
		w	25	1			1				2.71	.38	.35
pol-	elsewhere	p	31		1						2.23	.50	.46
		w	36						1		2.22	.50	.54
	after V	p	6		1	1					3.86	.16	.25
		w	18						1		2.22	.84	.75
	after nasal	p	18		1						2.23	.94	.95
		w	1	1					1		4.93	.06	.05
pai	elsewhere	p	58		1						2.23	.59	.66
		w	30					1			2.59	.41	.34
	after V	p	10		1	1					3.86	.22	.16
		w	53					1			2.59	.78	.84
	after nasal	p	75		1						2.23	.96	.93
		w	6	1				1			5.30	.04	.07
pi	elsewhere	p	131		1						2.23	.97	.96
		w	5							1	5.55	.03	.04
	after V	p	12		1	1					3.86	.84	.86
		w	2							1	5.55	.16	.14
	after nasal	p	41		1						2.23	.998	1
		w	0	1						1	8.26	.002	0

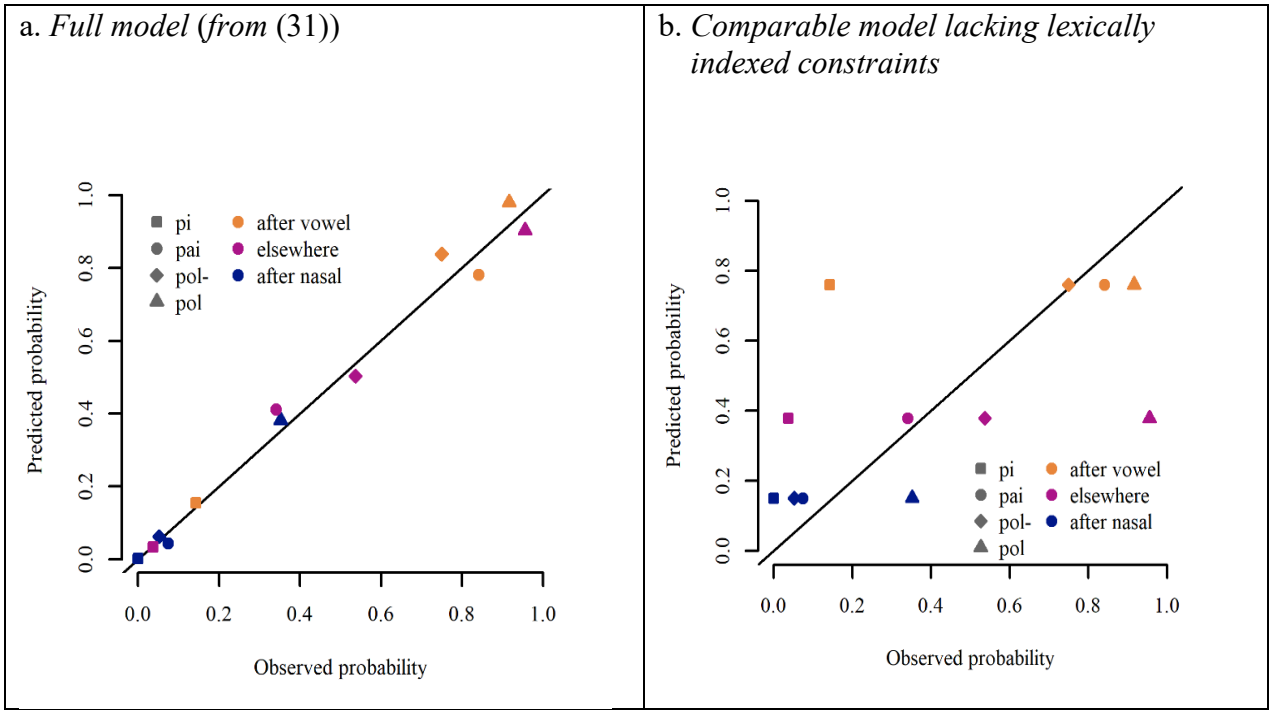
The tableaux in (31) also contain the weights and predicted probabilities gained from fitting with the Solver using the log likelihood, as detailed in §2.6.4 above. The detailed calculations can be viewed in **K5.xlsx**.

The reader may have noticed that the weight of Ident(son)$_{pol}$ came out as zero. This turns out to be no cause for alarm: since every [w] violates precisely one member of the Ident(sonorant) family, it turns out that all the matters is the *relative* weights within this family. For discussion, with the relevant math, see §4.3.

2.7.1 Assessing the K5 analysis with scatterplots

With K5, we have enough data that it makes sense to visually compare the observed and predicted values in a scatterplot. We compare the model probabilities of each candidate with the fraction of observations that that candidate receives (that is, observed cases divided by total cases for that input.) The left side of Figure 4 shows observed values for the lenition candidates on the *x* axis, and the corresponding predicted values on the *y* axis. For purposes of comparison, the graph also contains the line $y = x$. If the model's fit were perfect, every point would be on this line. Indeed, we can see that all the points are very close to the line, indicating a good fit of our model to the data.

Figure 4: Comparing "observed" with predicted values for MaxEnt model of K5



For purposes of comparison, we also give on the right the scatterplot for a model that is similar, but which replaces the four lexically-indexed constraints with one single IDENT(sonorant) constraint (see **K5.xlsx**/K5_noIndexation for weights and computations). For this model, adherence of the points to the $y = x$ line is patently less close; it really did make sense to split up IDENT(sonorant) into morpheme-specific versions. For more on model comparison and the use of scatterplots, see Chapter 5.

2.7.2 K5 vs. real Khalkha

Although our K5 dataset is close to real Khalkha, we briefly point out a few important differences, referring to Fuller (2024). The biggest difference is that in Khalkha, /p/ can actually surface three different ways: as [p], as [w], and as null. This necessitates a richer system of Faithfulness constraints. A second difference is that we have simplified the nature of the phonological perturbers. In real Khalkha, curiously, they seem to form non-natural classes

(Mielke 2005). Lenition is dispreferred after nasals [m], [n], and [ŋ] (as in K5) but also after [ɮ] and [w]. The lenition-preferring environment is not postvocalic, but the set consisting of velars and uvulars: [g], [ɢ], [x], plus the diphthong [ui]. With Fuller, we hope that further research will shed light on why these sounds pattern together.

2.8 Visualizing the K5 pattern with shifted sigmoids

At this point we seek to understand the K5 analysis more deeply, particularly how the constraint weights result in the observed patterns of candidate probability. This foreshadows the theoretical discussion of MaxEnt to be given in Chapter 9. We caution the reader that the discussion of this section gets more technical, and readers will be forgiven if they skip ahead to §2.9. They will, however, miss a theoretical point that might be worth returning to later on.

2.8.1 Examining the effects of the weight of one single constraint

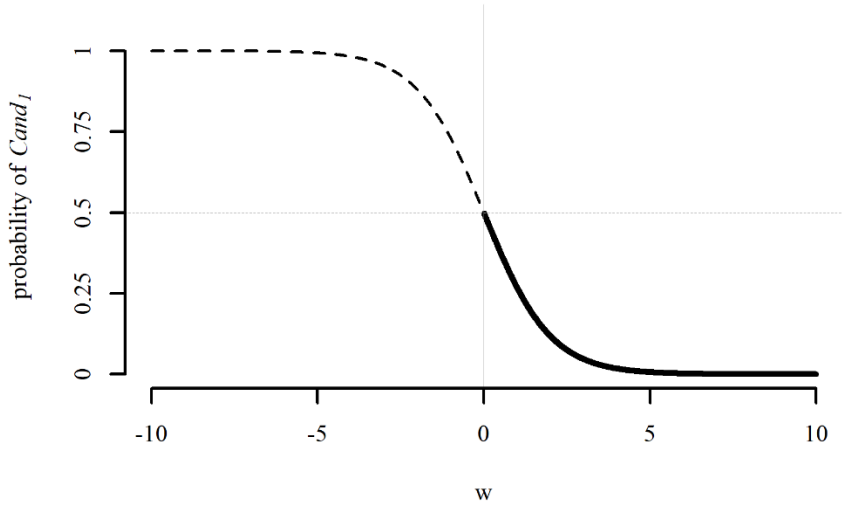
We first examine the relationship between a constraint weight and candidate probability. For this purpose imagine an extremely simple one-constraint grammar in which there are two candidates, one that violates the single constraint ℂ, and one that has no violations at all; hence tableau (32):

(32) *A minimal tableau for illustrating the effect of a constraint weight*

		ℂ **w**
Input	*Cand*$_1$	
	Cand$_2$	1

From this tableau, we can use the MaxEnt formula to compute the probability assigned to the candidates — this time, not a single value, but rather a whole set of values; p as a *function* of **w**. Skipping the actual computation, it emerges that the probability of *Cand*$_1$ is: $P(Cand_1) = e^{-\mathbf{w}} / (1 + e^{-\mathbf{w}})$. This function can be plotted as a graph, with **w** on the horizontal axis. Seeking full mathematical generality, we allow **w** to go negative, even though negative weights are not commonly used in analytical practice. Since these values are nonstandard, we plot them with a dotted line. The graph is given in Figure 5.

Figure 5: How P(Cand1) varies with **w** *in grammar (32*



This curve is often called a **sigmoid**, meaning S-shaped. Looking at Figure 5 in detail, we can observe the following. (a) At **w** = 0, the probability of *Cand1* is 0.5. This makes sense, because this is the equivalent of a constraintless grammar, and such a grammar should assign 50-50 probability to the two candidates. (b) The sigmoid curve is **monotonic**, in that higher values of **w** always result in lower values for *P*(*Cand1*). (c) All differences in **w** will result in differences in *P*(*Cand1*). However, as we increase **w** going outward from zero, we observe diminishing returns: each additional unit of weight has successively less influence. (d) Probability never actually reaches zero, though the value of *P*(*Cand1*) for high values of **w** becomes infinitesimally small.

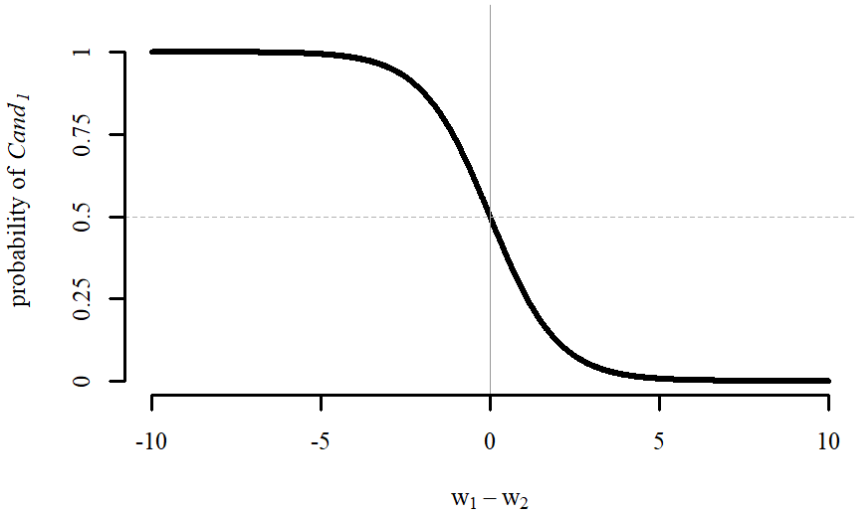
Let us next consider a more realistic example, involving constraint conflict. We suppose that instead of one constraint ℂ we have two constraints $\mathbb{C}_1$ and $\mathbb{C}_2$, which are opposed to each other. A sample tableau for such a case is in (33).

(33) *A minimal tableau for illustrating the effect of conflicting constraints*

		$\mathbb{C}_1$ $\mathbf{w_1}$	$\mathbb{C}_2$ $\mathbf{w_2}$
Input	*Cand1*	1	
	Cand2		1

When we apply the MaxEnt formula to this tableau, it turns out that the only thing that matters is the *difference* between the two weights. We can plot a sigmoid with $\mathbf{w_1} - \mathbf{w_2}$ on the *x* axis (the equation is $P(Cand_1) = 1 / (1 + e^{w_1 - w_2})$), and it will look just like Figure 5, except that the meaning of the *x*-axis has changed. We also refrain from dotting the negative part of the line, since it is perfectly possible for $\mathbf{w_1} - \mathbf{w_2}$ to be negative.

Figure 6: How P(*Cand*$_1$) *varies with* $w_1 - w_2$



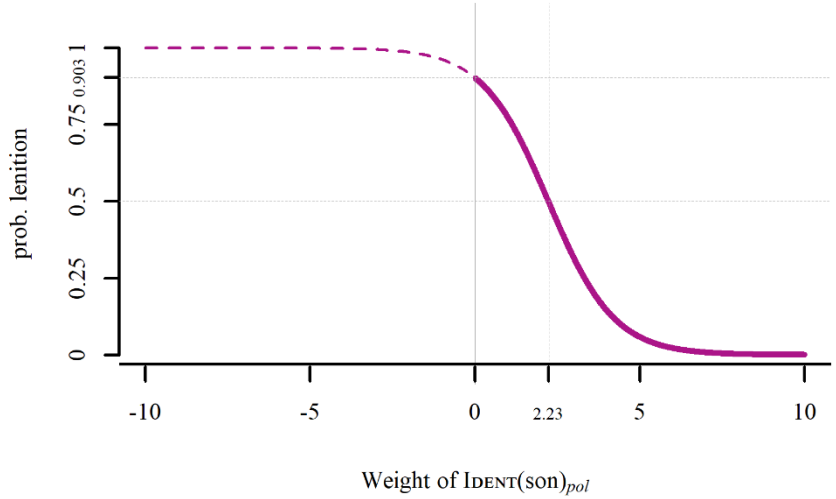
Let us return from these abstract cases and ponder the possibility of sigmoid curves in K5. To this end, consider the first input in the tableaux in (31), where the output probabilities are determined exclusively by the two conflicting constraints *p and IDENT(son)$_{pol}$ — indeed, by the difference in the weights assigned to these constraints. We repeat only the essentials of this tableau in (34), removing the constraints violated by neither candidate. In the full tableau, IDENT(son)$_{pol}$ had a weight of zero, but here we are interested in what would happen if it had different weights; hence the variable *w* in (34).

(34) *Tableau for K5 pol + elsewhere*

word	**context**	**candidate**	*p 2.23	IDENT(son)$_{pol}$ **w**
pol	elsewhere	faithful [p]	1	
		lenited [w]		1

Applying the MaxEnt in the same way as before (yielding P(w) = 1 / ($e^{(w - 2.23)}$ + 1)), we can plot yet another sigmoid, showing how the probability of the lenited [w] candidate depends on the weight of IDENT(son)$_{pol}$

Figure 7: Sigmoid curve: how P(w) varies when weight of IDENT(son)$_{pol}$ is changed



As can be seen, if the weight of IDENT(son)$_{pol}$ is set very high, lenition is predicted to be very unlikely. If IDENT(son)$_{pol}$ is given a weight identical to that of *p, namely 2.23, the rate of lenition is predicted to be 50%; this is indeed the point where the curve crosses the 50% line. As the weight of IDENT(son)$_{pol}$ approaches zero, the probability of lenition goes up; with $w_{\text{IDENT(son)}}$ = 0, it is 90.4%. Lastly, were we to permit negative weights ("it is bad to be faithful"), we could generate a lenition probability as close to 1 as we wish; this region is shown in the dotted portion of the sigmoid.

In more complicated cases we would actually want to plot the difference in *Harmony* between the two candidates, rather than the different in weights; see for instance §3.3 below. In Figure 7, the two approaches come out the same because there is exactly one violation per candidate.

2.8.2 Perturber constraints create shifted sigmoids

Let us now consider what sort of sigmoid patterns we get when we introduce perturber constraints into the analysis. Here, we will plot two sigmoids: one for inputs that violate the perturber constraint, and one for inputs that do not. What emerges when we do this — and this follows from the MaxEnt math — is that the two sigmoids are *identical and spaced apart by the weight of the perturber constraint*. Following Magri (2025), we will call this configuration **shifted sigmoids**.

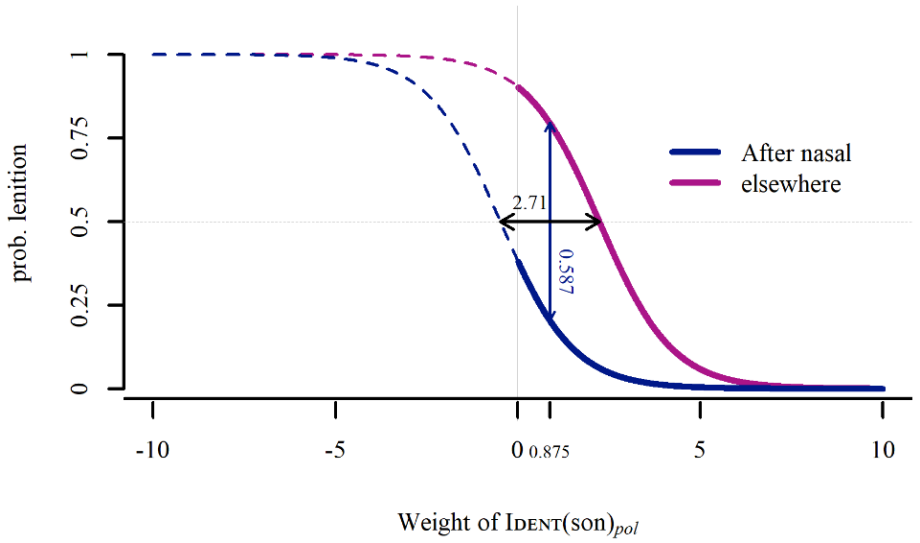
To illustrate shifted sigmoids, we expand our example from (34), this time including *N[+cont], which acts as a perturber constraint, lowering the probability of lenition after nasals.

(35) *Tableaux for K5 pol + elsewhere and pol + nasal*

Word	**context**	**candidate**	*N[+cont] 2.71	*p 2.23	IDENT(son)$_{pol}$ *w*
pol	elsewhere	faithful [p]		1	
		lenited [w]			1
pol	after nasal	faithful [p]		1	
		lenited [w]	1		1

Again, we explore the consequences of varying the weight of IDENT(son)$_{pol}$, but this time plotting a sigmoid for each input; this is shown in Figure 8. Here, the purple sigmoid corresponds to inputs with the "elsewhere" environment, and the blue sigmoid corresponds to inputs in the postnasal environment. As shown by the arrow, the two sigmoids are spaced apart on the horizontal axis by 2.71 units, which is, in fact, the weight of *N[+cont].

*Figure 8: An example of shifted sigmoids: identical sigmoids spaced apart by the weight of *N[+cont]*

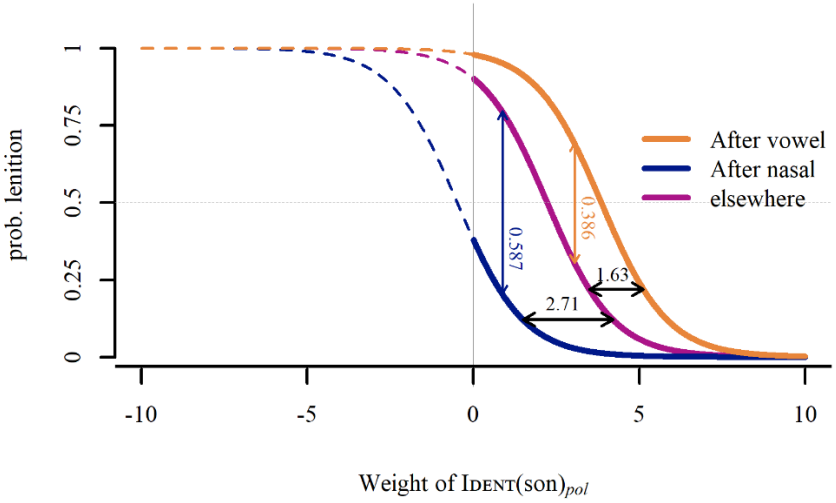


Of course, on the *vertical* axis, the distance between the two sigmoids varies considerably from point to point along the x axis. In the present case, the maximum distance is 0.587, occurring where the weight of IDENT(son)$_{pol}$ is 0.875 (see the blue vertical line in Figure 8). The vertical distance approaches zero for ever higher or lower weights of IDENT(son)$_{pol}$. What this means is that *the effect of a perturber constraint is contextually determined*: it is always maximal at the point halfway between where the two sigmoids cross 50%, and is ever smaller going off in either direction.

This contextual variation in the effect of a perturber is a fundamental prediction of MaxEnt, and is repeatedly referred to in analytical work. To reiterate: perturbers are predicted to be strong in their effects when they fall within the middle of the family of related sigmoid curves, but weak on either periphery.

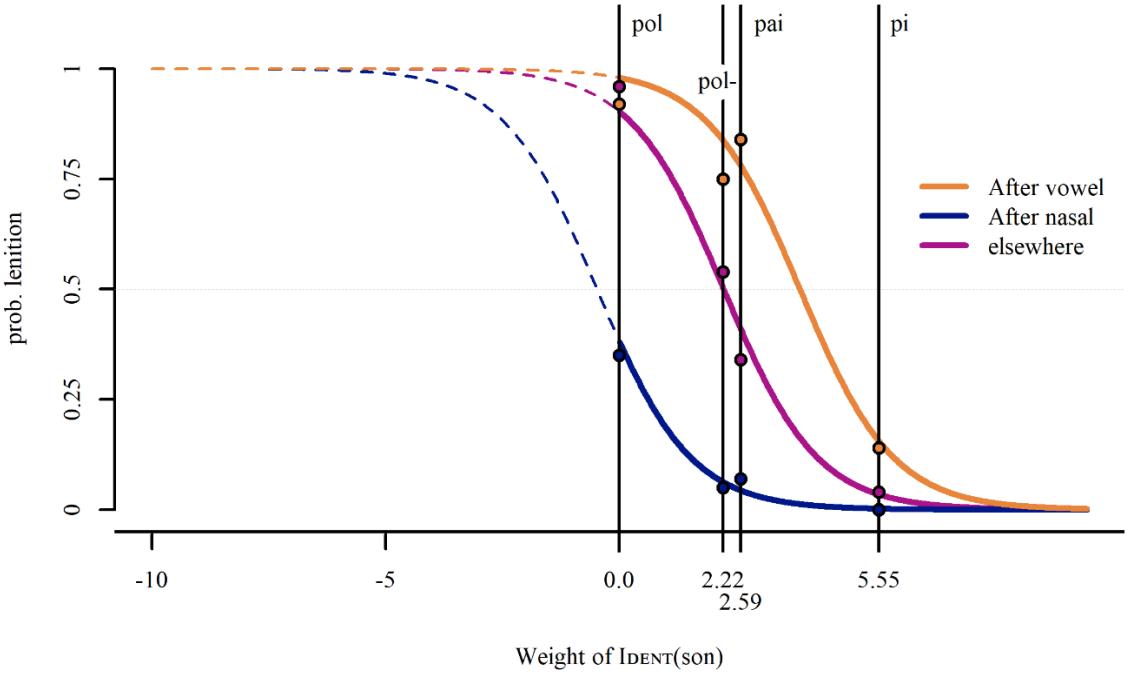
We consider next the case in which there is more than one perturber constraint. Here, there will be three sigmoids, one for the baseline condition and a shifted sigmoid for each perturber. For instance, in Figure 9 we add a second perturber from K5, the post-vocalic environment, violated by the constraint POSTVOCALICLENITE. This environment favors lenition postvocalically, so its sigmoid is displaced rightward.

Figure 9: Sigmoid curves for K5 with two perturbers



So far, we have dealt with a hypothetical case in which the weight of IDENT(son)$_{pol}$ is varied freely for illustrative purposes. We next attempt to better understand the actual K5 analysis using the same technique. The sigmoid curves will be identical to those of Figure 9, but we will mark with a vertical line the actual fitted weight for IDENT(son)$_{pol}$, 5.55. We will also include vertical lines marking the weights of the other three morpheme-indexed IDENT(sonorant) constraints, namely IDENT(son)$_{pi}$, IDENT(son)$_{pai}$, and IDENT(son)$_{pol\text{-}}$. The points at which these vertical lines intersect the sigmoid curves represent the predictions of our MaxEnt grammar. The figure also includes dots that show the (imagined) empirical values for the corresponding points in the K5 data. If our fitted model is a good one, then these dots should fall fairly close to the intersection points of the corresponding vertical lines and sigmoids.

Figure 10: Assessing the K5 model using shifted sigmoids



We see that the data points match fairly well with their predicted values. The largest errors are evidently those for *pol-* and *pai* in the postvocalic context, which deviate from the sigmoid by are 0.09 and 0.06, respectively. The data points for *pi* are far enough to the right to illustrate the fundamental principle just given: the effect of a perturber constraint is weaker on the periphery of the local weight range.

The use of annotated shifted-sigmoid diagrams like Figure 10 provides a useful complement to the use of scatterplots like Figure 4 above in evaluating an analysis, for it indicates how the predictions emerge from the constraints and their weights.

Beyond its role in assessing the accuracy of the analysis, the graph in Figure 10 makes a far more general point, namely the very fact that the data can be fitted with a set of shifted sigmoids. This property of MaxEnt is discussed in more detail in §9.3 below.

2.9 Very high weights and the approximation of zero

We return for a moment to K4, a language in which some outcomes are obligatory (20). For instance, lenition in K4 is completely blocked in the post-nasal environment. In our MaxEnt analysis, this outcome is attributed to a powerful constraint, *N[+cont], whose best-fit weight is 12.35. For similar reasons, the weight of IDENT(sonorant)$_{\text{content}}$ is also very high, 13.37.

These weights must be interpreted with care. First, we note that in fact there is *no* weight assignment that will fit the empirical value of zero exactly. To see this, consider the stage of the MaxEnt calculations given in (16b), where eH is calculated using eH = $e^{-\mathcal{H}}$. No matter how high is the value of $\mathcal{H}$ (resulting from high constraint weights), $e^{-\mathcal{H}}$ will always be greater than zero; this is in the very nature of the exponential function. The higher the weight we assign to

*N[+cont] and IDENT(sonorant)$_{content}$, the closer to zero will be the probability assigned to zero-frequency candidates like *[ʃit-t͡ʃʰəx-sə**n** **w**ai-sən] and *[ɔrʊɮ-əx **w**ɔɮəm-t͡ʃtʰɔi]. But the value will never actually reach zero; we can only approximate it. The ideal weight for *N[+cont] and IDENT(sonorant)$_{content}$ is, in this sense, infinity.[13]

In the real world, weights are computed by an algorithm that labors to find ever better approximations to the observed distribution. The algorithm will always include one or more parameters that tell it when to stop trying.[14] The values obtained for constraints like *N[+cont] and IDENT(sonorant)$_{content}$ are the values that have been computed up to this stopping point.

There are two lessons here. One is a simple practical one: for some highly-weighted constraints, it is sensible for the analyst not to obsess on the exact weights, even where they differ substantially. All such values are more or less arbitrary stopping points for the search algorithm, the points where its settings lead it to accept the current (tiny) predicted value of zero-frequency candidates as close enough. Perhaps surprisingly, in MaxEnt, the difference in weights between (say) 20 and 50 can be very unimportant, whereas the difference between 1 and 2 might be quite important. We discuss this issue further in §4.3.

A second lesson goes a bit deeper. Anyone who uses MaxEnt must be prepared to accept an infinitesimally small number as the practical equivalent of zero. For instance, our K4 grammar assigns the probability 3.2×10^{-7} to the bad candidate *[ɔrʊɮ-əx **w**ɔɮəm-t͡ʃtʰɔi] and 8.9×10^{-7} to the bad candidate *[ʃit-t͡ʃʰəx-sə**n** **w**ai-sən]; these were shown earlier in (18) for convenience as "≈ 0". The difference between these numbers and zero is smaller than the measurement error for almost all data sets.

Lastly, we point out the class of constraints for which the "infinite weight" syndrome arises: they are the constraints that *prefer a winner*, but never *prefer a loser*, in the terminology of Prince (2003b). In classical Optimality Theory, such constraints are the ones that can always safely be ranked at the top of the grammar. In MaxEnt, they have idealized infinite weights.

2.10 Some Excel tips

Since this is a textbook, we offer a brief section on how one can use Excel more quickly and reliably to carry out MaxEnt calculations.

2.10.1 Avoiding weight crashes

The fact that in some cases a constraint's ideal weight is infinite can raise practical problems in Excel. The Solver's weight-fitting algorithm will halt after a certain amount of computation, often landing on a weight that is high enough to give a good fit to the data. However, in some cases the Solver will fit a constraint weight which predicts probabilities so low that Excel cannot represent them. In this case, the algorithm will return an error message, and cells which cannot

[13] Readers with calculus will recognize here a *limit*: no matter no low a value δ we pick for P(*[ʃit-t͡ʃʰəx-sən **w**ai-sən]), there is always some value ε such that if $w_{*N[+cont]} > \varepsilon$, then P(*[ʃit-t͡ʃʰəx-sən **w**ai-sən]) < δ.

[14] To alter the stopping parameter in the Excel Solver, click Options, GRG Nonlinear, and change the value in the box labeled Convergence.

be represented will display #NUM! instead of actual values. A very simple spreadsheet that creates such a crash is given in (36).

(36) *A MaxEnt spreadsheet that creates an error in Excel*

		Freq.	C1						
			w = 708.3	H	eH	Z	p	ln p	LL
Input1	$Cand_1$	1	0	0	1	~1	~1	~0	~0
	$Cand_2$	0	1	708.3			2.45×10^{-308}	−708.3	

The hand-entered weight in (36), 708.3, is about as high as it can go. If we increase it to 708.4, the tiny probability for $Cand_2$ gets rounded down by Excel to zero. This yields a #NUM! error message in the *ln p* column, since zero has no logarithm.

We discuss a principled way for dealing with this problem in §4.5 below, but here give an ad hoc solution that does not require much additional calculation. Specifically, it is possible in the Solver simply to impose a hard upper limit on weights. A maximum of 50 is often sufficient to produce an accurate analysis while avoiding crashes. To implement such a limit, do the following:

(37) *An ad hoc procedure for limiting weights*

- In the Solver window, click "Add" to the right of the box labeled "Subject to the Constraints".
- In the left hand field, select the cells which contain the weights.
- Select "<=" for the middle drop-down menu, and type "50" in the right hand field.

The Solver will then give the best-fit answer available, subject to this weight limit: weights that otherwise would be unmanageably large will be limited to 50. For specific problems you may find that an upper limit greater than 50 is necessary, or a limit lower than 50 is sufficient.

2.10.2 Another method for calculating Z

A common source of error in a MaxEnt spreadsheet, often hard to detect, arises in the calculation of the Z values. These must be obtained by summing the eH values (cf. (16b)) for *exactly* the set of candidates for a particular input. Where each input has the same number of candidates, it is not hard to avoid error, but if the inputs can have different numbers of candidates it is easy for a slip of the mouse to create an error. Here is a safer, if more verbose method.

First, you must include (perhaps just adding it in) a column in which the input is specified for every row. For instance, for the first two inputs of (23) we need a column like column B in (38).

(38) *A method for greater reliability in calculating Z*

Excel column names

row #	**B**	C	D	...	H	I	J	K	**L**	M
1	**Example**			...	Id(son)	Postvoc Lenite	Harmony	eH	**Z**	p
2		output	count		4.00	2.00				
3	**g poɫən**	P	290				= SUMPRODUCT (E$2:I$2, E3:I3)	=EXP (-J3)	**=SUMIF(--EXACT(B3,B:B) *K:K)**	=K3/L3
4	**g poɫən**	W	60		1		= SUMPRODUCT (E$2:I$2, E4:I4)	=EXP (-J4)	**=SUMIF(--EXACT(B4,B:B) *K:K)**	=K4/L3
5	**ʊ pai-**	P	10			1	= SUMPRODUCT (E$2:I$2, E5:I5)	=EXP (-J5)	**=SUMIF(--EXACT(B5,B:B) *K:K)**	=K5/L5
6	**ʊ pai-**	W	50		1		= SUMPRODUCT (E$2:I$2, E6:I6)	=EXP (-J6)	**=SUMIF(--EXACT(B6,B:B) *K:K)**	=K6/L5
...	**...**	...	...	...	...	...	...	...	**...**	...

Once you have a column in place that repeats the inputs, then the reliable calculation of Z can be based on the formula `=SUMIF(--EXACT(A4,A:A)*J:J)`,[15] which can be entered once into the first cell of the Z column (in (38), cell A4), then copied down the entire column. This formula takes the subset of J (the eH column) where A is equal to the value in A4 (or whatever row you are on), even if it is not contiguous, and sums them. We recommend using this method when you have different numbers of candidates for different inputs, making it dangerous to compute all the Z calculations with separate copy-paste moves. It is also helpful if you plan on experimenting with changing the numbers of candidates for each input.

2.10.3 Troubleshooting

Often a first effort to solve a phonological problem yields blatantly poor results, such as a bad fit to the data for some or all inputs or nonsensical constraint weights (such as zero for a patently necessary constraint). We offer some hints for troubleshooting.

The problem may be with the linguistic analysis; for techniques for fixing shortcomings in this area see §4.3 and §5.2. However, the problem may lie with an incorrect Excel implementation. The following checks may be helpful for such troubleshooting, or indeed, you might find it useful to do them before you ever click the Solve button.

(a) Check that all the constraint violations in your spreadsheet have been correctly assigned.

(b) Visual spot-check: Check a few Harmony values at the beginning, middle and end of your spreadsheet by hand, if feasible. The most common error here is that the Harmony column has left out a constraint; D3:G3, for example, when it should be D3:H3. If you set all the weights by hand to the value 1, it becomes easier to do this check.

[15] The simpler expression `=SUMIF(A4,A:A,J:J)` would also work, but we recommend against it because it groups together inputs that differ only in capitalization.

(c) If you did *not* use the method given in §2.10.2, then hand-check your Z values to make sure that they do indeed sum over all the candidates for the same input.

(d) Check each input's probability distribution, to make sure it sums to 1. For small spreadsheets, you can highlight the candidates for the input with your mouse or tracker and check the sum at the bottom of the Excel screen. To be more thorough, you can use the following formula in cell M4 (in the case of (38)), and copy it down the column: `=IF(SUMIF(A:A,A3,L:L)=1, "", "!")`. Then, search for "!" in that column, which will indicate an output probability distribution that did not sum to 1. If you find an error it could mean that, perhaps, your Z scores were not calculated correctly, or perhaps that your exponentiation is off – maybe you didn't include the negative sign.

(e) Finally, go through each of your constraints and manually change the weight. You should see the probabilities in column L change for all inputs in which some candidate violates that constraint. (You can also check if the log likelihood cell changes.)

(f) It is especially helpful to carry out the above spot checks when you add a new input, candidate, or constraint to your spreadsheet; this is a frequent source of error.

(g) To obtain consistent results when revising your analysis, it is recommended that you reset the weights to zero before launching the Solver. The benefit of this is that you are more likely to obtain a minimum-weight solution, in which useless constraints will stand out because they are assigned a weight of zero (§5.3.3). For a more principled way of finding the minimum-weight solution, see §4.5.

(h) In larger simulations, the termination criterion employed as default by the Solver may not permit the most accurate feasible solution to be found. In our experience, clicking the Solver button a second time often offers modest improvement in log likelihood. It seems unnecessary ever to click a third time.

2.11 More MaxEnt software

We mention other forms of software that can implement a MaxEnt grammar and find the best-fit weights. Many of these perform other tasks as well.

The **MaxEnt Grammar Tool**, is available at brucehayes.org/MaxentGrammarTool/. A number of papers in the literature have used this software to implement a MaxEnt grammar. While it works perfectly well, we currently favor the Excel/Solver combination over it, since Excel renders more transparent the work that has been done.

`Maxent.ot` is a software package usable in the well-known statistical analysis system R. For documentation, see Mayer et al. (2024). It requires the user to have some fluency in writing R scripts, and thus perhaps is less suitable for beginning work. A key advantage of `maxent.ot` is that it lets you embed MaxEnt optimization into larger routines, which is useful if you need to optimize many times; indeed we use it here for this purpose in §5.3.6 and §7.5. We suspect Excel

can handle slightly larger problems and is a bit faster (on a single run). We also find it easier to invent new model types (for instance: one component feeding another) using Excel.

These are not the only software packages available for implementing MaxEnt, and new ones are published with some regularity. We briefly present a few here that we have not personally explored in detail, but which look promising. The web addresses for these software packages are given in the bibliography under their citations. **Fonology** (Garcia 2026), integrates lexical information for English, Spanish, French, Italian, and Portuguese, along with several helpful tools for dealing with transcriptions, features, and syllables. Pater and Prickett's (2022) python script (Prickett 2025) includes a few different optimization algorithms, and can learn hidden structure (see §7.7.1). Staubs' (2014) **hgr** learns MaxEnt with hidden structure in R. The well-known phonetics program **Praat** (Boersma et al. 2026) also includes many algorithms for probabilistic constraint-based linguistics, including MaxEnt and some of the related theories discussed in §9.5. **Phonology-Lab** (Dai 2026) is web-based, and can be used immediately without any software download or programming expertise.

Finally, we mention our own software. Hayes's **OTSoft** (Hayes et al. 2026) provides algorithms for several different stochastic frameworks and is used below for various purposes. Moore-Cantwell's **Gradual Lexicon and Phonology Learner (GLaPL)** (Moore-Cantwell 2026) learns MaxEnt grammars gradually, with an algorithm similar to that covered in §7.7.2, and incorporates lexical specificity in the form of lexically-indexed constraints and lexical listing.

2.12 Where to examine published analyses using MaxEnt

We have not tried to form a complete bibliography of research that has made use of the MaxEnt framework, but various sections of this book provide listings that might be useful for purpose of finding model analyses to emulate. See in particular the discussion of frequency matching in §3.5, the discussion of data-gathering in §5.5, and the studies cited in the broad review (pan-linguistics) of Chapter 8.

Chapter 3: Using MaxEnt to study lexical frequency-matching

An important empirical advance in phonology occurred in the early years of this century when investigators began to study inconsistent patterns in the lexicon: for instance, cases where certain forms undergo a phonological process and others don't. The two key findings were that exceptionality is *statistically patterned*, and that such patterning is internalized by native speakers, as shown by their ability to extend the pattern to novel words. We give a review of the studies that collectively made this research finding in §3.5, after we have covered a particular example. MaxEnt offers grammars that can describe such patterns of lexical exceptionality, as well as the ability of the native speakers to internalize them.

3.1 Modeling example: Tagalog nasal substitution

Tagalog is an Austronesian language spoken widely in the Philippines. For background on this language see Schachter and Otanes (1972). The focus here is on a phonological process of Tagalog known as **nasal substitution**; our treatment here is based almost entirely on the research (including MaxEnt analysis) of Kie Zuraw (Zuraw 2000, 2010; Zuraw and Hayes 2017).

3.1.1 The basic data pattern

Nasal substitution is a process triggered by the placement before the stem of a prefix that ends in the phoneme /ŋ/. In such cases, the underlying /ŋ/ of the prefix *merges* with the following stem-initial obstruent, producing an output sound that is a single nasal, taking the place of articulation of the vanished obstruent. For example, if the stem-initial segment is /p/, then the output for /ŋ+p/ will be [m], taking the place of articulation of the /p/ but the nasality of the /ŋ/; thus /pa**ŋ-p**igha'tiʔ/ → /pa**m**igha'tiʔ/ 'being in grief'. More generally, the following changes occur.

(39) *The changes made in Tagalog Nasal Substution*

a. *Bilabials*: /ŋp, ŋb/ → [m]
b. *Alveolars:* /nt, nd, ns/ → [n]
c. *Velars*: /ŋk, ŋg/ → [ŋ]
d. *Glottals*: /ŋʔ/ → [ŋ]

In what follows we will focus on the six stops [p, t, k, b, d, g]; omitting the treatment of [s] and [ʔ]. As a result the counts and analysis will differ in minor ways from the sources consulted.

Tagalog nasal substitution is a classic case of exceptionful phonology: although forms that have undergone nasal substitution are abundant, there are also many forms that do *not* display nasal substitution. For these forms, the output differs little from the input; the only thing that happens is a garden-variety place assimilation, as follows: /ŋp/ → [mp], /ŋb/ → [mb]; /ŋt/ → [nt], /ŋd/ → [nd]; and /ŋk, ŋg/ remain unchanged.

Zuraw (2000:19) emphasizes lexical exceptionality by listing pairs in which the same underlying sequence yields different outcomes. In the data below, "RED" stands for an abstract

morpheme that triggers CV reduplication. Normally a particular word will have just one acceptable pronunciation (either substituted or non-substituted), though a few words permit variation between the two options.

(40) *Similar forms that undergo, or do not undergo, nasal substitution*

a. [**p**igha'ti?]	'grief'
/pa**ŋ-p**igha'ti?/ → [pa**m**igha'ti?]	'being in grief'
but	
[**p**o'ʔok]	'district'
/pa**ŋ-p**o'ʔok/ → [pa**mp**o'ʔok]	'local'
b. pag-'**t**uloj]	'staying as guest'
/ka-pa**ŋ-t**uloj-an/ → [kapa**n**ulu:jan]	'fellow lodger'
but	
[**t**a'boj]	'driving forward'
/pa**ŋ-t**a'boy/ → [pa**nt**a'boj]	'to goad'
c. [**k**am'kam]	'usurpation'
/ma-pa**ŋ-k**am'kam/ → [mapa**ŋ**am'kam]	'rapacious'
but	
[kalis'kis]	'scales'
/pa**ŋ-k**alis'kis/ → [pa**ŋk**alis'kis]	'tool for removing scales'
d. [mag-**b**i'gaj]	'to give'
/ma**ŋ-b**i'gaj/ → [ma**m**i'gaj]	'to distribute'
but	
[**b**ig'kas]	'pronouncing'
/ma**ŋ**-RED-**b**ig'kas] → [ma**mb**ibig'kas]	'reciter'
e. [**d**a'laŋin]	'prayer'
/ʔi-pa**ŋ-d**alaŋin/ → [ʔipa**n**a'laŋin]	'to pray'
but	
[**d**i'nig]	'audible'
/paŋ-**d**i'nig/ → [pa**nd**i'nig]	'sense of hearing'
f. [**g**in'daj]	'unsteadiness on feet'
/pa**ŋ**-RED-**g**in'daj/ → [pa**ŋ**iŋin'daj] [16]	'unsteadiness on feet'
but	
['**g**awaj]	'witchcraft'
/ma**ŋ**-RED-'**g**awaj/ → [ma**ŋg**a'gawaj]	'witch'

[16] It can be seen that the consonant that results from nasal substitution, [ŋ], shows up *both* in the reduplicative prefix and stem-initially; thus [ŋiŋi]. This intriguing interaction of nasal substitution with reduplication has been known since Bloomfield (1917); for one analysis see McCarthy and Prince (1995, §4.3).

Zuraw discovered these facts partly by working with Tagalog consultants, partly by consulting a dictionary, and partly by gathering her own corpus. For the latter purpose, she used what is now often called a **web-scraper**; a computer program that automatically searched the Internet for Tagalog-language web pages, and harvested numerous relevant words from them. Here, we will be working with a corpus of 866 words from the dictionary study; this is a subset of the corpus used by Zuraw and Hayes (2017).[17] Our selection from the corpus and our MaxEnt calculations may be found in the supplementary materials in **TagalogFull.xlsx**.

We begin with some basic data illustrating patterned exceptionality. In (41), we give the counts for nasally-substituted and non-substituted forms, based on the initial consonant of the stem.

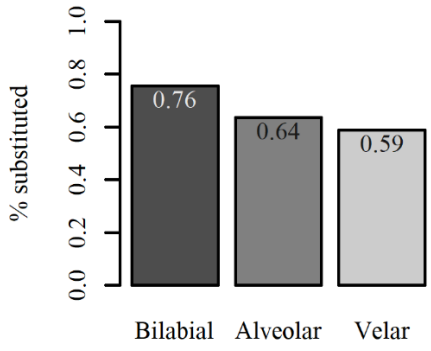
(41) *Tagalog nasal substitution counts by stem-initial consonant*

Stem-initial consonant	*Corpus counts per output type*	
	Not substituted	*Substituted*
p	11	168
t	23.5	129.5
k	18	142
b	86.5	134.5
d	58	12
g	82	1

In the table, fractional values reflect Zuraw's practice of counting forms that have word-specific free variation as 0.5 for each variant.

Within these data, we find two broad, intersecting generalizations. First, the rate of nasal substitution depends on *place of articulation* in a straightforward way: the "further back in the mouth" the stem-initial consonant is articulated, the less nasal substitution will occur. To show this, we replot the data of (41), suitably grouped and re-expressed as fractions, in Figure 11:

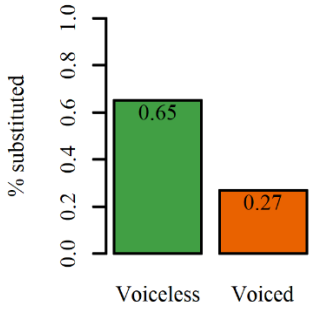
Figure 11: Tagalog nasal substitution: Further-back places have less substitution



[17] The corpus used in Zuraw and Hayes (2017) can be found online at https://muse.jhu.edu/article/670374.

A second, stronger generalization is that stems beginning with a voiceless consonant (here [p, t, k]) have a higher substitution rate than those that begin with a voiced consonant (here [b, d, g]), as shown in Figure 12.

Figure 12: Tagalog nasal substitution: voiceless consonants have more substitution



As we will see, the two factors interact in an additive way, such that [p] (voiceless and bilabial) has the highest substitution rate and [g] (voiced and velar) the lowest. The detailed interaction is shown in Figure 13, which shows that the place effect is quite a bit stronger for voiced consonants:

Figure 13: Tagalog nasal substitution: rates for six consonants plotted separately

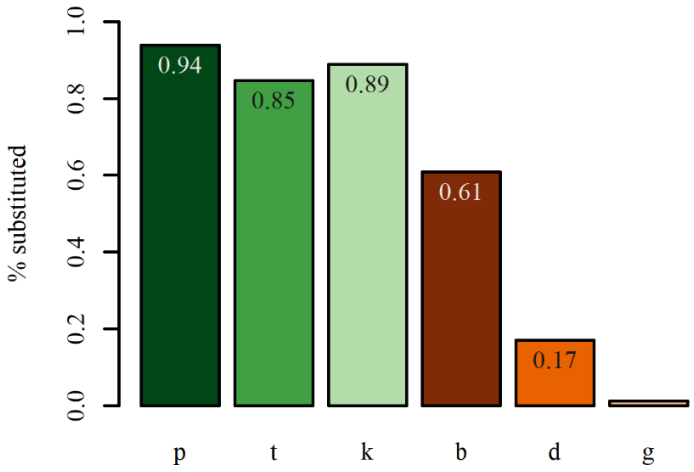


Figure 13 also shows that in this dataset /k/ unexpectly undergoes substitution a bit more often than the fronter consonant /t/. Later on, it will emerge that this is a statistical blip, understandable when we analyze more deeply; see §3.2.1.

3.2 An initial MaxEnt analysis

The premise of our analysis is that patterns like the one shown are part of the grammatical knowledge internalized by Tagalog speakers, and that it is appropriate to use MaxEnt to set up a grammar that predicts these patterns. We turn later (§3.4) to the question of whether these patterns are indeed internalized by native speakers.

To start, we will need a constraint that triggers nasal substitution. We follow Zuraw and Hayes (2017:502) in offering a straightforwardly descriptive constraint:

(42) NASSUB: Assess one violation for any *nasal* + *obstruent* sequence, where + is a morpheme boundary within a word.

The tendency of voiceless consonants to take substitution more often than voiced ones can plausibly be attributed to a widely documented constraint (Hayes and Stivers, 2000; Pater, 2004) that disfavors voicelessness after a nasal:

(43) *NC̥: Assess one violation for every sequence of a nasal and a voiceless obstruent.

In other languages, the same constraint is responsible for post-nasal voicing, pre-voiceless denasalization, or deletion of nasals (Pater 2004). For Tagalog, it will act as a perturber constraint, boosting the nasal substitution rate for stems with /p, t, k/, since in these cases retaining the prefix-final nasal would create a violation, as in [pa**mp**oˈʔok] in (40a).

As for place of articulation effects, we appeal first to a ban on syllables (or stems) that begin with [ŋ]; see for instance English, where there are no words like *[ŋɑp]. The WALS database (Anderson, 2013) suggests that this is a common pattern. Under this approach, forms like (40c) [ma-pa-**ŋ**amˈkam] are disfavored because the surface form has a stem-initial [ŋ]. The difference between alveolars and labials is harder to make sense of; for a possible explanation see Zuraw (2010:438-9). For simplicity we posit a constraint against initial [n]. The two constraints are given in (44) below.[18]

(44) *Constraints governing initial nasals*

a. *[$_{stem}$ ŋ: Penalize candidates where the stem begins with [ŋ].
b. *[$_{stem}$ n: Penalize candidates where the stem begins with [n].

As for a candidate set, in the interest of brevity we will provide only the two candidates that actually compete in real Tagalog, namely a nasally-substituted one (like /paŋ-poˈʔok/ → [pa**m**oˈʔok]) and a merely-assimilated one (like [pa**mp**oˈʔok]). A fuller analysis would include fully-Faithful candidates like [pa**ŋp**oˈʔok] (ruled out by a highly-weighted constraint banning non-homorganic nasal clusters), the deletion candidates [pa**ŋ**oˈʔok] and [pa**p**oˈʔok] (ruled out by highly-weighted MAX), and others.

The candidates, constraints, and violations needed so far are given in **TagalogSimple.xlsx**, which looks like this:

[18] The reader may have noticed that we have not provided any Faithfulness constraint violated by the substituted candidate. We will address this lacuna later on (§3.2.1, §4.3.2).

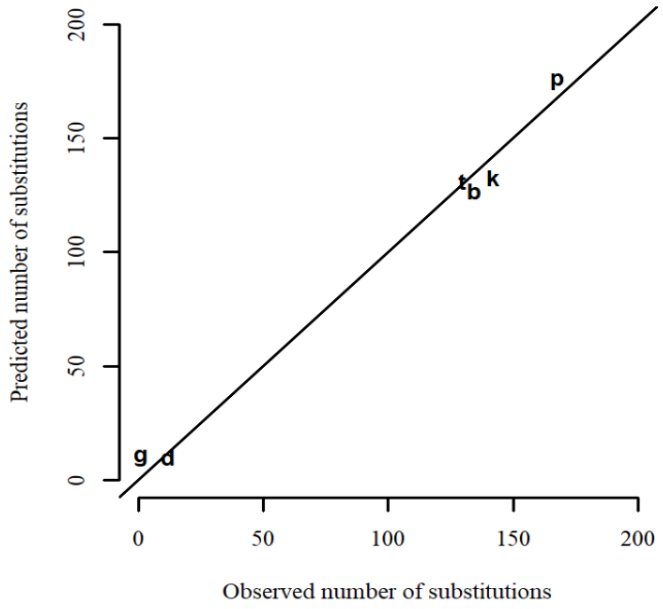
(45) *A MaxEnt tableau spreadsheet for Tagalog nasal substitution (simple version)*

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Consonant	Candidate		NAS SUB	*NC̥	*[n	*[ŋ						
2			**freq**	**0.30**	**3.57**	**2.07**	**2.28**	H	eH	Z	p	ln p	LL
3	p	mam[pX	11	1	1			3.87	0.02	1.02	0.02	-3.89	**-363.71**
4		ma[mX	168					0.00	1.00	1.02	0.98	-0.02	
5	t	man[tX	23.5	1	1			3.87	0.02	0.15	0.14	-1.95	
6		ma[nX	129.5			1		2.07	0.13	0.15	0.86	-0.15	
7	k	maŋ[kX	18	1	1			3.87	0.02	0.12	0.17	-1.78	
8		ma[ŋX	142				1	2.28	0.10	0.12	0.83	-0.19	
9	b	mam[bX	86.5	1				0.30	0.74	1.74	0.42	-0.86	
10		ma[mX	134.5					0.00	1.00	1.74	0.58	-0.55	
11	d	man[dX	58	1				0.30	0.74	0.86	0.85	-0.16	
12		ma[nX	12			1		2.07	0.13	0.86	0.15	-1.92	
13	g	maŋ[gX	82	1				0.30	0.74	0.84	0.88	-0.13	
14		ma[ŋX	1				1	2.28	0.10	0.84	0.12	-2.11	

The weights of the analysis were found using the same method employed in Chapter 2: we used the Excel Solver to maximize log likelihood (cell M3) by adjusting the constraint weights (cells D2-G2). These weights seem intuitively sensible: the fact that *[n has a positive weight reflects the lesser tendency of alveolars to undergo nasal substitution than bilabials; and the still greater weight of *[ŋ captures the fact that velars undergo nasal substitution least often (see Figure 11). The powerful effect of consonant voicing (Figure 12) is reflected in the high weight (3.57) of *NC̥. We also see that the markedness constraint NASSUB plays only a modest role; evidently a good deal of its work is done by *NC̥, whose violations overlap with it.

The fit of the analysis to the data comes out fairly well, as we can see with following scatterplot.

Figure 14: Scatterplot: Observed vs. Predicted counts for the simple Tagalog analysis



In Figure 14, we have chosen to display not probabilities but rather predicted counts, obtained by multiplying the probabilities for each candidate by the count of forms derived from that input; for further discussion of this method see §5.2.2.

3.2.1 Distinguishing different prefixes

Tagalog resembles Khalkha (§2.7) in that the various affixes that can trigger phonology differ in their propensity to do so. Here, there will be seven affixes under scrutiny; recall that RED stands for an abstract reduplication trigger.

(46) *The Tagalog prefixes examined*

a.	/maŋ-$_{other}$/	non-adversative-verbs
b.	/paŋ-RED-/	mainly gerunds
c.	/maŋ-RED-/	professional or habitual nouns
d.	/maŋ-$_{adv}$/	adversative verbs
e.	/paŋ-$_{noun}$/	various nominalizations
f.	/paŋ-$_{res}$/	reservational adjectives
g.	/mag-kaŋ-RED-/	verbs of accidental result

To illustrate the differences, we return to the same dataset analyzed in the previous section, only this time breaking down the data further according to what prefix is present.

(47) *Data for the full Tagalog simulation*

Consonant	*Prefix*	*sub/unsub*	*Consonant*	*Prefix*	*sub/unsub*
p	maŋ-$_{other}$	65/0	b	maŋ-$_{other}$	66/6
	paŋ-RED-	39/0		paŋ-$_{other}$	29.5/5.5
	maŋ-$_{adv}$	11/0		maŋ-$_{adv}$	6/6
	maŋ-RED-	18/0		maŋ-RED-	16.5/21.5
	paŋ-$_{noun}$	30.5/6.5		paŋ-$_{noun}$	15/30
	paŋ-$_{res}$	4.5/4.5		paŋ-$_{res}$	1.5/17.5
	mag-kaŋ-RED-	0/20		mag-kaŋ-RED-	0/20
t	maŋ-$_{other}$	39/0	d	maŋ-$_{other}$	7/3
	paŋ-RED-	22/0		paŋ-RED-	3/7
	maŋ-$_{adv}$	12/0		maŋ-$_{adv}$	0/12
	maŋ-RED-	19/0		maŋ-RED-	1/12
	paŋ-$_{noun}$	34/17		paŋ-$_{noun}$	1/17
	paŋ-$_{res}$	3.5/6.5		paŋ-$_{res}$	0/7
	mag-kaŋ-RED-	0/20		mag-kaŋ-RED-	0/20
k	maŋ-$_{other}$	74.5/0.5	g	maŋ-$_{other}$	0/13
	paŋ-RED-	25/1		paŋ-RED-	1/17
	maŋ-$_{adv}$	11/0		maŋ-$_{adv}$	0/9
	maŋ-RED-	20/2		maŋ-RED-	0/12
	paŋ-$_{noun}$	9.5/9.5		paŋ-$_{noun}$	0/27
	paŋ-$_{res}$	2/5		paŋ-$_{res}$	0/4
	mag-kaŋ-RED-	0/20		mag-kaŋ-RED-	0/20

It can be seen that the prefix complex /mag-kaŋ-RED-/ does not induce nasal substitution at all (Zuraw 2000:119). While the published studies do not include quantitative data for this prefix complex, we have included it in the interest of illustrating how a *non*-variation pattern could be derived in MaxEnt. To fit it in with the rest of the analysis, we adopt synthetic counts that seem reasonable; namely 20 to 0 non-substitution for each initial consonant.

Adopting counts based on both stem-initial consonant and prefix, we now have a richer data set embodying independent, intersecting principles: the inputs now cross-classify the six consonants /p, t, k, b, d, g/ with the seven prefixes of (46).

In our analysis we will deal with output frequency differences among affixes by setting up affix-specific Faithfulness constraints. These come from the UNIFORMITY family proposed by McCarthy and Prince (1995): such constraints militate against surface segments that are the surface reflexes of two distinct input segments, as in Tagalog /ŋp/ → [m]. The fully-explicit notation provided by McCarthy and Prince employs indices to make the changes clear; for instance we can write /p$_1$a$_2$**ŋ$_3$**-**p$_4$**o$_5$ˈʔ$_6$o$_7$k$_8$/ → [p$_1$a$_2$**m$_{34}$**o$_5$ˈʔ$_6$o$_7$k$_8$], where the double subcripting on the surface segment [m] indicates its status as the merged correspondent of /ŋp/. The general statement of UNIFORMITY is as given in (48).

(48) UNIFORMITY: Assign a violation for every surface segment in correspondence with two underlying segments.

For present purposes, we need particular versions of UNIFORMITY applicable to the various prefixes, as shown in (49).

(49) *Affix-specific* UNIFORMITY *constraints for Tagalog*

a. UNIF-/maŋ-$_{other}$/
b. UNIF-/paŋ-RED-/
c. UNIF-/maŋ-$_{adv}$/
d. UNIF-/maŋ-RED-/
e. UNIF-/paŋ-$_{noun}$/
f. UNIF-/paŋ-$_{res}$/
g. UNIF-/mag-kaŋ-RED-/

These are violated if at least one of the two merging segments is found within the prefix in question.

The new constraints of (49) are included in a substantial tableau, too large to print here but viewable in **TagalogFull.xlsx**. We used this spreadsheet to calculate the best-fit weights for the analysis, which were as in (50).

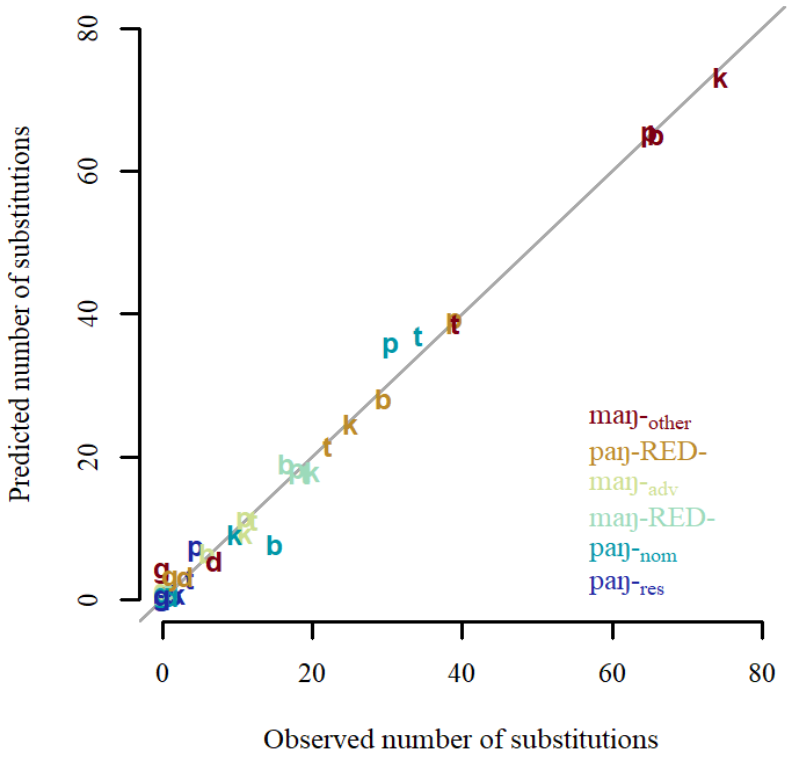
(50) *Constraints and weights for full Tagalog analysis*

NASSUB	2.28
*NC	4.65
*[n	2.10
*[ŋ	3.16
UNIF-/maŋ-$_{other}$/	0[19]
UNIF-/paŋ-RED-/	0.85
UNIF-/maŋ-$_{adv}$/	2.09
UNIF-/maŋ-RED-/	2.27
UNIF-/paŋ-$_{noun}$/	3.85
UNIF-/paŋ-$_{res}$/	5.69
UNIF-/mag-kaŋ-RED-	19.95

The resulting grammar maximizes log likelihood at −248.6, and can be assessed by the following scatterplot of observed vs. predicted counts.

[19] As before in K5, we find that one of the weights comes out as zero; this phenomenon is discussed below in §4.3.

Figure 15: Scatterplot for the full Tagalog analysis



The fit seems reasonably good, with the worst outlier being *paŋ-*$_{nom}$ + *b*, with observed count 15, predicted count 7.7. Examining the errors, it seems there are no obvious constraints that we might want to add in order to increase the accuracy of the analysis. For statistical significance testing, see §5.3.4.

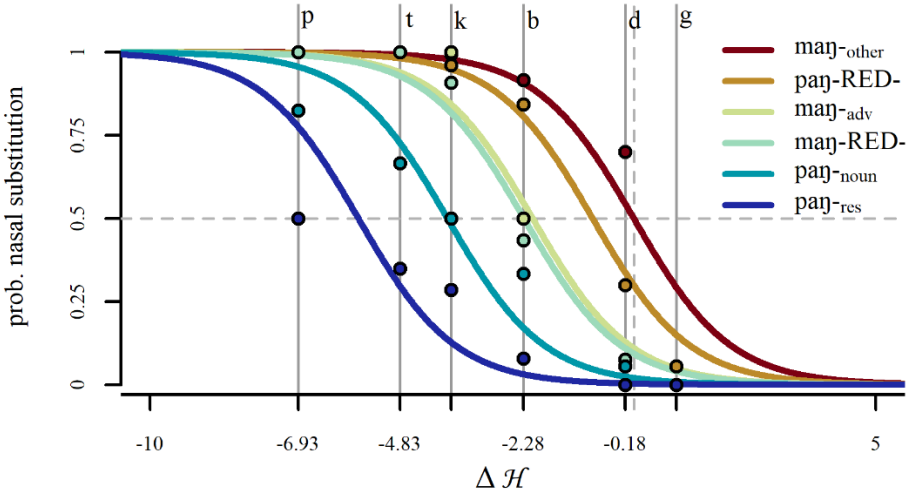
We saw a small anomaly in Figure 13, in that the phoneme /k/ undergoes nasal substitution somewhat more often (88.8%) than /t/ (84.6%), despite the general trend that further back place features undergo substitution less. We can see now that k is not an outlier once we take the behavior of the various individual prefixes into account. In fact, looking back at (47), we can see that /k/ just happens to occur more often than /t/ with *maŋ-*$_{other}$, a prefix that strongly prefers nasal substitution, while /t/ occurs more often than /k/ with *paŋ-*$_{noun}$, which prefers substitution somewhat less. This full model correctly attributes the greater rate of substitution in /k/ forms to the affixes they contain, rather than to /k/ itself.

3.3 Shifted sigmoids for Tagalog

Recall that for K5 in §2.7 above, we attempted to diagnose the probability pattern more closely by plotting the data against the family of sigmoid curves generated by the MaxEnt analysis. We do the same here for the Tagalog analysis, taking the opportunity to generalize the procedure somewhat. For K5, the horizontal axis portrayed the weights of a single constraint family, IDENT(sonorant)$_{affix}$. Here, we let the horizontal axis represent simply a difference in harmony between the substituted and the unsubstituted candidate, calculated using all constraints except the perturbers. In the present case, this will be the summed harmony contributions of

NASSUB, *NÇ, *[n, and *[ŋ. The perturber constraints will be the six affix-specific UNIFORMITY constraints, leaving out UNIF-/mag-kaŋ-RED since its weight is so high (19.95) that the rising part of the curve would fall far to the left of the actual graph, reflecting the fact that no consonant of Tagalog exhibits nasal substitution with this affix. Hence, we are plotting a fairly ambitious set of six shifted sigmoids. As before, the observed data values are included as dots, and model error can be assessed as the vertical displacement of these dots from the intersection of the relevant sigmoid with the vertical line affiliated with the perturber.

Figure 16: Shifted sigmoids for Tagalog: phonological constraints as baseline



We find that, to a rough extent, the data do seem to be accommodated by a family of parallel sigmoids. As with K5, the arrangement of curves and data point reveals how constraints interact in MaxEnt: more extreme values for baseline harmony cause the effects of prefix to be diminished, that is, squeezed into a more limited range.

3.4 Frequency-matching grammars and lexical listing

Having developed a grammar that approximately matches the Tagalog lexical frequencies, we must address an important question: should we expect *individual words* to be pronounced varyingly, in probabilities matching the grammar? Empirically, this is not the case: individual words usually have fixed pronunciations, often specific not to the stem, but to the combination of stem and prefix. To give one example (Zuraw 2000:33), the stem /biˈgaj/ 'gift' resists nasal substitution when the prefix is /paŋ-/: [pa**m**-**b**iˈgaj] 'gifts to be distributed'; but undergoes the process when prefixed with /paŋ-/+RED: [pa-**m**i-**m**iˈgaj] 'act of giving away'.

In fact, the probabilistic grammar we have developed describes the aggregate, not the individual case. Zuraw's view (2000, 2010) is that language learners acquire more knowledge than a simple problem-set analysis ever involves: not only do they memorize a large number of polymorphemic lexical entities (such as the cases just mentioned), but they are also aware of the statistical preponderance of the various output patterns.

Zuraw's claim is testable, since it is possible to query native speakers on how they might pronounce imaginary *novel* stems. Her own empirical work (2000, 2010) involved a rating study (modeled on the "wug test" of Berko, 1958). In this study, Tagalog speakers rated novel forms in which made-up stems either underwent substitution or not. She found that the speakers gave ratings that aligned with the frequency patterns of the Tagalog lexicon. For example, a nonword beginning with /p/ would get a high rating for substitution, a non-word with /g/ would get a low rating for substitution, and a nonword beginning with (say) /b/ or /k/ would get an intermediate rating. Zuraw's finding, which by now has been replicated in many other languages (see next section), will be referred to here as **frequency-matching**.

Frequency-matching may seem somewhat puzzling because it looks redundant: given that speakers already need to memorize individual words (like p*a-**m**i-**m**i* ˈ*gaj* and *pa**m-b**i* ˈ*gaj* above), why should they also bother to represent the general statistical pattern? In fact, there is good reason that language learners should go to trouble of learning these patterns. For instance, a number of prefixes that trigger nasal substitution are relatively productive, and the Tagalog speaker must figure out how to apply the process when a word is new. Beyond this, Tagalog speakers must be able to recognize, and morphologically parse, words they have never heard before; or words from other speakers whose pattern of nasal substitution is different from their own. For these purposes, an accurate internalized "theory of the Tagalog lexicon" would be helpful.

Thus, we make the working assumption that speakers learn a MaxEnt grammar that expresses the statistical patterning of the lexicon — even where the data are idiosyncratic and require memorization. This theory is redundant, and needs to be so, in order match the empirical findings.

It remains to establish just what is the mechanism for memorizing the behavior of individual lexical items. A popular approach is to index the Faithfulness constraints to lexical items; see e.g. Pater, (2000); Becker, (2009); Moore-Cantwell and Pater (2016). Zuraw (2000, 2010) offered a straightforward alternative: the surface forms of the relevant words are memorized. This claim is in line with psycholinguistic work supporting the lexical listing of regular forms (see e.g. Baayen et al., 2002; and Zuraw et al., 2021).

3.5 Frequency-matching more broadly

Zuraw's study was one of the first to establish lexical frequency-matching as an ability of native speakers. Since then, numerous studies have found similar results. For in-depth overviews of the literature, see Ernestus (2011), Shaw and Kawahara (2018), Pierrehumbert (2022), and Alderete and Finley (2023). Here, we offer a more detailed literature review than elsewhere, since the existence of frequency-matching is the very basis for using a probabilistic framework like MaxEnt.

We discuss six specific areas. In all of them, the research strategy is similar to Zuraw's: assess a corpus for its patterns and then test the productivity of the patterns with native speakers.

The first cases are like what Zuraw studied, namely frequency matching of patterns of phonological alternation. These include Eddington (1996) on Spanish diphthongization; Albright and Hayes (2003) on English past tense formation, Pierrehumbert (2006) on velar softening in English; Zhang et al. (2006) and Zhang et al. (2009) on Taiwanese tone sandhi, Hayes and Londe (2006) as well as Hayes et al. (2009) on Hungarian backness harmony, Jun (2015) on Korean /n/ insertion, Jun and Albright (2017) on Korean consonant alternations in verbal paradigms, and Xu (2025) on Egyptian Arabic vowel patterns.

A closely related case is allomorphy, the choice among lexically listed alternatives for the same inflectional category. Frequency matching for the forms of the Italian infinitive is demonstated by Albright (2002), for allomorphy for Hebrew plural suffixes by Becker (2009), for Russian preposition alternations by Linzen et al. (2013), for Russian diminutive allomorphs by Gouskova et al. (2015), and for the English comparative construction (*more* vs. *-er*) by Gouskova and Ahn (2025).

Another case where the phonological grammar makes a choice is in stress patterns. In many languages stress is unpredictable but there are strong statistical regularities; in choosing a stress pattern for a novel word, experimental participants have been observed to frequency-match these regularities; for English see Guion et al. (2003), Olejarczuk and Kapatsinski (2018), Domahs et al. (2014), and Moore-Cantwell (2016, 2020); similar findings have been obtained for Spanish by Eddington (2004), for German and Dutch by Domahs et al., and for Portuguese by Garcia (2017).

Another area of frequency-matching in morphophonology is a curious phenomenon first studied by Ernestus and Baayen (2003): in cases of phonological neutralization (e.g., both /rad/ and /rat/ map onto [rat]), speakers show an ability to guess the underlying representation of a form, using phonological cues. These guesses tend to match the statistics of the lexicon, so that URs that are more frequent for a particular phonological context get guessed more often for nonce forms that match that context. Several studies have found the same effect in other languages: Korean (Kang, 2003; Jun, 2010; Jo 2024a), and Turkish (Becker et al., 2011). For further discussion see §8.2.4 below.

In Chapter 2, we studied another form of frequency matching, namely that of free variation in outputs. Here, the frequency matching is carried out by language learners as they learn to reproduce the variation pattern observed in their own language community; thus, the four Khalkha speakers studied by Fuller (2024) ended up producing a similar, suffix- and phonology-specific pattern for /p/ lenition, presumably from exposure to this pattern as produced by older speakers. Similar cases are discussed in Coetzee and Kawahara (2013), Smith and Pater (2020), and Breiss et al. (2024). These results are just some of many, most of which are found in the discipline of classical sociolinguistics: free variation patterns are generally dialect-specific in their details, indicating that young learners are able to attune their idiolects to the frequencies of the ambient language. For discussion and specific examples see Labov (1994:578-580.)

The existence of frequency matching in syntax is supported by Szmrecsanyi et al.'s (2017) study of how the pair of English dative constructions (*Mary gave a book to John* / *Mary gave John a book*) are frequency-matched by speakers of four dialects of English (American, British,

Canadian, New Zealand). All four dialects are quantitatively different, implying that the experience of growing up in these different countries leads children to learn to match the local mix of dative constructions. For further discussion on frequency matching in syntax, see §8.6.3.

3.5.1 When is linguistic frequency not matched?

Frequency matching is not always observed. First, language learning appears to be subject to a number of biases, which can lead them to end up with a grammar that deviates from the patterning presented in the lexicon. We discuss a variety of these learning biases in §7.3.

Second, children appear to behave differently from adults with respect to frequency matching behavior generally. For example, Hudson Kam and Newport (2005) have argued for the view that young children, unlike adults, have a generality bias and tend to match the most frequent pattern only, ignoring less frequent variants. Yang (2016) argues that acquisition is governed by a tolerance threshold, above which children will not frequency-match but will adopt a simple rule, or try to memorize everything; for a critical discussion of this approach, see Jarosz et al. (2025).

In general, we feel that the studies cited above demonstrate that the end-state of language acquisition does involve frequency matching. We note that any frequency-matching grammar necessarily involves substantial detailed knowledge, which plausibly requires many years of experience to develop. At minimum, learners must know enough words to establish a statistical pattern before they can learn it.

Chapter 4: The MaxEnt Math

4.1 Studying the MaxEnt formula

"Probability is nothing but common sense reduced to calculation."

— Pierre-Simon Laplace (1814)

In this section, we adopt the classical Laplacean view of probability, with the goal of providing a closer intuitive interpretion of the MaxEnt formula. De Morgan (1847:172) was following Laplacean views when he wrote: "By degree of probability, we really mean, or ought to mean, degree of belief." Hájek (2023) amends this by requiring that we are referring to *rational* belief; that is, the degree of belief that would be held by a maximally rational being possessing the relevant facts. The grammar we have hypothesized for Tagalog expresses the native speaker of Tagalog's rational basis for holding a particular gradient belief about what will be the outcome for a given input. Because this rational basis is constructed from the violation pattern, it extends to all input forms that share the same violations. In essence, the constraint violations form the evidence to justify the gradient belief.

Let us now take apart the MaxEnt formula, thinking of it as a way of mapping from evidence to rational gradient belief. The goal is to show that at each step, the procedure given by the formula is intuitive and sensible. We repeat the formula below for convenience.

(51) *Maxent formula (verbose version)*

$$p(x) = \frac{e^{-\sum_{i \in Con} w_i \times v_i(x)}}{\sum_{y \in candidates} e^{-\sum_{i \in Con} w_i \times v_i(y)}}$$

*w*here w_i is weight of the *i*th constraint
$v_i(y)$ is the number of times candidate *y* violates the *i*th constraint
Σ_i denotes the sum over all constraints

We can break this formula down into its component parts, and more easily see what each component contributes to the overall behavior of the model.

4.1.1 Weights

Constraint weights appear in the formula as w_i, and express in numerical terms the idea that constraints differ in strength. A constraint with a high weight will generally force candidates which violate it to have low probabilities. A constraint with a low weight will affect the outcome relatively little, and a constraint with a weight of zero will not affect the outcome at all. For more on the interpretation of constraint weights, see §4.3.

4.1.2 Harmony

Weights combine with violation scores to determine each candidate's Harmony. Harmony, calculated in (51) as $\sum_{i \in Con} w_i \times v_i(x)$, has a negative relationship to probability: the higher the Harmony, the less probable the candidate.

Since MaxEnt relies on the calculation of Harmony, it is a member of the family of frameworks collectively known as HARMONIC GRAMMAR (Legendre et al., 1990; Smolensky and Legendre, 2006; Potts et al., 2010; Pater, 2016; Boersma and Pater, 2016; Hayes, 2017). The simplest, nonstochastic version of Harmonic Grammar was discussed above in §2.5.1.

The use of Harmony as the basis for the MaxEnt framework, as opposed to the strict ranking principle employed in Optimality Theory, has important consequences for the predictions it makes: the probability predicted for a given candidate is the consequence of *all* of its constraint violations, whereas in Optimality Theory, the choice between two candidates is made solely by the highest ranking constraint that distinguishes them. As Jäger and Rosenbach (2006) put it, theories with Harmony make the prediction of *cumulativity*, as defined below: multiple violations of a constraint count more than one violation, and different constraints can "gang" together to counter the effects of a more powerful constraint.

(52) *The two forms of cumulativity* (Jäger and Rosenbach 2006)

a. **Counting cumulativity:** Multiple violations of the same constraint penalize a candidate more than a single violation.

b. **Ganging cumulativity:** Violations of two constraints penalize a candidate more than violation of either one singly.

In his foundational work on probability theory, Jaynes (2003) suggested that one tenet of sound inductive reasoning is that *no relevant data be ignored*. In this respect, classical Optimality Theory is out on a limb, for indeed its decision procedure (see e.g. Prince, 2003a) frequently does ignore data: specifically, during winner-selection, when all losing candidates have been ruled out by higher-ranking constraints, any further constraint violations of the winner are ignored. The discussion of perturber constraints in §2.6.1 constitutes evidence, we feel, that classical OT's strategy of ignoring relevant data is not correct; the very architecture of OT makes it impossible to accommodate perturbers.

Excursus: OT-style decision making in other domains

Classical Optimality Theory finds an echo, perhaps, in the research of Gigerenzer (e.g. Gigerenzer 2001, Gigerenzer and Gaissmaier, 2011) who argues that people often make choices using single salient facts rather than by weighing all the evidence. This model is meant to apply chiefly to cases where data are scarce and speed is essential (Gigerenzer 2001); Gigerenzer and Gaissmaier cite the case of sending a patient to intensive care. In contrast, the decisions involved in learning a grammar are very different: the language learner has a whole childhood's worth of data to go on, with years to ponder and weigh the evidence. Beyond this, existing research suggests that people really do blend a variety of evidence when they make unconscious linguistic choices, as shown by the discussion of perturbers in §2.6.1.

The concept of Harmony permits a terser statement of the MaxEnt formula; instead of (51), we write (53).

(53) *The MaxEnt formula stated in terms of Harmony ($\mathcal{H}$)*

$$p(x)= \frac{e^{-\mathcal{H}(x)}}{\sum_{y \in candidates} e^{-\mathcal{H}(y)}}$$

4.1.3 eH, evidence, and nearness to certainty

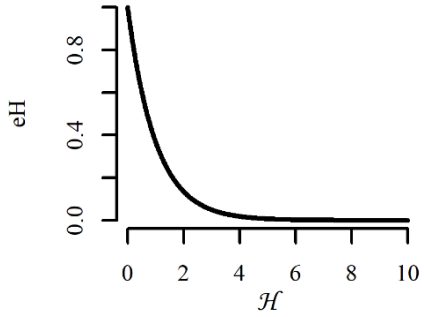
Having calculated Harmony, we can now calculate eH: $e^{-\mathcal{H}(x)}$. This expression appears as the numerator in the MaxEnt formula as given in (53). We arrive at eH by first negating Harmony, and then exponentiating it using *e* as the base.

The reason for negating the Harmony score is simply to ensure that it will act as a penalty; if Harmony is large, then eH will be small; and smaller eH values ultimately correspond to smaller probabilities.[20]

The use of exponentiation can be made intuitive if we consider how this function "warps" Harmony space. This is demonstrated in Figure 17.

[20] Recall that under the sign conventions we adopt (§2.5.2), Harmony is always positive or zero.

Figure 17: eH values change quickly with low Harmony, slowly with high Harmony



We consider Harmony to be an aggregated measure of evidence: the combined and weighted testimony from all the constraints, that a particular candidate should not be the winner. The consequence of exponentiation is that *added* evidence has effects that are not linear. Specifically, when we have little or no evidence about a candidate, a little more Harmony makes a large difference in eH (and, ultimately, probability). But if we are already close to certainty, then additional Harmony matters very little. For example, in Figure 17, when we move Harmony from 0 to 2, eH changes a great deal, by more than 0.8. But if we start Harmony off at 8 and move it up to 10, the change in eH is almost imperceptible.

We suggest that this mathematical pattern matches up well to human intuition and experience. As a spur to the imagination, we offer a simple example. Let the novel information be that the celebrated football player Lionel Messi has just pulled a hamstring and cannot play; a lesser player will have to take his place. We place Messi in a match between his national team Argentina and evenly-matched Brazil, and assume that there is only a little time left to play, say, two minutes. The following chart provides, we suspect, a reasonable approximation to normal human intuition, for three scenarios.

(54) *An illustration of the varying effect of evidence on belief*

New evidence: Messi has just pulled a hamstring and cannot play.

Scenario	*Change in* P(*Argentina will defeat Brazil*)
Argentina and Brazil are tied 1-1.	$0.5 \rightarrow 0.4$
Argentina leads Brazil 3-0.	$0.9995 \rightarrow 0.9993$
Argentina trails Brazil 0-3.	$0.0005 \rightarrow 0.0003$

The point is that Messi's injury could hardly matter much when we are already almost certain who will win (it's very rare for three goals to be scored in two minutes, no matter who is playing). When we are not certain about the outcome (game tied), Messi's absence matters much more.

It is not difficult, we feel, to dream up scenarios similar to this one, and we suspect that the pattern seen — and hence the MaxEnt conversion of $\mathcal{H}$ to eH — embodies what we think of as common sense.

The concept of eH permits a still-terser statement of the MaxEnt formula, as in (55).

(55) *The MaxEnt formula stated in terms of eH*

$$p(x) = \frac{eH(x)}{\sum_{y \in candidates} eH(y)}$$

4.1.4 Computing Z and probability

In order to convert eH into a probability, we must ensure that the probabilities for all candidate outputs for a single input sum to 1. This is what the denominator in (55) does: we create a normalizing value Z by summing the eH values across candidates. The crucial consequence of this is that the probability of a candidate depends not just on its own eH, but on that of its competitors; that is, a candidate gets a lower probability when it must compete with strong rivals.

4.1.5 Local summary

In sum, the MaxEnt formula defines a system in which constraints vary in their strength, where the meaning of strength is "ability to lower the predicted probability of violating candidates". The contributions of the various constraints are pooled, reflecting the idea that no relevant evidence should be ignored. The conversion of harmony to eH means that the same bit of novel evidence matters most when the decision is still in doubt, and least when the decision is close to certain. Finally, the division by Z expresses the commonsense notion that an outcome is less probable when it competes with strong alternatives. In sum, MaxEnt is a particular embodiment of the claim made by Laplace, quoted at the start of this chapter.

4.2 Relationship of MaxEnt to classical Optimality Theory

For every grammar formulated in classical Optimality Theory, there exists a close MaxEnt equivalent (Eisner, 2000; Manning, 2003). The trick in finding it is to reexpress all cases of OT constraint domination as differences in weight that are sufficient to assign essentially-zero probabilities to the losing candidates.

To illustrate this, we return to the non-stochastic toy language K1, whose OT tableaux were given in (5). We give here an alternative set of MaxEnt tableaux that generate closely equivalent outcomes.

(56) *Re-expressing the OT analysis of (5) in MaxEnt*

	IDENT (son)$_{cont.}$	*N [+cont]	*p	IDENT (son)				
/mʊntəg **p**oɮ-ən/	120	120	30	0	$\mathcal{H}$	eH	Z	p
faithful [p]			1		30	9.36×10^{-14}	≈ 1	9.36×10^{-14}
lenited [w]				1	0	≈ 1	≈ 1	≈ 1
/ʃit-t͡ʃʰəx-sə**n** **p**ai-sən/								
faithful [p]			1		30	9.36×10^{-14}	9.36×10^{-14}	≈ 1
lenited [w]		1		1	120	7.67×10^{-53}	9.36×10^{-14}	8.19×10^{-40}
/ɔrʊɮ-əx **p**ɔɮəm-t͡ʃʰɔi/								
faithful [p]	1		1		30	9.36×10^{-14}	9.36×10^{-14}	≈ 1
lenited [w]				1	120	7.67×10^{-53}	9.36×10^{-14}	8.19×10^{-40}

The constraint weights in (56) essentially recapitulate the domination relations portrayed earlier in §2.1: the relation *p ≫ IDENT(sonorant) is expressed by a weight difference in (56) of 30 (= 30 − 0) between these two constraints. The domination relation IDENT(sonorant)$_{content}$ ≫ *p is expressed by a weight difference of 90 (120 − 30) between these two constraints, and the same holds for *N[+cont] > *p. The resulting probabilities for the loser candidates of the OT analysis are all below 10^{-13}, a value that we have decided (§2.9) to accept as the practical equivalent of zero. For more technical discussion of how to reexpress an OT grammar as a weight-based grammar, see Prince (2003a).[21] For another example of an OT grammar rendered in MaxEnt, see §7.7.1.

The upshot of the present discussion is that MaxEnt can be thought of not so much as a replacement for classical OT, but as a probabilistic extension of it; any result derived in classical OT could in principle be derived in MaxEnt as well. Given the great body of empirical material that has been addressed in classical OT, this is important.

4.3 Interpreting constraint weights

4.3.1 How weights determine probability: log odds in MaxEnt

In §4.1.1, we gave the simple intuitive meaning of weights: the higher a constraint's weight, the greater its effect in lowering the probability of candidates that violate it. However, constraint weights actually have a very specific relationship to the probabilities they predict. We illustrate

[21] Prince's discussion relates to non-stochastic Classical Harmonic Grammar, described above in §2.5.1. To cover MaxEnt, the weight differences Prince calculates must be augmented to cover a large interval corresponding to what we are willing to accept as the equivalent of strict ranking (Eisner 2000).

this relationship here, with the intent that the reader can use it to develop an intuitive sense of what the real-number value of a constraint weight means.

Consider a very simple tableau, with two candidates and one constraint ℂ.

(57) *MaxEnt tableau: one constraint, two candidates*

		ℂ **w**	$\mathcal{H}$	**eH**	**Z**	**p**
Input	*Cand₁*		0	1	$1+e^{-\mathbf{w}}$	$\frac{1}{1+e^{-\mathbf{w}}}$
	Cand₂	1	w	$e^{-\mathbf{w}}$		$\frac{e^{-\mathbf{w}}}{1+e^{-\mathbf{w}}}$

The important thing here is that *Cand₂* incurs exactly one more violation than *Cand₁*. In this case, the ratio of the probabilities of *Cand₁* and *Cand₂* — what a gambler would call the *odds* — is completely determined by the weight of ℂ. We used the MaxEnt formula (computations given in the right columns of (57)) to calculate these odds for various values of **w**, shown in (58). As can been seen, the value of the odds increases non-linearly with **w**. We find it useful in the intuitive understanding of MaxEnt weights to keep some approximation of this table in our head.

(58) *Probability ratios (odds) of Cand₁ to Cand₂, as determined by* ***w****, in scenario* (57)

w	*Odds*
0	1.00
0.1	1.11
0.2	1.22
0.5	1.65
1	2.72 = *e*
2	7.39
5	148.41
10	22026.47

A weight of 0 yields an odds of 1, or '1 to 1', meaning that *Cand₁* and *Cand₂* are equally likely. A small weight like 0.1 has only a subtle effect, whereas a weight of 10 renders the odds overwhelmingly in favor of *Cand₁*. The probability of *Cand₂* will be empirically indistinguishable from zero at this weight.

The table in (58) is easily generalized: as stated it expresses the difference in harmony resulting from one single constraint, but the same results would hold for *any difference in harmony* (call it $\Delta\mathcal{H}$), whether arising from one constraint or several. In the Excursus box, we give a simple proof that justifies this claim. It tells us that $\Delta\mathcal{H}$ is equal to the log of the odds of two candidates whose Harmonies differ by $\Delta\mathcal{H}$, or in mathematical notation, $\ln(O)=\Delta\mathcal{H}$.

Excursus: The log-odds principle

Here, we show that any two candidates for the same input whose Harmony values differ by the quantity $\Delta\mathcal{H}$ will be assigned a probability ratio O (short for "odds") whose natural log is $\Delta\mathcal{H}$. In the proof below, the term "log odds" is used to mean "log of the odds".

(59) *The log-odds principle*

- Let P(*Cand*$_1$) and P(*Cand*$_2$) be the probabilities assigned by a MaxEnt grammar $\mathbb{G}$ to two candidates for the same input.
- Let O be the odds assigned by $\mathbb{G}$ to these candidates: O = P(*Cand*$_1$)/P(*Cand*$_2$)
- Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the Harmony values assigned by $\mathbb{G}$ to these candidates.
- Let $\Delta\mathcal{H} = \mathcal{H}_2 - \mathcal{H}_1$, the difference in harmony between the candidates.

Then:

$\ln(O)=\Delta\mathcal{H}$

In other words, the harmony difference is equal to the log odds.

Proof:

$O=\frac{P(cand_1)}{P(cand_2)}$	Definition of O
$=\frac{e^{-\mathcal{H}_1}/Z}{e^{-\mathcal{H}_2}/Z}$	Plugging in the MaxEnt formula
$=\frac{e^{-\mathcal{H}_1}}{e^{-\mathcal{H}_2}}$	Multiply denominator and numerator by Z
$=\frac{e^{-\mathcal{H}_1}}{e^{-(\mathcal{H}_1+\Delta\mathcal{H})}}$	By definition of $\Delta\mathcal{H}$
$=\frac{e^{-\mathcal{H}_1}}{e^{(-\mathcal{H}_1-\Delta\mathcal{H})}}$	Distribute the negative sign
$=\frac{e^{-\mathcal{H}_1}}{e^{-\mathcal{H}_1}\times e^{-\Delta\mathcal{H}}}$	$x^{a+b} = x^a \times x^b$
$=\frac{1}{e^{-\Delta\mathcal{H}}}$	Divide denominator and numerator by $e^{-\mathcal{H}_1}$
$= e^{\Delta\mathcal{H}}$	$\frac{1}{x^{-a}} = x^a$
and so:	
$O = e^{\Delta\mathcal{H}}$	
$\ln(O)=\ln(e^{\Delta\mathcal{H}})$	Take the natural logarithm of both sides.
$\ln(O)=\Delta\mathcal{H}$	$\ln(e^x) = x$

The log-odds principle has important consequences for MaxEnt analysis. Notably, in many analyses, the weights of individual constraints cannot be fully determined, since all that really matters is the $\Delta\mathcal{H}$ values for each pair of candidates. There are various examples of this kind; a very simple one is the case where *Cand1* violates the two constraints $\mathbb{C}_1$ and $\mathbb{C}_2$, and *Cand2* violates none. Here, any weighting of $\mathbb{C}_1$ and $\mathbb{C}_2$ in which **w1** and **w2** have the same sum will produce the same result; see (60).

(60) *Tableau: Any set of* ***w1*** *and* ***w2*** *whose sum is the same yields the same result*

		$\mathbb{C}_1$ **w1**	$\mathbb{C}_2$ **w2**	$\mathcal{H}$
Input	*Cand1*	1	1	**w1+w2**
	Cand2	0	0	0

Another simple case in which an infinite number of weightings will yield the same result also involves two candidates and two constraints, but with the violations arranged in a diagonal pattern, as shown in (61). For this case we provide a worked-out example, **DifferentWeightsYieldSameOutcome.xlsx**. The violation scheme here is abstract, but is instantiated in analysis over and over by a Markedness constraint and an opposing Faithfulness constraint.

(61) *Two constraints and two candidates with 75/25 probability*

			$\mathbb{C}_1$ **2**	$\mathbb{C}_2$ **0.901**				
		or:	**1.099**	**0**	$\mathcal{H}$	eH	Z	p
Input	*Cand1*	0.75		1	0.90	0.41	0.54	0.75
	Cand2	0.25	1		2	0.14		0.25

We find that if w_1 is held constant at 2, then the weight for $\mathbb{C}_2$ that yields the 75/25 probability difference is 0.901. But as the Log-Odds Principle tells us, *any* pair of weights that differs by the same amount — in this case, 1.099 — will yield the same answer; the reader can easily verify this by changing the weights in the spreadsheet.

Among the pairs of weights that differ by the value 1.099 will be what we listed in the third row of (61), namely **w2** = 0 and **w1** = 1.099. This looks perhaps puzzling, because $\mathbb{C}_2$ seems to have "dropped out of the grammar" — by the MaxEnt math it can have no influence on the outcome. We will cover this phenomenon in more detail below, because in our experience, the appearance of surprising zeroes is an ordinary occurrence in MaxEnt analysis — even for constraints that phonologists intuitively feel belong in the grammar. We have already seen this happen in K5 (§2.7) and Tagalog (§3.2.1). In the next section we take a closer look at the Tagalog case.

Before moving on, we note that the effect of a constraint weight on log odds must be interpreted in the context of the rest of the analysis. A constraint $\mathbb{C}$ with weight **w** = 1.099 will

always reduce the probability of a violating candidate relative to a non-violating one by a factor of 3. In the example above, this was a reduction from 0.75 to 0.25. However, in cases where the candidate competition is dominated by other, highly weighted constraints, then the reduction may be very small, say, from 0.000000075 to 0.000000025. Obviously only the first case matters empirically. We have seen such diminishing effectiveness of constraints before; see discussion in §2.8 and §4.1.3.

4.3.2 Zero weights in practice

Having studied the two-constraint case in (61), we can now see the Tagalog case as simply a more elaborate instance of the same thing. Corresponding to $\mathbb{C}_1$ is NASSUB, corresponding to $\mathbb{C}_2$is the *whole family* of UNIFORMITY constraints. We can see this for the simple case of stem-initial /b/, shown in (62).

(62) *Illustrative tableaux for Tagalog stems starting with /b/*

			NASSUB	UNIF-/maŋ-other/	UNIF-/paŋ-RED-/	UNIF-/maŋ-adv/	UNIF-/maŋ-RED-/	UNIF-/paŋ-noun/	UNIF-/paŋ-res/
prefix	**cand.**	**obs.**	3.28	0	0.85	2.09	2.27	3.85	5.69
maŋ-other	unsub	6	1						
	sub	66		1					
paŋ-RED-	unsub	5.5	1						
	sub	29.5			1				
maŋ-adv	unsub	6	1						
	sub	6				1			
maŋ-RED-	unsub	21.5	1						
	sub	16.5					1		
paŋ-noun	unsub	30	1						
	sub	15						1	
paŋ-res	unsub	17.5	1						
	sub	1.5							1

Crucially, any one candidate with substitution violates precisely one of the six UNIFORMITY constraints. It should be clear that, just as in (61), we can add an arbitrary amount to NASSUB, and the same amount to every member of UNIFORMITY, and this addition will make no difference to the output probabilities, by the same reasoning — each input, viewed by itself, will have the "diagonal" violation pattern that we saw in the two-constraint example. As with the two-constraint example, a weighting algorithm such as the one used in the Excel Solver, which attempts to keep the weights low, will likely assign a zero to the weakest of the UNIFORMITY weights, which happens to be the weight for UNIF-/maŋ-other/.

Seeking to understand the zero weight further, we observe that this analysis, like all generative grammar analyses, is intended as a *fragment*, covering but a small area of the vast system internalized by the native speaker. We believe that the zero weights obtained for Tagalog above are indeed the result of an incomplete analysis. For simplicity we deployed an analysis with only two candidates per input. But there are other candidates that violate UNIFORMITY. Consider for instance, a candidate like [ãmigaj] from underlying /ma**ŋ-b**iˈgaj/; where the [ã] corresponds to both the input m and the input a (in the formal notation of McCarthy and Prince (1995), this is /m$_1$a$_1$**ŋ$_3$-b$_4$**iˈgaj/ → [ã$_{1,2}$**m$_{3,4}$**iˈgaj]).[22] Since the segments /m/ and /a/ have been phonologically merged into [a], and since they belong to the prefix named "maŋ-$_{other}$," this candidate incurs a violation of UNIF-/maŋ-$_{other}$/. Since the candidate is impossible in Tagalog, we seek an analysis in which it gets a probability of near-zero. This will only be the case if UNIF-/maŋ-$_{other}$/ is given a very substantial weight, with NASSUB and the remaining UNIFORMITY constraints comparably boosted.

In light of such cases, we recommend careful attention in cases of zero weights: if the system is considered more broadly, does it turn out that the zero-weighted constraint has other uses? These often can be demonstrated by adding candidates or inputs to the analysis.

This said, we add that sometimes a constraint is fit with a weight of zero simply because that is the actual best-fit weight for the training data. A schematic case of this is shown in (63).

(63) *A case in which a weight must be zero to fit the data*

			$\mathbb{C}_1$ **1.099**	$\mathbb{C}_2$ **0**	$\mathcal{H}$	eH	Z	p
Input$_1$	*Cand$_1$*	0.75			0	1	1.33	0.75
	Cand$_2$	0.25	1		1.10	0.33		0.25
Input$_2$	*Cand$_1$*	0.75			0	1	1.33	0.75
	Cand$_2$	0.25	1	1	1.10	0.33		0.25

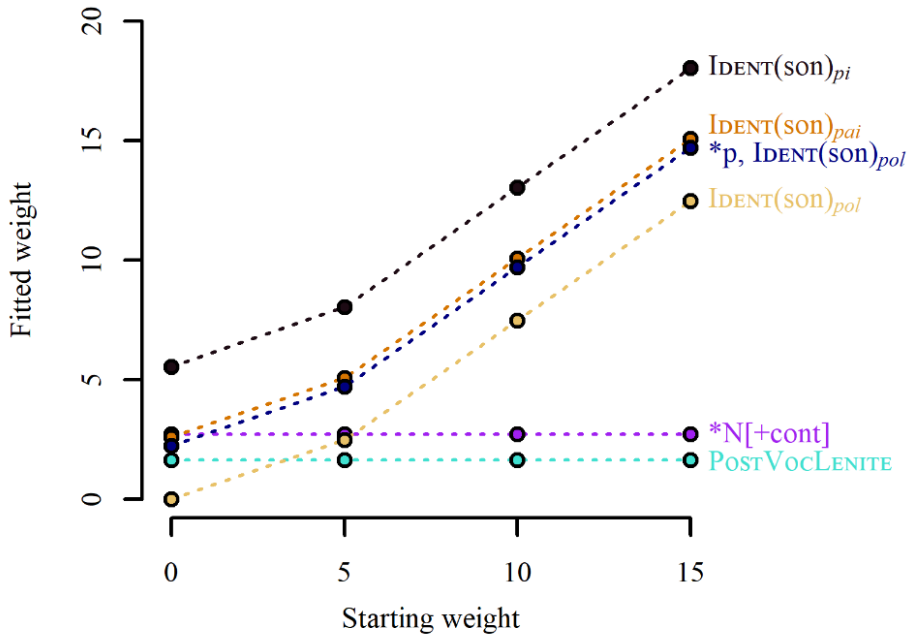
Here, any non-zero weight assigned to $\mathbb{C}_2$would result, incorrectly, in the assignment of different output probabilities to *Input$_1$* and *Input$_2$*.

How can the analyst tell whether a constraint's weight must be zero or not? We suggest two methods. First, in Solver, one can fix the weight of a constraint in advance at a non-zero value (see Figure 29 for how), run the Solver to fit the remaining constraints, and see if the same model fit can be achieved. For (63), if we fix the weight of $\mathbb{C}_2$ at 1 and solve, the log likelihood worsens from −1.12 to −1.17. The predicted probabilities are a worse match than those shown in (63) as well, with 0.66 predicted for *Cand$_1$* of *Input$_1$*, but 0.84 predicted for *Cand$_1$* of *Input$_2$*. In contrast, for the Tagalog case, setting the weight of UNIF-/maŋ-$_{other}$/ to 1 and re-solving results in exactly the same model fit as in the tableau in (62).

[22] The reader may object that [ãmigaj] could be ruled out independently by a ban on nasalized vowels, which are not part of the Tagalog phoneme inventory. But the very same point could be made with [amigaj], in which non-nasal [a] is also the result of coalescence.

A second method is to run the Solver using a few different starting weights. This can serve as a useful diagnostic not just of zeros but of the model structure overall.[23] For instance, we tried fitting weights to our K5 model (§2.7) starting all constraints out with a weight of 0, 5, 10, or 15. All of the resulting models were equally accurate (log likelihood = −267.54), but the constraint weights often varied depending on their starting point. For the perturber constraints *N[+cont], and POSTVOCALICLENITE, the weights did *not* vary, as shown in Figure 18. Evidently, these weights must have their original fitted values to match the data. But the remaining constraints responded to varying starting points by varying in parallel; for them is it is the *differences* among weights (specifically, pairwise between *p and each IDENT constraint) that matters.

Figure 18: How the weights for K5 emerge with different starting points



We find it helpful to examine the "elsewhere" inputs in our K5 tableaux ((31) above, p. 35) in light of this patterning — these show the pairwise violations that imply a constant weight difference.

To conclude this section, we return briefly to our earlier discussion (§4.6) of the differences between a MaxEnt grammar and a statistical analysis; we here consider zero weights from this perspective. In both (61) (schematic case) and (62) (Tagalog), one of the parameters of the model is — in statistical terms — redundant; you can predict the output patterns just as well leaving one parameter out. If we attempted to fit these models using statistical software such as R, we will get an error; in R the program provides an error message stating that the model is not 'identifiable', meaning that there is no unique solution — just as described above. Thus, it is yet another difference between MaxEnt OT and statistical testing is that it is only in the former that redundant constraints are tolerated. Of course, even though there are good scientific reasons to

[23] A note of caution: this procedure will not work if one is using a prior, as discussed in §4.5.

employ redundant constraints, it pays for analysts always to be aware of such redundancies when they interpret their weights.

4.4 The math of weight-fitting

In Chapter 2, we introduced the use of Excel Solver for fitting weights. Here we go into the math in greater depth, covering first the essential notion of likelihood, then covering the results that make trustable weight-setting possible within MaxEnt.

4.4.1 More on likelihood

As covered in Chapter 2, we use log likelihood as a metric for the accuracy of a MaxEnt analysis, and as the basis for finding the best-fit weights for the data. We demonstrated how to calculate and use log likelihood in §2.6.4, but we now go into more detail about why log likelihood is a sensible metric of accuracy.

For intuitive purposes, it is best to start not with log likelihood, but with likelihood per se. As noted above, the likelihood of the data is the probability that our data would be observed if our model is correct. The idea is that if we are comparing the behavior of several models on a single dataset, the best one is the one that makes the actual observed data the most likely.

Likelihood is calculated by building up from the probabilities that the model assigns to single data points. For a simple case, we return again to K3, from §2.4. To recall the facts, in K3 /p/ lenition is blocked after nasals, inapplicable in stems, and applies 73% of the time in all other cases; the data are repeated below in (64). Earlier, we developed a three-constraint MaxEnt grammar to cover these data, whose constraints and weights are re-listed in (65).

(64) *Hypothetical data for K3*

UR	SR	count
/N-p/	N-p	212
	N-w	0
/C-p/	C-p	300
	C-w	110
/C-$p_{content}$/	C-p	720
	C-w	0

(65) *MaxEnt grammar for K3*

Constraint	*Weight*
IDENT(son)$_{content}$	20
*N[+cont]	20
*p	3
IDENT(son)	4[24]

In order to understand likelihood better, let us begin with a single data point, an instance of the observation [mʊntəg **w**oɮən] from underlying /mʊntəg **p**oɮ-ən/ (see (18)). The grammar we developed for K3 assigns a probability of 0.27 to this outcome. According to this grammar, then, the likelihood of our observation is exactly 0.27.

Suppose we expand our data set to two forms, of which the second is an instance of [mʊntə**g** **p**oɮən], without lenition, from the same input. Our grammar assigns a probability of 0.73 to [mʊntəg **p**oɮən] and of 0.27 to [mʊntəg **w**oɮən]. Here, the likelihood of the data for this expanded data set is $0.73 \times 0.27 = 0.1971$. This is because the probability of two independent events is their product (e.g., the probability of two heads in a double coin flip is $0.5 \times 0.5 = 0.25$). Observe that there is another grammar that would better match this tiny dataset: it would assign 0.5 probability to both [mʊntəg **p**oɮən] and [mʊntəg **w**oɮən]. Its likelihood would be higher: 0.25.

Scaling up further, suppose now a still larger subset of the original data, as in (66):

(66) *Pseudo-data for K3 (likelihood illustration)*

/mʊntəg **p**oɮən/	→	[mʊntəg **p**oɮən]	15 tokens
/mʊntəg **p**oɮən/	→	[mʊntəg **w**oɮən]	5 tokens

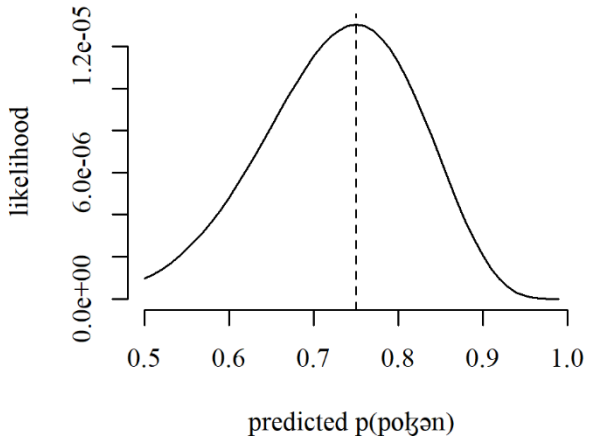
Probability assigned to [mʊntəg **p**oɮən] by grammar (65):	0.73
Probability assigned to [mʊntəg **w**oɮən] by grammar (65):	0.27

Again, we multiply out the probabilities for all 20 data, using exponentiation to save space. This turns out to be $0.73^{15} \times 0.27^{5} = 1.278 \times 10^{-5}$.

Once again, there is an intuitively better grammar available: we find weights such that P([mʊntəg **p**oɮən]) is exactly 0.75 (matching 15/20) and P([mʊntəg **w**oɮən]) is exactly 0.25 (matching 5/20). The likelihood for this grammar turns out to be 1.305×10^{-5} — a bit higher, reflecting our intuitions. More generally, we can calculate the likelihood of the data for a whole range of values of P([mʊntəg **p**oɮən]), showing how the maximum is indeed at 0.75. This is seen in Figure 19:

[24] Recall that our weights for K3 were fit by hand; this is why they are whole numbers.

Figure 19: How the likelihood of data (66) depends on P([mʊntəg **p**oɮən]*)*



The graph offers some insight into the problem that likelihood addresses. If we were to propose a very high probability for [mʊntəg **p**oɮən] (right side of graph), we would make this, the most common outcome, more likely — the likelihood for just the [mʊntəg **p**oɮən] data points would be close to 1, which is good. But the better we do at predicting the [mʊntəg **p**oɮən] data points, the worse we do at predicting the [mʊntəg **w**oɮən] data points. The likelihood criterion allows us to find the best balance across all of our data, which matches the overall 0.75/0.25 observed distribution.

The general procedure for calculating likelihood for any size of dataset is given in (67).

(67) *Procedure for calculating likelihood*

a. Find the probability that each observed datum is assigned by the analysis. For data types (identical violation profile) that occur multiple times in the training data, count the *n* tokens of this type as *n* separate instances.
b. Multiply these probabilities together. This product is the likelihood.

Expressed as a formula, likelihood is defined as in (68).

(68) *Likelihood formula*

$$\text{Likelihood} = \prod p_i^{n_i}$$

where

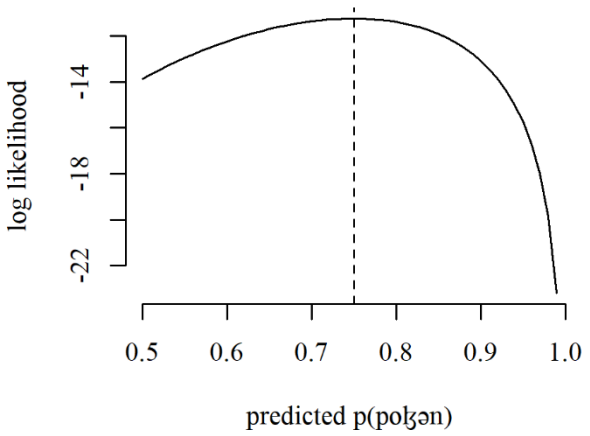
n_i is the number of tokens of output *i*
p_i is the probability the model assigns to output *i*
Π denotes multiplication

Having covered likelihood, it is straightforward to extend the concept to log likelihood, which is what we actually use for purposes of weight-fitting. When we scale up to larger cases we will find that, the actual likelihood — although fully meaningful as a measure of model fit — becomes extremely small. For example, the likelihood of the K4 data in (20) under our final

analysis is 5.31×10^{87}. As we move beyond K4 (which is still just a toy example), to cases of realistic size, the software we use will no longer be able to represent the likelihood and will round it off to the (useless) value of zero. We can overcome this problem by introducing logarithms: the *log likelihood* of a dataset under an analysis is the log of the likelihood.

For purposes of setting the weights, log likelihood will serve just as well as likelihood: because log likelihood increases *monotonically* with likelihood. Hence, whatever weights maximize log likelihood will also maximize likelihood. To show this, the graph of Figure 19 is replotted below in log-likelihood terms, showing that we get the same maximum point.

Figure 20: Log likelihood yields the same maximum point as likelihood



Thus, using logarithms makes the numbers more tractable, but otherwise works very similarly to likelihood and yields the same answer. In (69) we give the math for calculating log likelihood.

(69) *Deriving equivalent calculations for log likelihood*

$\text{Likelihood} = \prod p_i^{n_i}$	from (67)
$\log(\text{Likelihood}) = \sum \log(p_i^{n_i})$	because $\log(x \times y) = \log(x) + \log(y)$
$\log(\text{Likelihood}) = \sum n_i \log(p_i)$	because $\log(x^y) = y \times \log(x)$

As can be seen, whereas in calculating likelihood we dealt with *n* instances of a datum by taking its probability to the *n*th power, with log likelihood we multiply by *n*. Whereas for likelihood we dealt with multiple types of data by multiplying, for log likelihood we sum them. These are the calculations we have already been doing on the spreadsheets given in Chapters 2 and 3. For example, in (24) on p. 28 above, n_i corresponds to column D, $log(p_i)$ to column N, and the computation itself is carried out by the `SUMPRODUCT()` function in cell `O2`.

Excursus: Interpreting log likelihood values

The log likelihood of a model is not directly interpretable as a number; it only is useful as a basis of comparison. Indeed, log likelihood values will be very different across simulations because it responds directly to the size of the data set. But even though log likelihood is

arbitrary, it is possible to calculate a couple of anchor points that can give us a rough idea of how well an analysis is doing.

First, if we set all constraint weights to zero, we can get a baseline likelihood, essentially the equivalent of saying there is no grammar at all: the candidate probabilities for each input come out the same. We will call this the **zero-model**. Obviously, zero models are uninformative and receive very low log likelihood. We can use them as a lower anchoring point.

At the other end of the scale, we can construct an equally absurd grammar that obtains the highest possible log likelihood for a given data set. For instance, one could construct a model that includes a constraint of the form INPUT X MUST HAVE OUTPUT Y for all combinations of X and Y. Such a model is often called a **saturated model**.

Note that, although a saturated model is likely to have a high log likelihood, the value will *not* be as high as zero. A zero log likelihood will arise only when there is no variation in the data, and the grammar matches the data perfectly. In other words, there is a difference between perfectly matching every single data point and perfectly matching a probability distribution; it is the latter goal that we are generally aiming at in this book.

The saturated model does not actually need to be constructed; rather one can substitute "empirical probabilities" for predicted probabilities in the log likelihood equation. By empirical probability, denoted E_{ij}, we mean the fraction of the observed cases with input *i* in which the output is candidate *j*. Using this concept, we can calculate the log likelihood of a saturated model as in (70).

(70) *Formula for log likelihood of a saturated model*

$$\sum n_{ij} \log(E_{ij})$$

where

Σ denotes summation over all candidates for all inputs
n_{ij} is the number of times candidate j for input *i* is observed
E_{ij} is defined as above.

A sample spreadsheet that performs these calculations for our K5 analysis is given in the Supplementary Materials as **K5_zero_and_saturated.xlsx**. Below, we give for three models discussed in this book, the log likelihoods of the zero model, the model we proposed, and the saturated model.

(71) *The log likelihood of three models, as bounded by the random and saturated models*

	Zero model	*Our model*	*Saturated model*
K5 (§2.7)	−482.4	−267.5	−261.5
Tagalog (§3.1)	−600.3	−248.6	−221.2
Turkish (§6.2)	−31199.9	−25317.3	−24410.9

We see here that in the absence of comparison, log likelihood is meaningless; the Turkish simulation just happens to have far more data points than K5 or Tagalog.

A final note: in calculating log likelihood we need to consider the possibility that the calculation might crash due to the presence of zeros. First if the *observed* frequency is 0, we needn't worry: the log of the predicted probability will be multiplied by zero, adding zero to the total log likelihood. On the other hand, if any *predicted* probability were truly zero we wouldn't be able to calculate log likelihood — the log of zero is undefined, so the total log likelihood would also be undefined. It doesn't help to go back to plain (non-logged) likelihood, because we would wind up multiplying all other probabilities by zero, yielding a likelihood of zero no matter what the models' predictions were for other candidates. Fortunately for us, MaxEnt never generates true zero as a predicted probability, only probabilities very close to zero, so this issue will generally not arise. It is, however, a tricky issue for other theories of probabilistic phonology (see §9.5) which do generate actual zeros.[25]

4.4.2 Convexity and finding the maximum likelihood weights

Having shown that maximizing log likelihood is a sensible way to find an accurate grammar, we turn to the problem of finding a set of weights that achieves this maximum. To facilitate understanding how these weights are found, it is helpful to consider the problem in geometric terms. Imagine a simple two-constraint grammar, along with a dataset it is intended to model.

(72) *Toy grammar for illustrating weight computation*

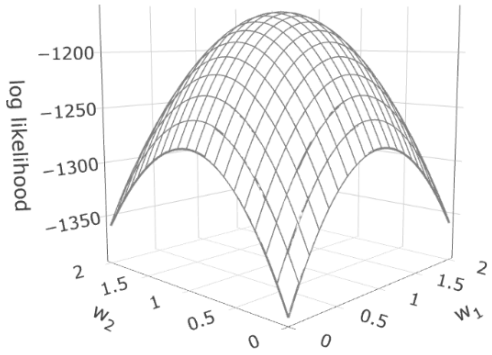
		obs.	$\mathbb{C}_1$ **w_1**=1	$\mathbb{C}_2$ **w_2**=1	$\mathcal{H}$	eH	Z	p	ln p	LL
$Input_1$	$Cand_1$	731			0	1	1.368	0.731	−0.313	−1164.5
	$Cand_2$	269	1		1	0.368		0.269	−1.313	
$Input_2$	$Cand_1$	731			0	1	1.368	0.731	−0.313	
	$Cand_2$	269		1	1	0.368		0.269	−1.313	

As can be seen, $Input_1$ and $Input_2$ behave identically with $\mathbb{C}_1$ and $\mathbb{C}_2$ respectively.

To visualize the weight computation (as carried out, for instance, by the Solver) we can make a three-dimensional graph, with the two constraint weights on the *x* and y axes, and the log likelihood corresponding to every possible combination of weights appearing on the vertical *z* axis. A graph of this type is shown in Figure 21.

[25] One way of dealing with this is simply to convert all zero model probabilities to some very small value; Zuraw and Hayes (2017) use 0.001.

Figure 21: 3D graph relating log likelihood to the weights of a two-constraint grammar



Here, the log-likelihood surface forms a dome, whose highest point (altitude −1164.5) occurs where $\mathbf{w_1}$ and $\mathbf{w_2}$ have their optimum values, in this case $\mathbf{w_1} = 1$, $\mathbf{w_2} = 1$. With visualizations like this, one can construe the problem of finding the best weights as a hill-climbing problem: keep going uphill until you can't any longer. Of course, when the number of constraints is higher than two, then the hill becomes many-dimensional and impossible to visualize, but the computation is essentially the same.

The efficient ascent of hills like in Figure 21 is a standard problem in computer science, and many algorithms exist, including, for instance, the one in the Excel solver (Generalized Reduced Gradient Method; Lasdon et al., 1974). Generally, the algorithms proceed iteratively, leaping up the hill in stages, each calculated by a formula that defines the algorithm (the **objective function**, discussed in §4.5). In larger problems you can see the stages proceeding one by one at the bottom of your screen when operating the Excel Solver.

For ordinary linguistic analysis, we can sensibly just pick some widely-used algorithm; it should not be a matter of linguistic theory to invent new hill-climbing algorithms. However, an encouraging fact about hill climbing in MaxEnt is that the search space is very well-behaved: there is never more than one peak, so that for any algorithm, it suffices simply to keep going uphill; you will never get stuck in a so-called local maximum. Another way this is often put is to say that the log-likelihood function for MaxEnt is **convex**. A proof of convexity for MaxEnt is given in Della Pietra et al. (1997), and explained in textbook form by Jurafsky and Martin (2026: §4.6). As we will see in §9.5, not all learning systems for constraint-based grammar come with a comparable guarantee of success, and this has been taken as a possible advantage of MaxEnt over similar theories.

Excursus: Slope as a means of effective weight-fitting

As noted in the main text, search algorithms for MaxEnt generally climb the hill of log likelihood in a series of uphill leaps (see Hayes and Wilson 2008: Figure 2 for a visual example). Just what direction to leap and how far is part of the art of algorithm design. To make the leaps effective, it is useful to know the *slope* of the hill at any one point, by which we mean both the uphill direction and the steepness. Calculus gives a specific meaning for slope: it is the vector of partial derivatives of the objective function against the constraint weights (the **gradient**).

As it turns out, the slope of the MaxEnt log likelihood at any point P can be expressed by a very simple formula: it is the number of Observed violations of the constraint in the data, minus the number of Expected violations (O − E). The latter can be calculated from the predicted probabilities of all candidates using the set of weights defined at P.

Many search algorithms, including the one used in the Excel Solver, use the slope to calculate what would be an effective size and direction for each uphill leap. Note that at the top of the hill, the slope is zero, and so O = E for all constraints. In other words, every constraint is violated in the aggregate exactly as often as the optimized grammar says it should be.

For more details on the use of slope in weight-fitting, including a proof of the "O – E" theorem, see Jurafsky and Martin (2026: Chapter 5).

4.4.3 Search spaces with no unique maximum

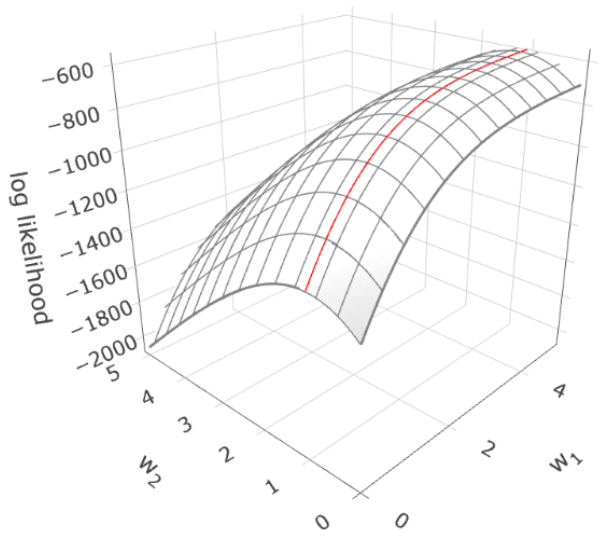
The discussion above is somewhat idealized: while the MaxEnt search space is convex in the mathematical sense, it is not always a pretty dome; and the exceptions relate to two scenarios we have already discussed. Consider first the case where one of the best-fit weights has an infinite value (§2.9). This can be intuitively understood if we draw the hill for this scenario, based on the tableaux in (73). These tableaux are just like (72), except that Input_2 yields exclusively $Cand_1$ as its output, necessitating an idealized weight of infinity for $\mathbb{C}_2$.

(73) *Toy grammar to illustrate weight computation with infinite ideal weights*

		obs.	$\mathbb{C}_1$ $\mathbf{w_1}$=1	$\mathbb{C}_2$ $\mathbf{w_2}$=∞	$\mathcal{H}$	eH	Z	P	ln p	LL
Input_1	$Cand_1$	731			0	1	1.368	0.731	−0.313	−582.262
	$Cand_2$	269	1		1	0.368		0.269	−1.313	
Input_2	$Cand_1$	1000			0	1	1	1	0	
	$Cand_2$	0		1	∞	~0		~0	−∞	

The corresponding search-space hill is given in Figure 22.

Figure 22: Hill of search for a two-constraint grammar, one weight infinite



With regard to weight $\mathbf{w_1}$, the hill is the same as in Figure 21, since for any value of $\mathbf{w_2}$, the highest log likelihood will fall where $\mathbf{w_1} = 1$. But for weight $\mathbf{w_2}$, the hill forms a ridge, extending off to infinity, with ever-decreasing slope along the line of the ridge, where log likelihood asymptotes to its maximum value of −1164.5 at (1, ∞). The search space is still convex in the mathematical sense, but it never attains a single best-fit value.

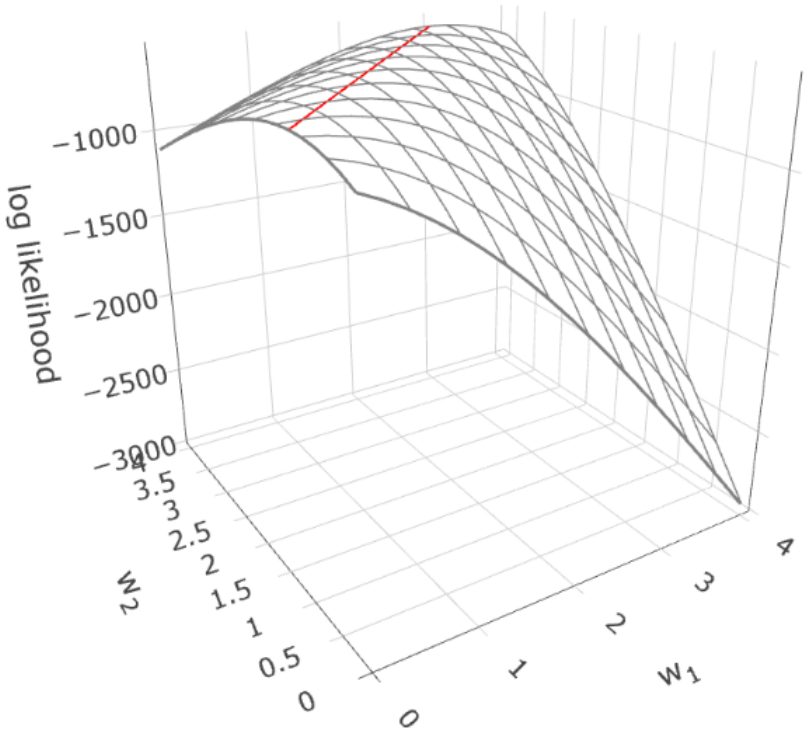
A second case not resembling a simple dome arises from a point we made earlier in §4.3: often there is not one unique setting of the weights that achieves maximum likelihood, because what actually matters is the sum or difference of the two weights. For the case of constant difference, a representative tableau is given in (74).

(74) *A grammar where only the difference between weights matters*

		obs.	$\mathbb{C}_1$ $\mathbf{w_1} = 1.5$	$\mathbb{C}_2$ $\mathbf{w_2} = 0.5$	$\mathcal{H}$	eH	Z	p	ln p	LL
Input_1	$Cand_1$	731		1	0.5	0.606	0.830	0.731	−0.313	−582.2
	$Cand_2$	269	1		1.5	0.223		0.269	−1.313	

Tableau (74) is similar to (61) above; a perfect fit will be achieved as long as weight $\mathbf{w_1}$ is precisely 1 higher than $\mathbf{w_2}$. When we visualize the search space (Figure 23), we see a flat ridge, marked in red, corresponding to the set of maxima where $\mathbf{w_2} = \mathbf{w_1} + 1$.

Figure 23: Hill of search for (74)



In addition to our values of (1.5, 0.5), the infinite ridge also includes the points (1, 0), (2, 1), (3, 2), (4, 3), and so on. Any position along the ridge yields the same, optimized, log likelihood. An adequate search algorithm can be trusted to climb up to the ridge, but where along the ridge it ends up will depend on the starting point and particularities of the algorithm.

4.5 An introduction to priors

In MaxEnt, weight fitting is often carried out with the use of a **prior**. In Chapter 7, we discuss the use of priors as a theoretical tool; here we simply introduce the concept, focusing on the problem raised by infinite weights. Let us return to example (73), where the best-fit weight of the second constraint is infinite. We repeat those tableaux here, in (75).

(75) *Tableaux from (73) (repeated)*

		obs.	$\mathbb{C}_1$ **w_1=1**	$\mathbb{C}_2$ **w_2=∞**	$\mathcal{H}$	eH	Z	P	ln p	LL
Input$_1$	*Cand$_1$*	731			0	1	1.368	0.731	−0.313	−582.262
	Cand$_2$	269	1		1	0.368		0.269	−1.313	
Input$_2$	*Cand$_1$*	1000			0	1	1	1	0	
	Cand$_2$	0		1	∞	~0		~0	−∞	

Recall that the goal here is to assign to constraint $\mathbb{C}_2$ a weight that is high enough to approximate the 1000-0 frequency distribution, without crashing our weight-fitting software. Whatever procedure we follow, it should also be able to weight $\mathbb{C}_1$ in a way that matches the frequencies of Input_1 correctly.

In §2.10.1 we suggested the practice, when using the Excel Solver, of placing an upward limit of 50 on constraint weights, with precisely problems like this in mind. This suffices as a rough-and-ready strategy, but it is not ideal, since the limit of 50 is arbitrary. In some cases, we may indeed need weights to be higher than 50, while in other cases a much lower weight could achieve essentially the same results. What we really want is a way to tell the Solver, "make these weights as high as you need to get close to correct predictions for the data, but no higher".

A **prior** is a more reliable method for limiting constraint weights which works in the general case, though it takes more effort to implement. To define a prior, let us first introduce a simple notion: the **objective function**, which is any computed value that we maximize (or for some functions, minimize) for purposes of fitting the weights of a model, such as a MaxEnt model. So far, the objective function we have been using has been log likelihood. This objective function values nothing but accuracy. Here we will introduce a slightly different objective function that sacrifices a bit of model accuracy to keep constraint weights at a reasonable level.

We start with a very simple prior: we sum the constraint weights, as in (76). This sum is then subtracted from the log likelihood to form the objective function.

(76) *An objective function with a simple prior (sum of all weights)*

a. $\text{Prior} = \sum_{i \in constraints} w_i$

b. Objective function = log likelihood − prior

The inclusion of the prior disfavors high constraint weights. This is because it is subtracted from log likelihood, which is maximized, to form the objective function. Implementing it with the Solver, we obtain the weights $\mathbf{w_1} = 0.995$, $\mathbf{w_2} = 6.907$. These are lower than what we fitted without the use of a prior, which were $\mathbf{w_1} = 1.000$, $\mathbf{w_2} = 18.094$. Notice that it is $\mathbf{w_2}$, whose ideal value is infinite, that responds more strongly to the prior.

The prior defined in (76) can be considered as implementing an *a priori* preference that weights be close to zero. The prior defines the point at which an increase in accuracy of model fit is assumed to not be "worth" a continued increase in constraint weights.

4.5.1 The Gaussian prior

The prior given in (76) is meant as a simple illustration the concept. A more commonly used prior is the **Gaussian** or **L2** prior, given in (77) below.

(77) *Calculating a Gaussian prior*

a. Square every constraint weight.
b. Sum the values obtained in (a) for all constraints.

c. Divide this sum by a user-chosen value. For reasons to be given, this value is expressed as $\mathbf{2\sigma^2}$.

Expressed as a formula, this is:

$$\sum_i \frac{w_i^2}{2\sigma^2}$$

As before, the prior is subtracted from the log likelihood to obtain the objective function.

(78) *Objective function incorporating a Gaussian prior*

$$\text{Objective} = \log \text{likelihood} - \sum_i \frac{w_i^2}{2\sigma^2}$$

Excursus: Terminology and detailed formation of the Gaussian prior

It is natural to wonder why the prior in (78) is called "Gaussian"; and also why the formula includes a non-obvious scaling method (why $2\sigma^2$, when just σ should suffice?). The two questions turn out to have the same answer. The MaxEnt objective function is normally defined in the domain of log likelihood. This is done only for computational convenience, and we can just as well switch it back to simple likelihood (§4.4.1); the function *exp*() switches it back. Doing this, we find that the contribution to log likelihood from the Gaussian prior, for any one constraint *i*, is $\exp(w_i^2/2\sigma^2)$. This expression is, in fact, the standard formula for a Gaussian distribution (the familiar "bell curve") of a random variable w_i, with mean zero and standard deviation σ. This is why the Gaussian prior is so-called, and it also gives a rationale for dividing by $2\sigma^2$: $\exp(w_i^2/2\sigma^2)$ establishes a *prior probability distribution* for the weight of constraint *i*, which has its maximum at zero, and a standard deviation of σ.

In Chapter 7, we will find it useful to work with more complex priors that give each constraint its own preferred weight, designated as μ_i. For these priors, the probability distribution becomes as in (79):

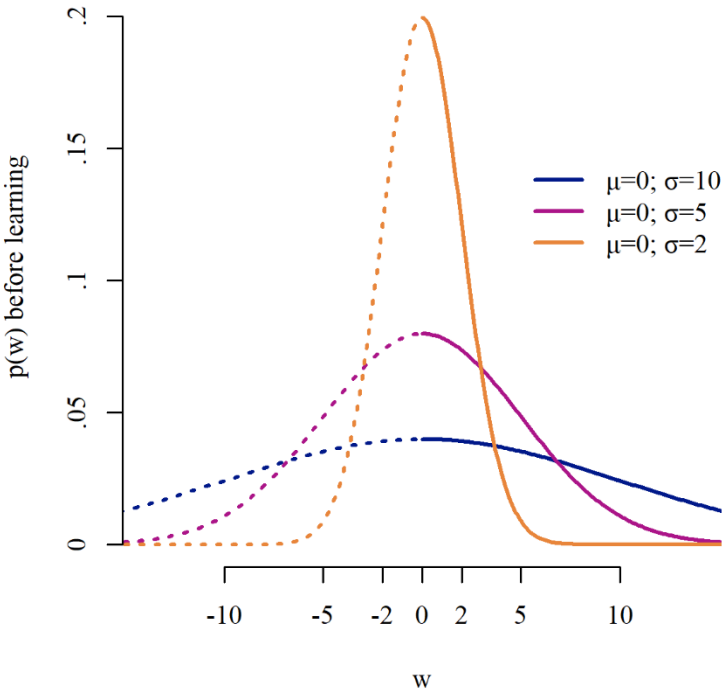
(79) *Gaussian prior for assigning preferred weights to individual constraints*

$$\sum_i \frac{(w_i - \mu_i)^2}{2\sigma^2}$$

where μ_i is the preferred weight of the *i*th constraint

The parameter μ expresses the mean of the Gaussian curve, and σ expresses its standard deviation. Three different Gaussian priors are illustrated for a one-constraint grammar in Figure 24; these priors all prefer the weight of the constraint to fall at or near μ, which is always 0, with strictness governed by σ, which is assigned the values 2, 5, and 10. As before, dotted lines are used for negative weights.

Figure 24: Three Gaussian priors



Note that for smaller values of σ, the main body of the Gaussian becomes narrower, expressing a stronger a priori preference for a weight close to 0.

We turn now to the implementation of a Gaussian prior in Excel. The spreadsheet **AvoidingInfiniteWeights.xlsx** is illustrated in (80). The formulas (rather than values) of the crucial cells are shown in `Courier`. The weights given in cells D2-E2 are in fact the weights that are delivered by the Solver when we set the weights under this arrangement.

(80) *A spreadsheet illustrating the use of a Gaussian prior*

	A	B	C	D	E	F	G	H	I	J	K
1				C1	C2						
2			Weight:	**0.995**	**5.25**	H	eH	Z	p	ln p	Log Lik.
3			Squared:	=D2^2	=E2^2						
4	Input$_1$	*Cand$_1$*	731			0	1	1.370	0.730	-0.3147	**-587.5**
5		*Cand$_2$*	269		1	0.995	0.370	1.370	0.270	-1.3094	
6	Input$_2$	*Cand$_1$*	1000			0	1	1.005	0.995	-0.0053	
7		*Cand$_2$*	0		1	5.245	0.005	1.005	0.005	-5.2504	
8	=SUM(D3:E3)	(77a,b): Sum the squares of the weights.									
9	1	We chose this value for sigma.									
10	=2*A9^2	Square and double.									
11	=A8/A10	(77c): Divide A8 by A10, yielding the Gaussian prior.									
12	=K4	Log likelihood (repeated)									
13	=A12-A11	Subtract A11 from A12, yielding the objective function (−601.8)									

As before, fitting with a prior results in lower weights. Here the best-fit weights are $\mathbf{w_1}$ = 0.995, $\mathbf{w_2}$ = 5.25; without a prior the best-fit weights are $\mathbf{w_1}$ = 1.000, $\mathbf{w_2}$ = infinite.[26] The lowering effect depends in general on the value of σ, with higher σ lowering the weights less.

In (80), we set our prior with a fairly low value of σ (that is, 1, shown in cell A9) in order to strongly favor low weights. With this low σ, we can see that priors do sacrifice some model accuracy; in this case *Cand$_1$* for Input1 was assigned a probability of 0.730 instead of 0.731; and *Cand$_2$* for Input2 was assigned 0.995 instead of 1.000. Had we used a larger value of σ, these discrepancies would have been smaller. At the other end of the scale, a tiny σ value like 0.0000001 would force the weights almost to zero, so the candidates would come out nearly equiprobable − a null grammar. Even a strong prior can be overcome by large amounts of data pointing to a non-zero weight for a given constraint; thus an appropriate value of σ depends on the size of the dataset.

The choice of σ also depends on the use that is being made of the prior. In the present context, all we are trying to do is avoid infinite weights, and for this purpose even a very weak prior (with a very high σ, like 100,000) will suffice. In Chapter 7, we will encounter a different use of priors in which the choice of σ is a theoretically informed decision.

For why we might favor a Gaussian prior vs. the simpler prior in (76), see the following excursus box.

[26] Infinity is indeed the best-fit weight for $\mathbf{w_2}$ without a prior, but the Solver actually comes up with 18.09. There is evidently something in its algorithm, not diagnosed by us, that sometimes prevents it from running lemming-like towards a fatal unbounded-weight crash. That the Solver algorithm does sometimes leap over the cliff was shown above under (36), where the use of a prior (or weight limit) is truly essential for weight-fitting to succeed.

Excursus: The L1 prior

The simple prior we discussed initially in (76) is an approximation of the so-called "L1" prior, in which the weights are not squared; in fact the Gaussian prior is sometimes called the "L2" prior precisely because weights are squared. The standard version of the L1 prior includes an absolute value sign in order to penalize weights that deviate far from the constraint's preferred weight, μ_i, in either direction. It also includes a multiplicative constant α, so the prior is can be made stronger by adopting a larger value for α.

(81) *The L1 Prior*

$$\alpha \times \sum_i |w_i - \mu_i|$$

When would we want to use an L1 vs. an L2 prior? The squaring used in the L2 prior has the effect that weight tends to be shared across constraints that overlap in their effects; L1 in contrast tends to concentrate weight on as few constraints as possible. The sharing tendency of the L2 prior is used to interesting effect by Martin (2011) to explain the so-called "leakage" effect, where (for instance) constraints that are in principle applicable only word-internally also end up having modest phrasal effects.

4.5.2 Priors and replicability

Adopting a modest prior helps the *replicability* of an analysis, because it generally forces a unique weighting, circumventing the difficulties shown in §4.5 above. To illustrate this, we re-ran our example (73) using different software, the `maxent.ot` R package of Mayer et al. (2024). The results, compared with Excel, are shown in (82).

(82) *Increasing replicability by using a prior*

	Weight of $\mathbb{C}_1$	*Weight of* $\mathbb{C}_2$	*P(Input$_1$, Cand$_1$)*	*P(Input$_2$, Cand$_1$)*
No prior:				
Excel with Solver	0.9996978	20.302	0.730992000	1.00000000
`Maxent.ot`	0.9997019	18.094	0.730999970	0.99999999
Difference	**0.0000041**	**2.208**	**0.000007970**	**0.00000001**
Gaussian prior, σ = 1				
Excel with Solver	0.9946501	5.245190	0.730005430	0.99475483
`Maxent.ot`	0.9946498	5.245181	0.730005373	0.99475479
Difference	**0.0000003**	**0.000009**	**0.000000057**	**0.00000005**

Since under a prior, the two software packages are optimizing the exact same problem, their answers come out extremely similar, far more similar than without the prior (compare boldface

rows). Without the prior, it appears `maxent.ot`, in its default settings, tolerates a bit more imperfection in model fit and so stops with $\mathbf{w_1}$ at 18.094.[27]

We find the use of the prior here to be comforting in a different sense: the solution to the optimization problem is mathematically well-defined, even for tricky cases where the best-fit weight is infinity, or where many weight settings achieve the same likelihood. Without a prior, the solution a particular algorithm produces is unprincipled, in the sense that (in its details) it depends on arbitrary aspects of the weight-setting algorithm.

Excursus: Gaussian priors ensure dome-shaped search spaces

When we employ a Gaussian prior, the search spaces corresponding to Figures 22 and 23 above lose their anomalous character and appear instead as dome-shaped. The reason is that the surface corresponding to the Gaussian prior is itself a dome, specifically an inverted paraboloid. When this surface is added to the ridgeline shape of Figure 23, for example, it "deflattens" the objective function, providing a unique maximum. The slope of the paraboloid along either axis turns ever more sharply downward as we move outward from the point of origin. This means that in cases like Figure 22, no matter how weak the prior (how high σ is), it will eventually create a downward slope for the combined objective function, again yielding a unique maximum.

All this is in effect a geometric restatement of the point just made: Gaussian priors help make the choice of weights less arbitrary in scenarios when many weightings fit equally well or almost equally well.

4.6 Relationship of MaxEnt to multinomial regression and log-linear models

MaxEnt OT inherits both its name and its math from a class of statistical models, now generally referred to as **log-linear models** or as **multinomial logistic regression**. As the purpose of this section is to clarify the distinction between the two, we will here be using "MaxEnt OT" to refer to the linguistic model covered in this book, and "log-linear models" to refer to the class of statistical models. For training in the use of multinomial logistic regression, we refer the reader to Field (2026:ch. 19) and Jurafsky and Martin (2025:ch. 5).

The point at hand is that despite having identical math, MaxEnt OT and log-linear statistical models are very different in their intellectual goals. Statistical models are a general tool used to assist scientific inquiry; they are designed to assess a set of predictors for how much of the variation in a given data set they can account for. Because the goal of a statistical analysis is to understand a specific dataset as thoroughly as possible, it is most informative when it is created bespoke for the specific dataset. In contrast, MaxEnt OT aspires to model how human minds represent phonological generalizations. To be useful for this purpose, it must do more than model a single dataset; it must also be responsible to phonological theory and typology.

[27] Thanks to Connor Mayer for providing this explanation.

We note in passing that the close relationship between MaxEnt OT and log-linear models means that statistical methods developed for log-linear models can be adapted to MaxEnt OT models. Examples of this include the use of random effects to model lexical exceptionality (Zymet 2018; Hughto et al. 2019; Linzen et al. 2013) and lexical classes (Shih 2018), the use of priors to model learning biases (Wilson 2006 and Chapter 7 below); and the calculation of confidence intervals on weights (Storme 2026).

In the sections below, we discuss a few cases in which best practices for modeling in linguistics differ from best practices for statistical modeling. We give practical advice addressing these issues in §4.6.3.

4.6.1 Choosing predictors vs. choosing constraints

When fitting a log-linear model to data, predictors should be chosen so as to cover all factors that could possibly influence the data, including "nuisance" variables such as trial order, the side of the screen on which a button press is made, and so on. Only predictors which logically cannot influence the data can be safely left out of the analysis, such as number of syllables in a study with only two-syllable words.

Constraints in MaxEnt OT are chosen by different criteria: the constraint set embodies a phonological theory, which makes substantive predictions about what properties phonological systems can be sensitive to and hence about what phenomena can occur in the world's languages. A particularly effective research method in OT has been to compute the **factorial typology** of the constraint set; that is, the set of outcomes obtained under all possible constraint rankings; see e.g. Kager (1999, §1.7); and for empirical examples Gordon (2002) and Kaun (2004). It is possible to consider the typological implications of a constraint set in MaxEnt as well, but the structure of that analysis is somewhat different; see Anttila and Magri (2018). In addition to substantive claims about exactly what is in the universal constraint set, OT theorists also sometimes address limits on the formal character of the constraints; see e.g. Eisner (1997b) and McCarthy (2003).

Another difference is that most regression models rely on an assumption of independence: that predictors are not highly correlated with each other. For discussion of the problems that arise when a statistical analysis is carried out with correlated predictors, see Çetinkaya-Rundel & Hardin (2024, §25.2). In contrast, MaxEnt OT analysis often employs constraints whose violation profiles are correlated. For example, there may be good theoretical reasons to include in the same analysis both a constraint penalizing all voiced obstruents (*VOICEDOBS), and a constraint penalizing just the phone [g] (*g) even though they are strongly correlated (since every time *g is violated, so is *VOICEDOBS). Such choices are made for the reason given above, namely that constraints are part of a general theory.

4.6.2 Choice of candidate set

Part of the basic training given to OT analysts (see e.g. McCarthy 2008) is in the choice of a candidate set sufficient to provide a true test of the analysis. The constraint set and weights in (MaxEnt) OT are responsible not just for predicting probabilities over possible outcomes, but also for explaining why each input does *not* surface in any impossible ways. Beyond this, OT

analysts must also be skilled in selecting the appropriate *inputs* for the analysis; often an analysis can be shown to fail because it makes wrong predictions for inputs not considered.

In contrast, statistical models are designed to fit actually-observed data; for any given experiment, the impossibility of cases not tested is simply assumed in advance, as part of the experimental design (though in principle it could be checked in subsequent experiments).

4.6.3 Using both statistical models and MaxEnt models together in research

The purpose of a log-linear statistical model is not to be a theory of phonology, but rather to assess a set of predictors for how much of the variation in a given data set they can account for. In contrast, a MaxEnt OT model is a model of the grammar internalized by the native speaker, as diagnosed by their outputs, intuitions, and experimental behaviors. Thus, if a program of investigation includes multiple experiments or other probes, ideally they should all be modeled with a single MaxEnt grammar, though each is subjected to it own statistical analysis. Indeed, even if one is doing just one experiment, it may make sense to provide both a statistical analysis of the experimental result and a hypothesized MaxEnt grammar describing the speakers competence; for papers that have done this see Hayes et al. (2009), and Moore-Cantwell (2020).

To sum up this section, we have suggested that despite the identical math used, statistical analysis of experiments and MaxEnt modeling of grammars are distinct enterprises. A statistical analysis can check the influence of nuisance variables and can use ad hoc reformulations of the phonological principles to avoid correlation between predictors. A MaxEnt theoretical model can address all plausible inputs and candidates (including those not tested), can integrate the results of several experiments, and can attempt to make sense of the data in terms of a particular linguistic theory that covers what a constraint can be like. This both tests the theory and renders the research responsible to typology. When we subject experimental results to statistical testing, we are seeking reassurance that they can be used as reliable building blocks for the construction of a phonological model.

4.7 History and etymology of "MaxEnt"

The first linguists to use the math of MaxEnt for theoretical purposes were Rousseau and Sankoff (1978), where the math was used to implement a probabilistic version of the phonological theory of Chomsky and Halle (1968). In this "variable rules" approach, each rule comes with a little MaxEnt grammar with just two candidates, "Apply rule" and "Don't apply rule". The framework proved influential and inspired a large literature, notably work that employed Sankoff's Varbrul software.

A version of MaxEnt more recognizable from the perspective of this text appeared in the early work of Paul Smolensky (Smolensky 1986). Interestingly, Smolensky at the time had no particular interest in language, but was using "knowledge atoms" (for us, constraints) as ingredients of a general theory of cognition. His article includes recognizable versions of weights, Harmony (Smolensky's own term), eH, and Z. It is widely cited in cognitive science.

MaxEnt OT arose in the early years of this century when computer scientists noticed that mathematical tools that they were already using could be bolted onto OT to form an effective

stochastic theory. Among these scholars were Johnson (2002), Eisner (2000), Manning (2003), and most influentially, Goldwater and Johnson (2003), who gave the framework its name and applied it effectively to previously-studied examples.

In the Introduction to this book, we noted that we are discontented with the term "MaxEnt" but use it in obedience to tradition. In the present context, we observe that our "MaxEnt" is actually the combination of two things: a calculation system for deriving candidate probabilities from weights and violations, and a system for setting the best-fit weights. Only the second of these has any actual connection with maximizing entropy, and indeed this system is compatible with other stochastic frameworks as well (see §9.5 below). In reading discussions of MaxEnt in the literature, it is helpful to identify which of these two aspects of the approach the author is addressing. For detailed discussion of how the concept of maximum entropy (as used in information theory) is applicable to MaxEnt, see Jäger (2007).

Chapter 5: Research technique

MaxEnt is a tool for research, and in this chapter we cover three aspects of linguistics research that bear on MaxEnt: first, the use of graphing and statistical methods to assess a MaxEnt analysis (§5.1-§5.3); second, the choice and deployment of constraints in the development of an analysis (§5.4), and lastly a brief survey of the methods used to gather data that can be analyzed in MaxEnt (§5.5).

5.1 Assessment of analysis in MaxEnt

In traditional generative linguistics (e.g. starting with Chomsky 1957, 1965; Chomsky and Halle 1968), analyses are evaluated based on whether they *work* (cover the relevant set of empirical data) or *don't* (justifying revision or reconceptualization). When studying probabilistic phenomena, however, it is hard to draw a clear line. If in your data, a certain form occurs 82% of the time, but your MaxEnt analysis predicts that it should occur 80% of the time, is that an analysis that 'works'? What about if your analysis predicts 75%, or even 65%? Does it make a difference if your 82% number comes from a corpus of several million entries, or from a fieldwork dataset with just hundreds? In this chapter, we suggest several methods that can address such questions, which are traditionally included under the term **model fit**. We suggest an approach that is both quantitative (how good is the fit between the MaxEnt analysis and the data?), and qualitative (what mistakes does the MaxEnt analysis make? What forms is it good at, or bad at?).

In the first sections of this chapter, we cover some basic methods that can help answer questions like the above: first graphing and visualization (§5.2), then statistical testing (§5.3).

5.2 Graphing and visualization

Although formal statistical methods are a gold standard for assessing models, experts also emphasize (e.g. Baayen, 2008: p. x) the need for visualizing and sorting data in a way that engages our intuition and qualitative thinking. Beyond this, there are instances where graphs can reveal essential patterns that are outright missed by standard statistical analysis; two famous instances are "Anscombe's quartet" (Anscombe 1973) and "Simpson's paradox" (Simpson 1951).[28] Lastly, when it comes to presentation of MaxEnt OT work, we suggest that audiences will generally appreciate the chance to apprehend the pattern visually before the statistical analysis is presented to them. Data visualization is of course an extensive field, for which introductory texts include Tufte (2001) and Cairo (2013).

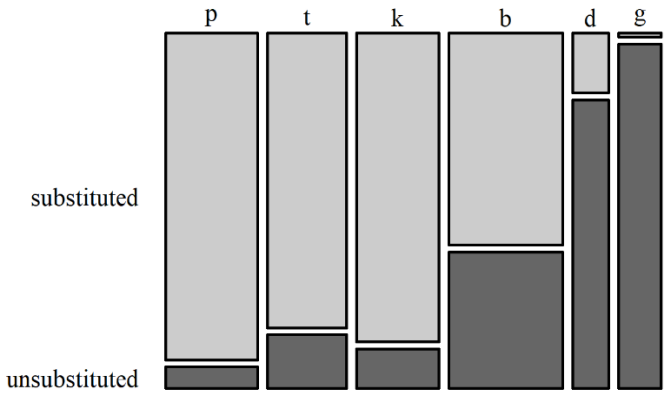
5.2.1 Mosaic plots

Before moving on to discuss the visualization of model fit specifically, we will mention here a type of plot which can be particularly useful to convey the type of data patterns used in MaxEnt

[28] Simpson (1951) makes its point without using graphs, but a helpful visual example of a Simpson-case can be found at this site: https://nastengraph.medium.com/why-you-should-always-visualize-your-data-first-18b8f432bc14, from Anastasiya Kuznetsova.

analysis: the **mosaic plot**. A mosaic plot is a bar graph in which the width of the bars indicates the number of data points contributing to each bar. For example, Figure 25 is a mosaic plot corresponding to the simple bar plot given in Figure 13 to illustrate the data pattern of Tagalog nasal substitution. In Figure 25, the width of each bar conveys the number of forms contributing to each observed substitution rate. The bar for 'b' is thus wider than the others, reflecting the greater number of 'b' forms in the corpus.

Figure 25: A mosaic plot version of Figure 13

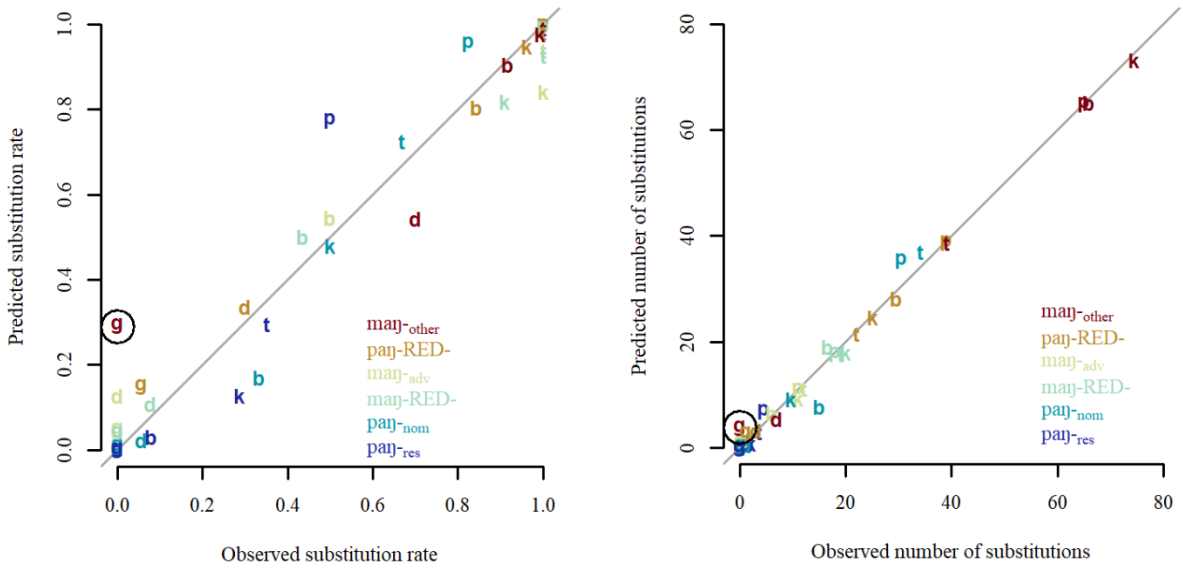


5.2.2 Visualizing model fit using scatterplots

For the purpose of comparing the predictions of a MaxEnt model to the data, we have already suggested using scatterplots, comparing the predicted frequencies of rival candidates with their attested frequencies. To make such scatterplots, we must first resolve a minor issue of scaling. The numbers to be plotted will consist of model probabilities (predictions) and empirical counts (observations). To render them comparable, we must either divide the empirical counts by total counts, converting them to estimated probabilities, as we did for K5 in Figure 4, p. 36; or we multiply the model probabilities by total counts, converting them to predicted counts. For the Tagalog substitution rates modeled in Chapter 2, we show the outcome of both methods compared below in Figure 26.

Figure 26. Two types of scatterplot: fractions vs. counts

a. Observations converted to fractions *b. Model probabilities converted to counts*



In this case, the analysis looks much cleaner using counts rather than probabilities. The reason is that some data types are represented by many tokens, and some by very few in the original data. Observed probabilities based on just a few data points are highly vulnerable to random variation. For instance, the worst-looking outlier point in Figure 26a, circled, is the one at (0, 0.29); measured in terms of counts, this turns out to be a misprediction of just 3.8 cases where zero were observed. On the scale of Figure 26b, this difference appears as a rather modest error — one that could easily be due to sampling error in a data set of this size. More generally, modeling mismatches that likely arise from small numbers will get crowded into the lower left corner of scatterplots like Fig. 26b, even when they loom large in scatterplots like Fig. 26a.

Here are some normative practices for creating scatterplots. First, if the modeling involves just two plausible candidates per input, it makes sense to plot just one of them, as we have done for Tagalog. The remaining points are redundant and would result in a more cluttered scatterplot. Second, it is sensible to scale the axes to be of equal length, since comparable entities are being compared. Third, it is useful to add in the diagonal line $y = x$, indicating the locus where prediction is perfectly accurate; this assists the eye in spotting the data points that deviate most from model accuracy. Lastly, we recommend that observed values be placed on the horizontal axis, and predicted on the vertical axis. This allows for more intuitive evaluation of a point's deviation from the $y=x$ line (we can say that the point "needs to move up" or "needs to move down").[29]

An important use of scatterplots is in trying to improve a tentative analysis, perhaps by adding further constraints. When we look at a scatterplot, the eye's attention is drawn to the

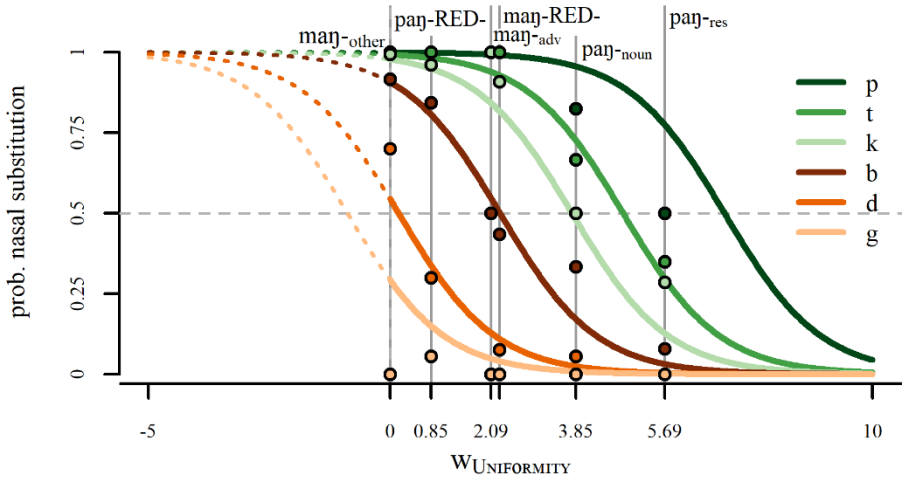
[29] Our advice goes against the published counsel of Piñeiro et al. (2008); their advice is based on the interpretation of slope and intercept values in a linear regression, which is not applicable to probability spaces.

outlier points. These can have various interpretations, concerning which we give our own intuitive judgment. (a) When the observed value is zero and the predicted value substantially above zero, this is a serious failure: in traditional generativist terminology, this is *overgeneration*. (b) Likewise, when the predicted value is zero and the observed value well above zero, this is also a serious failure; this time of *undergeneration*. (c) Cases where the data point is well off the diagonal, but not close to zero on either axis, also indicate defects of analysis, though perhaps not as serious, and may conceivably be due to random factors rather than being systematic errors. Obviously, the inspection of outliers from this point of view will be assisted by having your software annotate the points with data labels, as we have done in Figure 26 above. Examining the outliers will often suggest new constraints, which are best chosen (§4.6.1) based on our knowledge of the data pattern, language typology, and previous theoretical work.

The search for outliers can also be done non-visually by adding columns to the analytic spreadsheet. In Columns U and V of **Tagalog_testing.xlsx\find_outliers**, we have given the value obtained by subtracting the predicted value (as probability, then as counts) from the observed value. We have autoformatted this column to show up in pink when the error is greater than 0.01 (overgeneration) or less than −0.01 (undergeneration). We also found the worst errors by using the Excel MAX() and MIN() functions.

Our sigmoid complexes (e.g. Figure 16) can be thought of as another type of scatterplot, one which illuminates the way the analysis derives the observed result. We note here that the choice of baseline constraints vs. perturbers is an analytic decision, not part of the grammar itself, so sometimes it may be fruitful to make a different selection about what constraints count as perturbers. For instance, in constructing Figure 16, we took the baseline to be the constraints governing stem-initial consonants and the perturbers to be the prefix-based UNIFORMITY constraints. But we could equally well have let the UNIFORMITY constraints form the baseline and let the constraints governing stem-initial consonants act as the perturbers; this would yield the shifted sigmoids in Figure 27.

Figure 27: Shifted sigmoids for Tagalog: UNIFORMITY *constraints as baseline*



It can be useful to plot both ways, since sometimes there will be points that stand out as outliers more in one graph than the other. For example, the overgeneration for [g], /maŋ-$_{other}$/ and for [p], /paŋ-$_{res}$/ are more visible in Figure 27 than in Figure 16.

5.3 Statistical assessment of MaxEnt grammars

Much of modern statistics can be classified as either **hypothesis testing** or **model selection** (Burnham and Anderson 2002:3). Since our goal is to find an appropriate grammatical analysis of language data, it is the latter approach that is primarily applicable for MaxEnt OT modeling.

5.3.1 Why correlation should not be used to assess model fit

We begin by disposing of one method of model evaluation that may seem obvious but may not actually be appropriate for use — the familiar **correlation coefficient** (*r*; see e.g. Field 2026:ch. 7), as applied to observed vs. predicted probabilities. There are several reasons not to use correlations in this context.

First, the requirements for using correlation in statistical analysis are not met (the *r* statistic assumes that the data are continuous and normally distributed).

Second, in linguistic analysis, we seek a model that matches the observed frequencies exactly — but it is possible for predictions to be perfectly correlated with the data without matching exactly. For example, a set of predictions that fall in the range 0.25 to 0.75 could correlate perfectly with a set of observed values that fall between 0 and 1; the correlation could be the maximum of 1, but the model would nevertheless be imperfect, making errors as large as 0.25. For a worked-out example of this type, see **PerfectCorrelationButBadSlope.xlsx**.

Lastly, a high correlation can conceal the effects of a bad analysis, in which the analyst has failed to notice a generalization. In **CorrelationFailsToDetectMissedGeneralization.xlsx** we give paired analyses for a simple system of vowel devoicing: vowels devoice finally, never in stressed syllables (which are always initial), otherwise only after voiceless consonants. Note that this is a nonstochastic pattern. The constraints needed, with fitted weights, are given in (83).

(83) *Toy example illustrating how correlation values can be deceptive*

	Constraint	*Characterization*	*Weight*
a.	*VOICELESS STRESSED	Penalize a voiceless vowel in initial syllable	100.0
b.	*VOICED FINAL	Penalize a voiced vowel in word-final position	55.0
c.	AGREE	Penalize a voiced vowel after voiceless consonant	44.9
d.	*VOICELESS VOWEL	Penalize voiceless vowels	22.6

(84) *Tableaux: full analysis*

	voicing:	**freq.**	*VOICELESS STRESS 100.0	*VOICED FINAL 52.0	AGREE 44.9	*VOICELESS VOWEL 22.6	*p*
/ˈpipipi/	+ + +			1	3		
	+ + −				2	1	
	+ − +			1	2	1	
	+ − −	1			1	2	~1
	− + +		1	1	2	1	
	− + −		1		1	2	
	− − +		1	1	1	2	
	− − −		1			3	
/ˈbibibi/	+ + +			1			
	+ + −	1				1	~1
	+ − +			1		1	
	+ − −					2	
	(etc.)		1	1		1	

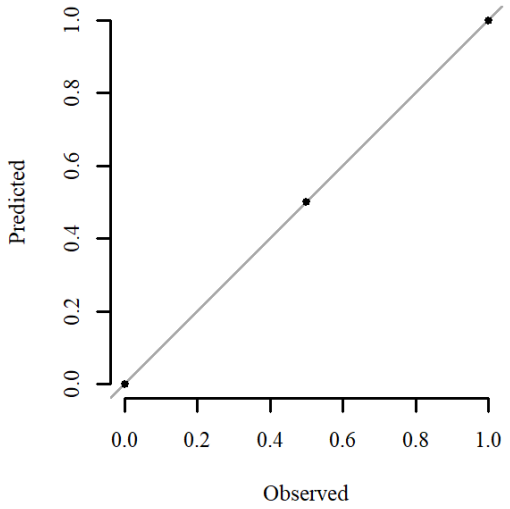
This analysis turns out to match the data almost perfectly, and its value for *r* accordingly is 1. This full analysis may be compared with a problematic alternative in which a careless linguist has failed to notice the post-voiceless generalization embodied in (83c), and therefore has not included AGREE. For this analysis, the tableaux entries are distinguished solely by syllable count, and forms like [ˈpipi̥bi̥] and [ˈpibibi̥] act as members of a lexically varying population, much as in Tagalog /pa**ŋ-p**ighaˈtiʔ/ → /pa**m**ighaˈtiʔ/ vs. /pa**ŋ-p**oˈʔok/ → [pa**mp**oˈʔok], given earlier in (40a). This miscontrual of the data is readily modeled in MaxEnt, as seen in (85).

(85) *A tableau for the grammar that is missing* AGREE

	voicing:	**freq.**	*VOICELESS STRESS 100.0	*VOICED FINAL 39.6	*VOICELESS VOWEL 0	*p*
'σ σ σ	+ + +			1		
	+ + −	0.5			1	0.5
	+ − +			1	1	
	+ − −	0.5			2	0.5
	− + +		1	1	1	
	− + −		1		2	
	− − +		1	1	2	
	− − −		1		3	

This analysis, while obviously wrong, also has a correlation of 1. In the scatterplot below, the data point at (0.5, 0.5) represents the second and fourth tableau rows of (85); namely the outputs that the bad analysis wrongly attributes to lexical variation.

Figure 28: A deceptive scatterplot: pooled predictable cases treated as random variation



This example, while intended to illustrate a point about correlation, also embodies a point that any linguist looking at variation needs to keep in mind: there always a risk of assuming that data reflect variation when in fact they reflect yet-undiscovered generalizations.

5.3.2 Assessing models with log likelihood

Having disposed of correlation, we now turn to a statistic that is more trustable than correlation for evaluating MaxEnt models: it is the very same log likelihood statistic that we have been using to set the constraint weights (§2.6.4). In the present context we note that the good analysis of the toy example above (given in (84)) yields a log likelihood of -1.46×10^{-9} (essentially zero, the best possible value); and the shortcomings of the bad analysis in (85) are reflected in its log likelihood, which is −5.54. For the illustrative example **PerfectCorrelationButBadSlope.xlsx**, given earlier, we calculate that the model that achieves a perfect match receives a much higher log likelihood (−440.4) than the model that that only achieves perfect correlation (−531.6).

Log likelihood can be used as a simple and intuitive way to compare different analyses of a dataset. Often a more appropriate comparison is to use not log likelihood per se, but a statistic derived from log likelihood, such as the Akaike Information Citerion, calculated as $2 \times LL - 2 \times n$, where n is the number of constraints. The AIC appropriately penalizes analyses that achieve high log likelihood through the use of many constraints. For extensive discussion of the use of the AIC in modeling, see Burnham and Anderson (2002).

5.3.3 Assessing the usefulness of particular constraints I: zero weights

We turn next to the statistical assessment of the individual constraints that comprise the analysis. In many cases, we first learn that a constraint is of no help in an analysis when we fit the weights and find that the best-fit weight of our constraint is zero. We discussed zero weights earlier in §4.3, covering cases where the zero arises from multiple equally-effective weightings, or reflects failure to locate appropriate inputs and candidates. Here we address zero weights more akin to what we studied in (63), where the zero is meaningful (if the weight is set any higher the

analysis becomes less accurate). In such cases, zero weight might be taken as a clear indication that the constraint does not belong in the grammar.[30]

When a constraint receives a best-fit weight of zero, it often emerges that the constraint is useless because its work is better done by others constraints already present in the grammar. For a case in which this has arisen in the literature, see Hayes and Minkova (in press, §6.4). We give a further simple example here, which requires a forward reference to the analysis of Turkish vowel harmony in Chapter 6. In brief, the constraint $*\begin{bmatrix}-\text{back}\\+\text{round}\end{bmatrix}\begin{bmatrix}+\text{back}\\-\text{round}\end{bmatrix}$ ("avoid front rounded followed by back rounded) is quite plausible in the Turkish context, but when we try adding it to the grammar we are proposing, it is fitted with a weight of zero —the constraints of the existing grammar already do all of the work of this constraint. For details, see Chapter 6 and the spreadsheet **AUselessConstraintForTurkish.xlsx**.

The usual reason that the weight for a useless constraint comes out *exactly* zero is that the analyst has chosen to forbid negative weights. For instance, if we re-weight the augmented Turkish constraint set with negative weights allowed, we find that the true best-fit weight of our useless constraint is a tiny negative number: −0.07.[31] The random fluctuations in almost any real data set mean that non-meaningful constraints get fitted with a weight of not exactly zero, but close to it; these get truncated to zero when negative.

5.3.4 Assessing the usefulness of particular constraints II: the Likelihood Ratio Test

Even among constraints that receive positive weights, skepticism is often still appropriate. A constraint can have a non-zero weight and yet serve no useful purpose in the analysis. Often such constraints are redundant, as discussed in the previous section; even high-weighted constraints can be redundant if they cover very few data. In this section, we present a statistical test, the **likelihood ratio test** (Wasserman 2004), that can serve to detect such cases.

The likelihood ratio test is based on model comparison: we consider a full model that includes the constraint ℂ we are interested in, and compare it with a minimally different model that omits ℂ. The minimal comparison serves as an assessment of the contribution that ℂ makes to the model fit. Specifically, we want to know whether the improvement in log likelihood yielded by including ℂ is "worth" the addition of the extra constraint to the analysis. The procedure is given in (86).

(86) *Testing constraint* ℂ *for significance with the Likelihood Ratio Test*

a. Calculate the log likelihood of the data under a MaxEnt grammar containing ℂ, with optimized weights.

[30] If the constraint set is taken to be universal, the question of removing constraints does not arise; in this section we will assume parochial constraints.

[31] We remind the reader of table (58), p. 69, intended to assist intuition about weight magnitudes. Violators of the useless constraint are about 7% more common than expected otherwise, a nonsignificant difference ($p = 0.59$ by the Likelihood ratio test, next section).

b. Form a second grammar by deleting ℂ from the first grammar, and reoptimize the weights.
c. Subtract the log likelihood of the second grammar from the log likelihood of the first, yielding a positive number *x*.
d. Multiply *x* by two: this value follows a chi square distribution.
e. Look up the *p*-value associated with 2*x* in the chi square distribution, with one degree of freedom (because one constraint was removed).

The test yields a *p* value, indicating the probability that the log likelihoods would differ as much as they do if the constraint ℂ was actually not useful in analyzing the data. Low values (a threshold can be assigned, such as 0.001) can be taken to mean that the observed data are much more likely under the analysis that contains ℂ than under the analysis that does not. High values indicate that the data is just as likely or nearly as likely without ℂ as with it. For discussion of the concept of *p*-values, and the perils associated with them, see Field (2026:ch. 3).

In **Tagalog_testing.xlsx**, we illustrate how this computation can be carried out for the model of Tagalog nasal substitution presented in Chapter 3. The table in (87) gives excerpts from the crucial columns. The boldface values in Column W were all computed separately and pasted in as values (so they wouldn't change on subsequent runs). On the right side, Column W is repeated, showing how each cell was calculated.

(87) *Application of the Likelihood Ratio Test to the Tagalog analysis*

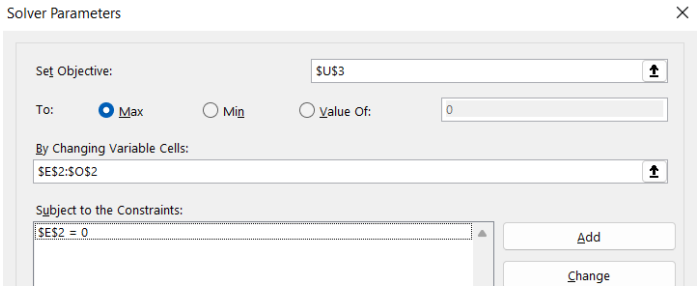
	W	X
3	**-248.61**	**Full model**
4		
5	**-296.13**	**Omit NASSUB**
6	47.52	difference
7	95.04	twice difference
8	1.87E-22	*p*
9		
10	**-428.67**	**Omit *NC**
11	180.06	difference
12	360.12	twice difference
13	2.65E-80	*p*
14		
15	**-273.62**	**Omit *[n**
16	25.01	difference
17	50.02	twice difference
18	1.52E-12	*p*

Column W as formulae
(calculated in U3, then pasted)
(calculated in U3, then pasted)
=W$3-W5
=2*W6
=CHIDIST(W7,1)
(calculated in U3, then pasted)
=W$3-W10
=2*W11
=CHIDIST(W12,1)
(calculated in U3, then pasted)
=W$3-W15
=2*W16
=CHIDIST(W17,1)

Cell W3 gives the log likelihood obtained when all 11 constraints of the full analysis are employed, as calculated with the Solver; this is about −248.61. W5 gives the log likelihood obtained when we remove NASSUB from the grammar. This calculation is most easily done by using the Solver Parameters box to require that the weight of the tested constraint be zero; under

MaxEnt this is the same as removing it. The image in Figure 29 shows what this box looks like when we impose this condition (in the full spreadsheet, E2 is the cell containing the weight of NASSUB):

Figure 29: Solver Parameters box: zeroing out weight of NASSUB



The log likelihood value returned by this calculation, −296.13, was pasted into cell W5, as seen above in (87). Next, cell W6 of (87) performs the subtraction needed to assess the improvement in log likelihood created when NASSUB is included in the grammar, namely (−248.6) − (−296.1) = 47.52. Cell W7 doubles this value, obtaining 95.04. Lastly, cell W8 uses the CHIDIST() function to return a *p*-value, 1.87×10^{-22}. Performing the same procedure for all the constraints, we obtain the values in (88) (*ΔLL* denotes change in log likelihood).

(88) *Significance testing by Likelihood Ratio Test of the Tagalog nasal substitution analysis*

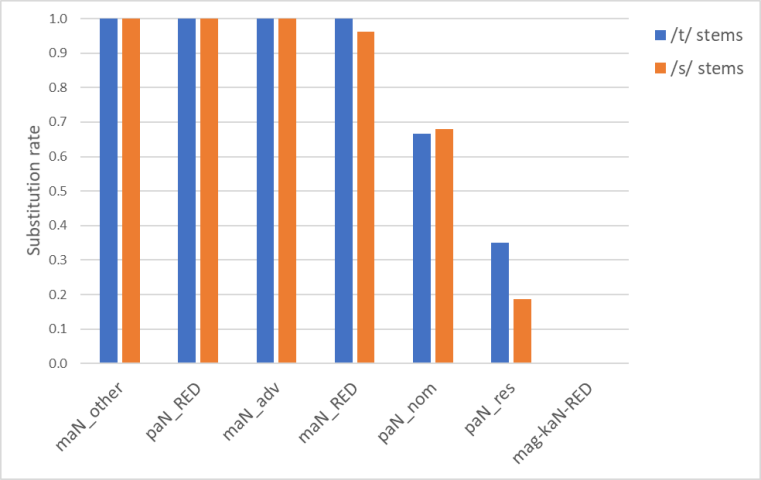
Constraint	*ΔLL*	*Weight*	*p*
NASSUB	47.52	2.28	1.87×10^{-22}
*NC	180.06	4.65	2.65×10^{-80}
*[n	25.01	2.10	1.52×10^{-12}
*[ŋ	54.04	3.16	2.59×10^{-25}
UNIF-/maŋ-$_{other}$/	0	0	1
UNIF-/paŋ-RED-/	2.24	0.85	0.034
UNIF-/maŋ-$_{adv}$/	9.95	2.09	8.18×10^{-06}
UNIF-/maŋ-RED-/	19.70	2.27	3.46×10^{-10}
UNIF-/paŋ-$_{noun}$/	75.63	3.85	9.15×10^{-35}
UNIF-/paŋ-$_{res}$/	73.51	5.69	7.77×10^{-34}
UNIF-/mag-kaŋ-RED-	257.60	15.15	4.66×10^{-114}

We remark briefly on these values. For UNIF-/maŋ-$_{other}$/, the *p* value of 1 is expected; its weight is zero for a mathematical reason (§4.3.2), and a zero-weighted constraint cannot make any difference to the log likelihood of the grammar. For UNIF-/paŋ-RED-/, we obtained a *p* value of 0.034. This makes only a weak case that the constraint is useful; were we to omit UNIF-/paŋ-RED-/, forms with the /paŋ-RED-/ prefix would be expected to behave just like forms with /maŋ-$_{other}$/ – which is not far from the truth. The other values look very impressive, but it is important to remember that *p*-values don't mean "probability that the analysis isn't true" – they only mean the probability that the weight could be as high as it is by mere accident. Most proposed analyses compete with alternative analyses – which might also receive impressive *p*-values.

5.3.5 An example of failing the Likelihood Ratio Test

Consider next an instance in which the likelihood ratio test strongly suggests that a constraint is not appropriate for inclusion in the grammar, even though it bears a positive weight. This arises when we expand the dataset for Tagalog to include stems starting with /s/, as in /ma**ŋ**-ˈ**s**ulat/ → [ma-**n**uˈlat] 'to write professionally'. In the original research by Zuraw, the /s/ stems are lumped together with the /t/ stems, for good reason. At the theoretical level, the constraint set employed does not distinguish /t/ and /s/ stems, as any voiceless alveolar obstruent would yield [n] as its mutated form; and the effects of underlying consonant are modeled at the surface level with the constraints *[n and *[ŋ. In addition, the data pattern for stems beginning with /t/ and /s/ is strikingly similar. Figure 30 shows the substitution rates for these two consonants with all seven of the prefixes studied here.

Figure 30: Substitution rates for /t/ and /s/ stems across seven Tagalog prefix types



The one case where the two columns for /t/ and /s/ diverge in any appreciable way is for the prefix /paŋ-$_{res}$/, where the data in Zuraw's corpus are sparse — only 10 /t/ stems and 8 /s/ stems. So the idea of employing a grammar that treats /t/ and /s/ stems the same is very appealing here.

Just for pedagogical purposes, however, we instead analyzed a synthesized data set, in which the number of instances of nasal substitution for /t/ is artificially boosted from the real Tagalog numbers by a factor of 1.5; this increases the chance that a /t/-/s/ difference might be statistically confirmed. We also needed to revise the analysis so that it makes a distinction beween /t/ and /s/ stems; specifically, a constraint that penalizes nasal substitution only when the source segments are /ŋ+s/, not /ŋ+t/. We will not dwell on what this constraint might be, and will simply name it literally as *[n] FROM /s/; it penalizes all UR-SR correspondences of the form [ŋs] ~ [n].

In the spreadsheet **TagalogWithSDistinct.xlsx** we introduce *[n] FROM /s/ into a grammar that retains all of the previous constraints. Setting the weights for this new grammar in Excel, we found that *[n] FROM /s/ received a very modest weight, 0.25. Including it improves the likelihood of the grammar from −292.749 to −292.501, a difference of only 0.247, which corresponds to a very unimpressive *p*-value in the likelihood ratio test of 0.48. Such a result might well be greeted with a sense of relief: it means we can drop from the grammar a rather

complex and ill-justified constraint, secure in the knowledge that whatever it did for us could very easily have been the result of the particular sample of Tagalog that we used.

5.3.6 Likelihood ratio testing in R

The practice of checking each constraint for significance with the Likelihood Ratio Test is one in which one can proceed more efficiently using the `maxent.ot` package in R, mentioned above in §2.11. In particular, you can set up a program loop to test the constraints one by one, start the program, and then go play with your cat. A sample script for this purpose can be found in the supplementary materials: **likelihood_ratio_test_loop.R**, which opens **TagalogFullForR.txt**.

5.3.7 Cross-validation

Large models, especially in computer science, are often checked by the method of **cross-validation**. The goal is to see if they are "overfitted," meaning that so many constraints have been included that the model is achieving an artificially inflated accuracy by fitting random fluctuations in the data. The method works as in (89).

(89) *N-fold cross validation (n = 5)*

a. Divide the data at random into five sets.
b. Omit Set 1, and refit the weights to the data of Sets 2-5.
c. Find the predictions of the refitted model for the data of Set 1.
d. Repeat steps b-c for Sets 2-5.
e. Combine the new predictions for the five subsets and calculate log likelihood, comparing with the original value.

The idea behind this scheme is that in no case is the prediction for a particular data point obtained by using that data point in the training data, so the data don't get to "fit themselves." It can be appropriate to report these cross-validated weights and their associated log likelihood, instead of results obtained by fitting the whole data set at once; this is done by Wilson and Gallagher (2018) in their MaxEnt analysis of Quechua phonotactics.

We suggest using cross-validation to check analyses that use large numbers of constraints, especially morpheme specific ones. However, when a model consists of just a few highly general constraints, as the example models in this book do, we follow the view of Goldwater and Johnson (2003: §4.3), who suggest that cross-validation is unnecessary in such cases: the generality of the constraints already assures that the model cannot be overfitted. As they put it: "the learning algorithm doesn't see words, it only sees patterns of violations," and these patterns are often rather few.

For extensive discussion of cross validation, including instructions on how to perform it in the `maxent.ot` R package, see Mayer et al. (2024).

5.4 On model-building

Generative grammarians have for decades, without using the term, been **model-builders**: they seek an abstract and formalized structure that plausibly represents the internalized knowledge (as diagnosed in various ways) of the speaker. In the context of MaxEnt, it seems helpful to consider issues of model-building as they arise more generally in the sciences. A useful text on model building which we have consulted is Burnham and Anderson (2002).

A key aspect of modeling is that we are seeking not so much accuracy as understanding. Just as a perfect map of a region would be an exact full-size copy of the region itself, and therefore completely useless as a map (Korzybski, 1933), models of linguistic phenomena are useful not because they are perfectly accurate, but because they represent an understanding of the data that has been simplified in some way so as to provide a meaningful answer to a question which the researcher raises. The goal of the analyst, therefore, is not necessarily a constraint set which predicts a perfect match for a given dataset, but rather a constraint set which answers (or comes as close as possible to answering) a specific research question.

5.4.1 Building a model from a set of potential constraints

A grammatical model submitted for publication often emerges from a gradual process, with iterated improvement in a series of models. Often, comparisons of a favored model with lesser-performing alternatives is informative. In this section we describe ways of building up an effective model, starting with a set of candidate constraints. The key element is using the likelihood ratio test to guide model construction.

A good starting point for building up a grammatical model is to consider a **constraint space** consisting of the constraints that might in principle be applicable to the problem at hand. This space can be relatively restricted and theoretically informed: containing for example all previously proposed constraints pertaining to a specific problem (as in Hayes et al. 2012); or it can be expansive and exploratory, as in Hayes and Wilson (2008), who consider all strings of feature matrices up to a length of three, or the "complete conjunctive constraint sets" employed in Moreton et al. (2017).

The selection of constraints from a constraint space must be informed by both the specific data at hand and the analyst's knowledge of theory. Algorithms exist (§4.4) to determine the best weights for a set of constraints and a data set, but no algorithm exists that can reliably determine which constraints should be considered in the first place. This is true not just for constraints in MaxEnt, but for parameters in any model. Because we are investigating questions whose answers are not yet known, we have only clues regarding what factors might ultimately be relevant, and we must always entertain the possibility that we might be missing an important one.

For MaxEnt analyses specifically, we recommend beginning with a thoughtful pre-theoretic understanding of the data at hand, particularly of what are the crucial generalizations and their possible causes. Having done this, the next step is to consult existing theory. Analytic work in linguistics normally adopts an explicit theoretical framework, which in the context of OT/MaxEnt will include specific hypotheses, grounded in typological study, about what a

constraint can be. Indeed, the broader task of theory-comparison can be made more rigorous by trying multiple frameworks and letting each one dictate its own constraint space.

On the other hand, it is hardly a signed and sealed research result that all languages can be analyzed with constraints are already part of some typologically-supported set, and in some cases we must include parochial constraints simply to come anywhere near describing the pattern of the language data. The role of general versus language-specific constraints is likely to remain an open research question for some time; see §7.6.3 below.

5.4.2 Selecting constraints from a constraint space

Once a constraint space has been defined, then model selection takes a more concrete form: what subset of the constraints in the constraint space should be selected? We seek a set of constraints which are all useful in analyzing the data — each of which improve the model's fit over a model without that constraint. A sensible concrete goal is for the grammar we choose (constraint set and weightings) to be one in which (a) every constraint tests as significant, by our chosen criterion; (b) excluding grammars ruled out by (a), log likelihood is maximized.

Finding such a grammar may be a daunting challenge when the constraint space is anything other than quite small. This is because of the problem of combinatorics. If the full number of constraints being considered is n, then the number of constraint *sets* we would consider deploying as a grammar is 2^n. This can be a very large number for values of n commonly used in linguistics, so finding the best-performing possible grammar in which every constraint tests as significant may be beyond our computational capacity.

The problem is indeed not reliably solvable, to our knowledge, but there are two heuristic methods (Walker, 2010:39) that can help to find something close to the best solution. These heuristics have different names in different sources.

Bottom-up method. Given a set of constraints $\{ C_1, C_2, \ldots C_n \}$, choose the one-constraint grammar $\{ C_i \}$ that, when optimally weighted, bears the highest log likelihood. Then keep adding new constraints, always picking the constraint that creates the greatest improvement in log likelihood. When there is no added constraint which creates a significant improvement in log likelihood, terminate and keep the already-chosen constraints as the final grammar.

Top-down method. Given constraints $\{ C_1, C_2, \ldots C_n \}$, form the grammar consisting of all the constraints, suitably weighted. Next, find the constraint that produces the smallest improvement in log likelihood and, *if* it fails to pass the likelihood ratio test, delete it, forming a smaller grammar. Continue until a grammar is obtained in which every constraint passes the likelihood ratio test.

In Hayes et al. (2012), both methods are applied to a constraint space with 88-constraints. Somewhat reassuringly, the grammars thus obtained are similar.

For discussion of potential pitfalls with the bottom-up and top-down methods see Field (2026:§8.10.1). Of the two methods, Field (p. 1039) recommends top-down for the avoidance of

"suppressor effects", in which a constraint that might have been helpful fails to make it in by accidents of the order of addition.

A recurring theme in selecting constraints from a constraint space is the important role of theory and previous research results. For instance, Field (p. 475) suggest that if one is using a bottom-up method, then "known predictors (from other research) into the model first". In the present context, this implies first adding the constraints that have a solid background in theory and typology, and only then the parochial constraints. Burnham and Anderson (2002) likewise lay heavy emphasis on favoring theoretically-motivated principles, and not letting a statistical search "fish" at random among a large set of dubious factors. In any case, both Walker and Field recommend reporting the full constraint space in journal publication.

5.5 On obtaining probabilistic data

We conclude the chapter with advice about obtaining the probabilistic data employed in MaxEnt analysis. The spreadsheets used in Chapters 2 and 3 were in fact the end point of extensive research and labor. Marshalling the data needed for a MaxEnt analysis is likely to involve considerable expertise in the language(s) of interest and in data collection, possibly including experimentation. This section offers a brief introduction to some of the available techniques, providing citations to sources from which in-depth training can be obtained.

5.5.1 Lexical corpora

We note up front that there exists a whole field of corpus linguistics, covered extensively for instance in Durand et al. (2014). We briefly cover a few types of corpora here. For further discussion related to corpora and phonology see Ernestus and Baayen (2011).

For studies of gradience in the lexicon, as in Chapter 3 above, it often suffices to analyze a lexical corpus (a list of words). Web search can help you with finding such corpora, often in the form of online dictionaries, such as the various Wiktionaries at www.wiktionary.org. In other cases, the only available list of words will be a hard-copy, printed dictionary. Hard-copy dictionaries can be reduced to digital form by optical character recognition ("OCR"), though further processing will be needed to isolate the entries from the definitions. For example, one such dictionary-based corpus, used for Shona phonotactics by Hayes and Wilson (2008), was scanned by the CBOLD project (http://www.cbold.ddl.cnrs.fr) from the print dictionary of Hannan (1959). An alternative to OCR, which for shorter dictionaries might be faster, is simply to use patience and hand-type every entry (e.g. Hayes and Wilson 2008, typing Dixon 1981), perhaps sampling every tenth (or n^{th}) entry.

The more ambitious dictionaries tend to include great numbers of words not known to the average native speaker, particularly in languages subject to rapid cultural change. Here, an effective control can be to ask/employ a native speaker to go through a dictionary and note which forms are familiar to him/her. This was the strategy, for instance, of the Turkish Electronic Living Lexicon (https://linguistics.berkeley.edu/TELL/TELLhome.html), used in Chapter 6 below for our analysis of vowel harmony in Turkish stems. The TELL speakers were familiar with about 17,500 of the 30,000 dictionary words that were submitted to their inspection.

The vowel harmony data for Hungarian in Hayes and Londe (2006) were also scrutinized individually by native speakers. These data involved 1130 stems, each notated for their harmonic class (back or front) by coauthor Londe, along with two other native speakers recruited by Dr. Londe. In the present day, many native speaker consultants are familiar with Excel and can patiently fill in appropriate blank columns with the forms used in their own speech.

When a dictionary provides no pronunciation entries, a phonologist will often face the need to convert an alphabetic orthography into phonemic transcription. For some orthographic systems, there exist downloadable conversion systems, or conversion can be done with a modest amount of user-created code. Suitable attention must be paid to digraphs (which create parsing problems; *ph* in *chophouse*) or neutralization of lost contrasts still spelled (Spanish *ll* and *y* are both /j/). In languages like English or French, with historical spelling, conversion by code is unreliable and can at best provide a rough draft for later hand-editing. An alternative to conversion is dictionary look-up, which however can be defeated by homographs (*lead*, *annex*); again, hand-editing will be required.

Often, lexical corpora are **lemmatized**, meaning that the inflected forms of a single stem are grouped together This permits, for instance, a study of the phonotactics that hold specifically of stems, as for Turkish in Chapter 6 below.

For many phonological applications, such as study of alternations, we need to study complete paradigms. Depending on the type of paradigm needed, dictionaries sometimes supply enough information to get started — listing out exceptional paradigm members with their stems, for example. Pedagogical materials for the language can also be helpful, such as the books of the *501* [*Language X*] *Verbs* series published by Simon and Schuster.

For some languages, though, the paradigm of any stem is so large that listing all the forms is impossible — for instance, Odden (1981:17) estimated that the verb paradigm of Shona has at least 16 trillion forms. In such cases we need to reduce the data into a form that we can cope with. For well-studied languages, we can turn to the method of **principal parts**, a technique widely used in language pedagogy. These are particular inflected forms that, if we know them, we can use them to predict all of the other inflected forms. For languages (e.g. Latin) where the principal parts can be shown to be reliable, we can focus our attention on just the principal parts, knowing that we will have in principle covered all of the data. For theoretical development of the notion of principal parts, see e.g. Finkel and Stump (2007) and Bonami and Boyé (2002).

For some purposes it may be helpful to *synthesize* full paradigms using software. This has the advantage of providing ample data, but runs the risk of missing irregular forms. For work with synthesized paradigms, see Hayes and Wilson (2008:413) and Cotterell et al. (2017).

5.5.2 Text and audio corpora

Text corpora can be gathered from books, newspapers, websites, and other written sources, or from transcripts of audio, such as transcripts of speeches or podcasts, or movie and television subtitles (e.g. Brysbaert and New, 2009). These can be useful to the phonologist for establishing usage frequencies of specific words and stems, but also for collecting data on phonological

processes that are reflected in the language's orthography. An example of such a pattern we have already seen is Tagalog nasal substitution, whose study was based on Zuraw's text corpus.

Audio corpora are of course more unwieldy, and tend to include fewer utterances than a text corpus. They also vary in terms of how closely they are transcribed — for example, the Buckeye Corpus (https://buckeyecorpus.osu.edu) contains phonetic transcriptions, but many corpora contain only orthographic transcriptions, as, for example, the TED talks used by Fuller (Chapter 2) for her Khalkha data. A helpful technique, used by Fuller, is to use the orthographic annotations to find instances of the phonological configuration of interest, and then consult the sound file to ascertain whether the process of interest applies.

Corpora for a wide variety of languages can be found online. Sites that host large collections of corpora can be a good place to start, such as the Endangered Languages Archive (https://elararchive.org/), the Hamburg Centre for Language Corpora (https://www.fdr.uni-hamburg.de/communities/hzsk/), or the Linguistic Data Consortium (https://www.ldc.upenn.edu/). Corpora vary widely in how they have been constructed, annotated, and cleaned (inspected for typos, non-word elements, uninterpreted numerals, and so on). It is important to carefully inspect a corpus's documentation to know these details.

Often, no corpus exists, especially for less well studied languages. A technique available in this case is **web scraping**, the process of extracting linguistic material by machine from Internet sources. Here, it is useful for one's scraping software to be able to identify the language of a web site, something which is in fact generally possible with a reasonable sample. For example, in Chapter 3 we noted that much of the data for Kie Zuraw's studies of Tagalog nasal substitution were gathered using a web-scraper that identified Tagalog websites by using a short list of known Tagalog words. Mayer (2021) improved his data quality (we suspect) for a study of Uyghur vowel harmony by limiting his search to the sites of two leading Uyghur-language newspapers; he found that his work was aided by already-programmed Uyghur scrapers which he could adapt to the task.

Text corpora can be reduced to lexical corpora by machine-collating the included words. The collation process will also yield frequency estimates. Creation of lexical corpora from text corpora may impose an accuracy cost, however, from combining the entries for homographs.

Lastly, corpora will vary in *style* depending on their source. If one can access bodies of archived speech, these will likely embody the vernacular style of the language, which may be quite different from what is used in books and newspapers. Jo (2024b) found quite different patterns of vowel harmony in Korean by examining data from various sources (historical Korean, newspapers, spontaneous speech, older speakers, younger speakers, wug tests); these data provide a clear diachronic picture of a phonological process gradually losing its productivity.

Once we are in the domain of style, we have ventured into the territory of sociolinguistics, a field which has long excelled in the quantitative study of corpus data; for references, see §2.2. For modeling based on style, see §9.6 below.

5.5.3 Field notes

For many languages, research by linguists is at an early stage. For such languages, work of the kind described here — or indeed, any sort of theoretical linguistics — cannot be done until a firm descriptive foundation has been achieved. This work will include establishing the system of contrasting phonemes and tonemes, compilation of a rudimentary dictionary, coverage of the system of inflectional contrasts, and coverage of the basics of word order and other syntactic structure. For fieldworkers, this foundational work is not only laborious but frequently poses its own intellectual puzzles and challenges, even before any issues of current theory get addressed.

Once a certain body of descriptive work is complete, it begins to be possible to study gradience. Suppose, for instance, that one were to compile all of the data files possessed by a single investigator. The data are likely to involve multiple areas of the grammar, but the words collected for diverse purposes may independently be usable for study of phonology. Thus, the data used by McPherson for her Tommo So reference grammar (McPherson 2013) and other works proved sufficient for a detailed quantitative study, including MaxEnt analysis, of the Tommo So vowel harmony system (McPherson and Hayes 2016).

5.5.4 Experimentation

The gathering of linguistic data from experimentation is a vast field, with considerable methodological content. Much of this content addresses gradience, for gradience is indeed the normal outcome for almost any experiment on language whose design permits it to be detected. For students interested in pursuing work in linguistic experimentation, we recommend coursework if possible, but some useful books to consult would include Solé et al. (2007), and Podesva and Sharma (2013). For the case of phonology, we can taxonomize the widely-used experiment types as follows.

An experiment that asks participants to provide inflected forms for novel stems is called a **wug test**, after an item from Berko's (1958) landmark study: the hypothetical noun *wug* [wʌg] is likely to have the plural *wugs* [wʌgz]. A wug test experiment tests for the productivity of both the morphological operation involved (here, English plural suffixation) and of whatever phonology is called forth by the relevant morphological operation (here, voicing agreement between stem and affix). The stimuli in a wug test can be either novel, linguist-created forms like *wug*, or existing stems that happen not to possess inflected forms of the kind being sought (e.g. for many English speakers the past participle of *stride*); the latter may be a more natural task for speakers (Kuo 2023a). Wug tests can also vary in whether they ask participants to volunteer a form, as in "Please tell me the past tense of *spling*," or to rate a form, as in "Please rate on a scale of 1 to 7 what you think of *splung* [splʌŋ] as the past tense of *spling*." Some publications on wug-testing that include implicit advice for practitioners include Moreton and Pertsova (2024) and Zuraw and Hayes (2017).

Another major type of test is the submission of novel forms to native speakers for ratings of phonotactic well-formedness, as in for example Greenberg and Jenkins (1964), Scholes (1966), Daland et al. (2011), and many others; see Chapter 6 below. While this kind of test is sometimes also called a wug test, we feel it is more precise to use the alternative term **blick test**, which derives from a thought-experiment in Chomsky and Halle (1965), who surmised (rightly, we

think), that English speakers would regard the nonce form *blick* [blɪk] as well-formed, comparable to existing *brick* [bɹɪk], but *bnick* *[bnɪk] as phonotactically impossible.

Both wug and blick tests can be carried out either on the real language spoken by the experimental participants, or on an imaginary, made-up language that they must learn on the spot; the latter are called Artificial Grammar Learning experiments (AGL). The two methods balance the advantages of naturalness/familiarity with the ability of the investigator to exercise full control over the linguistic pattern presented. AGL experiments often address issues of non-veridical learning discussed in §7.2 below. The literature on AGL is substantial; for review articles, see Finley (2017), Moreton and Pater (2012a, 2012b), and (Culbertson (2023).

5.5.5 Gathering data from multiple speakers

Probabilistic grammars, as with other work in generative grammar, are meant to model the internal linguistic system of a single speaker. When we are dealing with variation, though, variation in many datasets can arise both from within speakers and across speakers, and probably most often a mixture of both. Ideally, MaxEnt analyses should only analyze the variation that is within speakers. For this reason, it is important to check whether the variation in a dataset is between or within speakers.

This problem is a difficult one, but we offer a few suggestions. First, understanding the linguistic background of one's speakers is crucial. In finding consultants for fieldwork, it is often possible to ensure that the multiple speakers being studied speak closely similar varieties; experimenters often include a language background questionnaire as part of the participants' task and discard post hoc the responses not coming from speakers of the target dialect. Corpora may contain speaker labels along with demographic information. Once data have been gathered, it is helpful to scrutinize them in order to detect and systematize any differences between speakers that are found. Finally, in both testing and elicitation it is important to give each speaker multiple examples that are structurally parallel; to the extent that individual speakers vary in their responses to these parallel stimuli, we have evidence that we are seeing speaker-internal variation, not dialect variation.

Chapter 6: The study of well-formedness

6.1 A MaxEnt theory of well-formedness

Although Optimality Theory and its descendents have always focused primarily on mappings (input X is realized as output Y), this is probably not the most central formal problem in linguistics (though it is certainly very important for phonology). Rather, what linguists more often study is *inventories of well-formed entities*; e.g. the set of well-formed sentences (Chomsky 1957 *et seq.*), the set of well-formed words (Aronoff 1976 *et seq.*), or the set of well-formed verse lines (Halle and Keyser 1971 *et seq.*).

Within phonology, the area we have privileged here, the study of well-formed entities (particularly words) is often referred to as *phonotactics*. A responsible phonological description of a language characterizes what sequences of phonemes are legal. Chomsky and Halle (1965) give an overview, noting that native speakers can distinguish between existing words, such as *brick* [bɹɪk]; words that sound well-formed but happen not to exist, such as *blick* [blɪk]; and words that are ill-formed, such as **bnick* [bnɪk]. Experimental work on phonotactics, dating back to Greenberg and Jenkins (1964) and Scholes (1966), has always found that phonotactic judgments are not appropriately described in binary terms (as "well-formed" vs. "ill-formed"); rather, a broad class of forms receives intermediate ratings, along a wide range. [32] Gong and Zhang (2020) find evidence of gradience even in Mandarin, a language with a strikingly restrictive and "crystallized" phonotactic system. For further review, see Ernestus (2011:§3.2).

Here is an example, with a three-way contrast. The English initial cluster [tr] is perfectly ordinary, occurring in many words, and sounds fine when placed in a nonce word, e.g. [trɐp].[33] The initial cluster [dw] is exceedingly rare in English, albeit pronounceable: *dwell*, *dwindle*, *Duane*, and for some *dwarf*. Novel forms with [dw] (e.g. [dwɐp]) tend to sound odd. Yet [dw] is hardly as bad-sounding as a completely alien cluster like, say [rd], as in *[rdɐp]. This claim is supported, for instance, by the phonotactic ratings reported in Daland et al. (2011), where the highest ratings were obtained for clusters like [kl, fl, pl, fr, tr, gr], medium ratings for clusters like [ʃn, bw, vr, ʃm, dw, gw], and the lowest ratings for clusters like [rd, lm, dg, ln, lt, rg].

We describe here a MaxEnt treatment of gradient phonotactics developed in Hayes and Wilson (2008). This framework has a number of characteristics that distinguish it from OT and indeed from MaxEnt as otherwise practiced. The key idea is that gradient phonotactic well-formedness can be characterized as a *probability distribution over* GEN. As elsewhere, we set up the GEN function to consist of a set of all logically possible strings; which is in principle infinite

[32] Strikingly, Coleman and Pierrehumbert (1997) obtained gradient results even under an experimental condition that required participants to rate forms as an up-or-down choice; averaged across participants, the results came out quite similar to what was obtained when participants were permitted to rate on a scale.

[33] In our experience, when we think of a short, fully well-formed English nonce word, then look it up online, it turns out to have been adopted as a real word by a specialist speech community; this is true, for instance, of *trep* (and *blick*).

(see the Excursus Box). We assume an appropriate set of Markedness constraints, and weight them so as to assign a probability to every member of GEN.

Excursus: The infinite candidate set in MaxEnt

One of the boldest innovations of OT was to suppose that outputs are selected from an infinite candidate set GEN. Since human minds are obviously finite, this principle represents a strong idealization; for discussion see Prince and Smolensky (1993:215-216). These authors presciently noted that "in practice, computationalists have always proved resourceful," and indeed, subsequent research soon demonstrated that it is possible to set up implemented GEN functions that are infinite yet reliably searchable in finite time. The key idea is to represent GEN not as a list but as an abstract schema, often implemented with a finite state machine, as in Ellison (1994), Eisner (1997a) and Riggle (2009). Such work has not only provided reassurance on the tractability of infinite candidate sets, but also improved the reliability of hand analysis (Karttunen 2006). The approach of abstract schemata has been carried over to MaxEnt work, e.g. in Hayes and Wilson (2008).

In MaxEnt, infinite GEN raises an even more alarming prospect. We know that every candidate receives a positive value for eH; these values are summed (16c) to obtain the denominator Z. With infinite GEN, we have to worry about Z coming out infinite. This would in fact be catastrophic, since any number divided by infinity is zero, which means that all predicted probabilities would be zero, and the grammar would not distinguish between candidates.

Fortunately, it turns out that not all constraint sets give rise to this outcome. The key requirement is set forth by Daland (2015): roughly, there must be at least one constraint whose violations scale with the length of the candidates. *STRUC (Prince and Smolensky 1993:25) and DEP (McCarthy and Prince 1995:264) are constraints of this sort. Suitably weighted, such constraints guarantee that the eH values assigned to a set of candidates of ever-increasing length will follow a descending geometric series. If the weights are positive, then the series will sum to a finite number. For example, if $w_{*\text{STRUC}} = -\ln(0.9) \approx 0.109$, then each candidate in the series [a], [aa], [aaa], … will receive an eH that is 0.9 smaller than its predecessor. The sum that results ($0.9 + 0.9^2 + 0.9^3 + \ldots$) is finite (in fact, it equals 9). We suggest that the possibility of infinite Z is, at least logically speaking, a peril of which practioners should be aware, but in light of Daland's result, ordinary analytical practice probably remains safe.

This probability distribution *is* the phonotactic grammar's predictions for the language. Because GEN will necessarily be large, the probabilities assigned to possible words may be very small, but the difference between them can be very large proportionally. Specifically, probabilities that are vanishingly low correspond to ill-formedness; the highest values correspond to, essentially, perfection; and intermediate values are those for which the analysis predicts that the native speaker may feel ambivalent.

In such an arrangement, there is no role for an input form; only GEN matters. It follows that Faithfulness constraints, which govern the relation of inputs and outputs, likewise play no role.

Beyond this, the approach uses distinct tableaux structures for phonotactics and alternations: for phonotactics, candidates are grouped into a single tableau, whereas for alternations, there is a large set of tableaux, one for each distinct underlying representation. For discussion of the consequences of this bifurcated arrangement, see Hayes and Wilson (2008:§9.3), Jarosz (2011), Chong (2021), Do and Yeung (2021), Jo (2024), Jun et al. (2025), Xu (2026), and (Wang and Hayes to appear: §5.5).

The constraints of a Markedness-only phonotactic MaxEnt grammar can be weighted in the same way we covered above in §4.4, by using an algorithm (e.g. the Solver's) to set the weights to maximize the likelihood of the training data; here the actual words of the language's lexicon. In the training data, every existing word is given the value of one (in other words, we use type, not token frequency). Strings in GEN which do not correspond to existing words are given the value of zero. As before, the need to maximize the likelihood of observed data will minimize the likelihood of unobserved forms, resulting in low probabilities assigned to forms that are essentially impossible, insofar as this can be done under the constraint system assumed.

6.1.1 A toy example of phonotactic learning

These basic ideas are illustrated with the toy language L below, from the spreadsheet **ToyExampleForPhonotacticLearning.xlsx**. We let the vocabulary of L consist of 1066 randomly generated words, 26 of them taking the form CV, 1040 CVCV. Since all syllables are of the maximally unmarked CV type, both of the common constraints ONSET and *CODA[34] are inviolable in L. The vowels of L consist of [i, e, ɛ, a, ɔ, o, u], in equal numbers. We constructed the language to have five relatively frequent consonants, all obstruents, namely [p, t, c, k, s]; and an additional five infrequent voiced obstruents, [b, d, ɟ, g, z], such that overall, the voiceless obstruents outnumber the voiced by a factor of exactly 25 to 1; hence the voiced consonants can be regarded as somewhat aberrant phonotactically. The disparity between voiceless and voiced consonants will be modeled here with the constraint *VOICEDOBSTRUENT,[35] which should receive a substantial weight, but fall short of being all-powerful.

Potential words of L can be classified into three categories, based on their violations of ONSET, *CODA, and *VOICEDOBSTRUENT, as in (90):

(90) *Three categories of words in L*

Category	**Example words**
a. Typical words	[to], [pikɛ], [tusa]
b. Atypical words, with voiced obstruents	[zɛ], [pige], [tibɔ]
c. Impossible words, with codas or null onsets	*[a], *[mia], *[tap], *[ekto]

Having constructed the real words of our toy language, we needed to provide a useful approximation of the infinite GEN that the theory assumes. Since the words of L are at most four

[34] ONSET: Assign a violation for every syllable lacking an onset; *CODA: Assign a violation for every syllable with at least one coda consonant.

[35] *VOICEDOBSTRUENT: Assign a violation for every voiced obstruent.

phonemes long, we can approximate GEN with the collection of all 88,741 strings that can be formed from the 17 phonemes of L ([p, t, c, k, s, b, d, ɟ, g, z, i, e, ɛ, a, ɔ, o, u]) up to length four. This GEN, suitably annotated for which of its strings are actual words of L, is given in **ToyExampleForPhonotacticLearning.xlsx**. Four of the tableau lines from this file are shown in (91).

(91) *Sample tableau rows for the L simulation*

		*VCDOB **3.218**	*CODA **14.63**	ONSET **13.84**	$\mathcal{H}$	eH	Z	*p*
pikɛ	1				0	1	1361.5	0.000734
zɛ	1	1			3.22	0.040		0.000029
tap	0		1		14.6	4.43×10^{-7}		3.3×10^{-10}
a	0			1	13.8	9.78×10^{-7}		7.2×10^{-10}

We calculate the best-fit weights in the way already shown in previous chapters; they are as given in the second row of (91). From these weights, the spreadsheet calculates the probability of all members of GEN. Of these, the most important cases emerge as follows.

(92) *Probabilities assigned to representative words of L*

Word type	*Probability*	*Examples*
a. "Normal" words (violation free)	0.000734	[to], [pikɛ], [tusa]
b. Words with voiced obstruents	0.000029	[zɛ], [pige], [tibɔ]
c. Words violating *CODA	3.3×10^{-10}	*[tap]
d. Words violating ONSET	7.2×10^{-10}	*[a],*[ata], *[pia]

We see that our grammar for L basically achieved what we had hoped: impossible words, violating ONSET or *CODA, receive vanishingly low probability. The highest probability value, 0.000734, is awarded to violation-free words, and an intermediate probability, 0.000029, is given to words that violate *VOICEDOBSTRUENT. Hence the model is able to cover the three basic cases noted above, phonotactically (near-)perfect, intermediate, and bad. Larger-scale grammars with more constraints can in principle assign many different probability values across a broad range.

The grammar of (91) turns out to engage in *frequency matching* to the lexicon, in the same sense described earlier (§3.5) for phonological alternations. Recall that language L was deliberately designed so that voiceless obstruents would be exactly 25 times more frequent than voiceless ones. To a small margin of error, 25 is indeed the ratio of the probabilities that the grammar assigns to violation-free words like [ta], compared to words with one violation of *VOICEDOBSTRUENT like [da]; the actual ratio is 0.000734/0.000029 = 24.97. The same sort of frequency-matching is exhibited in the empirical simulations originally presented in Hayes and Wilson (2008): English onsets, Shona vowel harmony, and the full phonotactics of Wargamay.

Our grammar for L also displays *ganging*, in the sense of §4.1.2, such that (for example) words with two voiced obstruents receive a lower probability (1.18×10^{-6}) than words with just

one.[36] Ganging is likewise demonstrated thoughout Hayes and Wilson (2008), and in Breiss (2020), Breiss and Albright (2022), and Gong and Zhang (2020).

The key result of our simple L illustration is that the method of assigning a probability distribution over GEN using just Markedness constraints is capable in principle of serving as a model of gradient well-formedness, with a phonotactic grammar trained with a representative corpus of words from a language and assigning a range of probabilities to the members of GEN.

6.1.2 Phonotactic probability and frequency matching

As just shown, phonotactic grammars of the type under consideration are frequency-matching grammars, allocating probability to potential words according to how frequently their particular violation profile is attested in the language. This raises the empirical question of whether (as in other areas of language; §3.5), humans frequency-match in the domain of phonotactics. Such evidence would take the form of well-formedness intuitions. We have already mentioned the findings of Daland et al. (2011) above, where rare-but-existing clusters like [ʃn, bw, vr, ʃm, dw, gw] receive ratings lower than those assigned to perfectly-normal clusters, and higher than those assigned to unattested clusters. A few early studies that also found that people's phonotactic intuitions roughly track frequency were Coleman and Pierrehumbert (1997), Bailey and Hahn (2001), and Frisch et al. (2004). Subsequent work has found that acceptability tracks lexical frequency in onset clusters (for example Hayes and Wilson, 2008:382; Bárkányi, 2011), sonority and syllabification (Zellou et al. 2025), allophones in free variation (Sadeghi 2014), and consonant co-occurrence restrictions, including long-distance dependencies (Frisch et al., 2004; Koo and Oh, 2007; Jurgec and Schertz, 2020). A special case of interest for phonotactic frequency-matching is semi-predictable stress systems, such as is found English; such cases were discussed above in §3.5.

Just as with frequency-matching in alternations (§3.5.1), it is likely that there are many cases of deviation from a perfect match, often with the same causes. Notably, if the language has a lexicon that, for historical reasons, has an aberrant phonotactic pattern, then speakers' intuitions may reflect cross-linguistic markedness principles rather than just lexical frequency; for discussion see Iverson and Salmons (2005), Hayes and White (2013), Jarosz (2017), Jarosz and Rysling (2016), and Garcia (2019). The issue of how to incorporate such principles into MaxEnt modeling is taken up in Chapter 7.

Another issue is how the frequency-matching probabilities of the grammar match to speaker intuitions, as given in a Likert or similar scale; some sort of theory of scaling is needed. One approach (basically scaling participant responses to Harmony) is given in Boersma and Hayes (2001:§5.3); a similar idea is used by Hayes and Wilson (2008:§5.3.2), only with a scaling factor T ("temperature") that "unbends" curved distributions to achieve better model fit. In the domain of syntax, a theory relating probability to well-formedness judgments is put forth by (Lau et al. (2016). For how probability might be related to well-formedness judgments in syntax, see §8.6.6 below.

[36] The MaxEnt theory given here is only one contender in the general debate over how constraints gang in phonotactics; for discussion see {Citation}Breiss and Albright (2022) and Cabrera (2025).

6.1.3 The "no negative evidence" problem

A model of this kind also engages with some weighty and venerable issues in theoretical linguistics. Over the decades, linguists interested in learnability have put forth the view (Baker 1979, Dell 1981) that language learning is made harder by the **no negative evidence** problem. Human children are able to know what forms are legal, since they are presented with such forms, but there is said to be no straightforward way for them to know which forms are illegal; for general discussion see Bowerman (1988). The lack of negative evidence has sometimes been held to show that special mechanisms — especially, innate grammatical knowledge — are needed to make learning possible.

In the Optimality Theory research tradition, the no-negative evidence problem gave rise to a literature on "default rankings": certain types of constraints are posited to be given (as part of UG) an *a priori* ranking preference over others. For instance, it is posited that Markedness constraints are ranked a priori over Faithfulness constraints (see e.g. Smolensky 1996; for similar work in this vein see Gnanadesikan 2004, Hayes 2004, Prince and Tesar 2004, Becker and Tessier 2011, Jesney and Tessier 2011). When coupled with the Rich Base theory of phonotactics (Prince and Smolensky 1993:209), this principle forces a restrictive grammar, which becomes less restrictive only when the learning data force a different constraint ranking.

We feel that older discussions of the no negative evidence problem should be read with caution, because they were written before linguistics discovered how well certain algorithms, implemented as computational learning systems, are able to form restrictive grammars using only positive evidence. This is true of a number of learning models (see e.g. Regier and Gahl 2004, Hsu et al. 2013), of which we discuss only MaxEnt here. The key aspect of the MaxEnt system in this context is the use of a GEN function and learning by likelihood maximization. Specifically, when we maximize the likelihood of observed forms, we necessarily *minimize* the likelihood of the unobserved forms listed in GEN, since the probabilities sum to 1. This process can be seen in the schematic tableau of (91): the very high weights assigned to *CODA and ONSET represent (an approximation of) the maximum-likelihood solution: they help maximize likelihood by directing probability mass[37] away from the zero-frequency violating candidates like *[tap] and *[a] and toward the legal candidates. But there is no deep UG principle at work here; the calculations of the system are simply finding the most accurate grammar.

Note that to be able to fulfill the promise of no-negative-evidence learning, it is essential that MaxEnt be able to access a sensible system of constraints provided by linguistic theory. If to the contrary we allow utterly ad hoc constraints like DON'T BE OTHER THAN THIS LIST: {[[to], [pikɛ], [tusa], [zɛ], [pige], [tibɔ] …}, then the system will fail to generalize from the evidence and will be unable to distinguish (to use Chomsky and Halle's example) *blick* from **bnick.*

[37] "Probability mass" is used to designate the probability allocated across the full candidate set, summing to one.

6.2 Modeling example: Turkish vowel harmony in stems

We now apply MaxEnt modeling of phonotactics to a real-life example: the vowel harmony system of Turkish. In Turkish, suffixes agree in backness with stems. For example, the plural suffix appears as [-lɑr] or [-ler], depending on the stem's vowel: [utʃ-lɑr] 'tips' vs. [ev-ler] 'houses'. However, as in many languages with backness harmony, stems also overwhelmingly agree in backness internally. Just as Turkish uses [utʃ-lɑr], not *[utʃ-ler] as the plural of [utʃ], the relevant vowel sequences within Turkish stems also reflect this: for example the database we discuss in §6.2.1 below has 271 disyllabic stems with [u … ɑ], but only 44 with [u … e].

Often, the stem-internal system is not as strict as the affixal system; while the latter can be very strict indeed, stem harmony is sometimes more just a strong tendency, with substantial lexical exceptionality (Kiparsky 1973). A complete vowel harmony analysis should assess how harmony functions in both domains.

The weaker status of stem harmony is particularly noticeable in cosmopolitan languages, such as Finnish, Hungarian, and Turkish, which over time have acquired many disharmonic stems of foreign origin, while retaining their harmony system for affixes. For example, Turkish *muhit* 'neighborhood,' from Arabic, shows a [u … i] sequence; if this stem were to obey the principles of backness and rounding harmony that govern suffixes, it would have to appear not as [muhit] but as [muhut] (or, perhaps, [mihit]). Stem harmony often shows the sort of gradient patterning seen elsewhere in phonotactics: disharmonic stems may well be rather abundant, but they remain, intuitively, exceptional, for they are outnumbered by harmonic stems and are often felt, at least by the analyst, to be somewhat deviant.[38]

Stem harmony in Turkish was the subject of a thoughtful analysis in classical generative phonology by Clements and Sezer (1982). Working more than 40 years ago, these authors had no access to electronically-coded data nor to quantitative formal models, but they used careful scrutiny (probably, of a dictionary) and sensitive judgment to arrive at their conclusions.

Clements and Sezer first go through the basic generalizations of Turkish vowel harmony, familiar to many beginning students. Turkish has eight vowels that fit within a completely symmetrical featural arrangement, as in (93):

(93) *Turkish vowels*

	Front		*Back*	
	Unrounded	*Rounded*	*Unrounded*	*Rounded*
High	i	y	ɯ	u
Nonhigh	e	ø	ɑ	o

[38] Borrowing gives rise to sayable-but-deviant words in many languages and areas of phonology; for the case of English consonant clusters consider words like *Puerto Rico* (?[pw], from Spanish), *sphagnum* (?[sf], from Ancient Greek), and *tsunami* (?[ts], from Japanese).

The "classical" generalizations, which are well respected in suffixes, are as follows: (1) **Backness harmony**: vowels in the same word must agree in backness; (2) **Rounding limitation**: nonhigh vowels in non-initial syllables must be unrounded; (3) **Rounding harmony**: high vowels in non-initial syllables must agree with the preceding vowel in rounding. Together, these generalizations imply a sequencing chart like (94); we assume we are dealing with the first two vowels of a word; rows indicate the first vowel of a sequence, and columns the second, and "✓" indicates "legal."

(94) *Turkish vowel sequencing — idealized version*

		Second vowel							
		i	y	ɯ	u	e	(ø)[39]	ɑ	(o)
First vowel	i	✓				✓			
	y		✓			✓			
	ɯ			✓				✓	
	u				✓			✓	
	e	✓				✓			
	ø		✓			✓			
	ɑ			✓				✓	
	o				✓			✓	

As can be seen, this is a tight phonotactic system, in that any one initial vowel can be followed by just two of the eight possible vowels in the next syllable; and indeed, for this reason Turkish experts commonly group the suffixes into just two general classes, high and non-high. A high-voweled suffix like the copula *-dir* will have four allomorphs, [-dir], [-dyr], [-dɯr], and [-dur], depending on the backness and rounding of the preceding vowel. A non-high-voweled suffix like [-lɑr]/[-ler] can only have two vowels, owing to the Rounding Limitation, and they are determined by the backness of the preceding vowel.

What of stems? As suggested above, there are numerous non-harmonic stems, occurring typically in words borrowed from Persian, Arabic, and Western languages. Clements and Sezer (1982:222-225) offer a healthy supply of examples. Violating Backness Harmony are stems like *fiat* 'price', *hesap* 'calculation', and *haber* 'news'. Violating both Backness Harmony and Rounding Harmony are *bobin* 'spool', and *muhit* 'neighborhood'. Violating the Rounding Limitation is *tʃuroz* 'dried mackerel'.

Clements and Sezer suggest an interesting further generalization. Suppose we group the Turkish vowels into two sets, as follows. The first set consists of the familiar "unmarked" (in the pretheoretic sense) vowels /i, e, ɑ, o, u/[40] and the second set of the "marked" vowels /ɯ, y, ø/.

[39] Note that the columns for ø and o are empty - this is because as nonhigh vowels, the Rounding Limitation prevents them from occurring in the second syllable of a word.

[40] That these five vowels are unmarked is suggested by the fact that they greatly outnumber all others in the world's languages (Gordon 2016:51).

The phonological constraint the authors suggest is: *only unmarked vowels occur disharmonically*. This is not without precedent in the study of vowel harmony; notably, Suomi (1983) and Kaun (2004) have suggested the principle that "weak" triggers for a harmonizing feature more readily tend to induce spreading of that feature. Weak triggers are defined as vowels that manifest less clearly the harmonizing feature phonetically, and thus benefit most from the redundant cue obtained by spreading to neighboring vowels. F2 measurements in (Korkmaz and Boyacı (2018) indeed suggest that [ɯ] is the acoustically "poorest" back vowel of Turkish, and [ø] [y] the two poorest front vowels.[41]

Clements and Sezer suggest that exceptions to their more subtle marked-vowel principle are very rare in the lexicon, and that such exceptions tend to be repaired by some speakers (e.g. *mersørize* 'mercerized' is sometimes pronounced *merserize*). There is a further complication, however: according to Clements and Sezer the vowel sequences [i … y] and [y … i], though they include a marked vowel and are disharmonic, are nonetheless fairly normal, and should be considered as such.

Most strikingly, Clements and Sezer assert (p. 227) that, other than the marked-vowel generalization just stated, Vowel Harmony is *fully inactive within stems*; it should not be considered a synchronic generalization at all. In this section, we submit the Turkish vowel sequences to a form of analysis to which Clements and Sezer had no access, namely with an electronic data corpus and MaxEnt, in hopes of engaging more closely with the data pattern and providing a careful assessment of this interesting claim.

6.2.1 Turkish data

To start, we extracted the disyllabic stems from the Turkish Electronic Living Lexicon (Inkelas et al. 2000), discussed above in §5.5.1, and counted the vowel sequences they yielded, giving (95):

(95) *Turkish vowel sequencing — disyllabic stems of TELL*

		Second vowel							
		i	y	ɯ	u	e	ø	ɑ	o
First vowel	i	272	5	0	7	285	6	349	69
	y	58	131	1	32	182	2	64	7
	ɯ	11	2	177	0	11	0	149	3
	u	43	6	3	192	44	2	271	17
	e	589	18	1	97	492	5	345	42
	ø	1	70	0	1	131	4	8	2
	ɑ	546	15	517	219	279	16	936	130
	o	49	4	1	172	68	6	279	58

[41] Average measurements of F2 from their Table II: [o] 864 hz., [u] 923, [ɑ] 1135, [ɯ] 1507, [i] 2075, e [1795], [y] 1732, [ø] 1431.

Here, a white background marks canonically-legal sequences laid out in (94), a light grey background marks disharmonic sequences where one vowel is unmarked, and a dark grey background indicates the special "marked and disharmonic" sequences singled out by Clements and Sezer. We can see from the start that the overall characterization seems fairly sensible. The 16 canonically-legal cells have a mean count of 308.8, the 19 light-grey cells have a mean count of 90.4, and the 29 dark grey cells have a mean count of just 8.2.

A defect of such a simple form of data inspection is that it does not control for *overall* vowel frequencies. As can be seen in (96), which regroups the data of (95), the eight vowels of Turkish occur with substantially different frequencies.

(96) *Turkish vowel counts — disyllabic stems of TELL*

a	e	i	u	ɯ	o	y	ø
5059	3081	2562	1298	1053	965	728	258

Observe in particular that the marked vowels [ø], [y], and [ɯ] have the lowest, second-lowest, and fourth-lowest frequencies, which is a potential confound for Clements and Sezer's claim that they are less likely to occur in disharmonic sequences than the other vowels. If violating harmony is unusual, and being one of [y ɯ ø] is unusual, their combination would be very unusual even without a specific grammatical generalization targeting it.

6.2.2 MaxEnt analysis of the Turkish vowel sequencing

With these issues in mind, let us turn to the MaxEnt analysis. We seek to build a model which fits the relative frequencies in (95), and which can therefore evaluate Clements and Sezer's analysis using a large sample of Turkish words. Recall that Clements and Sezer believed that vowel harmony was active only in alternations.

In our simulation, we limited ourselves to a very simple GEN, namely the 64 logically possible vowel pairs that arise from combining the eight vowels of the Turkish inventory. The goal of the analysis is to match the probabilities assigned to each of the 64 pairs to the empirical disyllable frequencies given in (95).

For our Markedness constraints, we adopted versions that, glossing over various theoretical debates about vowel harmony, are fairly literal descriptions of the generalizations given by Clements and Sezer. Backness harmony can be treated straightforwardly with the constraint AGREE(back); it is violated by any sequence of consecutive vowels that disagree in their value for the feature [back]. For rounding harmony, we use the very similar AGREE(round). The constraint *ROLO, after Kaun (1995, 2000), implements the Rounding Restriction noted above; it penalizes the non-high rounded vowels {o, ø} when they occur in non-initial syllables.

In order to account for the pattern suggested by Clements and Sezer whereby disharmonic sequences are more acceptable when both of the vowels are from the unmarked class {a, e, i, o, u}, we adopt their constraint numbered (25), from p. 229 of their article. This constraint, which we will call AGREEIFMARKED, assigns violations for any vowel of the marked class{y, ø, ɯ} that disagrees with a neighboring vowel in backness or rounding; hence to {aø, ay, eɯ, eø, ey, ɯe,

ɯi, ɯo, ɯø, ɯu, ɯy, iɯ, iø, iy, øa, øe, oɯ, øɯ, øi, oø, øo, øu, oy, uɯ, uø, uy, yi, ya, ye, yɯ, yo, yu}. For our implementation of their proposal that [iy] and [yi] are exempted from this contraint, see below.

(97) *Constraints used in the MaxEnt analysis*

a.	AGREE(back)	Assign a violation for every pair of adjacent vowels which do not have the same specification for backness.
b.	AGREE(round)	Assign a violation for every pair of adjacent vowels which do not have the same specification for roundness.
c.	*ROLO	Assign a violation for every vowel that is both [+round] and [−high].
d.	AGREEIFMARKED	Assign a violation for every vowel in the set {y, ø, ɯ} which is adjacent to a vowel not sharing its specification for backness.

Above, we raise the possibility that any purported effect of AGREEIFMARKED might actually be an illusion, arising out of the combined effects of the simple AGREE constraints and the relative frequencies of unmarked vowels compared to marked vowels. In order to assess this possibility, we need a way to model the individual frequencies of each vowel. We adopt the strategy that is least elegant but forms the most effective statistical control, namely simply to add eight constraints, each one penalizing precisely one vowel: *i, *y, *ɯ, *u, *e, *ø, *ɑ, *o. The lack of elegance and generality of this vowel-by-vowel solution is permissible in this context, we believe, because it gives us the most stringent available test of the sequential vowel generalizations that are our real focus of interest.

We consider a series of four models, beginning with a primitive, eight-constraint model with just {*i, *y, *ɯ, *u, *e, *ø, *ɑ, *o}. This model, often called a "unigram" model, serves to control for vowel frequency. The next stage is to add to the unigram model the classical vowel harmony constraints AGREE(back), AGREE(round), and *ROLO, testing each with the Likelihood Ratio Test (§5.3.4). If each of these significantly improves our model fit, then we believe this calls into question Clements and Sezer's assertion that the classical principles are not involved in defining stem structure, but only in affix alternations. We next examine the effect of adding AGREEIFMARKED to the grammar, and evaluate the alternative hypothesis that the unigram constraints plus regular AGREE are sufficient to account for the rarity of disharmonic and marked stems. Finally, we will later add a further constraint, AGREE(y,i), which will serve to evaluate Clements and Sezer's claim that {iy, yi} are exempt from the ban on disharmonic sequences with a marked vowel.

The first four rows of the tableau for our unigram-only analysis are given in (98). This is extracted from the spreadsheet we used for the analysis, **TurkishVowelHarmonyInStems.xlsx**.

(98) *Sample partial tableau: phonotactic analysis of Turkish vowel sequencing*

sequence:	**freq.**	*a 0	*e 0.50	*ɯ 1.57	*i 0.68	*o 1.66	*ø 2.98	*u 1.36	*y 1.94
aa	936	2							
ae	279	1	1						
ai	546	1			1				
aɯ	517	1		1					
(60 more rows)	…								

As can be seen, the fitted weights align in inverse order to the frequencies of (96). We calculated weights with the Excel Solver, though in a case this simple we could have calculated it directly.[42] The full spreadsheet includes all 64 candidates, and constraints for all four of the models we mentioned above. The results of the modeling, including weights and Likelihood Ratio testing, are given in (99).

(99) *A series of MaxEnt models of Turkish vowel sequencing*

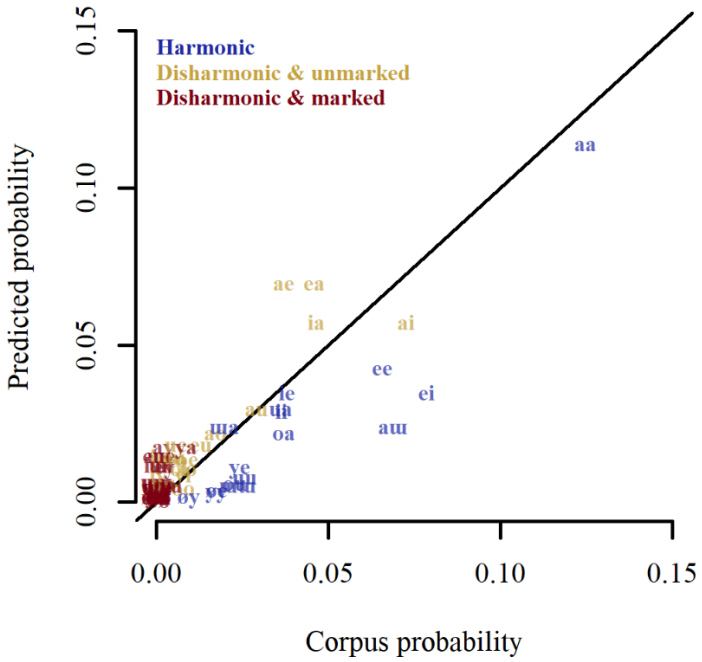
Model	*a	*e	*ɯ	*i	*o	*ø	*u	*y	AGREE(back)	AGREE(round)	ROLO	AGREEIFMARKED	AGREE(i, y)
a.	0.00	0.50	1.57	0.68	1.66	2.98	1.36	1.94					
b.	0.04	0.48	1.61	0.67	1.00	2.20	1.13	1.58	0.90	0.71	1.00		
								ΔLL: *p*:	1058.4 ≈ 0				
c.	0.00	0.36	1.16	0.55	0.99	1.60	1.10	0.97	0.73	0.35	0.98	1.38	
											ΔLL:	405.2	
											p:	$<10^{-170}$	
d.	0.00	0.36	1.16	0.54	1.00	1.62	1.11	0.96	0.74	0.34	0.98	1.35	0.33
												ΔLL:	3.1
												p:	0.012

Let us consider the performance of this series of grammars. To start, the unigram model (99a) faithfully reflects the vowel frequencies in the corpus — rarer vowels are penalized by constraints of higher weight. The model already helps distinguish the vowels of real Turkish words from random strings, as we can see this by the modest degree of adhesion of data points to the diagonal in the scatterplot below.

[42] The mapping from frequencies to weights works as follows: per (96), /a/ is the most frequent vowel at 5059 and its corresponding weight is zero. For every other vowel V, divide 5059 by the frequency of V, then take the log of this ratio to obtain the weight (Log Odds Principle, p.70).

Figure 31: Performance of the unigram model (99a) of Turkish vowel sequencing

a. Log likelihood: −26778.98
b. Scatterplot:



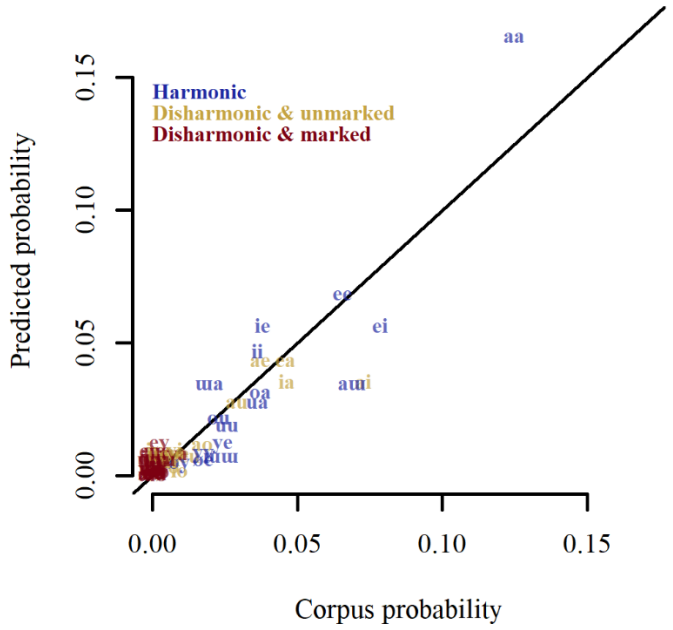
As can be seen, the model overpredicts the probability of exceptions to harmony with common vowels like [ae, ea, ia] (in beige), and underpredicts the probability of most harmony-obeying sequences (in blue).

Model (99b) embodies traditional linguistic wisdom on Turkish vowel sequencing, folding in AGREE(back), AGREE (round), and ROLO. We find that each of these three constraints passes the likelihood ratio test with a very low *p*-value.[43] It appears that, contra Clements and Sezer, the vowels of the stems in the TELL corpus do indeed obey the basic harmony principles to a degree far beyond what could ever arise by chance. Here are the basic results given as log likelihood and scatterplot:

[43] AGREE(back): ΔLL = 696.3, $p = 8 \times 10^{-305}$; AGREE (round): ΔLL = 260.2, $p = 4 \times 10^{-115}$; ROLO: ΔLL = 107.4, $p = 1 \times 10^{-48}$.

Figure 32: Performance of the classical model (99b) of Turkish vowel sequencing

a. Log likelihood: −25720.54
b. Scatterplot:

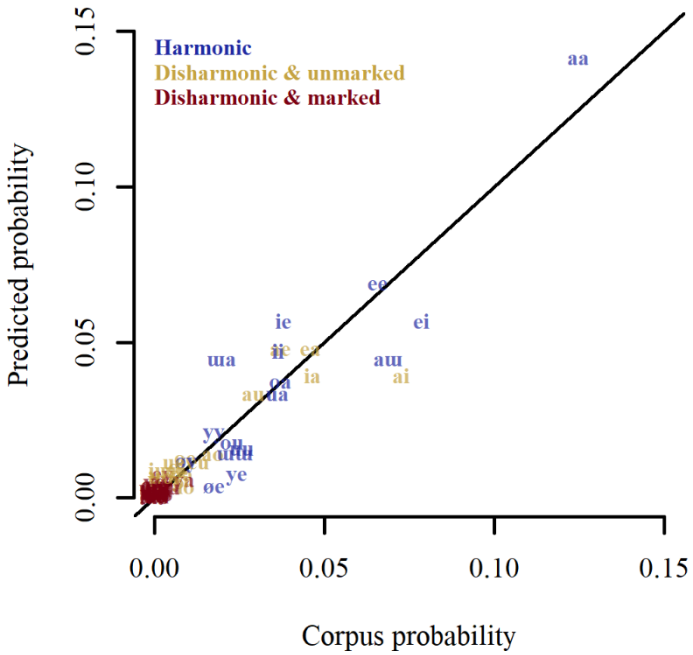


Note that the above/below asymmetry between harmonized and unharmonized sequence is here replaced by lesser, perhaps random variation.

Grammar (99c) adds the Clements/Sezer constraint AGREEIFMARKED, and we see that adding it to the grammar likewise yields a pattern very unlikely to be due to chance — hence support for Clements and Sezer's key idea.

Figure 33: Performance of the classical model (99c) of Turkish vowel sequencing plus AGREEIFMARKED

a. Log likelihood: −25315.34
b. Scatterplot



The improvement from adding this constraint is suggested by the significance value from the likelihood ratio test ($\Delta LL = 405.2$, $p < 10^{-170}$), and also, more subtly, from visual inspection of the scatterplot. The data points whose observed values are better predicted fall near the origin, since vowel sequences with marked vowels are less common than average.

We conclude with grammar (99d), which represents our effort to evaluate Clements and Sezer's decision to *exempt* the sequences involving {i, y} from the scope of AGREEIFMARKED. Some preliminary exploration we conducted suggested that, at least for the TELL corpus data, this exemption is *not* a good idea — to the contrary, there is a weak tendency for [i … y] and [y … i] to be avoided *more* than other disharmonic sequences.

To demonstrate this, we added to our model (99c) a constraint AGREE(i,y), that additionally *penalizes* [i … y] and [y … i], above and beyond their violation of AGREEIFMARKED. This constraint when fitted receives a positive weight, namely 0.33. The effect of the constraint is not at all strong ($\Delta LL = 3.1$, $p = 0.012$). But the key point is made by the fact that the weight is positive; the effect is reversed in direction to Clements and Sezer's original claim; so the TELL data provide no support for their claim.

Summing up, we suggest what this study implies as a evaluation of Clements and Sezer's important early work. First, their assertion that the traditional principles of Turkish vowel harmony are inactive in stems cannot be taken as correct, at least as a statistically valid

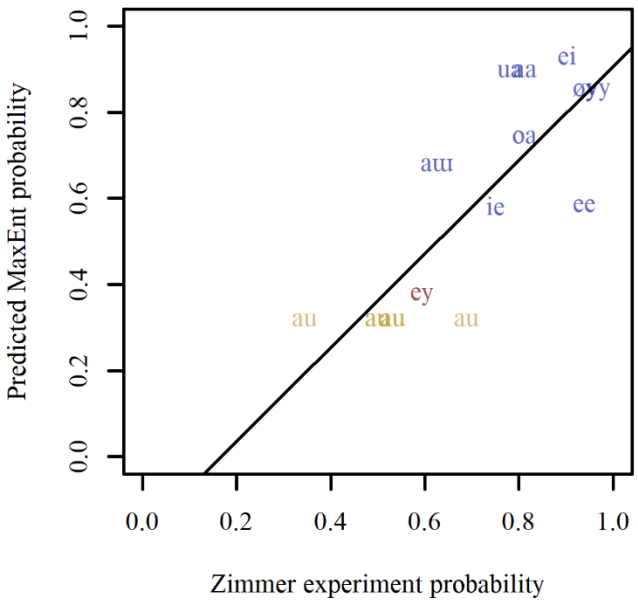
characterization of the disyllabic stem inventory of the lexicon; the effects of these traditional constraints are very strong indeed. Second, Clements and Sezer's guess that marked vowels are subject to stronger harmony constraints was an inspired one, which we find to be supported by the corpus data — except that they went on to complicate their grammar needlessly in providing a special exemption for {iy, yi} for which there turns out to be no corpus support.

6.2.3 Native speaker judgments

All this, of course, is subject to a large caveat: for scientific purposes we need to check the extent that these patterns are actually internalized by native speakers of Turkish. The data we need is a detailed ratings study of nonce stems with varying vowel sequences. Here, we briefly address a pioneering study by Zimmer (1969). Zimmer's experimental method was to give participants pairs of words that differed in their vowel sequences and ask them which word of the pair "sounds more like a word that might actually occur in Turkish (that might, for instance, be a Turkish word that you don't happen to know)". We report only the more carefully prepared Experiment 2. To give an example from this experiment, for the nonce-word pair [pemez]/[pemaz], 30 responses were to prefer [pemez], and 2 were to prefer [pemaz]. This is what we might expect, assuming the participants have internalized the principle of backness harmony.

In order to engage our models with Zimmer's findings, we interpreted the participant responses as relative frequencies. Thus, in the previous example, we compute a preference for the [ea] vowel pair as 30/(30+2) = 0.938. We modeled these preferences by doing the same calculation with the probability values assigned to these sequences by grammar (99c): P([e … e]) = 0.068, P([e … a] = 0.047, hence an [e … e] preference of 0.068/(0.068 + 0.047) = 0.590. The overall agreement between model and native intuition emerges as not too bad, as shown in the following scatterplot.

Figure 34: Matching predictions of Grammar (99c) to participant choice probability in Zimmer's Experiment 2



While we find this result moderately encouraging, in truth far more work needs to be done for us to feel that model (99c) has been seriously checked: a greater variety of stimulus types must be examined, the results of the experiment should be modeled in MaxEnt and the constraints tested for significance, and so on.

6.3 Scaling up: the UCLA Phonotactic Learner

In this chapter, we have "finessed" the basic point that GEN is infinite; for Turkish, we took the view that the system of vowel harmony in stems could be usefully understood simply by looking at two-vowel sequences, reducing an infinity to a simple list of 64 cases. We see real advantages to this approach, since it works transparently and efficiently on a spreadsheet. On the other hand, the strategy might be a bit of a gamble: might the Turkish system somehow involve constraints that require a three-vowel window, or constraints that relate vowels to adjacent consonants? This could not be tested with our simple, 64-member GEN.

It is indeed possible to perform MaxEnt phonotactic analysis on full strings. The original software for Hayes and Wilson (2008), the main source for this chapter, is available as in a version with a user interface, the "UCLA Phonotactic Learner", currently posted at https://brucehayes.org/Phonotactics/. The software uses finite state machines, following Eisner (2001, 2002), to manipulate the vast search spaces that arise when we look at the full segmental strings of words. The user can either formulate her own constraints (using a user-specifiable feature system) or let the software try to find the best constraints on its own. For understanding of how the program works, see Hayes and Wilson (2008) and also the program manual. For a fairly extensive grammar for English phonotactics created using the software, but with author-invented phonological constraints, see https://brucehayes.org/BLICK/.

For general use, we advocate a blend of work with small and larger GEN. The larger GEN, implemented with finite state machines, and machine-generated constraints, often has greater scope and security in the long run. Spreadsheet-size GENs are more manageable, and facilitate a close focus on theory and the research questions at hand; they also permit easier informal exploration of the data.

6.4 Comparison of MaxEnt with other phonotactic models

The standard approach to phonotactics adopted in Classical OT is non-stochastic; it classifies forms on an up-or-down basis as either totally legal or totally illegal. The key element is the principle of the Rich Base (Prince and Smolensky 1993: ch.9), which posits that GEN contains all logical possibilities. The phonological grammar neutralizes the forms of GEN down to a much smaller set of phonotactically legal forms. For example, to account for the fact that *[**pt**ɛrəˈdæktəl] is illegal in English, the Rich Base approach would set up a constraint ranking along the lines *MARKEDONSET ≫ MAX, such that an imaginary underlying form like /ptɛrəˈdæktəl/ would be mapped onto [tɛrəˈdæktəl] (*pterodactyl*). The latter, being the output of at least one derivation, is predicted to be phonologically legal (note that we could also derive it from /tɛrəˈdæktəl/). In this approach, there is nothing that can be said about intermediate forms like [dwɛp]; they must either *not* be derived, hence classified as illegal, or derived, hence classified as legal. The Classical OT approach does not aspire to be a model of intermediate well-formedness.

A model that is much closer to the approach of this chapter (see Breiss and Albright 2022, Cabrera 2025) lets every candidate form compete with the **Null Parse** candidate (Prince and Smolensky 1993:§4.3.1), which in the present context means essentially, "this form should not be in the language". Thus, *[ptɛrəˈdæktəl] is bad because it is better to violate *NULLPARSE than to tolerate the *MARKEDONSET violation of [pt]. Something like ?[dwɛp] might involve weights that assign nontrivial probabilities to both the faithful candidate and the Null Parse candidate. It is not hard to fit MaxEnt models using this framework, provided some way is given to fit them against corpus data; [44] and the results appear to be fairly similar to what we obtain with the method described here. For an implementation of the Turkish simulation in the Null Parse approach, see the spreadsheet **TurkishWithNullParse.xlsx**.

6.5 MaxEnt and the Observed/Expected statistic

The "Observed/Expected" statistic (Pierrehumbert 1993) is a computed value that has a venerable history in the study of phonotactics and continues to be used in current research. The statistic was employed by Pierrehumbert as part of her contribution to an important empirical research tradition covering the tendency of different consonant pairs to cooccur in a root; see

[44] The method we use is to set up two stages of calculation. At the first stage we use MaxEnt to calculate two-way probabilities, with each overt candidate competing individually against Null Parse. At the second stage, we sum all the overt-candidate probabilities (Z) and divide each by Z to obtain normalized probabilities. The latter sum to one, and so the input forms will compete against one another for probability mass in weight-fitting. The fitting is conducted in the same way described above, using log likelihood to achieve the best match of the normalized probabilities to the frequencies of the training data.

among other work McCarthy (1988), Pierrehumbert (1993), Frisch, Pierrehumbert and Broe (2004), Coetzee and Pater (2008), and Wilson and Obdeyn (2009). The key cross-linguistic pattern behind this work is that the more phonologically similar two consonants are, the less likely they are to cooccur in the same root. The main challenge is to find a metric of similarity that permits this generalization to be enforced accurately and consistently across languages.

Pierrehumbert devised the O/E statistic in the course of this research as a heuristic metric for assessing whether a particular pair of consonants is over- or under-represented in a data corpus. The statistic can easily be applied to other segment pairs, such as the vowel sequences of Turkish disyllables.

Here is how Observed/Expected is calculated. In our Turkish language sample, there are 2562 instances of [i] out of the total of 15004 vowels, so the probability of a vowel being [i] is 2562/15004 = 0.171. There are 1053 instances of [ɯ], so the probability of a vowel being [ɯ] is 0.070. From this we estimate the a priori probability of an [ɯ … i] sequence, namely the probability we would expect if vowels had no contextual dependencies. For this, we multiply 0.171 by 0.070, obtaining 0.0120. There being 7502 stems in the corpus, the expected number of [ɯ … i] stems is 0.0120 × 7502 = 179.8. But in fact there are only 11 instances of [ɯ … i]. The Observed/Expected value for this combination is obtained by dividing 11/70.8 = 0.155. This is well below one, and accords with our sense that [ɯ … i] is an unexpected (because disharmonic) stem type in Turkish. In contrast, if we calculate O/E for a typical harmony-respecting sequence like [ø … e], we get a value greater than one, in this case 4.95. We can also aggregate the counts for all pairs that violate a particular constraint and find the O/E value for a phonological constraint. For the familiar constraints of Turkish harmony these turn out to be 0.562 for AGREE(back), 0.342 for AGREE(round), and 0.653 for ROLO, again all well below one.

The Observed/Expected (O/E) statistic has widely served as a heuristic measure of over- and under- representation. However, Wilson and Obdeyn (2009) sharply criticize the method, suggesting that it leads to spurious results, and for purposes of evaluating under- and overrepresentation they recommend using MaxEnt instead. To be specific, when the best-fit weight of a constraint is positive, it means that (all else being equal), violators of the constraint are underepresented, and analogously for negative weights and overrepresentation. Further, a constraint may be submitted to further scrutiny using the Likelihood Ratio Test (§5.3.4). We here compare O/E with MaxEnt, focusing on the toy example employed by Wilson and Obdeyn.

In this simple language, the only consonants are /p/, /t/, and /k/. /t/ is statistically favored: 1/2 of all consonants are /t/. /p/ gets its equal 1/3 share, and /k/ receives the remainder, 1/6. Moreover, when the words of the language, which all contain exactly two consonants, are scrutinized, we observed a classical similarity-avoidance effect of the type studied in the literature cited above: words with identical consonants are 1/2 as common as would otherwise be expected. The ban on consecutive identicals is often called the "OCP" ("Obligatory Contour Principle," from Leben 1973).

The counts for this constructed dataset are found by multiplying out these frequency factors, then scaling up to create an imaginary wordlist of 5000 forms. So, for example, the cell for [p p] is obtained by starting with 1, then multiplying by 1/2 for the first [p], then 1/2 for the

second [p], then 1/2 for the OCP violation, then finally scaling up all values so that they total to 5000.

(100) *A toy dataset of 5000 words used by Wilson and Obdeyn for assessing O/E*

	p 1/3	t 1/2	k 1/6
p 1/3	345	1034	345
t 1/2	1034	776	517
k 1/6	345	517	86

where grey means OCP, multiply by 1/2

These pseudodata are analyzed by Wilson and Obdeyn using both MaxEnt and O/E.

Our implementation of their MaxEnt model (**PTKLanguageAfterWilsonAndObdeyn2009.xlsx**) has nine candidates with frequencies as in (100). There are four constraints, shown in (101) along with their best-fit weights and the p-values obtained from the Likelihood Ratio Test.

(101) *A MaxEnt model of the toy data in (100)*

			*p	*t	*k	OCP		
		w:	**0.41**	**0.00**	**1.10**	**0.69**		
		ΔLL:	120.2	0	716.9	215.3		
C1	**C2**	*p*:	3.1×10^{-54}	*See fn.*[45]	9.5×10^{-314}	1.3×10^{-95}	**predicted**	**observed**
p	p	345	2			1	344.8	344.8
p	t	1034	1	1			1034.5	1034.5
p	k	345	1		1		344.8	344.8
t	p	1034	1	1			1034.5	1034.5
t	t	776		2		1	775.9	775.9
t	k	517		1	1		517.2	517.2
k	p	345	1		1		344.8	344.8
k	t	517		1	1		517.2	517.2
k	k	86			2	1	86.2	86.2

The predictions of the model in (101) match the pseudo-observation counts with essentially zero error, and all three non-redundant constraints[45] test as significant. Intriguingly, we can use the MaxEnt formula to "reconstruct," as it were, the intention of Wilson and Obdeyn that went into the construction of the toy language. The fractions that were used in designing the table (1/2 for

[45] *t is the weakest of a family of constraints (*p, *t, *k), violated twice by any candidate, and so is redundant, matching the scenario discussed in §4.3.2 above.

/t/, 1/3 for /p/, 1/6 for /k/, 1/2 for OCP) can be obtained from the best-fit constraint weights by translating them back from the log-probability domain to the probability domain.[46] In short, every factor responsible for frequency variation in the chart is identified with complete precision.

What of O/E? The results of applying this method to the same toy language (second tab of the spreadsheet) are given below. What we see is a strange and misleading artifact, seen in the O/E values for the cells on the diagonal of (100).

(102) *O/E as calculated for the constructed data of (*100*)*

	p	t	k
p	**0.58**	1.29	1.05
t	1.29	**0.72**	1.17
k	1.05	1.17	**0.48**

As the boldface cells show, the OCP is supposed to be stronger for [k] than for [p], and stronger for [p] than for [t]; even though there is nothing in the process that designed the pseudodata that would indicate that this should be so. We can check this claim in the MaxEnt approach by adding constraints that penalize consonant-specific versions of the OCP, such as *OCP(t). Reassuringly, these come out with zero weights (**PTKLanguageAfterWilsonAndObdeyn2009.xlsx**, second tab).

We agree with Wilson and Obdeyn that comparisons like the one just given imply that it would be wise for the working linguist to avoid the use of O/E to assess phonological underrepresentation, and use MaxEnt (or some other well-founded statistical approach) instead.

6.6 Further developments and alternative approaches

Two papers address the need for greater formal power in constraints than was accommodated in Hayes and Wilson (2008). Berent et al. (2012) address the need for *variables*, requiring identity of features or segments. This article also improves the original Hayes/Wilson system by using "gain", an improved method for selecting constraints from a candidate space (§5.4). Gouskova and Gallagher (2020) address the need to learn non-local environments and propose an augmented MaxEnt-based learning system that can do this. An area where we think further development may be required is brought up by Wilson and Gallagher's (2018) MaxEnt analysis of Quechua phonotactics, which illustrates the need to accommodate the concept of "elsewhere" in allophones and perhaps beyond.

A diverse body of work covers phonotactics in ways that are different from the MaxEnt system described here, but share with MaxEnt a basis in probability theory and can generate gradient predictions. A traditional method from computer science is the *n*-gram model (Jurafsky and Martin 2026, ch. 3); which captures much detail but is prone to overfitting and cannot access

[46] For example, the weight of *t is 0.41 less than that of *p. In the probability domain, we have $e^{-0.41} = 0.667$, which is just the observed ratio of [p] to [t] used in designing (100). This is the Log Odds Principle (p. 70) in yet another context.

natural classes. Vitevitch and Luce (2004) offered a model, widely used by psycholinguists, whose key categories are "first segment of word, second segment, third segment", and so on. We see these categories as phonologically unrealistic and liable to lead to inaccurate modeling; for critiques see Hayes (2012) and Mayer et al. (2025). More phonologically sophisticated probabilistic approaches are offered in Goldsmith and Xanthos (2009), Goldsmith and Riggle (2012), Futrell et al. (2017), Mayer (2025), and Šegedin and Cohen Priva (2026).[47]

[47] The latter is a massive cross-linguistic study of the same research topic we cover here on a small scale for Turkish.

Chapter 7: Modeling phonological learning

7.1 Learning and learnability in MaxEnt

Our focus so far has been on doing good descriptive analysis, in the sense that we seek to construct grammars that accurately match the output of native speakers or the phonological patterns of the lexicon. Here, we turn to more ambitious modeling, related to issues of language learning and Universal Grammar. We will spend the majority of this chapter (§7.2-§7.5) covering a technique that tests possible innate biases for their role in learning. Later on, we cover two further specific issues that arise in MaxEnt learning: hidden structure (§7.7.1), and gradual processing of the data (§7.7.2).

We start with some discussion of learning and learnability in general terms. In recent decades, scholars have come to use computational modeling to pursue a traditional goal of generative linguistics, namely explaining how children learn the grammar of their native language, despite the limited time and data available and the noise and errors present in the learning data. Articulating this goal — "explanatory adequacy" — has been one of the major contributions of Noam Chomsky (1965 et seq.).

Computational modeling is, for many scholars, a particularly appealing way to study the acquisition problem in that it allows us to hypothesize fully explicit mechanisms, often incorporating hypotheses from linguistic theory, and test them out by simulating acquisition using datasets comparable to what infants and children encounter. We let the models process the data to learn a grammar, then probe the grammar in every useful way we can think of: does it actually match the pattern of the language; does it match the behavior of adult speakers in experimental studies; does it make the same errors as children during the course of acquisition? Such modeling has become an established research area; for a helpful overview see Jarosz (2019).

This research tradition can be dated back as least as far as the early work carried out in classical Optimality Theory, by Tesar and Smolensky (1993, 2000) and Tesar (2014), who developed the first OT algorithms that permitted part of the grammar (the constraint ranking) to be learned in response to input data. Tesar and Smolensky sought to argue in favor of OT on learnability grounds. Specifically, in the course of learning, when an OT grammar makes an error, it can be altered in a way that is "directed": a sensible response that will likely improve its accuracy. In the approach Tesar and Smolensky took, the response involves the demotion of "loser-preferring" (Prince's (2003b) term) constraints (later theories also allowed for promotion

of "winner-preferring" ones).[48] This situation stands in contrast to rule-based or parametric theories, where the path to fruitful grammar-modification in the face of error is less clear.[49]

After the initial appearance of Tesar and Smolensky's work, it was realized that switching to a probabilistic variant of OT would be advantageous — not just in being able to learn the gradient phenomena discussed in this book, but also in making the learning model robust to acquisition noise. The early OT learning algorithms were "brittle," in the sense that a single wrongly-heard form would serve to wreck accurate learning (for discussion see Boersma and Hayes, 2001:67). Such errors are in contrast rather innocuous in a probabilistic learner, since the effect of the occasional misheard forms will normally be overcome by the effect of the far more frequent correctly-heard forms.

7.2 Non-veridical learning and its interest for UG study

In this section we discuss the idea of "non-veridical learning", along with its theoretical consequences. We contrast two idealized cases. (a) A child who learns language acting as a perfect "inductive sponge," closely replicating the patterns presented to her in the ambient data; (b) a child whose grammar does not perfectly depict the patterning of the ambient data. We might say that the latter type of child carries out "non-veridical" learning; the grammar such a child learns does not perfectly describe the data.

It is not always easy to prove that learning is non-veridical. However, there are two areas where this seems feasible. First, there are the cases where experimental probing reveals an imperfect match to a **lexical** pattern, as in Becker et al. (2011, 2012), Hayes and White (2013), Jun (2015), Zuraw and Hayes (2017), Jarosz (2017), Jun (2015), and Gong and Zhang (2020), and. Second, there is imperfect matching of a pattern that was perfectly controlled by the experimenter in an Artificial Grammar Learning experiment (§5.5.4), as for instance in Wilson (2006).

Both of these sources of evidence indicate that non-veridical learning happens frequently. To be sure, some of the deviations from lexical or corpus data that experiments find reflect factors like inattention or other aspects of the experimental setup; see §9.1.4. But many deviations seem attributable to principles of language that can elsewhere be observed at work in the study of language typology. Thus, non-veridical learning has been taken to be a possible window into the study of the language faculty.

The way that UG impinges on language-particular grammar and on learning has evolved over the decades. "Hard" principles of UG, which are claimed to render particular grammars simply impossible, have repeatedly suffered counterexamples. Perhaps more promising, then, is an approach in which the grammar internalized by the learner is the combined result of the

[48] Indeed, the same impulse is pursued in MaxEnt, but even more so: computer scientists have honed their algorithms to make weight adjustments respond with great efficiency to the errors of the current-stage grammar, using the O/E theorem (Excursus Box, p. 82).

[49] For a review of the history of parameter-setting approaches see Prickett et al. (2019). These authors explore a promising avenue in which parameters are re-expressed using paired, opposed MaxEnt constraints.

learner's confrontation with the data along with the various **UG biases** — roughly, expectations about what language is like — that they bring to language learning. Learning biases are "soft UG"; they can be overridden by abundant learning data. Proposals about learning bias can be assessed empirically for their ability to explain nonveridical learning.

The most rigorous way to assess a proposed UG bias is through modeling: can we develop a learning model, incorporating UG biases, that when exposed to realistic learning data, learns a non-veridical grammar in the same way that children do? For purposes of such modeling, we follow Wilson (2006), who first adopted MaxEnt as a basis for it. The core of Wilson's proposal is to employ a prior (§4.5) to specify default weight values for particular constraints or constraint types; these values are overridden only in response to the learning data.

7.3 Some biases

The biases that have been worked with in modeling nonveridical learning are mostly ones familiar to phonologists, predating MaxEnt. We focus here just on biases where the system of Wilson (2006) is applicable.

7.3.1 Articulatory ease

First, the human articulatory system evidently favors certain outcomes, such as (a) non-extreme peripheral vowels;[50] (b) place-assimilated clusters (Jun, 2004), and (c) level tones on short vowels (Gordon, 2001; Zhang, 2004). These share the theme that speech production should conserve articulatory effort, in the form of (respectively) (a) extreme articulatory positions; (b) extra articulations, (c) rapid articulatory movements. Such preferences constitute the simplest and most intuitive form of "phonetic naturalness" bias. For efforts to incorporate such influence in the constraints of a formal grammar, see for instance (Kirchner 2001), Hayes (1999) and Flemming (2004). For evidence that articulatory ease (enforcing intervocalic lenition, or avoiding hiatus) can reshape paradigms over time, see (Kuo 2024a,b,c).

7.3.2 Paradigm uniformity

In phonology, there appears to be pressure to avoid alternation, giving a single pronunciation to each morpheme across contexts. This principle was proposed in Kiparsky (1971) and has been theoretically influential in the years since then. Alternation avoidance is often referred to by its alternative name, **paradigm uniformity** (a uniform paradigm lacks in alternation).

To give an example, Itô and Mester (1996) show that in conservative Tokyo Japanese, an allophonic process that replaces /g/ with [ŋ] in intervocalic position is optionally suppressed when it would alter the [g] in a compound that represents the initial segment of an independent word; thus /kagi/ → [kaŋi] 'key', but /niwa#geta/ → [niwaŋeta] or [niwageta] 'garden clogs'; cf. [geta] 'clogs'. A MaxEnt analysis of quantitative data for this phenomenon is given by (Breiss et al. 2022, 2026).

[50] See e.g. Becker-Kristal (2010), who finds that three-vowel systems normally have [ɪ, ʊ, ə]; dispersion (see Flemming 2004) is insufficient to force the use of [i, u, a].

The effect of paradigm uniformity can be gradient; the work of Breiss et al. just cited is one example, and another is found in a case we have already mentioned (§3.5), the pioneering frequency-matching study of Dutch Final Devoicing by Ernestus and Baayen (2003); we add here that, on top of the detailed, environment-by-environment patterning of frequency-matching that they found, they also report an overall preference for voiceless outputs greater than that found in the lexicon. Since each item was presented with a voiceless final consonant, and participants had to reconstruct either an underlyingly voiced or an underlyingly voiceless phone, one possible explanation for this effect is that participants are preferring to construct paradigms without alternations.

Paradigm uniformity can be gradient in a different sense, namely that phonetically extreme alternation is avoided more than phonetically modest alternations; see Steriade (2000), and for an influential formal approach Zuraw (2013, 2007). Zuraw's *MAP constraints are used by White (2017) to explain his experimental results (artificial grammar learning) on saltation (Hayes and White 2015), where participants are evidently biased to prefer short-distance phonological alternations like [d] ~ [ð] over longer distance ones like [t] ~ [ð]. This work is extended to show that the preference is manifested by infants (J. White and Sundara 2014) and can be addressed in MaxEnt modeling of the type explained below (J. White 2017).

7.3.3 Bias against complexity

It has long been suggested that language learners deploy a bias against complex hypotheses. Chomsky and Halle (1968) suggested that language learners choose, from among grammars equally compatible with the learning data, the one that expresses its generalizations with the fewest features. For discussion see Moreton and Pater (2012).

Expressed in this way, complexity avoidance is more or less the formalization of a preference for generality in grammar. Generality seems to be essential to explaining how language learners project their knowledge to novel forms not present in the data from which they originally learned their own language. A nice example (Halle 1978:301, citing Lise Menn) is the clear intuition of English speakers that the plural of German *Bach* [ˈbax] should be [ˈbaxs], not *[ˈbaxz] or *[ˈbaxəz]. If English speakers used a simple list of sounds ([p t k f θ]) to govern their choice of the [-s] plural allomorph, we could not explain this fact. Under complexity bias, English speakers learn and employ (productively) the one-feature natural class [−voice] to govern the choice of [-s].

Chomsky and Halle expressed their complexity bias as a requirement placed on the entire grammar, a scheme that has proven unimplementable to this day.[51] But it is much easier to imagine a complexity bias being imposed on individual constraints, as we will do below.

A good empirical case for complexity bias comes from Kuo's (2023a, 2023b) work on Seediq. A case of "prosodic correspondence", where inflected forms inherit their stressed vowel quality from a different vowel in the base, is generalized to cover a novel vowel ([a]) not attested

[51] Thoughtful efforts to realize Chomsky and Halle's vision of whole-grammar evaluation have been made by Jarosz (2006) in MaxEnt and by Rasin and colleagues in the Minimum Description Length approach; e.g. Rasin et al. (2021). These efforts, however, address small-scale examples.

in the lexicon. For example, [ˈbemup] ~ [buˈmep-a] is generalized to [ˈlamuk] ~ [luˈmak-a], even though there are no cases with [a] in the lexicon. In other words, a complex pattern of restricted copying is replaced by general copying of all vowels.

Becker et al. (2011) suggest a complexity bias principle stating that phonology values vowel-to-vowel or consonant-to-consonant processes over processes with vowel-consonant interaction, supporting it with data from an experiment on Turkish.

7.3.4 Attestation bias

It would make intuitive sense that lexical patterns that are supported by few examples might be taken less seriously by language learners. An extreme case is found in the English past-tense wug-test carried out by Albright and Hayes (2003), which deliberately included forms whose lexical "pattern" consists of one single word. The experimental participants were very reluctant to give the (imaginable) one-analog response; thus [ʃɔ] for [ʃi] (single model *see-saw*), [leɪm] for [lʌm] (*come-came*), and [zɛd] for [zeɪ] (*say-said*) all received very low ratings in the experiments. This behavior extends to other patterns with small counts: Albright and Hayes found that their best-fit model (not MaxEnt, which was not tried) included a Scope term (adopted from Mikheev 1997) which penalized patterns that are poorly attested.

Liang et al. (submitted) likewise suggest that certain phonological patterns of Catalan that (while exceptionless) are poorly instantiated in the lexicon may be extended less productively than lexically abundant patterns. A further recent demonstration of attestation bias is in Siah (in press), who shows that less-frequent lexical patterns receive lower ratings in a forced-choice experiment on echo reduplication in Malay. For Siah's modeling approach to attestation bias in MaxEnt see §7.6.1 below.

7.3.5 Other

The above four biases probably account for most phonological bias research today, but there are others. Becker et al. (2012) argue that Faithfulness constraints are *a priori* stronger in initial syllables, a principle that accounts for a striking disparity between the English lexicon and their wug-test results concerning the process of Fricative Voicing in English (e.g. *loaf ~ loaves*). Zhang et al. (2009) and others have suggested a learning bias against opaque process interaction. Culbertson et al. (2013) have offered experimental support, with MaxEnt modeling, for word-order bias in syntactic learning.

7.4 Modeling example: Hungarian vowel harmony

We turn now to the use of MaxEnt modeling to account for learning bias effects, using the method of Wilson (2006) on data from Hungarian vowel harmony.

As discussed in §4.4, the algorithms typically used to fit MaxEnt weights are *optimal*, in the sense that they find the best fit of constraint weights to input data based on mathematical principles such as likelihood maximization. However, should be clear from the previous section, humans do not always fit input data optimally. A good theory of language acquisition must explain both matches and mismatches to the input data. In the bias modeling paradigm that we

present here, we are able to probe mismatches of a very specific type: those in which certain specific constraints, or types of constraints, seem to require more evidence than others to achieve a given magnitude of weight. Thus we can test theories of UG when those theories are statable as preferences for particular weightings of particular types of constraints.

The vowel harmony system of Hungarian is in several ways an ideal area for modeling biased learning. The data pattern has been subjected to intensive generative phonological description and analysis, there is extensive lexical variation of the type examined in Chapter 3, there are large lexical databases available for assessing the variation quantitatively, and there are existing wug-test data covering the variation-related phenomena. Because we have access to these different kinds of data about the pattern, it is apparent that it is a case of mismatch between lexical frequency and wug-test results, of the type discussed in §3.5.1.) Hayes et al. (2009), hereafter "HZSL"), whose data we use here, claim that the mismatch can be explained via a "naturalness" bias, along the lines of those discussed in the previous section, which makes certain constraints easier to learn than others. Here, we describe the basics of Hungarian vowel harmony, the lexical pattern, and how is it both reflected and deviated from in wug testing. In the end, we will put forth a bias model that attempts to account for the deviations from lexical frequency-matching.

7.4.1 Description of Hungarian vowel harmony

The Hungarian vowel harmony system played an important role in early generative phonology and has been studied with increasing depth and detail over time; see e.g. Vago (1976), Siptár and Törkenczy (2000), Hayes and Londe (Hayes and Londe 2006), and Hayes, Zuraw, Siptár, and Londe (Hayes et al. 2009). Here, we will rely primarily on the study of Hayes, Zuraw, Siptár, and Londe (2009, hereafter HZSL). [52]

Backness harmony in Hungarian is a highly productive process;[53] the language abounds in suffixes, and most of them have front and back variants that alternate according to the principles of harmony. The harmony pattern depends on the following classification of vowels.

(103) *Hungarian vowels*

Back	[u, uː, o, oː, ɔ, aː]	abbreviated "B"
Front rounded	[y, yː, ø, øː]	abbreviated "F"
Front unrounded, often called "neutral"	[i, iː, eː, ɛ]	abbreviated "N"

To illustrate backness harmony, we can use examples with the dative suffix, whose behavior is representative. The dative has two allomorphs, back [-nɔk] and front [-nɛk], which are distributed according to the principles of harmony. To begin, whenever the rightmost vowel of a

[52] While we feel comfortable using the HZSL material for pedagogical purposes, we note that the gradient pattern of Hungarian vowel harmony has remained an active area of research, and readers interested in this area should consult later research. Notably, Forró (2013) confirms only one of the four unnatural effects noted below, and locates probabilistic affix-specific effects. Rebrus et al. (2023) make the striking suggestion that Hungarian vowel harmony works rather like inflectional classes, rather than phonology.

[53] There is also rounding harmony, restricted to certain suffixes; we will ignore it here.

stem is [+back] ("B"), the stem must take back suffixes. This generalization, illustrated below, is exceptionless.

(104) *Closest vowel back: back suffixes*

B**B**	[ɔblɔk-nɔk]	'window-dat.'
N**B**	[biːroː-nɔk]	'judge-dat.'
F**B**	[glykoːz-nɔk]	'glucose-dat.'

Likewise exceptionless is the generalization that if the closest stem vowel is front rounded ("F"), the stem must take front suffixes:

(105) *Closest vowel front rounded: front suffixes*

F	[yʃt-nɛk]	'cauldron-dat.'
N**F**	[sɛmøltʃ-nɛk]	'wart-dat.'
B**F**	[ʃoføːr-nɛk]	'chauffeur-dat.'

Moreover, front suffixes are obligatory when a front rounded vowel occurs separated from the end of the word by one or more neutral vowels:

(106) F + N**: front suffixes*

FN	[fyːsɛr-nɛk]	'spice-dat.'
FNN	[øːrizɛt-nɛk]	'custody-dat.'

The final two categories of stems are those in which a back vowel is followed by one or more neutral vowels (BN, BNN…), and those with only neutral vowels (N, NN, …). Stems in these categories fall into what HZSL call the **zones of variation**. In a zone of variation, individual stems vary in the kind of harmony they take; this is a fundamentally unpredictable property of these stems and thus reminiscent of the lexical variation covered in Chapter 2 for Tagalog. The possibility of variation in these zones is illustrated in (107).

(107) *The zones of variation in Hungarian vowel harmony*

Zone	*Stems taking front suffixes*	*Stems taking back suffixes*
a. BN stems	[domeːn-nɛk] ‘domain (on Web)-dat.’ [modɛl-nɛk] ‘design-dat.’	[poeːn-nɔk] ‘punch line-dat.’ [hɔvɛr-nɔk ‘pal-dat.’
b. BNN stems	[obɛliszk-nɛk] ‘obelisk-dat.’ [proteːziʃ-nɛk] ‘prothesis-dat.’	[poziti:f-nɔk] ‘positive-dat.’ [lɔbiliʃ-nɔk] ‘unstable-dat.’
c. N stems	[kɛrt-nɛk] ‘garden-dat.’ [tsiːm-nɛk] ‘address-dat.’	[hiːd-nɔk] ‘bridge-dat.’ [ʃiːp-nɔk] ‘whistle-dat.’
d. NN stems	[sɛmeːt-nɛk] ‘garbage-dat. [tɛpʃi-nɛk] ‘baking sheet-dat.’	[dereːk] ‘sturdy-dat.’ [férfi] ‘man-dat.’[54]

With a lexical database (we used HZSL’s, downloaded from https://brucehayes.org/HungarianVH/HungarianFromGoogleRevised.txt), it is possible to study the quantitative patterns present in these zones of variation — in particular, we can find perturbers that influence the distribution of back and front suffixes. These are discussed in the next two sections.

7.4.2 “Natural” perturbers

Study of the lexical database shows that BN stems take back harmony more often than BNN stems. The distinction is called the **Count Effect** by HZSL. Here are data from the corpus they used documenting the effect.

(108) *The Count Effect in Hungarian vowel harmony — BN and BNN stems*

Type	*Back*	*Front*	*Backness rate*
Stems ending in BN	245.7	121.3	0.67
Stems ending in BNN	11.6	51.4	0.18

A note on the non-integer counts given here: a number of stems in the zones of variation are “vacillators,” meaning they can optionally take either front or back harmony. Vacillator stems are counted fractionally, so that in forming the count, each backness value (+ or −) is incremented by a value between zero and one, corresponding to the fraction of forms in the database that the vacillator stem takes that backness value.

We see the Count Effect as plausibly involving a perturber, for which there is more than one possible responsible mechanism. One plausible account would rely on the greater distance of the triggering B vowel; the farther away, the less the backness influence. Here, we opt (following HZSL) for a different account, namely that in BNN there is a “double trigger” for frontness in the form of the NN sequence; for precedents and citations re. double triggers in phonology see HZSL p. 834. The relevant perturber constraint will be given in (111h) below.

[54] [dereːk] and [férfi] are actually “vacillators” in the sense defined in the next section; they take [-nɔk] with proportions of 0.623 and 0.797 respectively. There are no NN stems that take back suffixes 100% of the time.

Another perturber for Hungarian is based on what HZSL call the **Height Effect**, whereby stems ending in BN or BNN (on the average) take more front suffixes if their last vowel is [ɛ] than if it is [e:]; and likewise more front suffixes if their last vowel is [e:] than if it is [i] or [i:]; so, the lower the trigger, the more frequent the agreement. The Height Effect is very strong in the lexicon, as the database frequencies in (109) show.

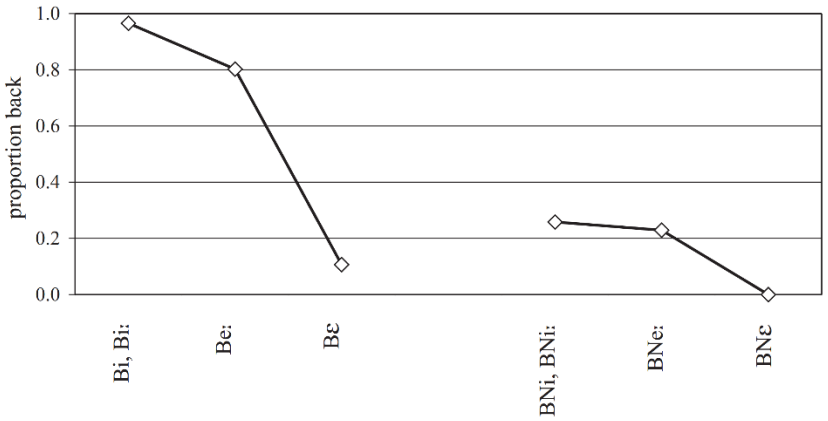
(109) *The Height Effect in Hungarian vowel harmony — BN and BNN stems*

Type	*Back*	*Front*	*Backness rate*
Stems whose rightmost vowel is [i] or [i:]	183.9	30.1	0.86
Stems whose rightmost vowel is [e:]	61.4	21.6	0.74
Stems whose rightmost vowel is [ɛ]	12.0	121.0	0.09

HZSL p. 833 adduce a natural basis for the Height Effect, namely the principle of "weak triggers" in vowel harmony (Suomi 1983; Kaun 2004) mentioned above for Turkish (§6.2): a vowel that only weakly manifests the harmonizing feature is more likely to spread the feature to neighboring vowels. Lower front vowels have lower F2 (Tarnóczy 1964), and are stronger triggers in this sense.

As the graph in Figure 35 indicates, the Height and Count Effects combine in a fairly consistent way.

Figure 35: The Height and Count Effects together



It can be seen that the Height Effect is weaker for BNN than for BN; we take this to be a normal consequence of the MaxEnt shifted sigmoid pattern (§2.8, §3.3): there is less room for the Height Effect to make big differences in backness rate when the Count Effect is already forcing low overall backness rates.

7.4.3 "Unnatural" perturbers

Equipped with a database of forms in the zones of variation, HZSL looked for further constraints. The search for constraints was carried out in a speculative vein, allowing for constraints that seem non-natural; indeed, to a phonologist, almost nonsensical. The search found four such constraints, all of which tested as significant against the lexical pattern.

(110) *Four unnatural vowel harmony constraints for Hungarian*

a. Prefer front suffixes when the stem ends in a bilabial noncontinuant ([p, b, m]).
b. Prefer front suffixes when the stem ends in a sibilant ([s, z, ʃ, ʒ, ts, tʃ, dʒ]).
c. Prefer front suffixes when the stem ends in a coronal sonorant ([n, ɲ, l, r]).
d. Prefer front suffixes when the stem ends in a sequence of two consonants.

For discussion of why these constraints are plausibly considered unnatural (no phonetic or typological basis), see HZSL, p. 837.[55]

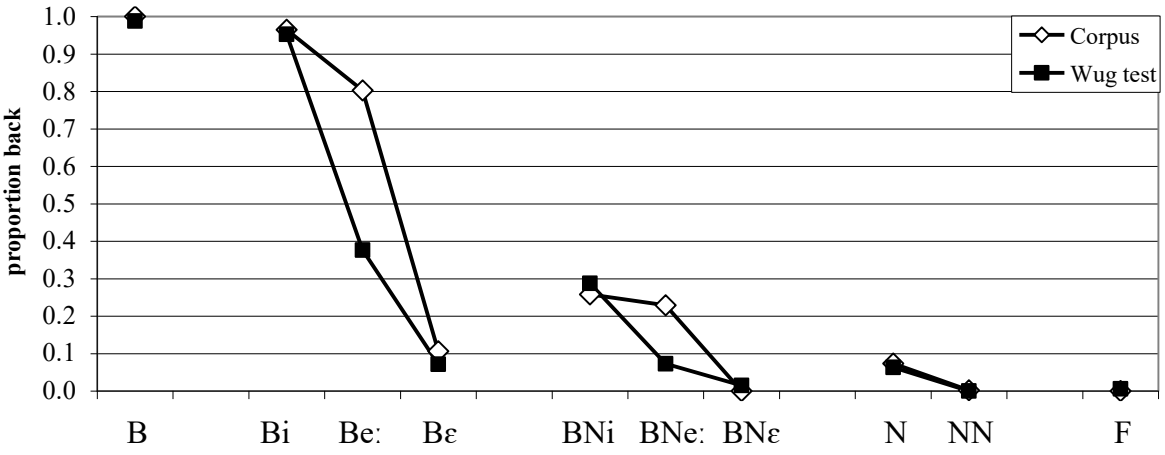
To summarize the Hungarian data pattern, there is a sort of hierarchy of simplicity: harmony as triggered by B and F(N) is straightforward and exceptionless; harmony with BN, BNN, and N is lexically idiosyncratic, and statistically observes the natural Count and Height effects. The harmony effects of stem-final consonants, though they test as significant in lexical data, seem puzzling and probably have no natural phonological basis.

7.4.4 Wug-test findings

There are two wug-tests that we rely on here. Hayes and Londe (2006) carried out a broader study that included exceptionless environments like after B or F vowels, as well as the "natural" perturber effects discussed in §7.4.2. The Hayes and Londe data are summarized in Figure 36 below; they were broadly replicated in HZSL's experiments.

[55] As perturbers these factors create shifted-sigmoid patterns; for an illustrative graph, see the Hungarian section of the "Gallery of Wug-Shaped Curves," https://brucehayes.org/GalleryOfWugShapedCurves.

Figure 36: Hayes and Londe's wug test: participant responses largely match lexical counts



We see both confirmation of the perturber effects (Count and Height), as well as a reassuring near-unanimity among participants for the forms that in the lexicon have invariant outcomes (that ones that are not in zones of variation).

HZSL's subsequent wug-test study ignored the invariant cases, taking them to be well-established, and instead focused on the puzzling consonant-perturber effects described in the preceding section. Remarkably, all four unnatural perturbers were shown in their experiment to have a significant effect on the participants' vowel harmony preferences; see HZSL, p. 844. We note that results of this sort constitute preliminary evidence against the view that unnatural, parochial constraints cannot occur in a phonological system. This is a finding made on occasion by various researchers over the years; see e.g. Blevins (1997), Hayes (1999), Albright and Hayes (2003),[56] and Pierrehumbert (2006).

The apparent synchronic viability of the unnatural constraints for Hungarian raises important theoretical issues. Do these bizarre constraints have the same status in phonology as the natural constraints that have a firm grounding in phonetics and typology? If so, this would be compatible with a view that in phonological learning all constraints are simply induced from the data, with no role for a universal CON component — this is the child as "inductive sponge," discussed above in §7.2.

HZSL take an intermediate position on this issue: both natural and unnatural constraints can be learned and deployed, but the unnatural constraints have a *weaker* status relative to the natural ones. In MaxEnt terms, this mean that relative to what would match the pattern of the data children hear, the unnatural constraints end up with lower weights than natural ones. For the statistical reasoning by which the authors arrive at this conclusion, see HZSL, §9.7.

[56] Albright and Hayes's featured example, backed by wug test results, is that every verb in English that ends in a voiceless fricative is regular.

We take up the same basic question here, but in this case apply the technique of biased-based MaxEnt modeling. The idea is that our model will be trained on representative Hungarian lexical data, but that the biases will result in lower weights for the unnatural constraints.

7.4.5 The constraint system employed

For the analysis of these patterns given here we adopt unaltered the constraints in HZSL and refer the reader to that work for full discussion of them. The constraints are listed in (111) below:

(111) *Hungarian vowel harmony constraints from HZSL*

Constraint	*Comment*
a. AGREE(back, local)	Assign a violation for every vowel that does not agree in backness with the closest preceding vowel. *Inviolable.*
b. AGREE(back, nonlocal)	Assign a violation for every vowel that does not agree in backness with *any* preceding vowel.
c. AGREE(front rounded, local)	Assign a violation for every vowel that does not agree in backness with the closest preceding vowel, if it is front and rounded. *Inviolable.*
d. AGREE(front rounded, nonlocal)	Assign a violation for every vowel that does not agree in backness with a preceding front rounded vowel anywhere in the word. *Inviolable.*
e. AGREE(front, local)	Assign a violation for every vowel that does not agree in backness with the closest preceding vowel, if it is front.
f. AGREE(non-high front, local)	Specific version of (e), limited to non-high triggers. *Perturber constraint for Height Effect*
g. AGREE(low front, local)	Specific version of (f), limited to low triggers[57] *Perturber constraint for Height Effect*
h. AGREE(double front, local)	Two-trigger harmony *Perturber constraint for Count Effect*
i. Front after [+lab,−cont]	Prefer front suffixes when the stem ends in a bilabial noncontinuant *Unnatural perturber constraint.*
j. Front after [+strident]	Prefer front suffixes when the stem ends in a sibilant. *Unnatural perturber constraint.*
k. Front after [+cor,+son]	Prefer front suffixes when the stem ends in a coronal sonorant. *Unnatural perturber constraint.*

[57] We treat [ɛ], the lowest front vowel of Hungarian as [+low]: in suffixes it alternates with a low vowel, and phonetically it is often rather lower than the conventional IPA symbol suggests. https://brucehayes.org/HungarianVH/

l. Front after CC	Prefer front suffixes when the stem ends in two consonants. *Unnatural perturber constraint.*
m. Back after [C_0i:C_0]	Prefer back suffixes when the stem is monosyllabic with the vowel [i:] *Included for completeness; not discussed here.*

We set up spreadsheets in which these constraints are used to analyze both the lexical data and the wug test data. In these spreadsheets, the candidates are defined for convenience by violation patterns; this is possible because any words that have identical constraint violations would be expected to pattern the same no matter what constraint weights. To reflect this, we represent the input forms with simple codes embodying the relevant structural properties; e.g. **BNe:R** encodes the word *acél* 'steel', which ends in a BN sequence, has the rightmost vowel [e:], and ends in a coronal sonorant ([l]).[58]

Here are three representive inputs from the lexical analysis spreadsheet; just five of the constraints are shown. The full tableau may be accessed as **HungarianLexiconModel.xlsx**; and the corresponding tableau for the wug test as **HungarianWugTestModel.xlsx**.

(112) *Extract from tableau: Hungarian analysis (lexicon)*

Input type	Example	Cand.	Freq.	AGREE (back, local)	AGREE (front local)	AGREE (nonhigh front local)	AGREE (back , nonlocal)	AGREE (front local)
BNN+Mid+ NoC	[a:be:tse:]	nak	1.9		1	1		1
		nek	6.1				1	
B+na+RCC	[sa:rɲ]	nak	266.9					
		nek	0.1	1			1	
BNNiB	[pantomim]	nak	2.7		1			
		nek	1.3				1	

Our spreadsheets replicate HZSL's results, in that the constraints of (111) receive closely similar weights and these weights suffice to achieve a reasonably close match with both lexicon and wug test results — assuming distinct sets of weights for these tasks. Below, we give the weights we found and significance figures for both lexicon and wug test analyses (the latter may be examined in full in **HungarianLexiconModel.xlsx**).

[58] The codes are as follows: BN, BNN = pattern of last vowels; High Mid Low = high of rightmost vowels (na = height not assessed in back vowels); *S* = stem ends in sibilant, *R* = coronal sonorant, *Cl* = consonant cluster, *B* = bilabial noncontinuant; NoC = no special final consonant.

(113) *Hungarian vowel harmony models for lexicon and wug test*

Constraint	*Lexicon*			*Wug test*		
	wt.	*ΔLL*	*p*	*wt.*	*ΔLL*	*p*
a. AGREE(back, local)	4.00	9.5	1.37×10^{-5}	*(not used)*		
b. AGREE(back, nonlocal)	5.35	213.6	6.80×10^{-95}	3.58	160.9	5.82×10^{-72}
c. AGREE(front rounded, local)	3.74	0.1	0.68[59]	*(not used)*		
d. AGREE(front rounded, nonlocal)	1.72	0.1	0.64[59]	*(not used)*		
e. AGREE(front, local)	1.64	5.8	0.0007	2.22	33.4	2.88×10^{-16}
f. AGREE(non-high front, local)	1.48	7.2	0.0001	0.95	16.1	1.43×10^{-8}
g. AGREE(low front, local)	2.99	27.6	1.13×10^{-13}	1.05	19.6	3.76×10^{-10}
h. AGREE(double front, local)	4.05	61.7	1.10×10^{-28}	1.94	50.2	1.25×10^{-23}
i. Front after [+lab,−cont]	2.46	6.4	0.0003	1.04	15.2	3.63×10^{-8}
j. Front after [+strident]	0.91	2.6	0.024	0.37	2.0	0.044
k. Front after [+cor,+son]	1.08	4.4	0.0031	0.43	3.0	0.015
l. Front after CC	1.75	6.0	0.0005	0.69	11.4	1.73×10^{-6}
m. Back after [C_0i:C_0][60]	2.37	7.6	0.0001	0.80	0.7	0.226

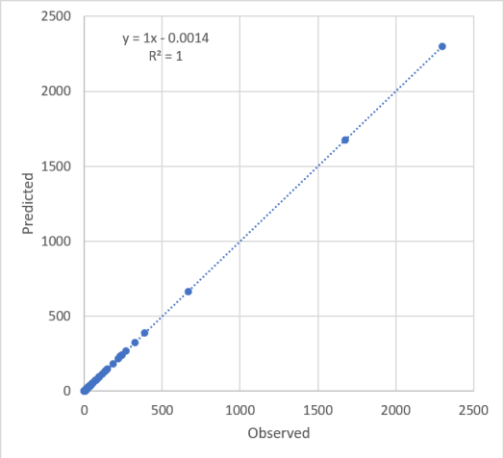
We also give illustrative scatterplots. The data consist of 8915 items (lexical study) and 1703 items (wug test), but what makes for an informative scatterplot is a plot based on subgroups of the data defined by sharing the same constraint violations, as defined above. With this pooling, there are just 127 data points for the lexical study and 80 for the experiment, each representing the averaged results for a larger set of words. We see in Figures 37a and 37b scatterplots thus defined provide a fairly good match to the data, particularly for the lexical forms:

[59] The two constraints with F triggers individually do not fail to test as significant, since they can take on most of each other's functions. Testing them together, we get $p = 0.043$. This, too, is a bit feeble, but it should be remembered that these constraints serve only to distinguish F and FN (exceptionless front harmony) from N and NN (near-exceptionless).

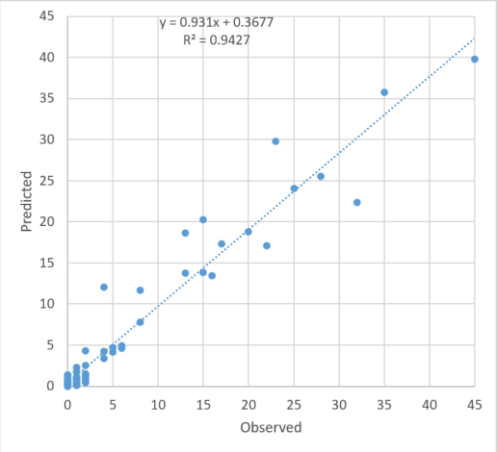
[60] Since this constraint ("Use back harmony after monosyllables with vowel [i:]") did not test significant in the wug test, we omit discussion of it here.

Figure 37: Scatterplots: best fit models for lexicon and wug-test, fitted separately

a. Lexicon (-nak counts)



b. Wug test (-nak counts)

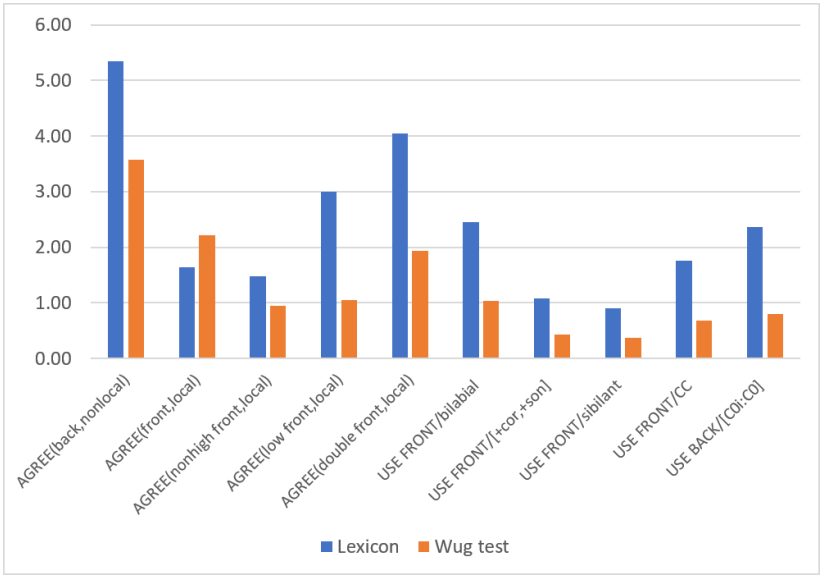


It seems, then, that our baseline models are working fairly well and it would be reasonable to move on to the key topic, which involves comparing them with each other.

7.4.6 Weight comparison

When we compare the two analysis, we can only consider the constraints shared by both; e.g. simple AGREE(back, local) cannot be compared because no relevant forms (B stems) were included in the wug test. That the two similations yielded rather different weight values is shown in Figure 38.

Figure 38: Comparing the weights found in fitting the same constraints to the Hungarian lexicon and wug-test data



The obvious trend is that wug test model has overall lower weights than the lexical model. We see several possible reasons for this. Perhaps language acquisition itself is conservative, and does not internalize phonological generalizations to the same degree that they are present in the lexicon — a sort of very general "skepticism bias." Another reason relates to attestation bias (§7.3.4), and is discussed in §7.6.1 below. A third possibility is that the lower weights reflect some kind of noise arising within the experimental procedure; e.g. inattention, mis-clicking, and so on. The procedures we follow below will not be able to distinguish these various possibilities, which are not our main focus in any event. We attend instead to the possibility that some of the constraints are learned more completely than others; this would be seen in their weights more closely approximating the weights that best fit the lexicon.

HZSL pursue this interest, conducting an elaborate Monte Carlo statistical test to find which constraints are "overlearned" or "underlearned". We refer the reader to HZSL for the the details of this test and report only the bottom-line conclusions in (114) below.

(114) *Degree of learning and bias properties of the HZSL constraints*

Constraint	*Lexi-con weight*	*Wug test weight*	*Ratio*	*Degree of learning*	*Natural-ness*	*Simplicity*
a. AGREE(back, nonlocal)	5.35	3.58	0.67	no bias	natural	simple
b. AGREE(front, local)	1.64	2.22	1.36	overlearned	natural	simple
c. AGREE(nonhigh front, local)	1.48	0.95	0.64	no bias	natural	complex
d. AGREE(low front, local)	2.99	1.05	0.35	underlearned	natural	complex
e. AGREE(double front, local)	4.05	1.94	0.48	underlearned	natural	complex
f. USE FRONT / bilabial ___	2.46	1.04	0.42	underlearned	unnatural	complex
g. USE FRONT / [cor,son] ___	1.08	0.43	0.40	underlearned	unnatural	complex
h. USE FRONT / sibilant ___	0.91	0.37	0.40	underlearned	unnatural	complex
i. USE FRONT / CC ___	1.75	0.69	0.39	underlearned	unnatural	complex
j. USE BACK / [C_0iC_0] ___	2.37	0.80	0.34	underlearned	unnatural	complex

The table gives the best-fit weights for the constraint set as weighted against both the lexicon and the wug-test results; only the constraints relevant to both are included. The fourth column gives the result of HZSL's statistical test for over- vs. under-learning.

The last two columns represent HZSL's theoretical speculations concerning why various constraints are over- or under-learned. These relate to the principles discussed above in §7.3. As noted earlier, the "crazy" constraints (114f-j) are felt to be crazy because they lack any plausible basis in production (§7.3.1), perception (§7.3.2), or typology. Thus, if there is a general bias in the language faculty to disfavor unnatural constraints, this might be used to explain the underlearning of these constraints. In contrast, we have already defended (§7.4.2) the phonological suitability of the constraints of (114a-e), all of the AGREE family. The distinctions just drawn appear in the sixth column of table (114).

There is a further principle that might be appealed to here: *complexity bias*, discussed in §7.3.3. HZSL note the following: "We think it is useful to make a distinction between harmony in its raw form—backness harmony, skipping across neutral vowels—and the more subtle

modulations of it seen in the height and count effects" (p. 853). The "pure" principles of single-feature agreement are suggested by HZSL as being favored in UG by their simplicity. While we do not work out a specific complexity metric here, it seems plausible to say that the two single-feature agreement constraints are simple, and the other are complex; this classification appears in the final column of table (114). HZSL leave as an open issue what forms of bias are truly evident in their data.

In what follows, we use a bias-analysis technique based in MaxEnt to try to explore these issues. Specifically, we will set up a model in which the system encounters the lexical data, but because of its biases learns constraint weights that to a fair degree ending up matching the wug-test data. This procedure is, in a way, an attempt to match real life, in which human speakers digest data in childhood, form a grammar, and end up with intuitions that can be examined and assessed experimentally. Our conjecture is that positing some combination of complexity bias and naturalness bias as part of language acquisition can be used to account for the differences between the lexical patterns and wug-test outcomes in the HZSL study.

7.5 Modeling learning biases in MaxEnt: the use of priors

In §4.5, we introduced the notions of **objective function** and **Gaussian prior**. To review, the objective function is the value that is either maximized or minimized in fitting the weights of a model with a data corpus. The formula for the simplest Gaussian prior was given in (77) (p. 85); we sum the squares of the weights and divide by $2\sigma^2$, where σ is a value chosen by the analyst and is inversely related to the prior's "strength" (meaning, the tendency to shift the weights away from the values they would receive in priorless fitting). This prior is subtracted from log likelihood to form the objective function. The goals of using a prior as set forth in §4.5 were modest: we sought to avoid crashes resulting from ideally-infinite constraint weights, as well as ensure replicability of the calculations.

Wilson (2006) realized that the Gaussian prior will have far more substantial effects if it is enriched to apply on a constraint-by-constraint basis. For instance, suppose for every constraint *i* there is a value μ_i, the value that the constraint "prefers" to have in the absence of any learning data.[61] Then the Gaussian prior would be calculated by adding up the squared deviations of the constraints from their preferred μ values, as in (115) below.

(115) *Gaussian prior for assigning preferred weights to individual constraints*

$$\sum_i \frac{(w_i - \mu_i)^2}{2\sigma^2}$$

where μ_i is the preferred weight of the *i*th constraint

Of course, the choice of the values for μ_i should not be capricious, but theory-driven. In his original analysis Wilson (2006) relied on the concepts of the p-map (Steriade 2000) and *MAP constraints (Zuraw 2010, 2013) to set his priors using experimental data. Here, we will follow

[61] For the possibility of σ_i, see §7.6.2 below.

the method of assigning the constraints to a small number of classes and assign a particular μ value to each class.

A constraint-specific Gaussian prior will cause the final fitted grammar to diverge in its predictions from the frequencies apparent in the learning data. Each weight will in general receive a weight that is a compromise between what would be obtained from simple fitting of the data and matching the μ value posited for its constraint type, in a UG-based prior. The resulting learning bias is appropriately described as "soft" in the sense that it can be overridden by large amounts of data — thus, we expect to see more of an effect of the prior where relevant data are scarce in a language, and less of its effect when data are abundant (see §7.6.1 below).

7.5.1 Analyzing Hungarian with cross-dataset fitting

In order to apply the above ideas to our Hungarian problem, we adopt the specific regime given in (116).

(116) *A regime for biased learning in the Hungarian example*

a. **Training data**: the lexical data corpus employed by HZSL.

b. **Testing data**: assess degree of match to HZSL's wug-test data.

c. **Constraint taxonomy**: Simple, Complex, and Unnatural, as defined in (114). (Note that every constraint that is Unnatural is also Complex.)

d. **Values adopted for σ and μ_i**: see (117)-(119) below. The various values for μ are assigned on the basis of the constraint taxonomy.

We also need a plausible criterion of success for the analysis. The key is to use *the log likelihood of the wug-test data*, even though the grammar is *trained* on the lexical data. In what follows, we will use the term **cross-dataset fitting** to describe this procedure. The use of cross-dataset fitting is intended to achieve our modeling task: explaining how language learners come up with grammars that deviate in intelligible ways from the data pattern that they encountered during language acquisition.

To start, we offer a spreadsheet to illustrate the process, **CrossFittingHungarianVowelHarmony.xlsx**. Here, the user may set μ and σ values for biased fitting of the weights of the lexical dataset grammar (worksheet LexicalData). The fitted values are automatically tranferred to the worksheet WugTest, where they are used to derive probabilities for the wug-test results; and the outcome is assessed with log likelihood and a scatterplot. This spreadsheet is only for purposes of illustration, since it does not accomplish the harder task, which is finding the actual σ and μ values that yield the best cross-fitting outcome. The latter can be thought of as "meta-fitting"; we want the σ and μ's that, when we use them to fit the lexicon weights, maximize the likelihood of the wug test data.

For this purpose we needed to move beyond our pedagogically-transparent spreadsheets and perform the computations in the `maxent.ot` R package, described above in §2.11. The R-using

reader (or any curious reader) is invited to open the R script **BiasModelingDemo.R**, which we have commented thoroughly to make its workings as transparent as possible.[62]

The method used in the script to find the optimal values for σ and various μ is a simple one known as **grid search** (see e.g. Dufour and Neves 2019).[63] Suppose at first that we assume all μ values to be zero, and searched solely for a best-fit σ. In grid search, we would fit the constraint weights over and over, with σ set at (say) .01, .1, 1, 2, … 10; then keep the value of σ that works best. This can be refined a bit: if the best value for σ turns out to be the maximum or minimum on our list, (here, say, 10), we can search a second series, like (10, 11, … 20), to check if the best value of σ falls outside the confines of our earlier list. Further, if the best value for σ turns out to be (say) 5, we could refine our result by searching the interval (4.1, 4.2 … 5.9). Lastly, the whole process can be scaled up to search for multiple parameters at once, simply by consecutively examining all possible combinations of the chosen parameter values. It should be clear that grid search is not an efficient method, but since we are only searching for at most four parameter values here, it will suffice.

Before we report our findings, it is sensible to discuss what range might we expect to find for the wug-test log likelihood values. First, we cannot hope to do better than the log likelihood we would get when we fit the constraint weights against the wug-test data themselves, which turns out to be −695.7 (see (113) above). Second, we can establish a "no help from the prior" likelihood value as a sort of worst-case scenario: this is the likelihood of the wug test data in a model trained on the lexical data without bias. This turns out to be −830.7. Hence we are seeking a bias regime that learns a grammar that improves substantially on −830.7, ideally approaching −695.7.

7.5.2 Results of the grid search

For purposes of significance-checking it is helpful to work out the model in stages, following the "bottom up" method covered in §5.4.2. First, we adopt the simplest hypothesis, that experimental partipants value *all* constraints less highly than their lexically-fitted values, as discussed in §7.4.6 above. This scenario can be modeled with just two parameters, σ and constraint-invariant μ. Our two-dimensional grid search yielded the following values, providing the best predictions for the wug-test data.

(117) *Best-performing values for the bias parameters: a model that makes no distinctions among constraints*

a. σ	0.13
b. μ for all constraints	1.0
Log likelihood for the wug-test data:	−714.2

[62] The scripts read the files **Rinput_corpus.txt** and **Rinput_wug.txt**, also in the supplementary materials.

[63] For an alternative to grid search, see Culbertson et al. (2013).

Given the worst-case and best-case scenarios outlined above (−830.7 and −695.7 respectively), the value −714.2 seems already to represent fairly good performance.

Next, we implement a model that includes a soft UG bias in favor of simple constraints. We do this by allowing two distinct values for μ, one for simple constraints and one for complex (see (114) above; recall that we classify all of our unnatural constraints as being also complex). So in this case the grid search is looping through target values for three parameters.

(118) *Best-performing values for the bias parameters: a model that distinguishes between simple and complex constraints*

a. σ	0.14
b. μ for Simple constraints	2.4
c. μ for Complex constraints	0.9
Log likelihood for the wug-test data:	−705.85

Reassuringly, the best-fit μ value for simple constraints comes out appreciably above the best-fit value for complex ones, reflecting our hypothesis that simple constraints are valued more highly. We also see that the log likehood of the model is somewhat higher than that obtained in (117). A likelihood ratio test on this improvement (−714.2 → −705.85, one additional degree of freedom) yields a *p*-value of 0.00004. We conclude that a theory positing a soft UG bias for simple constraints is able to some degree to account for differences in the preferences of Hungarian speakers relative to the lexicon of the language.

What about naturalness? Here, our grid search sought the best-fit values for a four-parameter model, which adds to (118) a distinct μ for unnatural constraints. Since all of our unnatural constraints are also complex, we expressed the new parameter as "delta μ," perturbing downward the μ for a complex constraint just in case it is also complex; in a sensible world, delta μ should be negative, and indeed in the best-fit values, it was, as shown in (119).

(119) *Best-performing values for the bias parameters: a model that distinguishes between simple, complex, and unnatural-complex constraints*

a. σ	0.14
b. μ for Simple constraints	2.6
c. μ for complex constraints	1.0
d. Downward perturbation of μ for unnatural constraints	−0.22
Log likelihood for the wug-test data:	−705.22

Although assigning distinct treatment to unnatural constraints does make an improvement in the log likelihood of the model, this improvement (−705.85 → −705.22) is exceedingly small, and the value of *p* for a likelihood ratio test comes out at 0.26. There is no clear basis, at least as emerging from these modeling efforts, to support the view that Hungarian learners respect a distinct between the natural and unnatural complex constraints of (114114).

We draw two conclusions from this exercise in cross-dataset fitting. First, the biases we imposed do indeed move the modeling results based on the lexicon to a pattern closer to what (according to the experiment) reflects native speaker intuition; achieving our first-stage goal.

We also think it potentially important what did *not* work; that is, the use of a separate μ term for naturalness added nothing significant beyond what can be attributed to general bias and to complexity. In other words, it appears that HZSL were right to worry that what their work was really accessing was a bias against constraint complexity, not naturalness. There would have to be additional, yet ungathered evidence to make the case that Hungarian speakers are influenced by naturalness as well as complexity in their harmony judgments.

For some further assessment on the significance of our result see discussion in §7.6.3 below.

7.6 Further topics in biased learning

7.6.1 "Overwhelming the prior" and attestation bias

A fact to bear in mind when doing work with priors is that in the objective function we have been using, the likelihood term scales up with the number of data examined, but the Gaussian prior stays the same. If one chooses to switch to a larger batch of data, the effect of the prior will diminish, and indeed the end point of this process would be for the prior hardly to matter at all. This effect is called "overwhelming the prior" (see e.g. Feldman (2015). In practical terms, it means it is necessary for the investigator to adopt a rescaled prior if the dataset is expanded.

An aspect of "overwhelming the prior" that may turn out to be important is that, when we examine the *different types of input* in a training set, the more abundantly-attested cases can overwhelm a uniform prior more than the sparsely-attested cases; e.g. an input where the winner wins over its rival 1000 to 0 is likely to get a higher weight for the constraint affecting it than an input where the winner wins over its rival 5 to 0. Hence the use of uniform bias term might be a good way to address attestation bias, discussed above in §7.3.4. We give a schematic simulation of this in (120) below.

(120) *Frequent inputs overwhelm the prior more easily than infrequent ones*

			C1	C2	
		weights:	6.63	2.29	*p*
Frequent input	Winner	1000			0.999
	Loser	0	1		0.001
Rare input	Winner	5			0.908
	Loser	0		1	0.092

1.807	likelihood
49.148	sum of squared weights
5	sigma (hand entered)
4.915	prior
−6.722	objective function

Here, priorless MaxEnt would fit both constraint weights very high and match the 1-0 probability for both inputs very closely. A suitable uniform bias (here, σ = 5) offers a bit of

skepticism for the infrequent candidate; that is, P(winning candidate) = 0.908 for the infrequent input, while P(winning candidate) = 0.999 for the frequent one.

In fact, this is exactly what Siah (in press) found in modeling attestation bias in Malay echo reduplication: in his simulations using a uniform bias term, an appropriate σ value improved model accuracy considerably, provided the data were presented in actual numbers, reflecting attestation, as opposed to mere proportions. Since we found a strong effect of a single, simple prior in the Hungarian simulation, it seems possible that attestation bias is playing a role here as well.

7.6.2 Constraint-specific sigmas

The pioneering work in soft-UG-bias MaxEnt modeling, Wilson (2006), adopted a somewhat different approach: instead of implementing UG principles as constraint-specific μ_i, Wilson set μ at a constant value and adopted constraint-specific σ_i values instead. This is not at all unreasonable; it simply says that the degree to which exposure to data inhibits the "climbing up" of constraint weights from a baseline of zero (equivalent to absence) varies from constraint to constraint, according to soft UG principles. White (2017) finds that it is possible to reanalyze Wilson's data with constraint-specific μ. We have not been able to find a workable model for our Hungarian example using constraint-specific σ. Further, Olson (2020), studying loanword adaptation in Cantonese suggests that only the constraint-specific μ method would work for grammars that characterize inventories, as in our Chapter 6.

7.6.3 Bias modeling: the long term

More than the other examples in this book, we feel that our account of learning bias here is a promissory note. The learning theory we describe does account for some of the deviations of the wug-test data from the lexical pattern, so there is some hope for a long-term strategy of modeling such data with soft UG biases. Moreover, we obtained a potentially useful result, namely that it appears unlikely that the unnatural constraints are devalued by Hungarian learners any more than what would be due simply to the fact that they are complex. Thus, we bring to bear a modest modeling result on the much broader debate (see Moreton and Pater 2012, Hayes and White 2013, Zheng and Do 2025) about the role of naturalness in phonological learning.

The work is a promissory note, however, because of its fundamental *circularity*: the values for μ_i and σ were obtained simply by finding the ones that yielded the best fit to the wug test data. This is hardly adequate for the long run. For the research strategy to be convincing, we need to establish (a) a clear criterion of constraint complexity; (b) a consistent strategy for translating this criterion into a prior constraint weights (μ_i). If these principles can be implemented, and if they consistently predict the lexicon-wug test differences *across multiple languages*, then we might say that we are on our way to an explanatory theory of soft UG bias. This is of course true of all hypotheses about UG: they usually need much broader checking than what has been done so far.

While our example uses arbitrary (data-fitted) μ_i and σ, this is not the case for all of the work done so far in this research tradition. Both Wilson (2006) and White (2017) are able to

appeal to external evidence to support their bias terms. In both cases, the key hypothesis was that of phonetic paradigm uniformity (§7.3.2), the tendency to avoid alternation that is phonetically too salient. Both researchers obtained their metrics of phonetic similarity from independent experimental work, which generated confusion matrices (pairwise error rates), from which can be inferred the phonetic similarity of pairs of sounds. Hence the charge of circularity does not loom as large for these studies. Even this work, though, awaits a general, language-independent theory for how the values for the crucial parameters μ_i and σ should be related to the phonetic similarity metrics.

Having stated these caveats, we conclude with an opinion: the game is worth playing, for it offers one of the few paths toward resolving some of the biggest current disagreements in phonological theory. To give one case, the framework of Evolutionary Phonology (Blevins 2004) attributes naturalness effects to diachrony, excluding them from the mind of the language learner. The approach of Substance-Free Phonology (Reiss 2017) gives a greater role to formal theory, but resolutely insists that appeals to phonetics, perception, or other forms of naturalness in phonology are ill-advised. If what we are suggesting, following Wilson's original idea, turns out to work in the long run, then the case that the language learner is aware of substance is strengthened: unnatural constraints are learned (that is, the data pattern is not a diachronic accident), and they are learned in a biased manner showing that the language learner is tacitly aware of substance.

7.7 Further topics in MaxEnt learning

7.7.1 Hidden structure

An important focus in phonological learnability is **hidden structure**. For present purposes we can define hidden structure operationally: if two surface phonological representations are formally distinct but are phonetically identical, then whatever aspect of representation distinguishes the two is hidden. For example, in metrical stress theory (Hayes 1995) it frequently arises that there are two distinct representations for the same phonetic stress pattern. Here is an example we will examine in more detail in Chapter 9. The Dutch word *locomotief* [ˌlokomoˈtiːf] 'locomotive' might be assigned either of the foot structures [(ˌloko)mo(ˈtiːf)] (begins with a binary foot, then a stray syllable) or [(ˌlokomo)(ˈtiːf)] (begins with a ternary foot). In the absence of further analysis and interpretation, these structures are phonetically equivalent, since they both place stress on the first and fourth syllables. As we will discuss in §9.5.2, analysts have opted for the first of these two structures, for it can be used to capture a distinction of optional vowel reduction (weak footed syllables reduce to schwa more readily than unfooted syllables). But it is only in light of this subsequent analysis that we can hope to resolve the hidden structure at issue; the feet themselves are inaudible.

The case of phonetically identical footings is only one of many, for it has been a major research activity of phonologists since the mid-1970s to propose hidden structures and explore the insights they might yield. Syllable structure (Kahn 1976 et seq.) is hidden, since in the absence of analysis, we have no a priori way to justify one of two possible syllabifications (compare, say, [ə.stɑ.nɪʃ] with [əs.tɑ.nɪʃ] for *astonish*). Moraic structure (Hyman 1985 et seq.) is hidden in that two different languages can moraify the same sequence differently (Hayes 1989).

Tier structure (Goldsmith 1976; A. S. Prince 1984) is hidden in that long segments and tonal spans can be assigned either multiply-linked or fully-sequential structure (see e.g. Hayes 1986 on "true" vs. "fake" geminates). The indices of Agreement-by-Correspondence theory (Rose and Walker 2004) are hidden. Lastly, from one point of view, underlying representations are also hidden. URs are, arguably, not the "deepest" form of lexical representation, since in lexically-listed allomorphy the same abstract morpheme can have more than one UR (Mascaró 2007). From this viewpoint, the relationship of a phonetic form like [sãĩn] with its deepest level of representation (the abstract morpheme "SIGN") is mediated by a hidden UR. In *SPE* (p. 234) this UR is claimed to be /sɪgn/;[64] in a concretist approach it could be /saɪn/. Either way, the choice is a matter for analysis and cannot just be read off the phonetics.

Looking beyond phonology, we note that all of morphological structure (morpheme membership of segments, bracketing) is hidden. The study of syntax requires hidden structure even at the most elementary level, and more advanced theories tend to adopt very rich conceptions of hidden structure.

We bring up hidden structure here because, as Tesar and Smolensky (2000) pointed out, it has important consequences for computational learning. The issue can be stated as follows: the child learning the phonology of her language hears only phonetic form, not structure, so the hidden structure must be *inferred*. In the MaxEnt context, we must devise a learning theory that establishes not just the correct weights, but also the hidden structures over which constraint violations are defined. The use of MaxEnt and similar probabilistic frameworks for learning hidden structure has given rise to a research literature; see (Jarosz 2006, 2013), Boersma and Pater (2016), and Lee et al. (forthcoming). For the specific case of learning URs, see Wang and Hayes (to appear) and numerous references cited there.

We cannot cover all this territory here, but for practical use put forth a simple method of hidden structure learning that suffices in some cases and runs on a spreadsheet. We will call it **cross-structure summation**.

Given a MaxEnt grammar, we let the predicted probability of a given phonetic form (assuming some UR) be the *sum* of the probabilities of all candidates that have that phonetic form, whatever their hidden structure may be. Thus, for the Dutch example above, the predicted frequency of the phonetic form [ˌlokomoˈtiːf] is the sum of the probabilities that the grammar assigns to the candidates [(ˌloko)mo(ˈtiːf)], [(ˌlokomo)(ˈtiːf)], and whatever other candidates might have [ˌlokomoˈtiːf] as their phonetic readout. In weighting the constraints, we seek only to match the frequency of the observable form. In the course of doing this, the computation will assign a precise probability to each candidate, e.g. [(ˌloko)mo(ˈtiːf)] and [(ˌlokomo)(ˈtiːf)].

In some cases, the model will "choose" a particular hidden structure, assigning it virtually all of the probability mass of the observed form. But it is also possible to arrive at a grammar that assigns substantial probability to more than one hidden structure for the observed form. Such a grammar fulfills the goal of completely matching native speakers' behavior, but psycholinguistic study is needed to determine if the variable hidden structure is actually

[64] The reason to posit /ɪg/ is to capture the alternation with *signature* [ˈsɪgnətʃɚ].

internalized by speakers.[65] For present purposes, we will assume a grammar is successful if it matches the correct observed forms.

Consider now an example. The stress pattern of Classical Latin is well known: in words of at least three syllables, stress falls on the antepenultimate syllable if the penult is light (e.g. [ˈsi.mu.laː] 'simulate-2nd sg. imp.') and on the penult if it is heavy (e.g. [a.ˈmiː.kus] 'friend'). In shorter words, stress is initial ([ˈka.nis] 'dog'). This pattern, which is reported to be invariant, has widely been analyzed with hidden foot structure. An adequate set of constraints, taken from Pater (2000), is given in (121); we include MaxEnt weights that suffice to derive the correct outcomes.

(121) *Constraints for Latin Stress*

a. FTBIN (weight 26.5)

   A foot must be a perfect moraic trochee (two moras, initial stress).

b. NONFIN (weight 50)

   Assess a violation if the final syllable is footed.

c. ALIGN RIGHT (weight 12.4)

   Assess a violation for every syllable that follows the main stress.

We also assume, without actual statement, some constraint that requires that words bear a stress, and a constraint that chooses the rightmost foot for main stress in words with more than one foot.

The tableaux in (122a) show in outline how (121) derives the right stress pattern.

(122) *Sketch tableaux for the Latin stress analysis*

		FTBIN	NONFIN	ALIGN R				
Input	*Candidate*	26.5	50.0	12.4	*H*	*eH*	*Z*	*p*
a. LLH	☞ (ˈLL)H			2	24.8	1.78E-11	1.78E-11	1
	LL(ˈH)		1		50.0	1.93E-22		1.09E-11
b. LHH	☞ L(ˈH)H			1	12.4	4.21E-06	4.21E-06	1
	(ˈLH)H	1		2	51.3	5.51E-23		1.31E-17
	LH(ˈH)		1		50.0	1.93E-22		4.58E-17
c. LH	☞ (ˈL)H	1		1	38.9	1.31E-17	1.31E-17	1
	L(ˈH)		1		50.0	1.93E-22		1.47E-05
	(ˈLH)	1	1	1	88.9	2.52E-39		1.93E-22

For (122a), [(ˈL L) H], with antepenultimate stress, we can obtain a well-formed (ˈL L) trochee and avoid footing the last syllable. This creates two violations of ALIGN RIGHT, but the latter

[65] For the case of syllabification, the psychologist Rebecca Treiman has carried out an extensive research program; one widely cited paper is Treiman (1985).

constraint is too weak to matter. For (122b) [L (ˈH) H], the only way to satisfy FTBIN and NONFIN is to make the heavy penultimate syllable into a monosyllabic foot. For (122c) [(ˈL) H], FTBIN must be violated, since the only way to make a binary foot would be *[L (ˈH)], which would violate stronger NONFIN. All of this is, in essence, classical OT constraint domination, rendered in MaxEnt as constraints with strongly different weights.

Our full analysis appears as spreadsheet **HiddenFootStructureInLatin.xlsx**. In the interest of rigor we included all 30 logically possible inputs (combinations of heavy and light syllables) up to length four, and every logical possible footing (excluding unfooted, and nested footings) for each input; 642 candidates total. What matters here is the way we used the Excel Solver to find the weights. Since the violations of FTBIN and NONFIN are determined by the hidden foot structure, we used cross-structure summation, which we will now explain. To begin, we note that the "phonetic" forms used here are weight sequences marked for position of the main stress, e.g. [ˈLLH] is the "phonetic" form of [(ˈLL)H]. For now we will ignore secondary stress, but see discussion below.

To illustrate the method of cross-structure summation, we quote the full spreadsheet for two representative inputs; for brevity rows and columns have been renumbered relative to the original format.

(123) *Learning Latin stress with cross-structure summation: a partial tableau set*

Key: A = Antepenultimate, P = Penultimate, F = Final

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Stress	FTBIN	NONFIN	ALIGNR	H	eH	Z	*p*	Surface	Sum *p*	Ln(Sum *p*)	LL
2			26.5	50.0	12.4								
3	Input: /LLH/												
4	(LLH)	A	1	1	2	101.2	1.1E-44	1.8E-11	0.000	F: 0	0.000	−24.51	**0.000**
5	(LL)H	A			2	24.7	1.8E-11		1.000	P: 0	0.000	−14.12	
6	(L)LH	A	1		2	51.2	5.5E-23		0.000	A: 1	1.000	0.00	
7	L(LH)	P	1	1	1	88.9	2.5E-39		0.000				
8	L(L)H	P	1		1	38.9	1.3E-17		0.000				
9	(L)(LH)	P	2	1	1	115.4	7.8E-51		0.000				
10	(L)(L)H	P	2		1	65.4	4.1E-29		0.000				
11	LL(H)	F		1		50.0	1.9E-22		0.000				
12	L(L)(H)	F	1	1		76.5	6.0E-34		0.000				
13	(LL)(H)	F		1		50.0	1.9E-22		0.000				
14	(L)L(H)	F	1	1		76.5	6.09E-34		0.000				
15	(L)(L)(H)	F	2	1		103.0	1.9E-45		0.000				
16	Input: /HHL/												
17	(HHL)	A	1	1	2	101.2	1.06E-44	8.43E-06	0.000	F: 0	0.000	−63.42	
18	(HH)L	A	1	0	2	51.2	5.51E-23		0.000	P: 1	1.000	0.00	
19	(H)HL	A	0	0	2	24.7	1.78E-11		0.000	A: 0	0.000	−13.09	
20	H(HL)	P	1	1	1	88.9	2.52E-39		0.000				
21	H(H)L	P	1	0	1	12.4	4.21E-06		0.500				
22	(H)(HL)	P	1	1	1	88.9	2.52E-39		0.000				
23	(H)(H)L	P	1	0	1	12.4	4.21E-06		0.500				
24	HH(L)	F	0	1	0	76.5	5.99E-34		0.000				
25	H(H)(L)	F	1	1	0	76.5	5.99E-34		0.000				
26	(HH)(L)	F	1	1	0	103.0	1.86E-45		0.000				
27	(H)H(L)	F	0	1	0	76.5	5.99E-34		0.000				
28	(H)(H)(L)	F	1	1	0	76.5	5.99E-34		0.000				

Most of this apparatus should be familiar from earlier chapters: columns A-I form a standard MaxEnt calculation of the probabilities of each candidate, where candidates are defined to include foot structure. For example, 7A L(LH) and 8A L(L)H, though they both depict penultimate stress, are not the same candidate for this purpose. What is new is in column K: clicking on cell K4 to inspect its formula, one will find that it sums the probabilities of all candidates for the first input that derive antepenultimate stress, which in this case would be cells I4:I6. This is what we mean by "cross-structure summation". The same computation is carried out for all the other observable forms in cells K5:K6 and K17:K19.

In Excel, cross-structure summation can be carried out, should you wish, simply by using the normal SUM() formula. For (123), we used a slightly more convenient approach based on the

SUMIF() formula. Specifically, the formula occupying cell K4 in our spreadsheet is "=SUMIF(D4:D15, L4, I4:I15)", which means, "find all the cells in the range I4:I15 for which the analogous value in column D matches cell L4; and compute their sum."

The remainder of the computations are just ordinary maximum-likelihood weight-fitting: we maximize the log likelihood of the data, construed as surface stress patterns. The empirical frequencies of these surface patterns are given in column J, the surface probabilities assigned by the model were formed by addition (as just described) in column K, the logs of these probabilities are in column L, and the log likelihood of the data is computed (with SUMPRODUCT(), as usual) in cell M4. The final grammar was obtained by running the Excel Solver on the full spreadsheet, maximizing M4 by adjusting the values of the constraint weights in C2:E2. The grammar works very close to perfectly, assigning essentially all the probability mass to the correct stress patterns for all 30 inputs.

It is instructive to see what the grammar does with respect to hidden structure. Since in this case we have pedagogically chosen to pay no attention to predictions about secondary stress, we see that the 1.00 probability assigned to penultimate stress for A16 /HHL/ is divided equally between candidates A21 [H(H)L] and A23 [(H)(H)L]; these differ in that the latter predicts a secondary stress on the first syllable. The probabilities of these two candidates come out equal because they have the same constraint violations. For certain other inputs, the surface probability mass is divided into four hidden structures; e.g. for /HHHL/, each of the candidates [HH(H)L], [H(H)(H)L], [(H)H(H)L], and [(H)(H)(H)L] is assigned a probability of 0.25.

In our limited experience, the method of cross-structure summation has proven sufficient for finding effective grammars where there is hidden structure. Two papers that use the method to navigate a thicket of possible footings are Moore-Cantwell (2020) for English stress and Katsuda (2025) for Japanese pitch accent.

Before concluding, we need to note a major issue in this area. Unlike in ordinary MaxEnt modeling, cross-structure summation is not guaranteed to find the best answer (highest log likelihood). The problem is related to the discussion of convexity in §4.4.2: when there is hidden structure, we lose the guarantee that the search space is convex. Our Latin example evidently is not prey to this problem, but other examples are. We mention here the case of "TS44," a fictional-but-plausible language invented by Tesar and Smolensky (2000). The stress pattern of TS44 is "penultimate except in light-initial disyllables." Given suitable training data, we find that cross-structure summation instead learns the trivial pattern "always penultimate". [66]

Dealing with local maxima is definitely a non-trivial problem; for instance, the OT learners posited by Tesar and Smolensky (2000) and Jarosz (2013), both computationally impressive, fail on TS44 for this reason. In dealing with local maxima, a "rough and ready" strategy is to try different starting weights (e.g., Hayes and Wilson 2008 use 1 everywhere), or even to assign starting weights randomly to each constraint. This allows the algorithm to begin search in a different part of the search space, and thus potentially avoid a local optima which it would get

[66] For diagnosis see Jarosz (2013:63-64), who notes that 60/62 of training forms have penultimate stress.

stuck in when the weights start at 0. The development of methods that can effectively search nonconvex spaces continues to be an active research area.

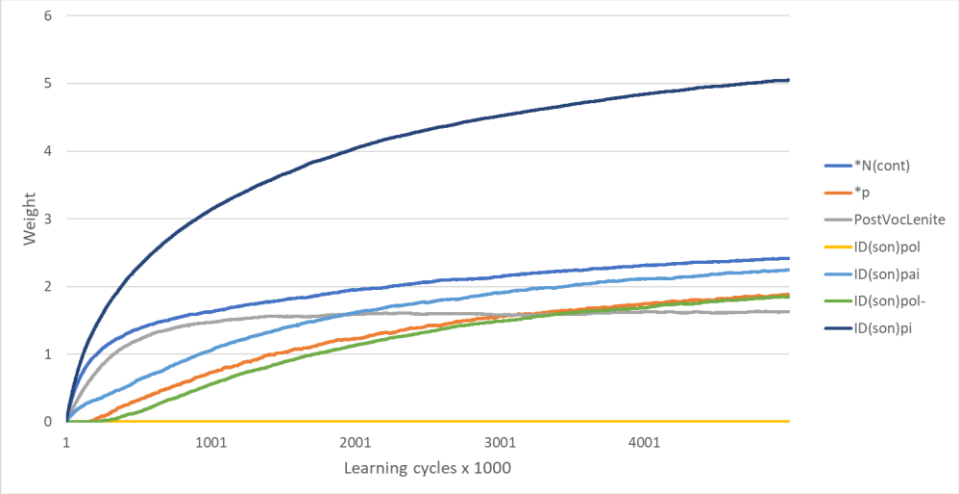
7.7.2 Modeling the child's course of progress

Following the conception given in §7.1, the generative linguistics of the future will leap out the cubicle: emerging from the conceptual stage to full implementation, it will combine the right theory of grammar, a realistic and implemented learning theory, and data corpora that closely match childhood, to learn grammars that, in the end, closely match the linguistic knowledge of native speakers as checked by elicitation and experiment. But this should not be the only research goal: the learning model should also closely match the *path of progress* made by infants and children as they gradually approach the adult grammar. Such progress can be monitored empirically by recording the children's productions and by checking their perceptions and intuitions in experimentation. For extensive coverage of this empirical work, see (Tessier 2016).

On the theory side, this research program requires models that make explicit predictions about the acquisition path. In particular, we could like to consult grammars that are in some sense intermediate along the route to full acquisition. However, the task of training these grammars have led scholars to ponder what sort of learning algorithm that it needed. The child encounters the data needed to construct her grammar incrementally over time. Yet the weighting algorithms that are normally used for MaxEnt require that all the data be present at once — they are "batch" algorithms. The rationale for using a batch algorithm is pointed out by Jäger (2007): the basis of weighting adjustment (§4.4.2) is to calculate the *overall* Observed and Expected violation counts for all the constraints, which gives the local slope on the hill of log likelihood, which points to the best move in the iterative improvement of the weights.

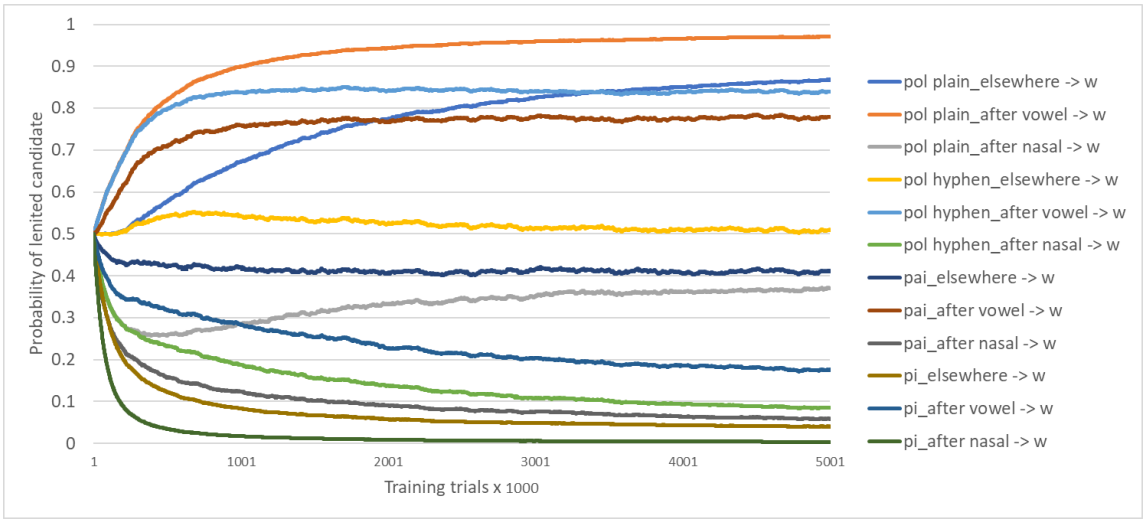
As Jäger showed, there is an alternative approach that works serially. The key is to take a *random sample* from the current output distribution for the input just encountered, and use the fact of whether it matches the currently observed output (a kind of micro-measure of Observed vs. Expected) to make adjustments in the weights. Jäger's procedure is an instantiation of the Perceptron Update Rule (Rosenblatt 1958), which also inspired the Gradual Learning Algorithm created for Stochastic OT (Boersma 1998). This procedure is less efficient than the batch method, but in time will also achieve the right weights within the convex search space. Jäger's algorithm is implemented in Praat (Boersma et al. 2026), in OTSoft 2.7 (Hayes et al. 2026), and in GLaPL (Moore-Cantwell 2026). These programs can be set to track and record the weights as they evolve toward the final, stable values that maximize likelihood, with the changes in weights that occur after exposure to each learning datum. So, for instance, using OTSoft we can see the weights for our K5 simulation in Chapter 2 evolving over simulated time, as shown in Figure 39 (five million learning iterations, initial weights set at zero).

Figure 39: The gradual approach of the weights to their final values for K5



Equipped with a time series like Figure 39, we can use the evolving weights to calculate the probabilities of the candidates themselves and plot these over time. For K5, the probabilities (of lenited [w] candidates) evolve as in Figure 40.

Figure 40: Approaching the observed frequencies for [w] with serial learning in K5



Graphs like 39 and 40 are employed widely in work that adopts the serialist conception of acquisition; see for instance Boersma and Levelt (2000), Jäger (2004), and Moore-Cantwell and Pater (2016).

In doing serialist work of this sort, there are theoretical choices that must be made that do not arise in batch learning. For instance, we need to establish a schedule of "learning rates," specifying the the size of the adjustments that are made in the weights in response to data (the simulation shown in the graphs used 0.001 throughout). Further, we must establish the initial

values for the constraint weights (defining the so-called "initial state"; here we simply adopted zero).

In OT it has been proposed (e.g. Smolensky 1996; Gnanadesikan 2004) that the initial state should play a crucial role in learning, notably with the goal of achieving restrictiveness by having Markedness constraints start out ranked above Faithfulness. In §6.1.3 we critiqued such proposals as superfluous, at least under an approach like MaxEnt that uses likelihood maximization — in such an approach the intended restrictiveness falls out simply from seeking an accurate grammar. But proposals about the initial state take on new life, perhaps, in the context of tracking the path of the weights along the path toward the final grammar, revealing learning biases as manifested in early, imperfect grammars. Note that as applied to MaxEnt, such initial weightings will *not* in general influence the stable solution achieved under complete learning. This is because the search space is convex and implies a unique solution, irrespective of the search procedure used (this is why we cover the mu-sigma method of imposing UG biases, in §7.5 above).

It should be noted that the conception of serial learning just illustrated involves an idealization unlikely to hold true in the general case; namely that the individual training items arriving at the child's ears will be interpretable on their own. For instance, there are cases where this is plainly not so; often, a surface form is fully meaningful for purposes of learning when the learner knows its only in the context of knowing its underlying form. Thus a Catalan-learning child who hears the word [san] 'holy-masc.' will learn only superficial things (e.g. a new lexical entry, or that [n] is a legal coda); but if the same child knows the underlying form /sant/ (cf. feminine [sant-ə]), then [san] becomes informative for purposes of learning the phonological process that drops [t] word-finally after [n] (Wheeler 2005:§7.2). Thus, in some sense, learning cannot be fully serial; a small "batch" of data needs to be appealed to; specifically, the masculine-feminine pair {[san], [santə]}. Indeed, Tesar and Smolensky (2000) suggested that the grammar must optimize over whole paradigms; and while Tesar (2014) does not optimize not over paradigms themselves, he does employ a stored buffer of inferred ranking information ("ERCs").

As noted earlier (§3.4), there is psycholinguistic evidence that paradigms are in principle available to language learners in phonological acquisition; for instance, lexical-decision experiments indicate that humans really do memorize a substantial number of paradigmatic forms — even in regular paradigms, where one might think memorization unnecessary; see e.g. Sereno and Jongman (1997), and Baayen et al. (2002). Memorization of regulars provides just the sort of "batches" (like {[san], [sant-ə]}) that are needed to drive batch learning.

For syntax and phrasal phonology, batch learning would require that the child memorize full sentences. Whether children actually do this is not currently known, but there are hints in this direction; see in particular Gurevich et al. (2010).

In batch learning, learning paths like Figs. 39 and 40 can be approximated by running the model repeatedly on a series of batches of data, intended to mimic the child's vocabulary as it expands over time. For an acquisition curve created by modeling on this basis (acquisition of the suffixes /-z/, /-ɪŋ/, and /-d/) see Breiss et al. (2025:23).

We anticipate that the experimental and modeling work of the future will shed light on the extent to which children use serial or batch learning during language acquisition.

Chapter 8: A survey of research areas to which MaxEnt grammars are applicable

8.1 Research areas within phonology

MaxEnt seems to us to be applicable to a large subset of the current research questions that face the field of phonology, as we illustrate in (124) below. We include both MaxEnt work and research deploying other stochastic constraint-based frameworks.

(124) *Some questions on the current phonological research scene*

a. The role of **naturalness**: are natural constraints favored by learners? The use of bias modeling to address this question was a major topic of Chapter 7. Beyond the modeling issue, we also seek to learn how the language learner comes to know what is natural (see e.g. Hayes, 1999; Steriade, 2000), and the precise character of the needed naturalness principles.

b. The right approach to forms of alternation that go beyond the capacity of the standard theory of Faithfulness constraints (McCarthy and Prince 1995): **opacity** (Kiparsky 1973) and **saltation** (Lubowicz, 2002; Itô and Mester, 2003; Hayes and White, 2015). For experimental work on saltation modeled in MaxEnt, see Wilson (2006) and White (2017).

c. How **paradigms** (both derivational and inflectional) influence phonological alternation (see §7.3.2). Here, there is a major bifurcation between grammatical-architecture theories using levels (Kiparsky 1982 et seq., Bermudez-Otero 2018), and constraint-based theories using Output-Output correspondence (Benua 1997, Pater 2000). We think that MaxEnt work has provided support for the latter approach, in that it permits the modeling of cases in which the influence of a base form on a derived form is gradient; see Martin (2011), Breiss (2024) and Breiss et al. (2026). Inflectional paradigms involve particularly rich patterns, with bases occurring in multiple layers (e.g. Bonami and Boyé, 2002), and inflectional paradigms affecting derivational paradigms (Steriade, 2008; Steriade and Yanovich, 2015; Steriade (2026).

d. How syntactic structure (and many other factors) influences the patterning of phrasal phonology, particularly in the formation of phonological phrasing (Selkirk 1978 et seq.). Since phonological phrasing is well known to involve optionality (see e.g. Werner et al. 2022), this seems to us an area where application of gradient frameworks like MaxEnt might be fruitful. For classical OT work, see Nagahara (1994), Truckenbrodt (1999), and Elfner (2018). Gabriel (2010) uses Stochastic OT to model variable word order and variable prosodic structure together.

e. How children learn phonotactics and alternations, and the relation between the two forms of learning — are they all one grammar, as in classical OT, or separate systems, of which the first facilitates the learning of the second? See Jarosz (2006), Hayes and Wilson (2008), Martin (2011), Chong (2021), and Xu (2026).

f. **Hidden structure**, discussed in §7.7.1.

g. **Exceptional forms** and how they are learned together with the grammar. This topic invokes the challenge of how the grammar can be learned simultaneously with the lexicon, where exceptionality is widely thought to reside; see Zuraw (2000, 2010), Linzen et al. (2013), Moore-Cantwell and Pater (2016), Zymet (2018), Hughto et al. (2019), and Jarosz et al. (2025).

h. The role of **frequency** in phonology, a topic covered in general terms in work such as Bybee and Hopper (2001), Pierrehumbert (2000), and Phillips (2006). Two particular issues addressable with MaxEnt are attestation bias, discussed above in §7.3.4 and §7.6.1; and how applicability of a phonological process depends on word frequency (Coetzee and Kawahara 2013; Smith and Moore-Cantwell 2017; Kilbourn-Ceron et al. 2020; Jarosz et al. 2025).

8.2 Research areas related to phonology

At this point we move on to topics outside the core of traditional phonological research but certainly of interest to us. A theme for these areas is that MaxEnt could be said to "open doors," bring formalization and insight to areas previously hard to approach formally.

8.2.1 Loanword adaptation

The process of loanword adaptation involves conflicting principles: the adherence to the phonotactics of the borrower's native language (L1) goes against with the urge to render the source language (L2) faithfully. The Faithfulness principles for loanword adaptation might be relatively abstract in character, or related more closely to the phonetic form of L1 and L2 (Peperkamp and Dupoux 2003; Davidson and Shaw 2012). It is suggested by White (2003) that UG principles by themselves influence adaptation, independently of the particular restrictions of L1. Lastly, in written languages it seems clear that faithfulness to orthography rather then phonetic form plays an important role (Vendelin and Peperkamp 2006; Daland et al. 2015). For a recent literature survey see Smith (2025).

The competition between these various principles is often likely to involve variation in output forms. To give an anecdotal example, the authors of this book render Japanese *tsunami*, more or less at random, as [suˈnɑmi] or [tsuˈnɑmi]; the former prioritizes Markedness principles, the latter faithfulness to the Japanese pronunciation. Loanword adaptation study has sometimes been carried out focusing on just one of the factors affecting adaption, but we feel that establishing a blended model with distinct, conflicting constraint families is likely to offer more complete understanding. We illustrate with two recent MaxEnt studies.

In Glewwe (2021), covering Mandarin loanword adaptation, the effects of phonological form in the source word — for instance obstruent voicing or stress pattern — are demonstrably present in the set of existing loanwords, but they are subtle, and to prove their presence it is necessary to incorporate these constraints into a model that is dominated by stronger constraints; notably those that require use of the normatively-enforced "standard syllables." The example

illustrates a point made in §2.6.1: perturber effects are important but undetectable in a framework that cannot model them.

Katsuda (2025) addresses how pitch accent contours are assigned to English loanwords in Japanese. Previously established principles include foot-based Markedness constraints (Itô and Mester 2016) and faithfulness of Japanese accent location to English stress (Shinohara 2000; Kubozono 2006). Katsuda adds a novel kind of perturber, which invokes the L1 speaker's internalized theory of L2 phonology. This idea can explain "hyperforeignisms" (Janda et al. 1994), where L2 speakers construct phonological representations that, while probabilistically favored within L1, happen to be wrong for a particular word. Thus Japanese [ˈʃiatoru] for *Seˈattle* happens not to match the English pronunciation, but is a good guess in light of the English stress pattern as a whole. This interesting perturber would not be discoverable or justifiable without MaxEnt or a similar framework.

8.2.2 Lexical class assignment

Languages sometimes divide their vocabularies into **strata** of various types, including etymological, expressive, syntactic-class based, and even morphological classes (e.g. conjugation and declension classes). These have phonological consequences, which have long been a topic of research.

Etymology-based strata. These arise due to systematic borrowing from another language. Well-studied examples include the Sino-Japanese stratum of Japanese (McCawley, 1968; Itô and Mester, 1999), the Latinate and French strata of English (Chomsky and Halle, 1968; Dabouis and Fournier, 2022), and the stratum of fingerspelled English borrowings in American Sign Language (Brentari and Padden 2001). These strata can lend coherence to the morphophonological system, as in the Latinate system of English, where a set of affixes (such as *-ity* or *-ation*) are associated with particular phonological processes, such as Trisyllabic Shortening (e.g. *sane ~ sanity*). Strata affiliated with borrowings tend to have more permissive phonotactics than those associated with core 'native' vocabulary (see Pinta and Smith 2017), but the opposite can also hold true. An example again comes from the Latinate stratum of English, which Hayes (2016) analyzes using a MaxEnt system. In this system, a word like ?[wɛpəˈtʃeɪʃən] "*wepechation*," marked as Latinate because of *-ation*, sounds odd because it contains non-Latinate [w] and pretonic [tʃ].

Lexical classes related to grammar. Perhaps the most familiar case is grammatical gender, which in many languages involves more than just memorization; native speakers tacitly know how to assign gender to novel forms, following a variety of principles. Poplack et al. (1982) demonstrate that this process is probabilistic, even though individual loanwords are quickly lexicalized as invariant through tacit community consensus. (Their work anticipates the listing-cum-generation approach to lexical exceptionality covered above in §3.4.) The math used by the Poplack et al. appears to be that of MaxEnt (see §4.6), with weighted constraints favoring particular genders in particular contexts.[67] Similar models can capture the particular phonotactics of parts of speech (e.g., verbs different from nouns); for a survey see Smith (2011). Further

[67] For the most recent model for predicting German gender, Fedden et al. (2025), see §9.5.5.

afield, lexical classes can be extended to identifying the stems that take particular allomorphs (Becker 2009) or even to vowel harmony, as proposed by Rebrus et al. (2023).

Implementations. Formally, lexical class assignment can in principle be pursued with MaxEnt in two ways. The most direct approach is to set up a separate auxiliary grammar in which every surface word is affiliated with two candidates indicating stratal membership, like [+Latinate] and [−Latinate]; as in Hayes (2016, 2022). Another approach is to have separate phonotactics for the two strata; the classification of a word is based relative probability difference it receives as interpreted as a member of each of the two strata. The latter system is pursued by Gouskova et al. (2015) and by Magpale (2026).

8.2.3 Phonesthemes and sound symbolism

Probably all languages include within their vocabularies a set of phonologically expressive words, often said to contain "phonesthemes" (small phonological chunks that convey the expressive character). For instance, *snot* is an expressive word of English, with evaluative connotations not evoked by its neutral near-synonym *mucus*. *Snot* contains the phonestheme *sn-*, a short, vaguely-meaningful string present in other expressive words that involve the nose, like *snoot*, *snout*, *sniffle*, and *snort*. In some cases the expressive character of a word is held to be natural in some sense, as in the widespread affiliation of high front vowels with smallness. The expressive words of a language can be singled out and analyzed with phonological constraints in a MaxEnt system; see for instance Hayes (2022)(which covers the *sn-* phonestheme), Shih et al. (2018), and Kawahara (2020, 2021a, 2021b).

8.2.4 Probabilistic inference of underlying representations

Often, when a speaker first encounters a word, it appears in a paradigmatic form that has undergone a phonological neutralization. Thus, if the wug-word [xlop] is presented to speakers of Dutch (as in (n.d.; Kerkhoff 2007)), they cannot know in principle whether it is /xlob/, with the underlying /b/ realized as [p] by a process of Final Devoicing; or if it is rather /xlop/, with the UR the same as the surface form. Only when suffixed paradigm members with a vowel-initial suffix (where Final Devoicing is inapplicable) are provided, can a Dutch speaker determine the correct UR. However, as Ernestus and Baayen (2003) first found, there is evidence that speakers are willing to *guess* the UR, even before they have encountered these additional forms. These guesses, which are surprisingly accurate, are based on statistical phonological generalizations about the Dutch lexicon. For instance, according to Ernestus and Baayen's study of a lexical corpus, if an isolation stem in Dutch ends in [p], and the [p] is preceded by a short vowel, then the lexical probability that the UR will be /b/ is 0.22. If, in contrast, the final segment is [s], preceded by a long vowel, the lexical probability of the voiced UR /z/ is much higher, 0.82. Ernestus and Baayen's experimental work suggests that Dutch speakers are tacitly aware of these distinctions, because the guesses they provide on wug words are not random, but match the statistical patterns of the existing lexicon. Put in different terms, UR-guessing appears to be a frequency-matching operation, just like the lexical frequency-matching observed for Tagalog phonological alternations in Chapter 3.

Ernestus and Baayen try out several models to explicate the UR-guessing ability shown by their participants, among them a MaxEnt model of the type presented in this book. (It is not the best-performing model, but works fairly well.)

Subsequent work on other languages has confirmed the ability of native speakers to frequency-match when guessing URs; a set of references was given on p. 61 above. For an explicit formal theory of UR-guessing, see Kuo (2024).

A harder problem for linguists is to explicate the learning of underlying representations as it occurs in the first instance, during infancy and childhood. The problem is harder because URs must be learned in conjunction with the phonological system that derives the surface forms (which for adult UR-guessing is already known). This challenging problem has attracted a substantial literature; see e.g. Tesar (2014), Rasin and Katzir (2016), and for MaxEnt approaches Jarosz (2006) and Wang and Hayes (in press).

8.3 Metrics

Metrics is a discipline closely related to phonology that studies how phonological representations are employed to embody rhythmic patterns in poetry and song. The MaxEnt grammars suited to metrics are essentially similar to the phonotactic grammars studied in Chapter 6: we want to characterize (gradient membership in) a category consisting of the possible lines of iambic pentameter, possible folk song quatrains, and so on. We suggest that the use of MaxEnt is particularly useful in this domain because it permits statistical testing. For instance, Hayes et al. (2012) offer MaxEnt analyses of corpora of iambic pentameter from Shakespeare and Milton, with the constraints drawn from several decades of generative work on this meter. It emerges that not all constraints perform alike: many of them "pass the MaxEnt test" perfectly well, but others — including perhaps the most famous constraint, the "Stress Maximum Constraint" (Halle and Keyser 1966 et seq.) turn out to have zero weights in the best-fit models. Hayes and Schuh's (2019) study of Hausa verse is an example of treating a fairly intricate system with a very small GEN (cf. §6.2 above). Henriksson's (2022) MaxEnt study of Ancient Greek meters offers a novel way to control for the frequency of the words and phrases that are incorporated into the meter.

8.4 Phonetics

We refer here to the research tradition of generative phonetics, in which explicit formal grammars are created that map from surface phonological representation to the actual contours in physical and acoustic space that manifest speech concretely. This line of research dates from the 1980s with rule-based work by Pierrehumbert (1980) and others; but in more recent times has been extended fruitfully to cases in which phonetic realization involves the resolution (usually through compromise) of conflicting phonetic constraints, often the constraints specifying the ideal realization of neighboring segments. Key literature here includes Flemming (2001) and Flemming and Cho (2017), who employ Classical Harmonic Grammar (§2.5.1). A next step, still a pioneering area, is to acknowledge the widespread presence of free variation in phonetic form, and use MaxEnt grammars to model this variation with probability distributions. For two such efforts, see Lefkowitz (2017) and Hayes and Schuh (2019).

8.5 Morphology

Ryan (2010) uses MaxEnt constraints to analyze variable affix order in Tagalog. Our sense is that this language is unusual in its freedom of affix order; in most languages affix order is "crystallized," limiting the usefulness of gradient frameworks. Why morphology should be often, but not always, crystallized is an interesting question.

Morphology is more often stochastic, perhaps, when affix allomorphs compete with one another (see Becker (2009) for a Modern Hebrew case), or indeed where a morphological construction competes with a syntactic construction, as in English comparatives (*angrier* vs. *more angry*). A MaxEnt analysis of this case is given in Smith and Moore-Cantwell (2017).

8.6 Syntax

The literature applying MaxEnt and related ideas to syntax is not large, but we think it is informative to the extent that it establishes that syntax shows data patterns that are not essentially different from the rest of language in being suited to MaxEnt analysis. The discussion here partly updates the survey of Manning (2003), a work which concludes with the suggestion that syntacticians should try MaxEnt. The reader is also referred to Francis (2022) for extensive helpful discussion and literature review.

To start, we note that the syntactic theoretical mainstream, rooted for decades in the work of Noam Chomsky, seems to be the most prominent area of linguistic theory in which MaxEnt ideas are difficult to implement. Notably, the mainstream rejects the atomization of grammars into constraints (favoring different atomizations) and has for the most part adhered to the practice of modeling idealized, "either-or" data, rather than engaging with gradience. So it is perhaps unsurprising that the research we have been able to find that relates to MaxEnt syntax mostly falls outside the Chomskyan mainstream. Naturally we would be curious if in future work theoreticians could find ways to make mainstream syntax "MaxEnt-able."

In what follows we cover literature offering both analysis using MaxEnt itself, or in some cases, just parts of the MaxEnt math. We focus on instances in which syntactic phenomena pattern in ways suited to analysis in MaxEnt.

8.6.1 Gradience in syntax

As a preliminary point, we note that it would not make much sense to embark on a program for MaxEnt syntax unless syntactic competence is indeed gradient. That syntactic knowledge can be gradient is a view that is endorsed by some, but not all researchers in this area, and we discuss the issue in §9.1 below. For now, it suffices to note that the particular studies we cite here all offer explicit analysis of gradient data, either from experimentation or from corpus counts.

8.6.2 Harmony as a model of syntactic well-formedness

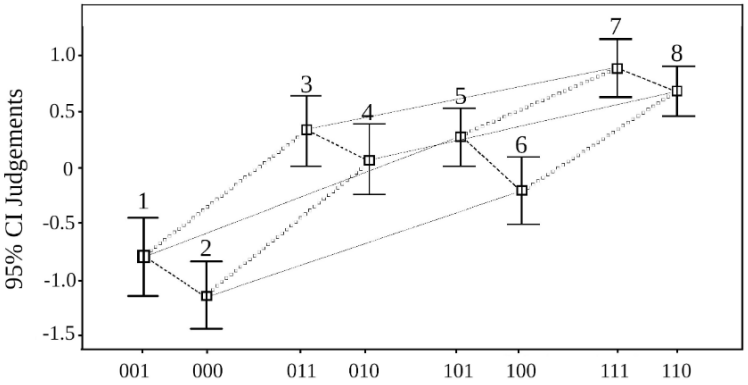
Syntactic judgments are often collected informally in the traditional context of the consultant session. Yet there is also an ongoing tradition of obtaining judgments adopting the

experimental methods employed elsewhere in the cognitive sciences (e.g. multiple participants, presentation of material by computer not experimenter, statistical testing); for an early review see Schütze (1996/2016) and a more recent update Francis (2022). The advantage commonly attributed to formalized experimentation is that it gives us more confidence in the validity of the often-subtle patterns of gradient judgments that are the subject of analysis; see e.g. Gibson and Fedorenko (2013).

What has emerged from work of this kind is a consistent pattern: *it is feasible to model gradient syntactic judgments using Harmony*. The crucial experimental findings are twofold. First, the syntactic constraints adopted turn out to have specific strengths, such that each constraint imparts a characteristic decrement of well-formedness when it is violated. Second, experimenters have found in modeling their data that there are ganging effects (§4.1.2), with increasingly worse degrees of well-formedness when more than one constraint is violated. These basic findings initially emerged from two distinct research programs of modeling and experimentation, one carried out by Keller and colleagues (Keller 2000, 2006; Sorace and Keller, 2005), the other by Featherston (2005b, 2019).

A striking graph from Featherston (2019), illustrating the points just made, summarizes the results from a ratings experiment on German multiple *wh*- questions, such as *Welcher Arzt hat dem Patienten welche Arznei gegeben*?, 'Which doctor gave the patient which medicine?' To model the intricate rating patterns obtained in his experiment, Featherston used three constraints, and the graph copied below as Figure 41 demonstrates the effects of combinations of constraint violations on participants' judgments. On the horizontal axis, "010" means "a sentence that violates the first and third constraints," and similarly for the other codes (000 violates all constraints, and 111 violates none).

Figure 41. Results from Featherston (2019) illustrating constraint strength and ganging



As can be seen, every constraint inflicts roughly the same decrement of well-formedness, no matter what the violations are of any other constraints; so for example the difference between 001 and 000 is the same as the difference between 111 and 110 and so on. The amount of the decrement is constraint-specific. Moreover, ganging is manifest: item 2, with three violations, is the worst, followed by the three items with two violations (1, 6, then 4), followed by the three items with just one violation (5, 3, 8), followed lastly by violation-free 7. These are just the patterns that we would expect if the participant's ratings directly reflect the Harmony values that would be assigned in a constraint-based harmonic grammar; e.g. a MaxEnt one.

Both Keller and Featherston worked at developing appropriate formal models to describe such findings, and arrive at solutions involving Harmony in some form. (Neither called it by that name, which when they began their work had not yet become widespread in the field). Harmony, by its very mode of calculation, embodies both differences in constraint strength and ganging effects.

The program of modeling ratings with Harmony has been extended further by An (An 2020; An et al. submitted), who using a 0-10 rating scale, again find it possible to achieve a good match to ratings data by using Harmony. Their modeling covers the results of eleven experiments on agreement in conjoined structures in French. As with their predecessors, the basic principles of varying constraint strength, ganging, and Harmony are shown to be closely related to well-formedness ratings. An et al. specifically relate their work to Harmonic grammar as practiced elsewhere in linguistics.

While in the studies just mentioned, Harmony in syntax has involved experimentally gathered well-formed ratings, it has also proven possible to diagnose the effects of Harmony in corpus data. Irvine and Dredze (2017) use MaxEnt and related Harmony-based models to study how the factors of topic and focus stochastically determine word order in a large naturalistic corpus of transitive sentences of Czech. The surface forms they attempt to predict with their modeling are the six logically possible orderings of subject, verb, and object (SVO, SOV, OSV, OVS, VSO, and VOS). Here again, the key findings of varying constraint strength and ganging emerge from the data. In this study, the MaxEnt formula (19) is shown to be effective in relating Harmony to observable data frequencies — the frequencies are best predicted not by Harmony per se but by the MaxEnt rescaling of Harmony.

8.6.3 Frequency-matching in syntax

In our discussion of MaxEnt in phonology, a key element has been the possibility of employing MaxEnt grammars to model the ability of speakers to form grammars that (modulo UG bias effects; Chap. 7), match the frequencies of the phonological types they encounter in the acquisition data; see §3.5 above. Does syntactic learning match frequency? There are two cases to consider.

First, there is the possibility that speakers fine-tune their syntactic systems to match the frequencies that prevail in their own speech community. Evidence in support of this view comes from the work of Bresnan and Ford (2010) and Szmrecsanyi et al. (2017), comparing the use of double-object and prepositional-phrase datives (*give Mary the book*, *give the book to Mary*) in different national varieties of English (American, British, Canadian, and New Zealand). No two dialects are exactly the same.

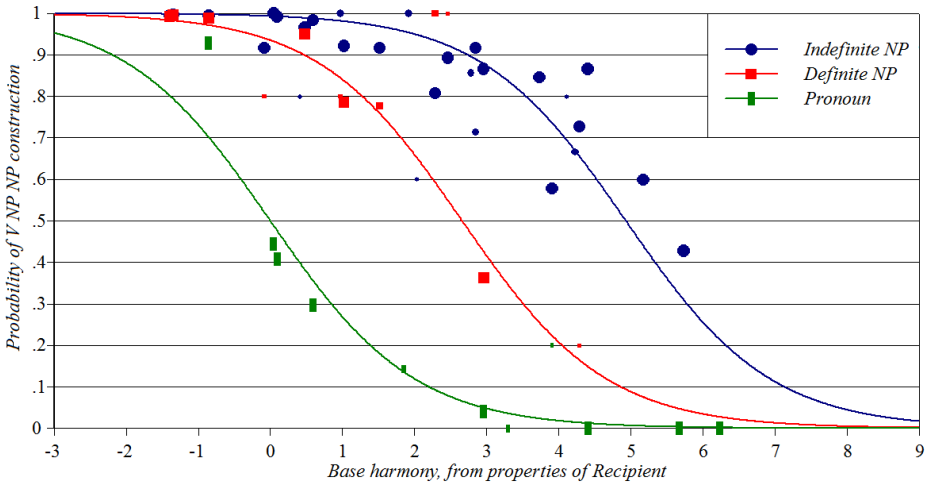
Second, there is the question of whether gradient well-formedness in syntax might be founded in frequency, as elsewhere in grammar (cf. §3.5, §6.1.2): some sentence types are accepted only reluctantly by native speakers because the evidence to support their grammaticality is so sparse in the acquisition data. Francis (2022:ch. 5) reviews several experiments that reveal a connection between corpus frequency and well-formedness ratings.

We see in this case the potential for future research. In the modeling efforts of Keller, Featherston et al., and An et al., discussed above, the authors adopted the expedient of directly fitting the experimental data, using regression to set the constraint weights so that the Harmony values would match the experimental findings. This is, of course, an idealization, since real children have no access to their parents' grammatically judgments. But since, as we have seen, MaxEnt learning can match the frequencies of grammatical patterns as they appear in a corpus, a more ambitious plan would to see if experimental results like that obtained by these authors can be modeled by training the constraint weights against representative corpus data. Thus we ask, what sort of modeling might be appropriate to project from the rareness of a syntactic pattern to its marginal acceptability, as measured experimentally? The methods covered in Chapter 7 above, including the possibility of learning bias, might be applicable.

8.6.4 Perturbers and shifted sigmoids in syntax

The work of Bresnan and Ford (2010) and Szmrecsanyi et al. (2017) just mentioned uses the MaxEnt math to make its key predictions, and uses a variety of different constraint types. Since the output choice being made is binary (*give X to Y*, *give Y X*), we can plot the corpus data to see if they fall into the shifted sigmoid patterns described here (§2.8, §3.3, §9.3), with baseline constraints and perturbers. This is done in Hayes (2022b), who plots the shifted sigmoid set (based on Szmrecsanyi et al. 2017) reproduced in Figure 42. Here, the separate sigmoids represent the status of the theme NP (role *X* in the formulae just given): indefinite full NP, definite full NP, or pronoun; and the horizontal axis depicts the Harmony value contributed as a result of the recipient NP (role *Y*). The latter constraints are based on definiteness, animacy and other factors.

Figure 42: Shifted sigmoids in syntax: probability of the V NP NP *outcome in dative sentences*



Further shifted sigmoid plots from this work may be viewed in Hayes (2022b).

Interestingly, the existence of shifted-sigmoid syntactic patterns for syntax was demonstrated long ago ago by Kroch (1989), working not with synchronic data but with diachronic patterns, wherein one syntactic realization of the same verbal intent comes to replace another over time. The individual changes, plotted in the probability domain against time, match

MaxEnt sigmoids. Moreover, there are perturbers, which can retard or encourage a change, in the diachronic curves the data affected by perturbers plot as shifted sigmoids, leading or trailing the other sigmoids in time. Formally, Kroch generates his sigmoids using the MaxEnt math. The key question here is how Kroch's findings might relate to synchronic patterns: perhaps there is some way that change occurs at a constant rate in the domain of constraint weights, rather than raw frequencies. For discussion, see Blythe and Croft (2012), Zimmermann (2017), and Hayes (2022).

8.6.5 Local summary

The preceding sections have cited evidence that syntax involves gradient well-formedness, variable constraint strength, constraint ganging, frequency matching, and shifted-sigmoid probability patterns. Based on this constellation of phenomena, we are led to wonder if an effective and comprehensive syntactic framework might some day be devised incorporating MaxEnt principles. We turn next to the issues that would arise in such an effort.

8.6.6 Some open issues for a future MaxEnt syntax

Architecture. It is not yet clear what sort of architecture would be appropriate for a MaxEnt syntax, though there are two general possibilities.

Most of the MaxEnt syntactic work we have seen uses the method described in our Chapters 2 and 3, in which surface syntactic structures are derived from some more abstract, deeper representation. Thus An et al. (submitted) assume surface-underspecified syntactic trees as their inputs (derived perhaps by a non-gradient grammar[68]), to which agreement features are added using a Harmonic grammar. Velldal and Oepen (2005), addressing a dataset of 1000 annotated English sentences, use as their base forms the semantic representations of Head-Driven Phrase Structure Grammar (HPSG; Pollard and Sag, 1994); their GEN component provides multiple candidate surface representations (often, dozens or even hundreds) to which a MaxEnt grammar assigns probabilities. The framework of Lexical Functional Grammar (Bresnan et al. 2015) provides underlying lexical-functional structures for which GEN can assign multiple surface-tree candidates. While to our knowledge no work in MaxEnt LFG has been done, it seem a natural combination, given that there is substantial work in LFG using classical Optimality Theory (for review see Bresnan (2000).

There is also the possibility of working in a way analogous to our Chapter 6: we would enumerate a vast set of possible syntactic forms and assign each one a well-formedness assessment in the form of a probability value, much as we described for phonotactics. The scheme put forth by Pullum (2013) might be taken as envisioning this scheme; and a variety of actual implementations (borrowed from the field of natural language processing, and not MaxEnt) are explored in Lau et al. (2016). We note that at some level of the syntactic system, something like this is necessary; for some languages there are cases of coherent and sensible meanings that are syntactically inexpressible, so that for example island constraint violations

[68] p. 11, "We assume some formal generative grammar that licenses a set of hierarchically organized and linearly ordered structures."

(Ross 1967) are the syntactic equivalent of phonological **bnick*; they must somehow be excluded outright.

Probability and grammaticality. The use of probability as a metric of grammaticality, as suggested for phonotactics in Chapter 6, faces two long-standing difficulties when we try to do the same for syntax. We suggest some responses here.

First, as beginners are taught, in any language there is an infinite number of possible sentences, since the syntactic system is recursive. It seems reasonable to think that a 10,000 word sentence beginning along the lines of *Alice believes that Mary knows that Fred thinks that …* is grammatical, even if it takes a great number of minor "hits" in the form of weak constraint violations along the line. Lau et al. (2016) offer a sensible proposal to address this issue: our grammaticality intuitions access not the absolute probability of a long utterance, but something closer to "weighted violations per word"; corresponding to the local debits of Harmony experienced by the listener as the sentence unfolds.

The second issue is the well-known "New York/Dayton" problem, put forth in Chomsky (1962). Chomsky compares the sentence *I live in New York* with the sentence *I live in Dayton*.[69] Obviously, the probability of a speaker uttering the first sentence is far higher than the probability of uttering the second, yet the syntactic probability of the two sentences seems intuitively the same (both are quite grammatical). So how do we keep frequency-matching from assigning a higher probability (hence greater grammaticality) to *I live in New York*?

The key to this problem is to recognize that in syntactic learning we are not trying to match the frequencies of the whole world, but just only those involved in the syntactic aspects of the data (Stefanowitsch 2005). For example, if we imagine a future system of MaxEnt syntactic learning (e.g. a greatly scaled-up version of what appears in Chapter 6), then, as always, the only way that any training sentence can impinge on learning is through its constraint violations. The *Dayton* and *New York* sentences almost certainly have identical violation profiles, since *Dayton* and *New York*, both proper nouns of location, have identical syntactic properties. Thus, hearing either of the two sentences should have precisely the same effect on setting the weights of the grammar. More generally, to the extent that *any* probabilistic syntactic learning system attends solely to syntactic principles, it will frequency-match the syntactic structures in the training data, not the details of everyday life.

8.7 Work blending syntax and phonology

A number of scholars have demonstrated an intriguing blend of syntax and phonology: where the languages offer a choice between two syntactic encodings of the same message, they tend to pick the one that avoids phonological markedness violations. For instance, Shih and Zuraw (2017) examine the syntactic choice in Tagalog between two word orders: *magandá-ng babáe* 'beautiful-LINKING MORPHEME woman' and *babáe-ng magandá* 'woman-LINKING

[69] A bibliographic detail, based on Stefanowitsch, fn. 1: in print Chomsky used *Nevada* (at the time quite underpopulated and a plausible example), not *Dayton*, which appeared in oral presentations preceding the printed version. Since it is *Dayton* that has entered the folklore of our field, we use it here.

MORPHEME beautiful', and find on the basis of corpus work that it is affected by phonological factors, notably the tendency to avoid violations of the well-known constraints *NÇ (Pater 1999) and *HIATUS. They use the MaxEnt math to evaluate a large set of such predictors. For similar work on English see e.g. Benor and Levy (2006), Anttila et al. (2010), Shih et al. (2015)Shih et al. (2015), and Breiss and Hayes (2020).

8.8 Semantics and pragmatics

AnderBois et al. (2012) demonstrate the interaction of two possible constraints, one involving linear order, the other grammatical relations, in determining the quantifier scope relations perceived by participants in an experiment. They analyze their results with logistic regression, the same math as MaxEnt (§4.6). Their data fall into a shifted-sigmoid pattern, displayed online at Hayes (2022).

Chapter 9: MaxEnt as a theory of language

This book has so far treated MaxEnt grammars as a sort of tool, a way to get coherent and explicit analyses of quantitative patterns using well-motivated constraint sets. However, MaxEnt is also interpretable as a theory, which makes predictions about what kinds of quantitative patterns can be internalized by language learners as grammars. In this chapter, we explore a number of these predictions.

9.1 MaxEnt as probabilistic grammar

MaxEnt is a probabilistic framework, one of several under current study (§9.5). For input-output mappings (Chapters 2-3), it generates not just one output but a probability distribution of candidates; and for well-formedness (Chapter 6) it does not designate strings as grammatical or ungrammatical, but provides a probability distribution over the strings in GEN, depicting gradient well-formedness. However, the assumption that grammar is probabilistic has repeatedly been called into question. This debate is so fundamental to MaxEnt that it seems worth addressing the issue, and giving our arguments. Our discussion is only a summary; for further discussion we refer the reader to work such as Clark (2005), Guy (2011), Francis (2022), Alderete and Finley (2023), Pierrehumbert (2022), and Magri and Flemming (2024).

9.1.1 Variation is ubiquitous

To begin the argument, we observe that gradient phenomena are found in many contexts, including the following cases.

- Variation in output forms, including variation conditioned by structural linguistic factors, has been documented in the field of variationist sociolinguistics since at least the 1960s; see references in §2.2.

- Any experiment that elicits well-formedness judgments is virtually guaranteed to obtain gradient outcomes. This has proven true both for phonology (see §3.5) and for syntax (see§8.6)

- Likewise, experiments eliciting productions of novel forms yield variable outputs. Where the patterns being tested are based on patterns in the lexicon, responses generally follow the pattern of their occurrence in the lexicon of the language, following the principle of frequency matching; we gave a literature survey in §3.5.

- Participants in Artificial Grammar Learning studies learn variable patterns, and produce gradient judgments and variable outputs which generally observe frequency matching; see references in §5.5.4.

Thus, for the scholar who seeks to claim that grammar is not gradient, there is a great deal of existing data for which alternative explanations must be found.

What alternative explanations exist? We suggest there are basically two. First, there is the view that variation is not present in the individual speaker's grammar, but rather represents dialect mixture. Second, there is the view that variation and gradience are part of performance, not competence. We address both views below.

9.1.2 Variation claimed to be only dialect variation

This view primarily arises for phonology, since its focus is where a single input gives rise to multiple outputs. The idea is that whenever we see more than one output, it is because the two outputs reflect distinct "dialects". If this were right, the researcher should in principle retreat a bit whenever variation is found, find the two dialects in the data from his/her speakers, and analyse each one separately, since a linguistic grammar can only reside within the mind of one single speaker.

But in fact, quite early on, variationist research established that variation is present *within the individual speaker*: each person varies in his or her own output forms, often in a continuous way that cannot reflect a binary distinction between dialects; see for instance Labov's (1973) careful series of "phonological portraits" of individual speakers of New York City English. Similarly, the original research by Fuller (2024) that stands behind our discussion of K1-K5 in Chapter 2 showed that the pattern of variation studied there is *not* the result of aggregating the data of four speakers, but is present for each speaker individually.

To our knowledge, efforts to attribute variation to "dialect mixture" (Chomsky and Halle 1968,[70] Halle and Vergnaud 1981) have typically been casual and have not invoked any of the established research methods of dialectology (e.g. community surveys, geographic mapping, and the Comparative Method; see Bynon 1977: ch. 4). Ringen and Heinämäki (1999) offer a strong response to Halle and Vergnaud's (1981) conjecture that variation in Finnish vowel harmony is to be attributed to dialect variation: not only is there extensive variation within individuals in the data they collected, but the dialect theory has nothing to say about how a subtle array of harmony preferences arises from interaction of multiple phonological factors, including perturbers.

To be sure, the data phonologists gather may well exhibit systematic differences between different consultants that may reflect authentic dialect differences. This is especially true if (for whatever reason) we are forced to blend data from speakers of different geographic, ethnic, or social class backgrounds. It is also possible that idiolect differences arise from random differences in the sample of generally-uniform input data that they receive, or from random individual traits. The disentangling of individual speaker variation from interspeaker variation is an ongoing research problem; for efforts to do this in individual cases see Hayes and Londe (2006:§4.4), Moore-Cantwell et al. (2024), and Liang et al. (submitted).

[70] The authors posit (p. 227) two English dialects, in one of which *auxiliary* (/uːksil+i+æːr+i/) surfaces as [ɔgˈzɪliˌɛri], and in the other [ɔgˈzɪljəri]; this follows from a particular detail of rule (118b). No geographic or social information is provided concerning the putative dialects, nor any information about how to analyze speakers who can say *auxiliary* both ways (and similarly for analogous words).

9.1.3 The same constraints govern variable and categorical phonology

A second way that gradience in grammar has been rejected is to say that grammar itself is categorical, and that all effects of gradience are to be attributed to "performance". This invokes a distinction — competence vs. performance (Chomsky 1965) — that has been the subject of debate for decades. We adopt here Chomsky's definitions (p. 2): competence is "the speaker-hearer's knowledge of his language" and performance is "the actual use of language in concrete situations". In present terms, performance is anything that influences linguistic production, perception, and interpretation, but is not part of the grammar, which constitutes the formal characterization of competence. In principle, performance factors might lead to gradient judgments or behavior in language use, even if the underlying grammar is categorical. We address here the specific idea that gradience is *always* the result of performance factors. This idea seems to be widespread in the field of linguistics, but is argued for relatively rarely in print. Some notable examples are Sprouse (2007) and (Du and Durvasula (2025).

We argue here that, contrary to this view, gradience in phonology, and language in general, often arises from the grammar itself. We will focus here on phonology specifically, but see Francis (2022) for extensive discussion of this issue in syntax.[71]

The primary empirical evidence for our view was given above in §2.6.1 and is covered here in more detail. The key generalization is (125):

(125) *The "same constraints" generalization*

> The constraints responsible for gradient patterns in some languages are frequently responsible for categorical patterns in others.

This point has been made many times in the literature, often by scholars unaware of previous mentions. For syntax see Featherston (2005), Bresnan et al. (2001), Clark (2005) and references cited there; for metrics see Hayes (2010); and for phonology see Hayes (2000), Guy (2011), and Kumagai and Kawahara (2018). In the terminology of this book, we would say that constraints that act as perturbers in one language often can be seen dictating categorial outcomes in others. In (126) we offer some specific examples.

(126) *Principles that are perturbers in some languages, determinative in others*

a. *Trigger for vowel harmony is low*: perturber in Hungarian (Hayes and Londe 2006), absolute in Murut (Prentice 1971).
b. *Avoid voiceless consonants after nasals*: perturber in Tagalog (Chapter 3), absolute in Ecuadorian Quechua (Orr 1962)
c. *Avoid onset* [ŋ]: perturber in Tagalog (Chapter 3), undominated in English

[71] Pullum (2013:535) has this to say about the performance explanation in syntax: "The competence/performance distinction is what [certain] linguists have relied on ever since 1965 as a promissory note for future bridges over the chasm between what grammars say and what linguistic experience is like." The context of his remark is a defense of constraint-based syntax as a way of addressing gradience.

d. *Avoid adjacent stresses*: perturber in French (Smith and Pater, 2020) and in multiple areas of English (Smith 2011, Hilpert 2008), Breiss and Hayes (2020); undominated in (e.g.) Pintupi (Hansen and Hansen 1969, 1978; Hayes 1995).

e. *Ban on triple consonant clusters:* perturber for English [t]-Deletion (Sankoff and Labov 1979); absolute in many languages, e.g. Yowlumne (Newman 1944).

g. The constraints that categorically govern person and voice in the active and passive sentences of Lummi are detectible as perturbers in English (Bresnan et al. 2001).

f. Constraints of poetic metrics frequently express tendencies in one language or idiom but are absolute in others (Hayes 2010).

This pattern seems to us as a strong argument for not dumping variation into a "performance-wastebasket." In order to get "performance" to account for all of variation, it would likely need to incorporate the very same constraint system as grammar itself.[72]

Below, we discuss some factors that we think really are performance, but, as will be seen, they cannot serve as a replacement theory for gradience in language.

9.1.4 What really does arise from "performance"?

While we do not believe that performance factors can completely explain gradience in language, they are no doubt present in the linguistic behavior that we use as data for the study of competence. Here, we present two specific types of performance factors that can be identified in phonology, and briefly argue against them as complete explanations for gradience.

The first such performance factor is perceptual errors. In studies where participants give acceptability judgments, there is reason to believe that participants are not always hearing what experimenters intended them to hear. Some of this misperception is random, but some of it is closely linked to the grammar of the language. In particular, listeners tend to perceptually 'repair' sequences which are not allowed in their language (Dupoux et al. 1999; Davidson and Shaw 2012; Berent et al. 2007), even perceiving vowels in sequences of illicit consonants (for example, English speakers may perceive [lbɪk] as [ləbɪk]). Kahng and Durvasula (2023) demonstrate explicitly that these perceptual repairs affect speakers' judgements. When their participants misperceived an illicit form as a licit one, they judged it to be more acceptable.

While such misperception undeniably occurs, we see no mechanism proposed in the literature whereby misperception could account for gradient well-formedness judgments. Indeed, it seems to us to make the wrong predictions: [lbɪk], perceived as [ləbɪk], should be perfect, better than a readily-perceived but imperfect [pɔɪk]; but e.g. Hayes and White (2013) found just the opposite. Moreover, experiments often find systematic gradient distinctions among forms that are perfectly perceptible; see e.g. Tagalog in Chapter 3 and Hungarian in Chapter 7.

[72] Perhaps uniquely, Newmeyer (2003), in a defense of nongradience, takes on the "same constraints" argument. His suggestion is that the raw functional principles that underlie constraints are active in performance and skew behavior in ways that make it look like non-gradient grammar. This thoughtful suggestion does not generalize to later research which shows that gradience can be the result of idiosyncratic, nonfunctional factors, as in our primary examples of Khalkh, Tagalog, and Hungarian (chs. 2, 3, and 7 respectively).

A second performance factor undoubtedly affecting variation is the size of planning units (Kilbourn-Ceron and Goldrick 2021; Tilsen 2012). In order for a phonological process to occur, both trigger and target must be present in the input to phonology. The idea is that variable application of certain processes can be attributed to variation in the formation of planning units, which in turn form the input to phonology. Consider the case of cross-word tapping in English, examined by Kilbourn-Ceron and Goldrick: If *get up* is planned as a single unit, and the input to phonology is /gɛt+ʌp/, then tapping will occur. If the two words are planned separately, /gɛt/ would surface as [gɛˀt], and /ʌp/ as [ʌp]. Since they are evaluated separately, the environment for tapping is never present. Factors like predictability, frequency, and speech rate all often influence the rate of application of variable phonological processes, and also plausibly influence the size and contents of planning units.

Planning unit size may account for some of the variation in phonological processes, but like misperception, it cannot account for all variation in production. Variation is not always at boundaries that are plausible edges of planning units; many cases of word-internal variation exist, for which the planning-unit size is inapplicable. This is true, for instance, for French schwa deletion (Smith and Pater, 2020), or the Hungarian and Tagalog cases discussed here.

9.2 Constraints have the same strength whether opposed or acting alone

With the above background in place, we turn to particular predictions made by MaxEnt as a theory of probabilistic grammar which distinguish it from other probabilistic theories (§9.5). We think an informative comparison can be made by considering cases where a constraint C sometimes acts in concert with other constraints (either cooperation or conflict), but in other cases is the sole relevant factor in determining the outcome. A very simple instance of this is given in (127).

(127) *Key pattern: Constraint C3 either interacts with C1 and C2, or acts alone*

		C1	*C2*	*C3*
Input 1	$Cand_1$	1		
	$Cand_2$		1	
Input 2	$Cand_1$	1		
	$Cand_2$		1	1
Input 3	$Cand_1$			
	$Cand_2$			1

In (127), the constraint C3 is set up to act as a perturber in influencing the outcome for Input 2: if its weight is greater than zero, C3 will perturb the frequency of $Cand_2$ for Input 2 downward, decreasing its probability relative to $Cand_2$ for Input 1. For Input 3, C1 and C2 are not applicable and constraint C3 acts alone. Here, $Cand_1$ and $Cand_2$ will receive probabilities ranging from 50/50 to 1/0 depending on the weight of C3.

Given this configuration of constraint violations, it is possible to construct target output frequencies, as we have done in (128), which impose impossible demands on MaxEnt constraint

weighting. Given the identical output frequencies for Input 1 and Input 2, C3 could be called a "failed perturber": to derive the observed frequencies it must be weighted at zero (or else near zero, if we are satisfied with an approximate result). In contrast, for Input 3, C3 acts alone. In this case, to derive unanimity for Cand 1, the weight of C3 must be infinite, or some high number of near-equivalent effect (§2.9). In other words, with MaxEnt there is a contradiction: a very low weight is needed for one purpose, a very high weight for the other. If we try to satisfy these requirements at the same time using the Excel Solver, as in (128), the results are distinctly disappointing; for details see **ConstraintActingAlone.xlsx**.

(128) *An attempt to match contradictory data frequencies in MaxEnt*

			C1	*C2*	*C3*	*MaxEnt best*
		observed	*0.76*	*0*	*1.51*	*fit predictions*
Input 1	*Cand*$_1$	0.5	1			0.319
	Cand$_2$	0.5		1		0.681
Input 2	*Cand*$_1$	0.5	1			0.681
	Cand$_2$	0.5		1	1	0.319
Input 3	*Cand*$_1$	1				0.819
	Cand$_2$	0			1	0.181

As can be seen, the best-fit weight 1.51 makes C3 far too powerful as a perturber for Inputs 1 and 2, and at the same time far too weak to achieve unanimity for Input 3. Similar results are obtained under different input frequencies that likewise impose a contradiction; e.g. 50/50, 49/51, 99/1.

None of this is surprising in light of the discussion earlier in §4.3.1: in MaxEnt, constraints cannot be turned off; they always have an effect, wherever their structural description is met. Indeed, when assessed in the domain of log probability, it is always the very same size effect. As the math tells us, it makes no difference whether a constraint acts alone or in conjunction with other constraints.

Turning now to alternative stochastic frameworks, we see a very different picture: for these frameworks, the frequency pattern of (128) is predicted to be entirely normal and can be easily matched, exactly or at least to a good approximation. For instance, the theory of partially ordered constraints (POC; §2.3) can achieve an exact fit, as tableau (129) shows.

(129) *Deriving the pattern of (128) in POC*

			C1	*C2*	*C3*	*predicted*
Input 1	*Cand*$_1$	0.5	*!			0.5
	Cand$_2$	0.5		*!		0.5
Input 2	*Cand*$_1$	0.5	*!			0.5
	Cand$_2$	0.5		*!	*	0.5
Input 3	*Cand*$_1$	1				1
	Cand$_2$	0			*!	0

Here, C1 and C2 form a unranked stratum, for which *Cand₂* wins as output for both Input 1 and Input 2 under the ranking C1≫C2, and *Cand₁* wins for both inputs under the ranking C2≫C1. Since C3 is in a lower stratum, outranked by both C1 and C2, it cannot affect the outcome. For Input 3, however, C3 is the only constraint that matters. Since *Cand₂* violates C3 and *Cand₁* does not, *Cand₁* always wins as the output of Input 3.

Other frameworks, reviewed below in §9.5, work similarly. Our calculations (carried out in OTSoft) indicate that Stochastic OT (§9.5.3) can match the frequencies of (128), as can the "exponential" version of Noisy Harmonic Grammar (§9.5.1; see Boersma and Pater, 2016:§3.5). A more traditional version of Noisy Harmonic Grammar, with the options of "truncation" and "trial cancellation" suggested by Hayes and Kaplan (2025),[73] cannot match the pattern of (128) perfectly, but the mismatch is modest, just a 5% error for Input1 and Input 2. For our calculations, see **ActingAloneStochasticOT.xlsx**, **ActingAloneExponentialNHG.xlsx**, and **ActingAloneClassicalNHG.xlsx**.

To summarize, we state in (130) a prediction that sharply distinguishes MaxEnt from alternative frameworks.

(130) *A prediction of MaxEnt*

A weak perturber constraint cannot be all-powerful in cases where it is unopposed.

What are the facts? At present, we do not know, but the authors of this book know of no cases that violate (130), which stands as a demonstration that the MaxEnt framework, all on its own, makes empirical predictions. A well-documented counterexample would be a strong impetus to switch to a distinct framework.

The example of (128) can be flipped on its head, with its point restated as follows: if a constraint has a high enough weight to make a difference in any simple case, then it must also act as a perturber constraint in more complex cases. Generalizing from this, we suggest the possibility that in real-life variation, *simple cases will be rare*: there will be usually multiple constraints present that may bear on an outcome, each of which plays a role elsewhere in the system, or perhaps has a non-zero default weight in UG (Chapter 7). In either case, they will not be assigned a zero weight, and should therefore act as perturber constraints. This claim has been made before in variationist research. Young and Bayley (1996) calls this the "principle of multiple causes," and suggests that "it is unlikely that any single contextual factor can explain the variability observed in natural language data."

9.3 Intersecting constraint families and Magri's "four from three" principle

The issue of MaxEnt as a predictive theory was taken up by Zuraw and Hayes (2017), who examined its behavior in the context of *intersecting constraint families*. The idea here is to identify in the constraint set two subsets A and B, such that the violation by a candidate of any constraint of subset A is independent of whether it violates a constraint of subset B. For instance,

[73] These are imposed to ensure that the property of Harmonic Bounding will be respected; see §9.5.2 below.

in the Tagalog example of Chapter 3, the constraints based on the stem-initial consonant form one family, and constraints regulating faithfulness on the basis of prefix choice form the other. Zuraw and Hayes observe that in MaxEnt, it is predicted that (within limits of random variation) it will be possible to model such data with a set of two or more shifted sigmoids, of which we have seen examples in Khalkha (Chapter 2) and Tagalog (Chapter 3). They also observe some rival theories, to be covered below (§9.5), that do poorly in modeling such patterns.

This result, obtained from three languages, led Hayes to search for more cases of intersecting constraint families, both in phonology and in other areas of linguistics including phonetics, syntax, semantics, and historical linguistics. The results of this survey are given in Hayes's (2022a) journal article and in the online "Gallery of Wug-Shaped Curves" (Hayes 2022b). The pattern of shifted sigmoids does indeed seem to be widely distributed in human language, and we continue to collect more examples.

These findings — essentially, from explorations of model fit — were characterized more sharply in terms of theoretical restrictiveness by Magri (2026). First, Magri developed a specific characterization of how MaxEnt behaves in these cases; specifically a principle that takes the form *from three, predict the fourth*. To illustrate this idea, we borrow Magri's example, which consists of a subset of our Zuraw-derived Tagalog analysis (Chapter 3). For reference, we start with a tableau for this miniature case (available as **TagalogStrippedDown.xlsx**). For clarity, gray shading is used to mark the constraint families that intersect: light gray for the constraints that adjudicate between the two consonants /k/ and /b/, and dark gray for the constraints that adjudicate between the prefixes maŋ-$_{other}$ and maŋ-$_{RED}$.

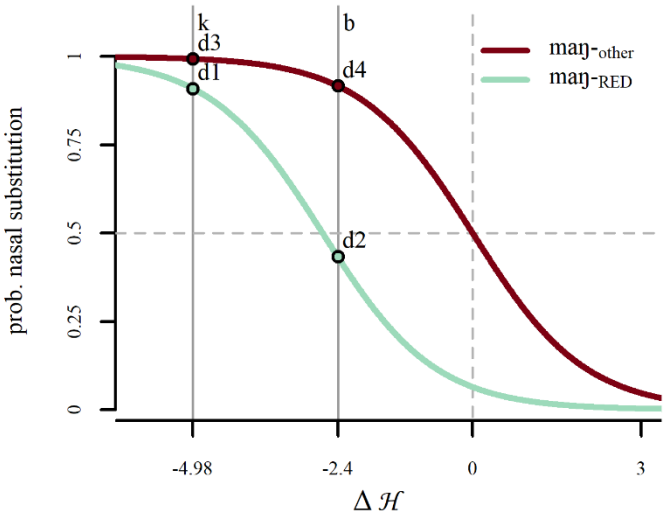
(131) *Tagalog stripped down to four inputs*

			NAS SUB	*NC̥	*[ŋ	UNIF-maŋ-$_{other}$	UNIF-maŋ-RED-			
		Freq.	**2.40**	**3.84**	**1.26**	**0.00**	**2.67**[74]	H	p	Obs.
k_maŋ-$_{RED}$ (**d1**)	unsub	2	1	1				6.24	0.090	0.091
	sub	20			1		1	3.93	0.910	0.909
b_maŋ-$_{RED}$ (**d2**)	unsub	21.5	1					2.40	0.566	0.566
	sub	16.5					1	2.67	0.434	0.434
k_maŋ-$_{other}$ (**d3**)	unsub	0.5	1	1				6.24	0.007	0.007
	sub	74.5			1	1		1.26	0.993	0.993
b_maŋ-$_{other}$ (**d4**)	unsub	6	1					2.40	0.083	0.083
	sub	66				1		0.00	0.917	0.917

In this tableau, inputs have either [k] or [b] as their stem-initial consonant, and contain either the prefix maŋ-$_{other}$ or maŋ-$_{RED}$. The key reasoning is best illustrated with a set of shifted sigmoids corresponding to this tableau, given in Figure 43.

[74] Note that the weights given here are not quite the same as those appearing in the tableau (xx) in Chapter 3, which are fitted to the full data set.

Figure 43: Shifted sigmoids for Tagalog, stripped down to four inputs



Magri's claim, expressed graphically, is that given three data points in a configuration like this, the position of the fourth is determined by the theory; e.g. if we know the substitution rates for the points labeled **d1**, **d2**, and **d3**, then the theory already tells us the value predicted for **d4**. The reasoning goes as follows. To start, the spot where the vertical "k-line" (stem-initial /k/) intersects the sigmoid line for the prefix maŋ-RED is determined by the data point **d1**; the intersection of straight and curved line must occur where the height of the sigmoid is at the empirically observed value of 0.909. Next, using the probability difference between **d1** and **d2** (b_maŋ-other) we can use the MaxEnt formula to calculate the degree of separation between the vertical k-line and the vertical b-line. This comes out to about 2.58, which is in fact the difference in the final fitted analysis between the weights of the opposed constraints *NC̥ and *[ŋ (tableau (131)). Next, we compare **d3** (k_maŋ-other) with **d1**, and with similar calculations obtain the degree of separation between the two sigmoids curve; this turns out to be 2.67, the difference between the weights of UNIF-maŋ_RED and UNIF-maŋ_ other.[75]

At this point, all the freedom we have in setting up the shifted sigmoids has been used up, and the predicted value for **d4** (b_maŋ-other), will necessarily fall wherever the vertical b-line intersects the sigmoid for maŋ-other. This is a consequence of the theory, and at least in this case, the theory is confirmed by the data.

As Magri shows, the ability to predict the fourth outcome from the other three is, in the general case, a mathematical consequence of MaxEnt under some fairly modest assumptions. These are that the competition is always binary (all other candidates must be excluded by strong constraints), and that there is a pattern of two independently violated constraint families, for which Magri provides a precise definition (§2.4).

[75] Nothing in this discussion determine the *overall* horizontal position of the shifted sigmoids, which may be freely chosen so long as the relative differences are respected (§4.3.1). The horizontal position actually chosen in Figure 43 is the one that minimizes the weights.

Magri's example is a vivid way of stating a more general point, which is that only certain frequency values for a quadruplet of inputs treatable with intersecting constraint families can be modeled accurately in MaxEnt. If the data are modelable at all, then (due to Magri's theorem) the "four from three" principle will hold good.

There are indeed a vast number of conceivable data sets that are unmodelable. To give one example, we offer the imagined data frequencies of (132) (boldface); full spreadsheet at **TagalogStrippedDownDesignedToFail.xlsx**, first tab.

(132) *A data set for Tagalog designed to fail*

		Freq.	NAS SUB	*NC̥	*[ŋ	UNIF-maŋ-other	UNIF-maŋ-RED-
k_maŋ-RED (d1)	unsub	**20**	1	1			
	sub	**80**			1		1
b_maŋ-RED (d2)	unsub	**80**	1				
	sub	**20**					1
k_maŋ-other (d3)	unsub	**80**	1	1			
	sub	**20**			1	1	
b_maŋ-other (d4)	unsub	**20**	1				
	sub	**80**				1	

For these data, with the constraints of (131), the best-fit model turns out to be the random one, with all constraints at zero, and the predicted substitution rates at 50/50 for all inputs — the 80 /20 balancing of the input frequencies means that no constraint can be assigned a positive value (to improve accuracy on one input) without imposing a worse accuracy penalty for some other input.

All this shows that the MaxEnt framework itself is a theory with empirical predictions at stake. Counterexamples like (132), and many others, should in principle be easy to identify in the real world.

The "four from three" principle established by Magri is only the starting point of his mathematical analysis of MaxEnt. He further inquires: what is the set of grammatical frameworks under which the predictions made for intersecting constraint families fall into the pattern of shifted-sigmoid curves, from which "four-from-three" follows? To this end, he identifies a property called "separability" that singles out the set of such frameworks, and moreover identifies properties that favor MaxEnt among this restricted set. Thus, if the "four from three" principle for intersecting constraint families is empirically correct, then MaxEnt is part of a rather small set of frameworks capable of capturing this generalization.

9.4 Do conjoined constraints subvert the predictions of MaxEnt?

The previous section reviews arguments that MaxEnt has empirical teeth, in the form of the "four from three" generalization. This claim must be assessed against the possibility that the predictions might be subverted by a widely employed device of Optimality Theory, **conjoined**

constraints (Smolensky and Legendre 2006, ch.14). A conjoined constraint C is created by composing together two existing constraints X and Y; C is violated only when both X and Y are violated in the same candidate (often a further locality condition is added: C is violated only when X and Y are both violated in the same segment, or the same syllable, etc.). As several scholars have point out (e.g. Farris-Trimble 2008, Potts et al. 2010, Pater 2016), conjoined constraints have an effect rather similar to that of adding up Harmony values in Harmonic Grammar. Potts et al. show that for the complex phonology of Lango, the use of Harmonic Grammar (in a non-gradient form) obviates the need to use constraint conjunction, as employed in the earlier analysis of Smolensky and Legendre. In light of this, it is natural to wonder whether conjoined constraints could be dispensed with entirely in MaxEnt (or indeed in other forms of Harmonic Grammar).

In the context of §9.3 above, serious issues of restrictiveness arise here. If we allow the use of conjoined constraints freely in MaxEnt, then the interesting predictions given there will evaporate. To give a simple example, we considered again the concocted data pattern (132), designed to illustrate the restrictiveness of MaxEnt, and added to the constraint set a single conjoined constraint: *[ŋ & UNIF-maŋ-$_{other}$ (full tableaux at **TagalogStrippedDownDesignedToFail.xlsx**, second tab). This constraint would be violated when nasal substitution simultaneously is triggered by the prefix maŋ-$_{other}$ and derives an output stem-initial [ŋ]. When we add this constraint to the analysis, the model fit goes from very bad to essentially perfect. Of course, this improvement might not be grounds for rejoicing, since many linguists would consider the theoretical cost to be excessive: what could be the basis of a constraint like *[ŋ & UNIF-maŋ-$_{other}$? More specifically, why should unrelated factors like prefix choice and stem phonotactics be bound together in determining probability?[76]

Precisely this issue is raised in Fukazawa and Lombardi's (2003) critique of conjoined constraints. They suggest that in the right theory there are, fundamentally, only unitary contraints. The cases where conjunction might seem sensible can be given deeper explanations by appeal to phonetics or other external principles.

To give a simple example, we ask if the ban respected in Japanese (other than in recent loans) on voiced-obstruent geminates (Kawahara 2006) should be treated as the conjunction of a ban on geminates *[+long] and a ban on voiced obstruents *[−sonorant, −voice]. Both conjuncts have an impeccable typological basis, and the ban on voiced obstruents has a plain phonetic basis in the difficulty of maintaining voicing in obstruents (Westbury and Keating 1986). But here, there is no good reason to attribute the effect to constraint conjunction, because, as Kawahara notes, the unitary constraint *[+long, −sonorant, −voice] itself has an especially strong phonetic motivation and arguably belongs in the constraint set on its own merits.[77] More generally, Fukazawa and Lombardi suggest that this is true in general of conjoined constraints, so the device of freely allowing conjunction should not be permitted in phonological theory. Perhaps

[76] The present discussion has nothing to do with Magri's demonstration that MaxEnt implies that "predict fourth from three" pattern, which is explicitly limited to cases of independent intersecting constraint families. The conjoined constraint belongs, in effect, to two families at once, exempting it from Magri's criterion.

[77] Kawahara's formal analysis relies on Faithfulness, rather than a Markedness ban on voiced geminates, but his phonetic study nonetheless gives strong support for the view that voiced geminates are difficult to utter: they are realized quite imperfectly, with extensive stretches of voicelessness.

these constraints occasionally give the appearance of being necessary, as in Smolensky's original work, but we suggest that this is likely true only in Classical OT, a theory that has no *inherent* means of making constraints gang together and must therefore adopt the conjoined-constraint strategy as a matter of necessity.

All this said, we can ask if there there any cases of constraint conjunction that cannot be rationalized within MaxEnt as involving a unified basis (phonetic or otherwise)? Shih (2017) makes a case for such a conjunction from a tonal harmony process in Dioula d'Odienné, and argues, backed by statistical testing, that a MaxEnt grammar with conjunction of not-obviously-related constraints outperforms a grammar with only simplex constraints. We are not fully convinced by her example; the predictions of the two grammars differ only modestly (p. 260), and the study is based only on lexical data, not native speaker judgments. The issues of whether conjoined constraints are really needed in a Harmonic Grammar remains, we think, open.

To summarize: we began this discussion with Magri's "four from three" principle, illustrating it as an instance of the restrictiveness of MaxEnt. We next considered how free use of constraint conjunction subverts the prediction, then expressed our reasons for skepticism about this mechanism. We think putative conjoined constraints are often better analyzed as simplex, when we appeal to typology and phonetic modeling, and putative cases of constraint conjunction should be taken with skepticism until they are shown to be productive with evidence from psycholinguistic testing.

9.5 If not MaxEnt, then what?

Of course it is sensible to explore alternatives, of which we mention a few in this section.

9.5.1 Noisy Harmonic Grammar

Noisy Harmonic Grammar (Boersma and Pater, 2016) is a family of frameworks in all of which simple Harmonic Grammar, described in §2.5.1, is rendered "noisy." More precisely, we imagine a series of "evaluation times" (Boersma 1998), at each of which the weights of the grammar are each slightly altered by a random amount (the "noise"), typically taken from a normal distribution of mean zero and modest, constant variance. It will be recalled that the winner in a Harmonic Grammar is the candidate with the lowest Harmony penalty; noise means that this winner can be different at different evaluation times. Over a large set of evaluation times, the process produces a statistical sample of the behavior of the grammar, which when large enough will closely approximate the probability distribution that the grammar generates.

Noisy Harmonic Grammar is illustrated below with the first input of our K3 language (§2.4), where we seek to derive the lenited [w] outcome with probability 0.72.

(133) *Lenition of /p/ in K3 under Noisy Harmonic Grammar*

a. *Tableau*[78]

	*p	IDENT(son)	Observed frequency
weights:	4.58	5.42	
a. /mʊntəg **p**oʒ-ən/ great become-NPST			
mʊntəg **p**oʒən	1		0.72
☞ mʊntəg **w**oʒən		1	0.28

b. *Sequence of evaluation times with random noise*

	$noise_{*p}$	$noise_{IDENT}$	w_{*p}	w_{IDENT}	*H([p])*	*H([w])*	*Winner*
	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	4.58	5.42			
1	0.03	0.04	4.58+0.03 = 4.61	5.42+0.04 = 5.46	4.61	5.46	[p]
2	0.81	2.41	4.58+0.81 = 5.39	5.42+2.41 = 7.83	5.39	7.83	[p]
3	−1.51	−2.75	4.58−2.75 = 3.07	5.42−2.75 = 2.67	3.07	2.67	[w]
4	−1.38	0.25	4.58−1.38 = 3.20	5.42+0.25 = 5.67	3.20	5.67	[p]
5	−0.49	−0.64	4.58−0.49 = 4.09	5.42−0.64 = 4.78	4.09	4.78	[p]
6	0.40	−1.69	4.58+0.40 = 4.98	5.42−1.69 = 3.73	4.98	3.73	[w]

The table (133)b) illustrates the process for the first six evaluation times. In the second and third columns, random noise values have been sampled from a normal distribution ($\mu = 0$, $\sigma = 1$). In the third and fourth columns, these random noise values are added to the base weights for *p and IDENT(son), namely 4.58 and 5.42. This creates a series of noise-perturbed weights, shown in the fourth and fifth columns. The next two columns calculate Harmony; since the constraints are violated just once, the values here simply repeat the noise-perturbed weights. The final column lists the winner, the candidate with the lower Harmony penalty. The outcome will vary according to the perturbed weights for that particular row. It can be seen that on 4/6, or 0.667, of the evaluaton times, the winner is [p]. When we ran 100,000 evaluations, we found that 72.1% of them yielded [p], quite close to the 72.0% percentage of the training data.

We have found in our own experience that this system works pretty well, achieving a good fit to data if the constraint set is adequate. Noisy Harmonic Grammar is quite capable of treating perturber effects (§2.6). In some versions (see e.g. Flemming (2021) it can derive the shifted-sigmoid pattern (§2.8), though the exact sigmoids it generates are not quite the same.[79] As a test, we redid our K5 and Tagalog simulations using Noisy Harmonic Grammar, and obtained very similar results; see **K5WithNHG.xlsx** and **TagalogWithNHG.xlsx**. More speculatively, we

[78] Weights were fit using OTSoft; for reasons given below we chose 5 as the starting weight for both constraints.

[79] The sigmoids of NHG are the cumulative distribution function of the normal distribution, not the logistic function of MaxEnt. Empirically, the difference is almost nil, since they have almost identical shapes (Hayes 2022, (Flemming 2021).

adapted Noisy Harmonic Grammar to the form of phonotactic analysis illustrated in §6.4 (that is, using the Null Parse candidate), and obtained results for Turkish very similar to those obtained in MaxEnt; see **TurkishModelWithNHG.xlsx**.[80] These results are all reports of informal analytic experience; careful mathematical comparison of NHG with MaxEnt is carried out by (Flemming 2021).

Here are brief comments on practical analysis in Noisy Harmonic Grammar. We find that it is possible to use the Excel Solver to compute weights, with the limitation that there be just two viable candidates for each input. The spreadsheets just mentioned demonstrate how this can be done. Use of Excel also permits priors and biased learning. For analysis with more candidates, the software packages Praat and OTSoft can be used.

We suggest setting up analysis to begin the weight search with all weights at 5 or above, as we did in (133)) above. The purpose is to avoid some tricky thickets that arise if we let the noise perturb weights below zero. This can obliterate harmonic bounding (see §9.5.2 below) or even reverse the effect of constraints, causing them to favor forms that violate them; see Hayes and Kaplan (2025). It is in fact often feasible to keep all weights high, since as in MaxEnt (§4.3), it is only the relative weighting of conflicting constraints that matters.

Smith and Pater (2020) and Flemming (2021) have conducted empirical tests of MaxEnt vs. Noisy Harmonic Grammar, using small, carefully-controlled data sets. The predictions of MaxEnt emerge as more accurate, though the margin is not large. Our own comparisons (K5, Tagalog, Turkish) are also inconclusive.

Noisy Harmonic Grammar comes in multiple varieties. In the classical version of the theory, described above, the part of the grammar that is perturbed is the constraint weights. But it is also possible to perturb the scores in the tableau cells, or the *violation* × *weight* products (tableau cells), or the harmony scores; for discussion see Hayes (2017). It is also possible to sample the noise from a non-Gaussian distributions; see next section.

9.5.2 Noisy Harmonic Grammar and the harmonic bounding controversy

Harmonic bounding has long been an important concept of Optimality Theory, appearing for instance in Prince and Smolensky (1993). We define it here as follows. In a particular candidate competition (same input), Candidate A *harmonically bounds* B if (a) no constraint prefers B to A; (b) at least one constraint prefers A to B. From the standard method for culling out losers employed in OT, it follows that B can never be the winning candidate no matter how the constraints are ranked; it will always be culled out before A is. Note that for this purpose, it doesn't matter if A turns out not to be the actual winner; either way, B will not be. This is illustrated in (134).

[80] NHG *cannot* be used for the primary method covered in Chapter 6, i.e. deriving all candidates from a single abstract input; the reason is that this method requires that positive probability be assigned to candidates that are harmonically bounded (§9.5.2).

(134) *A simple illustration of harmonic bounding*

		C1	C2
Input	A		1
	B	1	1

Here, there is no ranking under which B wins, since it is harmonically bounded by A.[81]

In MaxEnt, the harmonic bounding violation pattern is also associated with an empirical prediction, but it is a weaker, stochastic one: a harmonically bounded candidate *can never receive the highest probability*.[82] The reasoning comes from the same scenario as above with candidate A bounding B: the Harmony penalty of B cannot be less than that of A, and therefore B cannot be assigned a higher probability than A, and therefore B cannot have the highest probability. Note further that if we remove zero-weighted constraints from consideration in assessing harmonic bounding, B will receive a probability less than that of A (and hence B cannot have the maximium probability).

Moving on to stochastic constraint-based frameworks *other* than MaxEnt, there is yet another pattern: for most such frameworks, the probability assigned to harmonically bounded candidates is *exactly zero*. We discuss these frameworks below in more in detail; but note for now that the zero-probability prediction for harmonically bounded candidates holds true of POC (§2.3, §9.5.4), of Stochastic OT (§9.5.3), and of any of the versions of NHG (see below) that are normally used in actual analytic practice (since maintaining the harmonic bounding property is usually considered a reason to use NHG).

In particular cases, the difference in prediction that arise from adopting a framework that assigns zero probability to harmonically-bounded candidates can be dramatic. Consider for instance tableau (32), repeated here as (135).

(135) *A minimal tableau for illustrating the effect of a constraint weight*

		ℂ **w**
Input	$Cand_1$	
	$Cand_2$	1

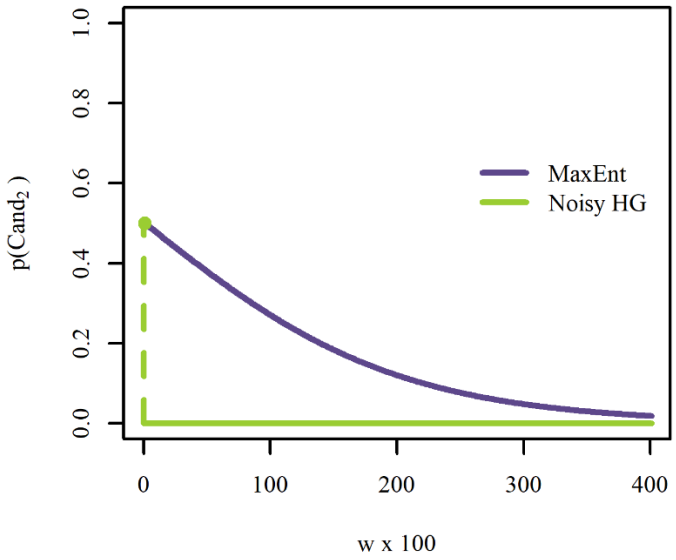
As before we consider happens when *w* is varied. In all the non-MaxEnt theories just mentioned that respect harmonic bounding, the probability assigned to $Cand_1$ is 0.5 when *w* is zero, else 0.

[81] There are more subtle cases of *collective* harmonic bounding: B is not harmonically bounded by any one particular candidate, but the other candidates are such that there is no ranking under which B can win; see Samek-Lodovici and Prince (1999).

[82] To be sure, it has been noted by Wilson and Obdeyn (2009: fn. 18) that it would be possible to set up a version of MaxEnt in which harmonically bounded candidates were simply pruned from the candidate set prior to evaluation and thus would receive a zero probability by fiat. To our knowledge this analytic possibility has yet to be explored.

In contrast, MaxEnt derives a smooth probability curve, specifically the right side of a sigmoid. The two behaviors are compared graphically in Figure 44.

Figure 44: Probability of Cand2 in (135), various weights, for MaxEnt and NHG



In short, NHG and similar frameworks show a discontinuity from 0 to any weight above zero - what we might call "hairtrigger" behavior; MaxEnt is smooth.

The concept of harmonic bounding can help us understand the prediction made by MaxEnt covered above in §9.2: in MaxEnt, weak perturbers cannot become all-powerful in cases where they act alone. This holds true because constraints in MaxEnt *always* have a constant effect, assessed in the log domain (§4.3.1). In the rival frameworks, however, the "acting alone" context is a special one because it is a harmonic bounding context, and harmonically bounded candidates lose absolutely, even if the weight of the constraint penalizing them is extremely small.

The interesting hairtrigger behavior of the frameworks that respect harmonic bounding can be related to our attempt above (§4.1) to characterize the MaxEnt math as a formalization of common-sense inductive reasoning. In the present case, a consequence of the MaxEnt math is this: "if you know from some other context that Factor X is weak, then do not use it to make hairtrigger decisions in contexts where it is only the applicable factor." We think this principle, too, would be generally thought of as common sense. In this sense, if NHG is true, then it is surprisingly, counterintuitively true. Though this does not rule NHG out, it does gives us pause.

The comparison of MaxEnt and NHG has attracted a substantial literature, much of it on the side of NHG; relevant work includes Jesney (2007), Hayes (2017), Kaplan (2018, 2021, 2024), (Anttila and Magri 2018; Magri and Anttila 2022, 2026), Flemming (2021), O'Hara (2022), Magri (2023, 2024), and Magri and Flemming (2024).

A detail that remains to be addressed is the special ways that Noisy Harmonic Grammar need to be set up to guarantee that it will assign zero probability to harmonically bounded candidates. As Hayes and Kaplan (2025) observed, the original version of NHG invented by Boersma and Pater (2016) actually does not respect harmonic bounding, the reason being that noise can send constraint weights into negative territory, making the constraints act as credits rather than penalties. The two ways this can be avoided are to truncate negative sampled weights to zero and cancel tieing trials (Hayes and Kaplan 2025), or more straightforwardly just sample from a truncated probability distribution that doesn't go into negative territory (Flemming 2021).

We conclude this section with an example that has been taken to argue in favor of NHG over MaxEnt, from Kaplan (2025). The empirical focus is an intriguing puzzle in Dutch phonology that has attracted research attention since the 1970s; as of now the literature includes Booij (1977, 1999), Kager (1989), van Oostendorp (1995a, 1995b), Nazarov (2016), and most recently Kaplan (2025).

Dutch has a number of words that begin with the syllable sequence *secondary stress*, *stressless*, *stressless*, and have identical vowels in the second and third syllables. This is true, for instance, of the word *locomotief* 'locomotive', for which the UR is something like /ˌlokomoˈtif/.[83] The fully-stressless vowels of this word can be optionally reduced to schwa, in an intriguingly asymetrical pattern. Kager (1989:309) gives the following intuitions about its pronunciation.

(136) *Variant surface forms of Dutch* locomotief

a.	[ˌlokomoˈtif]	acceptable, maximally faithful
b.	[ˌlokəmoˈtif]	acceptable, reduce second vowel
c.	[ˌlokəməˈtif]	acceptable, reduce both stressless vowels
d.	*[ˌlokoməˈtif]	"excluded"

To our knowledge, no one has ever done formal experimentation on words of this type, but two Dutch native-speaker linguists who have worked on the same problem (see van Oostendorp 1995a, 1995b; Booij 1999) seem comfortable with the judgments given.

We give an analysis roughly following that of Kaplan (2025). We suppose a general Markedness constraint, founded in articulatory effort (e.g. Crosswhite 2004) that penalizes full vowels. This constraint competes with position-specific Faithfulness constraints, banning the reduction of particular full vowels in particular contexts. The complete analysis would include a (here-unstated) highly-weighted Faithfulness constraint that protects full vowel quality in all vowels with primary or secondary stress; here we will just not bother to include candidates like *[ˌləkəməˈtif] that violate this constraint.

More crucial are the Faithfulness constraints that regulate the second and third vowels. If we assume Kager's metrical structure for *locomotief*, namely (ˌloko)mo(ˈtif), then the crucial second and third /o/ vowels differ in that the only the first one is footed. This, in turn, seems to be the

[83] We ignore the question of whether the stress pattern is predictable and hence excluded from the UR, which won't matter here.

basis for a noticeable durational difference: the footed /o/ is shorter than the unfooted one (see measurements in Slootweg (1988).

Chaining this result with the phonetically-driven theory of Faithfulness put forth in Steriade (2000) and Zuraw (2013, 2007), we would be justified in setting up two distinct Faithfulness constraints of the *MAP family. The third vowel of *locomotief* is protected by both of the constraints, first *MAP(o, ə)$_{\text{unfooted}}$, banning alternation of [o] and [ə] in unfooted syllables (following Kager 1989, 313), and second by the more general constraint *MAP(o, ə). Vowel 2, being footed, is protected only by *MAP(o, ə). The proposed tableau appears as in (137).

(137) *Tableau for vowel reduction in Dutch* locomotief

/ˌlokomoˈtif/	freq.	*MAP(o, ə)$_{\text{unfooted}}$	*MAP(o, ə)	*FULL VOWEL
[(ˌloko)mo(ˈtif)]	> 0			4
[ˌ(lokə)mo(ˈtif)]	> 0		1	3
[ˌ(loko)mə(ˈtif)]	0	1	1	3
[ˌ(lokə)mə(ˈtif)]	> 0	1	2	2

We assert, following Kaplan, that this pattern is not derivable as simple free variation in MaxEnt. To rule out *[ˌlokomaˈtif] completely, we would have to assign a very high weight to some constraint. This cannot be *MAP(o, ə)$_{\text{unfooted}}$, since this would give zero frequency to [(ˌlokə)mə(ˈtif)]. It cannot be *MAP(o, ə), since this would give zero frequency to every candidate but [(ˌloko)mo(ˈtif]. It cannot be *FULL VOWEL, since this would give zero frequency to every candidate but [(ˌlokə)mə(ˈtif)].

That there is no MaxEnt model with good fit is also suggested if we try to match a frequency distribution with equal frequencies for the three good candidates and zero for illegal [(ˌloko)mə(ˈtif)]: here we get an unwanted probability of 0.11 for this candidate; for the failed model see **DutchLocomotief_SimpleComparisonOfMaxEntAndNHG.xlsx**.

As Kaplan notes, Noisy Harmonic Grammar can provide an accurate model fit. Here, using OTSoft, we find that assigning weights of 5.30 to *FULL VOWEL, 4.70 to *MAP(o, ə), and 1.30 to *MAP(o, ə)$_{\text{unfooted}}$ will produce roughly equal frequencies for the three good candidates and zero for *[(ˌloko)mə(ˈtif)]; for details see the spreadsheet just noted. The key aspect of this account is the fact that NHG, suitably implemented, assigns zero probability to harmonically bounded candidates, and in tableau (137) the bad candidate [ˌ(loko)mə(ˈtif)] is indeed harmonically bounded (specifically, by [ˌ(lokə)mo(ˈtif)]). The picture looks, at first blush, like a good argument for NHG over MaxEnt, but we will be be returning to these data later on (§9.6).

9.5.3 Stochastic OT

Stochastic OT was invented by Boersma (1998) and first applied to phonology by Boersma and Hayes (2001). It works as follows. As in MaxEnt, every constraint comes affilated with a real number, though for reasons to be made clear it would not be appropriate to call it a weight; Boersma and Hayes call it a **ranking value**. As in Noisy Harmonic Grammar, there is a sequence

of evaluation times, each of which yields a single winning candidate. The method of evaluation for Stochastic OT is as in (138).

(138) *Finding the winning candidate in Stochastic OT*

At any given evaluation time:

a. Provide a separate random noise value for each constraint.
b. For each constraint, add the noise to its ranking value. The sum is what Boersma and Hayes call the **selection point**.
c. Sort the constraints by selection point, thus obtaining a total ranking.
d. Find the (unique) winner for this ranking following the procedure of OT.

The remaining step is the same as in Noisy Harmonic Grammar: sample over a large number of evaluation times, obtaining an estimate of the probabilities assigned by the grammar to the set of candidates.

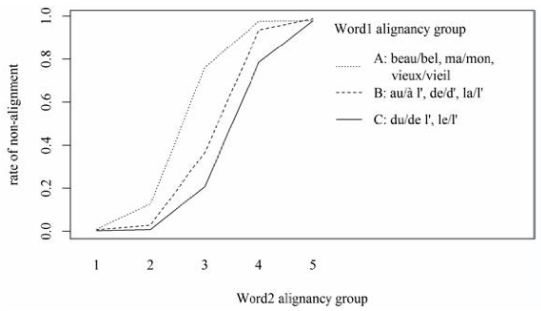
When Boersma invented Stochastic OT, he also invented an algorithm for fitting the ranking values to the data; it was called the "Gradual Learning Algorithm." It works by repeatedly running the grammar on the forms of the training data, checking for cases in which the grammar's guess did not match the observed outcome as derived (on that occasion) by the grammar. When such cases are found, the ranking values of winner-preferring constraints are boosted by a small amount, and the ranking values of loser-preferring candidates are lowered by a small amount.

Stochastic OT was the first stochastic constraint-based framework in which the parameters of the grammar could be trained by algorithm; and indeed the framework found fairly broad application in the field. Why do we not recommend it to analysts today?

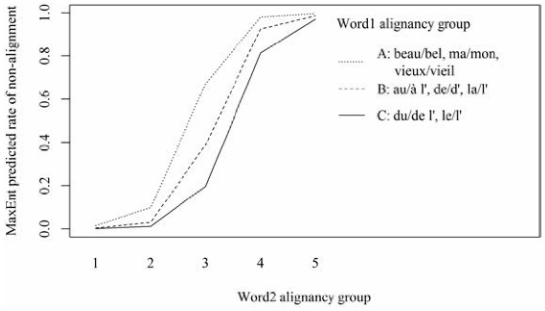
First, Stochastic OT often performs poorly in capturing the behavior of perturber constraints. Except when special, favorable circumstances are present, the model cannot capture perturber behavior, leading to poor model fit. For instance, here is a comparison of how MaxEnt and Stochastic OT fit the data in the analysis of French liaison given in Zuraw and Hayes (2017). We do not summarize the analysis here, which is treated in detail in that article; all that is needed here is to know that there is a shifted-sigmoid pattern in the corpus data the authors studied, and that this pattern is based on families of lexically-specific perturber constraints such as the ones we studied earlier in Chapters 2 and 3.

Figure 45: Model fit for the Zuraw/Hayes French liaison example: MaxEnt vs. Stochastic OT

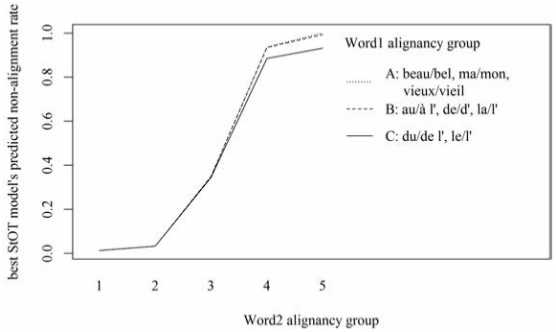
a. *Observations*



b. *MaxEnt*



c. *Stochastic OT*



It can be seen that the shifted-sigmoid pattern from the data (Figure 45a) is well fitted by the MaxEnt analysis, as in 45b. In contrast, in the Stochastic OT analysis (45c), the lines A, B, C that should be forming a shifted-sigmoid pattern emerge as almost superposed (two of the lines cannot even be distinguished.)

Zuraw and Hayes (pp. 512-514) provide extensive diagnosis of this failure. In brief, if we are to get perturber effects, the framework we use needs to let the perturber constraint have some effect *across the whole range of values* for the harmony or ranking values of the other constraints. In MaxEnt and other versions of Harmonic Grammar, this follows simply from the additive process of calculating harmony. But in Stochastic OT, a perturber can have an effect only over a narrow range, defined by the size of the evaluation noise. To the extent that (as we currently believe) shifted-sigmoid patterns are characteristically observed when there are intersecting constraint families, then Stochastic OT would not count as a good framework for probabilistic linguistics.

Beyond this, it appears that there is as yet no reliable algorithm for finding best-fit weights for Stochastic OT. The Gradual Learning Algorithm is in fact the same procedure (Perceptron Update Rule, Rosenblatt 1958) that works well for gradual learning of MaxEnt weights; see §7.7.2 above. However, in the latter case the search task is aided by the convexity of the search space. That Stochastic OT does not define a convex search space was shown by Pater (2008), who presented a case where a correct set of ranking values definitely exists, but the Gradual Learning Algorithm cannot find it. Subsequent efforts to repair the Gradual Learning Algorithm have so far been unsuccessful (Magri 2012; Magri and Storme 2020). The consequence is that when a particular constraint set seemingly fails to handle a particular dataset, we cannot know

for sure if the failure results from poorly chosen constraints, or rather whether this happens to be one of the cases in which the constraints are fine but we cannot discover the correct ranking values.

9.5.4 The Theory of Partially Ordered Constraints (POC)

POC was covered in §2.3 above. To recapitulate: the constraints are arrayed in a set of ranked strata, free ranking is posited within strata, and strict ranking across strata. The entire set of possible strict rankings compatible with the strata is computed, and the probability of a given candidate for a given input is the fraction of the total rankings that derive it.

Additional outputs from constraint cloning. While in §2.3 we suggested that only a few probability values can be generated by POC, strictly speaking this is not true; many more probability values can be generated by adding *clones* of the constraints. For instance, a stratum consisting of n copies of constraint A and m copies of B, each preferring distinct candidates, is capable of generating candidates dispreferred by A at a probability of $n / (n + m)$; hence of essentially any probability. To our knowledge, this cloning scheme has never been employed in the literature, which has limited itself to one copy of each constraint. The cloning idea seems to expand the POC theory in a fairly major way, and we cannot say at this point how the expanded theory might perform (e.g. in matching empirical data, or in giving rise to a suitable learning algorithm.)

The connection to Stochastic OT. Every grammar expressed in POC is, to a close approximation, expressible as a grammar in Stochastic OT. All we need to do is require that the constraints bear ranking values from a set like {… −60, −40, −20, 0, 20, 40, 60 …}, with wide spacing. We let each stratum be represented by one of these ranking values. Since it is exceedingly unlikely that two constraints with ranking values separated by 20 could receive selection points that reverse their base ranking, this means that the ordered set of ranking values essentially replicates the ranked set of strata. A consequence of this (Zuraw and Hayes 2017:515) is that the weighting algorithm invented for Stochastic OT, namely Boersma's Gradual Learning Algorithm, can be used on a rough-and-ready basis for POC; we simply fix the increment/decrement made in response to errors at 20.

Perturbers. Another consequence of the fact that the grammars allowed under POC is essentially a subset of that allowed in Stochastic OT is that the performance of the former theory in capturing perturber effects can be no better than that of the latter; and indeed in the cases studies by Zuraw and Hayes (2017), it is generally worse.

9.5.5 Neural networks

Neural networks rose to great prominence in the 1980s, particularly with the publication of Rumelhart, McClelland, et al. (1986), a book that presented the collective work of a leading research team. An important chapter in that volume was the one by Smolensky (1986), whose contribution, roughly, was to invent a way to mathematically characterize the activity of a neural network in terms of discrete 'knowledge atoms' and representations, a characterization that would later form the basis of Harmonic Grammar, with certain kinds of 'knowledge atoms'

being roughly equivalent to constraints. In this paper, probabilistic outputs of neural networks are computed using very similar equations to those presented in this book as MaxEnt grammar.

Smolensky's goal was transparency: to create a neural network whose actions could be understood by the investigator. Several decades later, neural networks received a massive infusion of computational resources (e.g. multiple layers, training data), and indeed environmental resources, creating the juggernaut "AI" that confronts our society today. As we write, a lively debate is being conducted by linguists and allied scholars about the significance and often-impressive performance of neural networks in dealing with human language; see e.g. Piantadosi (2024), Katzir (2023), and Cuskley et al. (2024).

For us, a key point is that, despite Smolensky's early efforts, mainstream neural networks continue to be quite undiagnosable in their behavior: the outsider observing linguistic outputs deriving from a network is in no better a position than the outsider observing linguistic outputs deriving from the computations of the human brain (Kodner et al., 2023; Bender et al., 2021). For this reason, we diffidently see a scientific role for a framework like MaxEnt, which while less powerful than a neural network, is eminently diagnosable: we know exactly how responses emerge from the configurations of the MaxEnt weights and constraint violations. From a scientific point of view, more powerful is not necessarily better. Our hope is that MaxEnt modeling can serve as an informed rough approximation of the computations carried out in the human brain, enabling a theory of language structure that is predictive and accurate, and also understandable by human observers.

9.5.6 Analogical models

We take the term "analogical model" to denote a system (here, just for phonology and morphology) in which outcomes are determined by consulting existing lexical entries online, returning a candidate that best matches these entries in some way. For instance, the analogical past tense model reported in Albright and Hayes (2003), based on Nakisa et al. (2001), chooses the past tense of *scride* by weighting the testimony of its phonetically-nearest neighbors, shown in Fig. 46; shown are the neighbors that have an [aɪ] → [oʊ] change, supporting the candidate *scrode.*

Figure 46: Some close models for scride in the English lexicon (Albright and Hayes 2003:131)

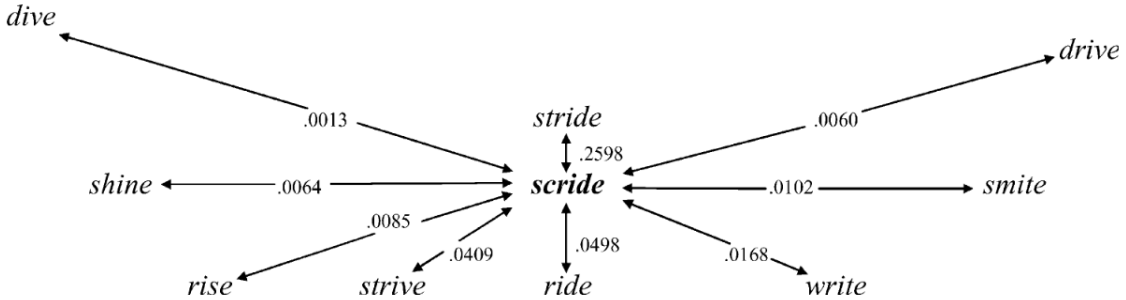


Fig. 1. Similarity of all [aɪ] → [o] forms to *scride*.

When the testimony of all the verbs weighed, it emerges that the most likely candidates are *scrided* (predicted 4.89 on a 1-7 scale), *scrode*, (4.73), and *scrid* (3.95). There are a variety of other analogical models that have been explored, for instance the tradition of Analogical Modeling of Language (Skousen et al. 2002).

We see some pluses and minuses to this approach. On the plus side, it has sometimes performed well in predicting experimental data; indeed Ernestus and Baayen (2003) find that some analogical models slightly outperform MaxEnt in predicting the results of their Dutch wug experiment. On the other hand, sometimes analogy is outperformed by more traditional systems that are based on actual learned linguistic generalizations: Albright and Hayes found that their analogical model was outperformed by their rule-based model: the analogical model suffered from its systematic use of "variegated similarity," in which it sought analogical forms on a haphazard, mixed basis.

We are also not convinced that every model billed as an "analogical" model really is one. For instance, in the well-developed Tilburg Memory-Based Learner (Daelemans et al. 2018), the researcher must pre-select the dimensions on which similarity is calculated, a procedure that inevitably will begin to steer the model toward generalizations that match what might be found in conventional linguistics.

Analogical effects are, in principle, experimentally detectible, and it seems likely to us that some of the responses obtained in wug-testing are indeed based on resemblance to particular lexical items. For instance, Baker and Smith (1976) obtained mostly antepenultimate stress for the wug word "*cinempa*," going against the trend in English for heavy penults to attract stress. As they note, this very likely is based on the close resemblance of *cinempa* to *cínema*. More systematic testing along these lines has been carried out by Prasada and Pinker (1993), Guion et al. (2003), and Shademan (2007). Effects like this have a consequence a research practice that is widely followed by wug-testers: a test meant to test phonological generalizations should avoid stimuli that are closely similar to existing words. The whole issue of how grammar interacts with local analogies of this sort remains an active research topic.

9.6 Faithfulness scaling: style, rate, frequency

In this final section we address not an alternative framework to MaxEnt grammar, but an augmentation of it that involves us in underexplored directions. What is needed is to expand the scope of phonological analysis to encompass **system-external factors** like style, rate, and frequency, in the manner proposed by Oostendorp (1995a,b), Boersma and Hayes (2001) and, in most detail, Coetzee and Kawahara (2013). In the approach we adopt from the latter paper, the Faithfulness constraints are *scaled* to reflect these factors. As we will see, the pursuit of these ends turns out to bear on the harmonic bounding controversy discussed above in §9.5.2.

Let us return briefly to the Dutch *locomotief* example discussed in §9.5.2. There, we presented the variation in idealized form, simply as a probability distribution over candidates. But the original describers of the pattern offered a more nuanced view of the data: these variants are related to speaking style. Thus Kager (1989) describes the four candidates as in (139):

(139) *Variant surface forms of Dutch* locomotief *— annotated for style (Kager 1989: 309)*

a. [ˌlokomoˈtif] "very formal"
b. [ˌlokəmoˈtif] "rather formal"
c. [ˌlokəməˈtif] "informal"
d. *[ˌlokoməˈtif] "excluded"

These judgments are not just Kager's but are also given by other native-speaker linguists, as noted above. We would be fully convinced by a corpus study, but the agreement among observers seen here suggests that we can take these data seriously.

Judgments of this sort can be related to the research literature connecting phonological variation to style. To illustrate, here are some very simple examples, which (as is necessarily the case) are tied to the culture one inhabits. We judge that *formal speech* might plausibly be used in reading the names of university graduates as they cross the stage to receive their diplomas. *Casual speech* is normal for relaxed interactions among close friends and family. A possible venue for speech that is *neither casual nor formal* is found in our own department when faculty and grad students engage in conversation over lunch. While this goes beyond the scope of this text, we note in passing that style controls more than just phonology; it affects syntax, phonetics, and lexical choice; and indeed it can govern what language, dialect, or code-switched mixture is employed.

Style is hard to study, since there is no precise metric for it; see Labov (1973:ch. 3) for one attempt to systematize by elicitation context. However, there is one simple technique by which linguists can get a handle on stylistic variation. Suppose we have data supporting the pattern given in (140).

(140) *Two phonological processes, differing in formality*

Process A: Optional, applies if style is $\geq x$ degree of casualness
Process B: Optional, applies if style is $\geq y$ degree of casualness, $y > x$

If this arrangement holds, we should *not* find surface forms in which B has applied but not A.

We suggest that Dutch *locomotief* may be an example of this kind. It takes a greater degree of casualness to reduce the third vowel of *locomotief* than it takes to reduce the second; hence reducing the third vowel while leaving the second vowel intact implies a contradiction of style level. *[ˌlokoməˈtif] is unsayable not because it is harmonically bounded, but because there is no style level that would support uttering it.[84]

We have a constructed a MaxEnt analysis under which this reasoning can be applied to the Dutch facts. Returning to the schematic tableau of (137), we set the weight of *FULL VOWEL at invariant 9.0, and the weight of *MAP(o, ə)$_{\text{unfooted}}$ at invariant 4.5 — this is treating unfooted status as a perturber, in the sense of §2.6. Let *s* be a parameter, similar to Coetzee and

[84] Obviously, style level can go up and down as one speaks, but we suggest it cannot change in just a few tens of milliseconds, the likely timing of the stressless syllables of *locomotief*.

Kawahara's scaling factors for lexical frequency, but expressing style. We let the value 1 denote the extreme of formal speech and 0 denotes the extreme of casual speech. The weight of the general Faithfulness constraint *MAP(o, ə) is set at 30 times *s*. In sum, the proposed weights for the analysis are as in (141):

(141) *A style-sensitive MaxEnt analysis of Dutch vowel reduction*

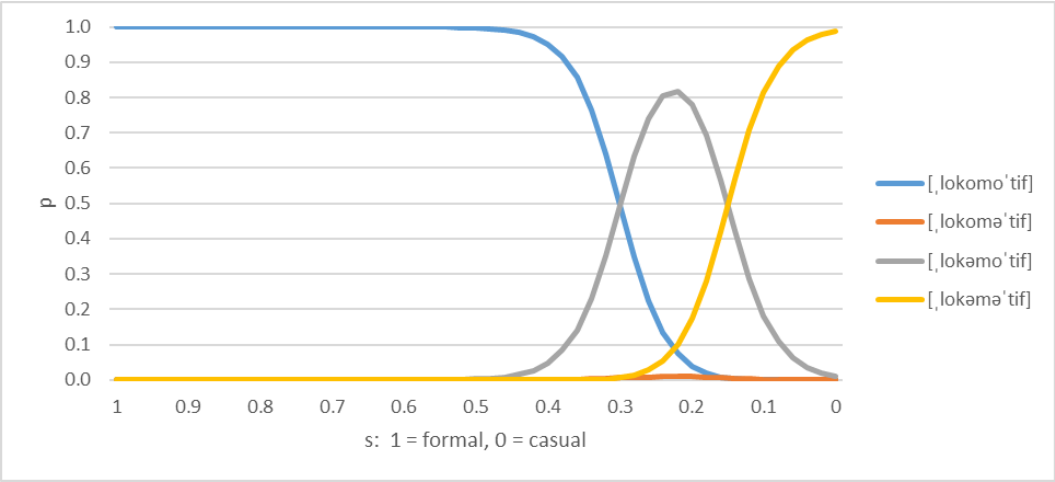
a. Let *s* occupy a continuum; 0 in highly casual speech, 1 in highly formal speech

b. *Schematic tableau*

/ˌlokomoˈtif/	*MAP(o, ə)$_{\text{unfooted}}$	*MAP (o, ə)	*FULL VOWEL	H	p
	4.5	30 × *s*	9.0		
[(ˌloko)mo(ˈtif)]			4	36	=exp(−H)/Z
[ˌ(lokə)mo(ˈtif)]		1	3	30*s* + 27	
[ˌ(loko)mə(ˈtif)]	1	1	3	4.5 + 30*s* + 27	
[ˌ(lokə)mə(ˈtif)]	1	2	2	4.5 + 60*s* + 18	

The predictions of the analysis can be examined by calculating the predicted output probabilties for whole range of values of *s* (we chose 50) ranging from 0 to 1; these can be seen in **DutchLocomotiefByStyle.xlsx**. In Figure 47 below, we show descending *s*, going from 1 to 0; intuitively this means increasing casualness from left to right.

Figure 47: How the probabilities of the candidates in (141) respond to changes in the style parameter s



We see that each of the candidates [ˌlokomoˈtif], [ˌlokəmoˈtif], and [ˌlokəməˈtif] gets substantial probability at some point of the style range, and that their zones of preference fall in the correct order on the style continuum. The ill-formed candidate *[ˌlokməˈtif] receives close to zero probability (its maximum is in fact 0.009). This seems to us a reasonable depiction of the intuitions given earlier by Dutch-speaking phonologists. The analysis thus implements the idea that *[ˌlokməˈtif] is ill-formed because there is no style level that would support uttering it

It is worth examining just a few more data here. Kager (1989: 303) and (Booij 1999: 134) suggest that every Dutch vowel has a certain *degree of reduceability*. In particular, /a/ is more reducible than /o/, /e/ is more reduceable than /a/; and high vowels are essentially not reducible. Strikingly, for a word like *abracadabra*, with the more reducible vowel /a/, the candidate [ˌabrakəˈdabra], analogous to *[ˌlokəməˈtif], actually sounds not too bad to native speakers. Thus *abracadabra* can be pronounced [ˌabrakaˈdabra], [ˌabrəkaˈdabra], [ˌabrəkəˈdabra], and, to some degree, as [ˌabrakəˈdabra].

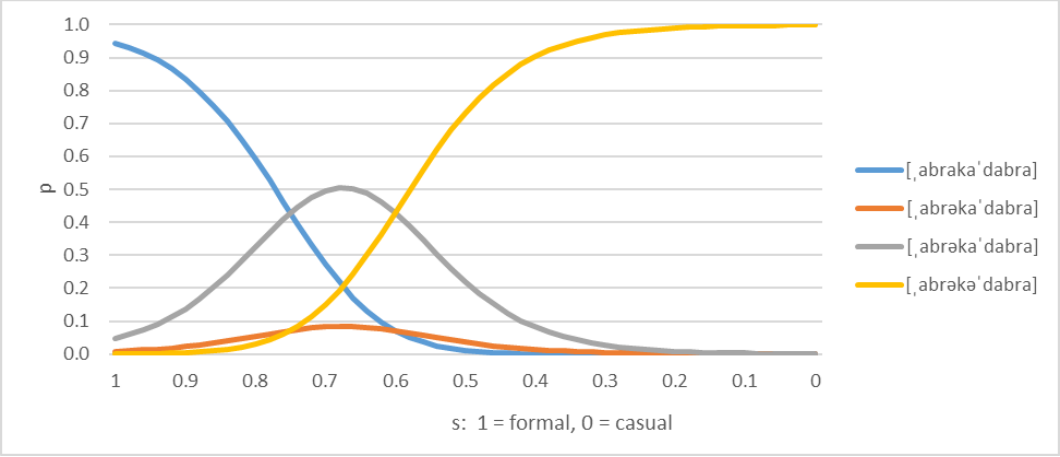
To model this, it is sensible break up the *MAP constraints into versions for distinct vowels, reflecting their observed differences in reducibility. The base weight (prior to multiplication by *s*) of each vowel-specific *MAP constraint is set up to correspond inversely to that vowel's reducibility; we obtain a reasonable outcome by picking the value 12 for *MAP (a, ə), reflecting the greater reduceability of /a/ relative to /o/ (for which the corresponding weight is 30). We also scale the unfooted version of each *MAP constraint so that it is directly related to the weight of the corresponding simple *MAP constraint for each vowel, fixing it at 0.15 times the value of the corresponding simple *MAP constraint (thus for /a/, 1.8 = 0.15 × 12). These choices are summarized in (142).

(142) *Extending the vowel reduction analysis to /a/*

/ˌabrakaˈdabra/	*MAP(a, ə)$_{\text{unfooted}}$	*MAP (a, ə)	*FULL VOWEL	H	eH	p
	1.8	12 × *s*	9.0			
[ˌabrakaˈdabra]			5	45	=exp(−H)	=eH/Z
[ˌabrəkaˈdabra]		1	4	12*s* + 36		
[ˌabrakəˈdabra]	1	1	4	1.8 + 12*s* + 36		
[ˌabrəkəˈdabra]	1	2	3	1.8 + 12*s* + 27		

Calculating the probability curves as before, we obtain Figure 48:

Figure 48: How probabilities for abracadabra *respond to changes in the style parameter* s



We note first that the curves are noticeably shifted leftward relative to Figure 47; this corresponds to the greater "reduceability" of /a/. Second, and importantly, the maximum probability for [ˌabrakəˈdabra], with second vowel full and third reduced, is 0.083, considerably above the corresponding value for [ˌ(loko)məˈtif)] (0.009). Thus, the analysis captures the generalization that the candidate with a reduced third but not second vowel is possible only for vowels that are overall more reducible. The remaining cases — consistent reduction for /e/, very rare reduction for high vowels — also follow from setting the weight of *MAP(V, ə) very low or very high, respectively.

The Noisy Harmonic Grammar analysis of Kaplan (2025), which inspired our work, relies on the property of NHG that it assigns zero probability to harmonically-bounded candidates, in order to rule out *[ˌlokəməˈtif]. But this creates trouble for *abracadabra*, for which the candidate analogous to *[ˌlokəməˈtif], namely [ˌabrakəˈdabra], has to win some of the time. Kaplan accomplishes this by adding an extra constraint, banning unfooted [a], that makes [(ˌabra)kə(ˈdabra)] no longer harmonically bounded. But this is a constraint that specifically penalizes unmarked structure. We feel that the addition of this constraint weakens his argument in favor of adopting a framework that assigns zero frequency for harmonically bounded candidates; the MaxEnt analysis needs no added constraint to derive [ˌabrakəˈdabra].

In sum, this section has outlined a frontier research area, constraint-based modeling of variation as influenced by style and other factors. The example we have presented is anecdotal, and raises two questions that seem at present rather difficult. First, how can we (or indeed the language learner) decide which constraints are responsive to style? Relatedly, what might be a suitable learning algorithm that could construct accurate scale-based MaxEnt grammars, given a suitable set of style-labeled data? The grammar constructed here was made without any algorithm, simply carrying out ad hoc explorations on a spreadsheet.

Looking more widely, there are external factors other than style that also influence varying phonological outputs: Coetzee and Kawahara's (2013) proposal, which inspired our analysis, covers word frequency, to which we may add speaking rate and perhaps also degree of intended clarity (speech for noisy channels; see e.g. Lindblom (1990), Turk and Shattuck-Hufnagel (2020). In the end, an adequate theory of phonology may need multiple parameters like *s*, each one mediating a separate external influence on phonological realization.

Regarding the harmonic bounding controversy, our (very tentative) position is that the math of MaxEnt may well be still a good choice for phonological grammars, and that for cases that look troublesome for MaxEnt in terms of harmonic bounding, an alternative analysis might well be sought in the analysis of style and other factors.

We have enjoyed the challenge of writing a textbook about MaxEnt in part because its status is relatively new and uncertain. We regard the propects of this and related approaches as promising and look forward to what future research will reveal.

References

Albright, Adam. 2002. "Islands of Reliability for Regular Morphology: Evidence from Italian." *Language* 78: 684–709.

Albright, Adam. 2008. "Explaining Universal Tendencies and Language Particulars in Analogical Change." In *Linguistic Universals and Language Change*, edited by Jeff Good. Oxford University Press.

Albright, Adam, and Bruce Hayes. 2003. "Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study." *Cognition* 90: 119–61.

Alderete, John D. n.d. "Structural Disparities in Navajo Word Domains: A Case for LEXCAT-FAITHFULNESS." *The Linguistic Review* 20 (2–4): 111–57.

Alderete, John, and Sarah Finley. 2023. "Probabilistic Phonology: A Review of Theoretical Perspectives, Applications, and Problems." *Language and Linguistics* 24 (4): 565–610.

An, Aixiu. 2020. "Theoretical, Empirical and Computational Approaches to Agreement with Coordination Structures." Ph.D. dissertation, Université Paris Cité.

An, Aixiu, Anne Abeillé, and Roger Levy. submitted. *Gradience in the Grammar of Agreement: The Case of French Coordination*.

AnderBois, Scott, Adrian Brasoveanu, and Robert Henderson. 2012. "The Pragmatics of Quantifier Scope: A Corpus Study." In *Proceedings of Sinn Und Bedeutung*, 1st ed., vol. 16.

Anderson, Gregory D. S. 2013. "The Velar Nasal." In *WALS Online*, edited by Matthew S. Dryer and Martin Haspelmath, v2020.4. Zenodo.

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21.

Anttila, Arto. 1995. "How to Recognise Subjects in English." In *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, edited by Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. Mouton de Gruyter.

Anttila, Arto. 1997. "Deriving Variation from Grammar: A Study of Finnish Genitives." In *Variation, Change, and Phonological Theory*, edited by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels. John Benjamins Publishing Company.

Anttila, Arto, Adams, Matthew, and Speriosu, Michael. 2010. "The Role of Prosody in the English Dative Alternation." *Language and Cognitive Processes* 25: 946–941.

Anttila, Arto, and Giorgio Magri. 2018. "Does MaxEnt Overgenerate? Implicational Universals in Maximum Entropy Grammar." *Proceedings of the 2017 Annual Meeting on Phonology* 5.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. MIT Press.

Baayen, Harald, Robert Schreuder, Nivja de Jong, and Andrea Krott. 2002. "Dutch Inflection: The Rules That Prove the Exception." In *Storage and Computation in the Language Faculty*, edited by Sieb Nooteboom, Fred Weerman, and Frank Wijnen. Springer.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Bailey, Todd M., and Ulrike Hahn. 2001. "Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?" *Journal of Memory and Language* 44: 568–91.

Baker, Robert G., and Philip T. Smith. 1976. "A Psycholinguistics Study of English Stress Assignment Rules." *Language and Speech* 19: 9–27.

Bárkányi, Zsuzsanna. 2011. "Gradient Phonotactic Acceptability: A Case Study from Slovak." *Acta Linguistica Hungarica* 58 (4): 353–91.

Bayley, Robert. 2002. "The Quantitative Paradigm." In *The Handbook of Language Variation and Change*, 1st ed., edited by J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes. Blackwell Publishers.

Becker, Michael. 2009. "Phonological Trends in the Lexicon: The Role of Constraints." PhD Thesis, University of Massachusetts Amherst.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. "The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternations." *Language* 87 (1): 84–125.

Becker, Michael, Andrew Nevins, and Jonathan Levine. 2012. "Asymmetries in Generalizing Alternations to and from Initial Syllables." *Language* 88 (2): 231–68.

Becker, Michael, and Anne-Michelle Tessier. 2011. "Trajectories of Faithfulness in Child-Specific Phonology." *Phonology*, no. 2: 163–96.

Becker-Kristal, Roy. 2010. "Acoustic Typology of Vowel Inventories and Dispersion Theory: Insights from a Large Cross-Linguistic Corpus." Ph.D. dissertation, UCLA.

Beckman, Jill N. 1997. "Positional Faithfulness, Positional Neutralisation and Shona Vowel Harmony." *Phonology* 14 (01): 1–46.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Smargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Conference on Fairness, Accountability, and Transparency (FAccT '21)* (New York), 14 pages.

Benor, Sarah Bunin, and Roger Levy. 2006. "The Chicken or the Egg? A Probabilistic Analysis of English Binomials." *Language* 82: 233–78.

Benua, Laura. 1997. "Transderivational Identity: Phonological Relations between Words." PhD Thesis, University of Massachusetts, Amherst.

Berent, Iris, Donca Steriade, Tracy Lennertz, and Vered Vaknin. 2007. "What We Know about What We Have Never Heard: Evidence from Perceptual Illusions." *Cognition* 104 (3): 591–630.

Berent, Iris, Colin Wilson, Gary F. Marcus, and Douglas K. Bemis. 2012. "On the Role of Variables in Phonology: Remarks on Hayes and Wilson (2008)." *Linguistic Inquiry* 43: 97–119.

Berko, Jean. 1958. "The Child's Learning of English Morphology." *Word* 14 (2–3): 150–77.

Bermudez-Otero, Ricardo. 2018. “Stratal Phonology.” In *The Routledge Handbook of Phonological Theory*.

Blevins, Juliette. 1997. “Rules in Optimality Theory: Two Case Studies.” In *Derivations and Constraints in Phonology*, edited by Iggy Roca. Oxford University Press.

Blevins, Juliette. 2004. *Evolutionary Phonology*. Cambridge University Press.

Bloomfield, Leonard. 1917. *Tagalog Texts with Grammatical Analysis*. With Alfredo Viola Santiago. University of Illinois.

Blythe, Richard A., and William Croft. 2012. “S-Curves and the Mechanisms of Propagation in Language Change.” *Language* 88 (2): 269–304.

Boersma, Paul. 1998. “Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives.” Ph.D. dissertation, University of Amsterdam.

Boersma, Paul, and Bruce Hayes. 2001. “Empirical Tests of the Gradual Learning Algorithm.” *Linguistic Inquiry* 32 (1): 45–86.

Boersma, Paul, and Joe Pater. 2016. “Convergence Properties of a Gradual Learning Algorithm for Harmonic Grammar.” In *Harmonic Grammar and Harmonic Serialism*, edited by John McCarthy and Joe Pater. Equinox Publishing Ltd.

Boersma, Paul, David Weenink, and Anastasia Shchupak. 2026. *Praat: Doing Phonetics by Computer [Computer Program]*. Released.

Bonami, Olivier, and Gilles Boyé. 2002. “Suppletion and Dependency in Inflectional Morphology.” In *Proceedings of the 8th International HPSG Conference*, edited by Frank van Eynde, Lars Hellan, and Dorothee Beermann. CSLI Publications.

Booij, Geert. 1977. *Dutch Morphology: A Study of Word Formation in Generative Grammar*. Foris Publications.

Booij, Geert. 1999. *The Phonology of Dutch*. Oxford University Press.

Bowerman, Melissa. 1988. “The ‘no Negative Evidence’ Problem: How Do Children Avoid Constructing an Overly General Grammar?” In *Explaining Language Universals*, edited by Edith A. Moravcsik. Basil Blackwell.

Breiss, Canaan. 2020. “Constraint Cumulativity in Phonotactics: Evidence from Artificial Grammar Learning Studies.” *Phonology* 37: 551–76.

Breiss, Canaan. 2024. “When Bases Compete: A Voting Model of Lexical Conservatism.” *Language* 100 (2): 308–58.

Breiss, Canaan, and Adam Albright. 2022. “Cumulative Markedness Effects and (Non-)Linearity in Phonotactics.” *Glossa: A Journal of General Linguistics* 7: 1–32.

Breiss, Canaan, and Bruce Hayes. 2020. “Phonological Markedness Effects in Sentence Formation.” *Language* 96 (2): 338–70.

Breiss, Canaan, Hironori Katsuda, and Shigeto Kawahara. 2022. "A Quantitative Study of Voiced Velar Nasalization in Japanese." *Proceedings of the 45th Annual Penn Linguistics Conference*.

Breiss, Canaan, Hironori Katsuda, and Shigeto Kawahara. 2024. "Paradigm Uniformity Is Probabilistic: Evidence from Velar Nasalization in Japanese." *Proceedings of the 39th West Coast Conference on Formal Linguistics* 39.

Breiss, Canaan, Hironori Katsuda, and Shigeto Kawahara. 2026. "Token Frequency Modulates Optional Paradigm Uniformity in Japanese Voiced Velar Nasalisation." *Phonology* 43: e7.

Brentari, Diane, and Carol A. Padden. 2001. "Native and Foreign Vocabulary in American Sign Language: A Lexicon with Multiple Origins." In *Foreign Vocabulary in Sign Languages*. Psychology Press.

Bresnan, Joan. 2000. "Optimal Syntax." In *Optimality Theory: Phonology, Syntax, and Acquisition.* Clarendon Press.

Bresnan, Joan, Asudeh, Ash, Toivonen, Ida, and Wechsler, Stephen. 2015. *Lexical-Functional Syntax*. 2nd ed. Wiley-Blackwell.

Bresnan, Joan, Shipra Dingare, and Christopher D. Manning. 2001. "Soft Constraints Mirror Hard Constraints: Voice and Person in English and Lummi." *Proceedings of the LFG01 Conference* (Stanford, CA) 13: 32.

Bresnan, Joan, and Marilyn Ford. 2010. "Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English." *Language* 86 (1): 168–213.

Broselow, Ellen. 1980. "Syllable Structure in Two Arabic Dialects." *Studies in the Linguistic Sciences* 10 (2): 13–24.

Brysbaert, Marc, and Boris New. 2009. "Moving beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English." *Behavior Research Methods* 41 (4): 977–90.

Burnham, Kenneth P., and David R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer.

Bybee, Joan, and Paul Hopper. 2001. "Introduction to Frequency and the Emergence of Linguistic Structure." In *Frequency and the Emergence of Linguistic Structure*, edited by Joan Bybee and Paul Hopper, vol. 45. Typological Studies in Language. John Benjamins.

Bynon, Theodora. 1977. *Historical Linguistics*. Cambridge University Press.

Cabrera, Marisabel. 2025. "One-Shot vs. Competitions Phonotactics in Modeling Constraint Cumulativity." *Proceedings of the Annual Meetings on Phonology*.

Cairo, Alberto. 2013. *The Functional Art: An Introduction to Information Graphics and Visualization*.

Casali, Roderic F. 1997. "Vowel Elision in Hiatus Contexts: Which Vowel Goes?" *Language* 73 (3): 493–533.

Cedergren, Henrietta J., and David Sankoff. 1974. "Variable Rules: Performance as a Statistical Reflection of Competence." *Language* 50 (2): 333–55.

Çetinkaya-Rundel, Mine, and Johanna Hardin. 2024. *Introduction to Modern Statistics*. 2nd ed. OpenIntro.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton de Gruyter.

Chomsky, Noam. 1962. "A Transformational Approach to Syntax." *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English* (Austin, TX), 124–58.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, Noam, and Morris Halle. 1965. "Some Controversial Questions in Phonological Theory." *Journal of Linguistics* 1 (2): 97–214.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. Harper and Row.

Chong, Adam J. 2021. "The Effect of Phonotactics on Alternation Learning." *Language* 97: 213–44.

Clark, Brady. 2005. "On Stochastic Grammar." *Language* 81: 207–17.

Clements, George N., and Engin Sezer. 1982. "Vowel and Consonant Disharmony in Turkish." In *The Structure of Phonological Representations (Part II)*, edited by Harry van der Hulst and Norval Smith. Foris Publications.

Coetzee, Andries W. 2000. "Post-Nasal Neutralization Phenomena in Tswana; More on *N... Constraints." Unpublished manuscript. University of Massachusetts, Amherst.

Coetzee, Andries W., and Shigeto Kawahara. 2013. "Frequency Biases in Phonological Variation." *Natural Language and Linguistic Theory* 31: 47–89.

Coetzee, Andries W., and Joe Pater. 2008. "Weighted Constraints and Gradient Restrictions on Place Co-Occurrence in Muna and Arabic." *Natural Language & Linguistic Theory* 26 (2): 289–337.

Coetzee, Andries W., and Joe Pater. 2011. "The Place of Variation in Phonological Theory." In *The Handbook of Phonological Theory*, 1st ed., edited by John Goldsmith, Jason Riggle, and Alan C. L. Yu, vol. 75. Blackwell.

Coleman, John, and Janet Pierrehumbert. 1997. "Stochastic Phonological Grammars and Acceptability." *Computational Phonology: The Third Meeting of the ACL Special Interest Group in Computational Phonology* (Somerset, NJ), 49–56.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, et al. 2017. "CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages." Paper presented at CoNLL-SIGMORPHON 2017. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*.

Crosswhite, Katherine. 2004. "Vowel Reduction." In *Phonetically Based Phonology*, edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge University Press.

Culbertson, Jennifer. 2023. "Artificial Language Learning." In *Oxford Handbook of Experimental Syntax*, edited by Jon Sprouse. Oxford University Press.

Culbertson, Jennifer, Paul Smolensky, and Colin Wilson. 2013. "Cognitive Biases, Linguistic Universals, and Constraint-Based Grammar Learning." *Topics in Cognitive Science* 5 (3): 392–424.

Cuskley, Christine, Rebecca Woods, and Molly Flaherty. 2024. "The Limitations of Large Language Models for Understanding Human Language and Cognition." *Open Mind: Discoveries in Cognitive Science* 8: 1058–83.

Dabouis, Quentin, and Pierre Fournier. 2022. "English Phonologies?" *Models and Modelisation in Linguistics*.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2018. "TiMBL: Tilburg Memory-Based Learner Reference Guide, Version 6.4."

Dai, Huteng. 2026. *Phonology Lab*. Released.

Daland, Robert. 2015. "Long Words in Maximum Entropy Phonotactic Grammars." *Phonology* 32 (3): 353–83.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrmann. 2011. "Explaining Sonority Projection Effects." *Phonology* 29: 197–234.

Daland, Robert, Mira Oh, and Syejeong Kim. 2015. "When in Doubt, Read the Instructions: Orthographic Effects in Loanword Adaptation." *Lingua* 159: 70–92.

Davidson, Lisa, and Jason Shaw. 2012. "Sources of Illusion in Consonant Cluster Perception." *Journal of Phonetics* 40 (2): 234–48.

De Morgan, Augustus. 1847. *Formal Logic, or, the Calculus of Inference, Necessary and Probable*. Taylor and Walton.

Dell, François. 1981. "On the Learnability of Optional Phonological Rules." *Linguistic Inquiry* 12: 31–37.

Della Pietra, Stephen, Vincent Della Pietra, and John Lafferty. 1997. "Inducing Features of Random Fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (4): 380–93.

Dixon, R. M. W. 1981. "Wargamay." In *Handbook of Australian Languages*, edited by R. M. W. Dixon and Barry J. Blake, vol. 2. John Benjamins.

Do, Youngah, and Ping Hwei Yeung. 2021. "Evidence against a Link between Learning Phonotactics and Learning Phonological Alternations." *Linguistics Vanguard* 7 (1): 20200127.

Domahs, Ulrike, Ingo Plag, and Rebecca Carroll. 2014. "Word Stress Assignment in German, English and Dutch: Quantity-Sensitivity and Extrametricality Revisited." *The Journal of Comparative Germanic Linguistics* 17 (1): 59–96.

Du, Naiyan, and Karthik Durvasula. 2025. *Psycholinguistics and Phonology: The Forgotten Foundations of Generative Phonology*. Cambridge University Press.

Dufour, Jean-Marie, and Julien Neves. 2019. "Finite-Sample Inference and Nonstandard Asymptotics with Monte Carlo Tests and R." In *Handbook of Statistics*, vol. 41. Elsevier.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler. 1999. "Epenthetic Vowels in Japanese: A Perceptual Illusion?" *Journal of Experimental Psychology: Human Perception and Performance* 25 (6): 1568.

Durand, Jacques, Ulrike Gut, and Gjert Kristoffersen, eds. 2014. *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

Eddington, David. 1996. "Diphthongization in Spanish Derviational Morphology: An Empirical Investigation." *Hispanic Linguistics* 8: 1–35.

Eddington, David. 2004. "Issues in Modeling Language Processing Analogically." *Lingua* 114: 849–71.

Eisner, Jason. 1997a. "Efficient Generation in Primitive Optimality Theory." *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (Madrid, Spain), 313–20.

Eisner, Jason. 1997b. "What Constraints Should OT Allow?" Paper presented at 71st Annual meeting of the Linguistic Society of America.

Eisner, Jason. 2000. Review of *Review of Kager: "Optimality Theory,"* by René Kager. *Computational Linguistics* 26 (2): 286–90.

Eisner, Jason. 2001. "Expectational Semirings: Flexible EM for Finite-State Transducers." In *Proceedings of the ESSLLI Workshop on Finite-State Methods in NLP (FSMNLP)*, edited by Gertjan van Noord.

Eisner, Jason. 2002. "Parameter Estimation for Probabilistic Finite-State Transducers." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (East Stroudsburg, PA), 1–8.

Elfner, Emily. 2018. "The Syntax-Prosody Interface: Current Theoretical Approaches and Outstanding Questions." *Linguistics Vanguard* 4 (1): 20160081.

Ellison, T. Mark. 1994. "Phonological Derivation in Optimality Theory." *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics* 2.

Ernestus, Miriam, and R. Harald Baayen. 2003. "Predicting the Unpredictable: Interpreting Neutralized Segments in Dutch." *Language* 79 (1): 5–38.

Ernestus, Mirjam. 2011. "Gradience and Caregoricality in Phonological Theory." In *The Blackwell Companion to Phonology, Volume IV: Phonological Interfaces*, edited by Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice. Wiley-Blackwell.

Ernestus, Mirjam, and R. Harald Baayen. 2011. "Corpora and Exemplars in Phonology." In *The Handbook of Phonological Theory*, 2nd ed., edited by John Goldsmith, Jason Riggle, and Alan S. Yu. John Wiley.

Farris-Trimble, Ashley. 2008. "Cumulative Faithfulness Effects in Phonology." Ph.D. dissertation, Indiana University.

Featherston, Sam. 2005a. "Magnitude Estimation and What It Can Do for Your Syntax: Some Wh-Constraints in German." *Lingua* 115 (11): 1525–50.

Featherston, Sam. 2005b. "The Decathlon Model of Empirical Syntax." In *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, edited by Stephan Kepser and Marga Reis, vol. 85. Mouton de Gruyter.

Featherston, Sam. 2019. "The Decathlon Model." In *Current Approaches to Syntax: A Comparative Handbook*, 1st ed., edited by András Kertész, Edith Moravcsik, and Csilla Rákosi, vol. 3. De Gruyter Mouton.

Feldman, Jacob. 2015. "Bayesian Models of Perception: A Tutorial Introduction." In *The Oxford Handbook of Perceptual Organization*, edited by Johan Wagemans.

Field, Andy. 2026. *Discovering Statistics Using R and RStudio*. London. Sage.

Finkel, Raphael, and Gregory Stump. 2007. "Principal Parts and Morphological Typology." *Morphology* 17: 39–75.

Finley, Sarah. 2017. "Locality and Harmony: Perspectives from Artificial Grammar Learning." *Language and Linguistics Compass* 11 (1): e12233.

Flemming, Edward. 2001. "Scalar and Categorical Phenomena in a Unified Model of Phonetics and Phonology." *Phonology* 18 (1): 7–44.

Flemming, Edward. 2004. "Contrast and Perceptual Distinctiveness." In *Phonetically Based Phonology*, edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge University Press.

Flemming, Edward. 2021. "Comparing MaxEnt and Noisy Harmonic Grammar." *Glossa: A Journal of General Linguistics* 6: 1–42.

Flemming, Edward, and Hyesun Cho. 2017. "The Phonetic Specification of Contour Tones: Evidence from the Mandarin Rising Tone." *Phonology* 34 (1): 1–40.

Forró, Orsolya. 2013. "Variation in Hungarian Backness Harmony: Aspects of the Investigation and Interpretation of the Synchrony and Diachrony of Variability." Ph.D. dissertation, Pázmány Péter Catholic University.

Francis, Elaine. 2022. *Gradient Acceptability and Linguistic Theory*. Oxford University Press.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. "Similarity Avoidance and the OCP." *Natural Language & Linguistic Theory* 22 (1): 179–228.

Fukazawa, Haruka, and Linda Lombardi. 2003. "Complex Constraints and Linguistic Typology in Optimality Theory." *The Linguistic Review* 20 (2–4): 195–215.

Fuller, Corinna Lee. 2024. "Word-Initial /p/ in Khalkha Mongolian: Variation in Connected Speech." M.A. Thesis, University of California, Los Angeles.

Futrell, Richard, Adam Albright, Peter Graff, and Timothy J. O'Donnell. 2017. "A Generative Model of Phonotactics." *Transactions of the Association for Computational Linguistics* 5: 73–86.

Gabriel, Christoph. 2010. "On Focus, Prosody, and Word Order in Argentinian Spanish: A Minimalist OT Account." *Revista Virtual de Estudos Da Linguagem (ReVEL)* Special edition n. 4.

Garcia, Guilherme. 2019. "When Lexical Statistics and the Grammar Conflict: Learning and Repairing Weight Effects on Stress." *Language* 95 (4): 612–41.

Garcia, Guilherme D. 2017. "Weight Gradience and Stress in Portuguese." *Phonology* 34 (1): 41–79.

Garcia, Guilherme Duarte. 2026. *Fonology: Phonological Analysis in R*. Released.

Gibson, Edward, and Evelina Fedorenko. 2013. "The Need for Quantitative Methods in Syntax and Semantics Research." *Language and Cognitive Processes* 28: 88–124.

Gigerenzer, Gerd. 2001. "The Adaptive Toolbox." In *Bounded Rationality: The Adaptive Toolbox*, edited by Gerd Gigerenzer and R. Selten. MIT Press.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. "Heuristic Decision Making." *Annual Review of Psychology* 62: 451–82.

Glewwe, Eleanor. 2021. "The Phonological Determinants of Tone in English Loanwords in Mandarin." *Phonology* 38 (2): 203–39.

Gnanadesikan, Amalia. 2004. "Markedness and Faithfulness Constraints in Child Phonology." In *Constraints in Phonological Acquisition*, edited by René Kager, Joe Pater, and Zonneveld. Cambridge University Press.

Goldsmith, John. 1976. "Autosegmental Phonology." Ph.D. dissertation, MIT.

Goldsmith, John, and Jason Riggle. 2012. "Information Theoretic Approaches to Phonological Structure: The Case of Finnish Vowel Harmony." *Natural Language & Linguistic Theory* 30: 859–96.

Goldsmith, John, and Aris Xanthos. 2009. "Learning Phonological Categories." *Language* 85: 4–38.

Goldwater, Sharon, and Mark Johnson. 2003. "Learning OT Constraint Rankings Using a Maximum Entropy Model." In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, edited by Jennifer Spenader, Anders Eriksson, and Osten Dahl.

Gong, Shuxiao, and Jie Zhang. 2020. "Gradient Acceptability in Mandarin Nonword Judgment." In *Proceedings of the Annual Meetings on Phonology*, edited by Hyunah Baek, Chikako Takahashi, and Alex Hong-Lun Yeung.

Gordon, Matthew. 2001. "A Typology of Countour Tone Restrictions." *Studies in Language* 25: 423–62.

Gordon, Matthew. 2002. "A Factorial Typology of Quantity-Insensitive Stress." *Natural Language & Linguistic Theory* 20: 491–552.

Gordon, Matthew. 2016. *Phonological Typology*. Oxford University Press.

Gouskova, Maria, and Suzy Ahn. 2025. "Sublexical Phonotactics and English Comparatives." In *UMOP 42: Papers in Celebration of the 50th Anniversary of UMass Linguistics*, vol. 42. University of Massachusetts Occasional Papers in Linguistics. GLSA, UMASS.

Gouskova, Maria, and Gillian Gallagher. 2020. "Inducing Nonlocal Constraints from Baseline Phonotactics." *Natural Language & Linguistic Theory* 38: 77–116.

Gouskova, Maria, Newlin-Łukowicz, and Sofya Kasyanenko. 2015. "Selectional Restrictions as Phonotactics over Sublexicons." *Lingua* 167: 41–81.

Greenberg, Joseph H., and James J. Jenkins. 1964. "Studies in the Psychological Correlates of the Sound System of American English." *Word* 20: 157–77.

Guion, Susan G., J. J. Clark, Tetsuo Harada, and Ratree P. Wayland. 2003. "Factors Affecting Stress Placement for English Nonwords Include Syllabic Structure, Lexical Class, and Stress Patterns of Phonologically Similar Words." *Language and Speech* 46 (4): 403–27.

Gurevich, Naomi. 2011. "Lenition." In *The Blackwell Companion to Phonology*, 1st ed., edited by Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, vol. 3. Blackwell Publishers.

Guy, Gregory. 2011. "Variability." In *The Blackwell Companion to Phonology*, edited by Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice. Wiley-Blackwell.

Guy, Gregory R. 1993. "The Quantitative Analysis of Linguistic Variation." In *American Dialect Research: Celebrating the 100th Anniversary of the American Dialect Society, 1889-1989*, edited by Dennis R. Preston, with John G. Fought. John Benjamins Publishing Company.

Hájek, Alan. 2023. "Interpretations of Probability." In *The Stanford Encyclopedia of Philosophy*, Winter 2023, edited by Edward N. Zalta and Uri Nodelman.

Halle, Morris. 1978. "Knowledge Unlearned and Untaught: What Speakers Know about the Sounds of Their Language." In *Linguistic Theory and Psychological Reality*, edited by Morris Halle, Joan Bresnan, and George A. Miller. MIT Press.

Halle, Morris, and Jay S. Keyser. 1966. "Chaucer and the Study of Prosody." *College English* 28: 187–219.

Halle, Morris, and Jay S. Keyser. 1971. *English Stress: Its Form, Its Growth, and Its Role in Verse*. Harper and Row.

Halle, Morris, and Jean-Roger Vergnaud. 1981. "Harmony Processes." In *Crossing the Boundaries in Linguistics*, edited by W. Klein, J. Hintikka, and S. Peters. Springer.

Hannan, M. 1959. *Standard Shona Dictionary*. St. Martin's Press.

Hansen, Kenneth C., and Hansen, Lesly E. 1969. "Pintupi Phonology." *Oceanic Linguistics* 8 (2): 153–70.

Hansen, Kenneth C., and Hansen, Lesly E. 1978. *The Core of Pintupi Grammar*. Institute for Aboriginal Development.

Harris, James Wesley. 1969. *Spanish Phonology*. MIT Press.

Hayes, Bruce. 1986. "Inalterability in CV Phonology." *Language* 62: 321–51.

Hayes, Bruce. 1989. "Compensatory Lengthening in Moraic Phonology." *Linguistic Inquiry* 20: 253–306.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press.

Hayes, Bruce. 1999a. "Phonetically Driven Phonology: The Role of Optimality Theory and Inductive Grounding." In *Functionalism and Formalism in Linguistics, Vol. I: General Papers*, edited by M.Darnell, E. Moravscik, M. Noonan, F. J. Newmeyer, and K. Wheatly. John Benjamins.

Hayes, Bruce. 1999b. "Phonological Restructuring in Yidiɲ and Its Theoretical Consequences." In *The Derivational Residue in Phonological Optimality Theory*, edited by Ben Hermans and Marc van Oostendorp. John Benjamins.

Hayes, Bruce. 2000. "Gradient Well-Formedness in Optimality Theory." In *Optimality Theory: Phonology, Syntax, and Acquisition*, edited by Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer. Oxford University Press.

Hayes, Bruce. 2004. "Phonological Acquisition in Optimality Theory: The Early Stages." In *Constraints in Phonological Acquisition*, edited by René Kager, Joe Pater, and Wim Zonneveld. Cambridge University Press.

Hayes, Bruce. 2008. *Introductory Phonology*. 1st ed. Wiley-Blackwell.

Hayes, Bruce. 2010. "Review of Fabb and Halle (2010) Meter in Poetry." *Lingua* 120: 2515–21.

Hayes, Bruce. 2012. "The Role of Computational Modeling in the Study of Sound Structure." Talk given at the Conference on Laboratory Phonology.

Hayes, Bruce. 2016. "Comparative Phonotactics." *Proceedings of the 50th Annual Meeting of the Chicago Linguistic Society* 50: 265–85.

Hayes, Bruce. 2017. "Varieties of Noisy Harmonic Grammar." In *Proceedings of the 2016 Annual Meeting on Phonology*, edited by Karen Jesney, Charlie O'Hara, Caitlin Smith, and Rachel Walker, vol. 4.

Hayes, Bruce. 2022a. "Deriving the Wug-Shaped Curve: A Criterion for Assessing Formal Theories of Linguistic Variation." *Annual Review of Linguistics* 8 (1): 473–94.

Hayes, Bruce. 2022b. *Gallery of Wug-Shaped Curves*. Personal Website.

Hayes, Bruce. 2022c. "The Coinages in Seuss." *English Language & Linguistics* 26 (3): 487–511.

Hayes, Bruce, and Aaron Kaplan. 2025. "Zero-Weighted Constraints in Noisy Harmonic Grammar." *Linguistic Inquiry* 56: 605–18.

Hayes, Bruce, and Zsuzsa Cziráky Londe. 2006. "Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony." *Phonology* 23 (1): 59–104.

Hayes, Bruce, and Donka Minkova. in press. "A MaxEnt Analysis of the Beowulf Meter." *Phonology*.

Hayes, Bruce, and Russell G. Schuh. 2019. "Metrical Structure and Sung Rhythm of the Hausa Rajaz." *Language* 95 (2): e253–99.

Hayes, Bruce, and Tanya Stivers. 2000. "Postnasal Voicing." Preprint, University of California, Los Angeles.

Hayes, Bruce, Bruce Tesar, and Kie Zuraw. 2026. *OTSoft*. V. 2.5. Released.

Hayes, Bruce, and James White. 2013. "Phonological Naturalness and Phonotactic Learning." *Linguistic Inquiry* 44 (1): 45–75.

Hayes, Bruce, and James White. 2015. "Saltation and the P-Map." *Phonology* 32: 267–302.

Hayes, Bruce, and Colin Wilson. 2008. "A Maximum Entropy Model of Phonotactics and Phontactic Learning." *Linguistic Inquiry* 39 (3): 379–440.

Hayes, Bruce, Colin Wilson, and Anne Shisko. 2012. "Maxent Grammars for the Metrics of Shakespeare and Milton." *Language* 88: 691–731.

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. 2009a. "Natural and Unnatural Constraints in Hungarian Vowel Harmony." *Language* 85 (4): 822–63.

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Cziráky Londe. 2009b. "Natural and Unnatural Constraints in Hungarian Vowel Harmony." *Language* 85 (4): 822–63.

Henriksson, Erik. 2022. "Greek Meter: An Approach Using Metrical Grids and Maxent." Ph.D. dissertation, University of Helsinki.

Hilpert, Martin. 2008. "The English Comparative - Language Structure and Language Use." *English Language & Linguistics* 12 (3): 395–417.

Hsu, Anne S., Nick Chater, and Paul Vitanyi. 2013. "Language Learning from Positive Evidence, Reconsidered: A Simplicity-Based Approach." *Topics in Cognitive Science* 5: 35–55.

Hudson Kam, Carla L., and Elissa L. Newport. 2005. "Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change." *Language Learning and Development* 1: 151–95.

Hughto, Coral, Andrew Lamont, Brandon Prickett, and Gaja Jarosz. 2019. "Learning Exceptionality and Variation with Lexically Scaled MaxEnt." *Proceedings of the Society for Computation in Linguistics (SCiL)*, 91–101.

Hyman, Larry. 1985. *A Theory of Phonological Weight*. Foris.

Inkelas, Sharon, Aylin Küntav, Orhan Orgun, and Ronald Sprouse. 2000. "Turkish Electronic Living Lexicon (TELL)." *Turkic Languages* 4: 253–75.

Irvine, Ann, and Mark Dredze. 2017. "Harmonic Grammar, Optimality Theory, and Syntax Learnability: An Empirical Exploration of Czech Word Order." arXiv:1702.05793. Preprint, Department of Computer Science, Johns Hopkins University.

Itô, Junko, and Armin Mester. 1996. *Correspondence and Compositionality: The Ga-Gyō Variation in Japanese Phonology*. Rutgers Optimality Archive No. 145.

Itô, Junko, and Armin Mester. 1999. "The Structure of the Phonological Lexicon." In *The Handbook of Japanese Linguistics*, edited by Natsuko Tsujimura. Blackwell Publishers.

Itô, Junko, and Armin Mester. 2003. "On the Sources of Opacity in OT: Coda Processes in German." In *The Syllable in Optimality Theory*, edited by Caroline Féry and Ruben van de Vijver. Cambridge University Press.

Itô, Junko, and Armin Mester. 2016. "Unaccentedness in Japanese." *Linguistic Inquiry* 47: 471–526.

Iverson, Gregory, and Joseph Salmons. 2005. "Filling the Gap." *Journal of English Linguistics* 33: 1–15.

Jäger, Gerhard. 2007. "Maximum Entropy Models and Stochastic Optimality Theory." In *Architectures, Rules and Preferences: Variations on Themes by Joan W. Bresnan*, edited by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Christopher D. Manning. CSLI Publications.

Jäger, Gerhard, and Anette Rosenbach. 2006. "The Winner Takes It All - Almost: Cumulativity in Grammatical Variation." *Linguistics* 44: 937–71.

Janda, Richard, Brian Joseph, and Neil Jacobs. 1994. "Systematic Hyperforeignisms as Maximally External Evidence for Linguistic Rules." In *The Reality of Linguistic Rules*, edited by S. Lima, R. Corrigan, and G. Iverson. John Benjamins.

Jarosz, Gaja. 2006. "Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory." Ph.D. dissertation, Johns Hopkins University.

Jarosz, Gaja. 2011. "The Roles of Phonotactics and Frequency in the Learning of Alternations." In *Proceedings of the 35th Annual Boston University Conference on Language Development*, edited by Nick Danis, Kate Mesh, and Sung. Cascadilla Press.

Jarosz, Gaja. 2013. "Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing." *Phonology* 30: 27–71.

Jarosz, Gaja. 2017. "Defying the Stimulus: Acquisition of Complex Onsets in Polish." *Phonology* 34 (2): 269–98.

Jarosz, Gaja. 2019. "Computational Modeling of Phonological Learning." *Annual Review of Linguistics*, no. 5: 67–90.

Jarosz, Gaja, Cerys Hughes, Andrew Lamont, et al. 2025. "Type and Token Frequency Jointly Drive Learning of Morphology." *Journal of Memory and Language* 144: 104666.

Jarosz, Gaja, and Amanda Rysling. 2016. "Sonority Sequencing in Polish: The Combined Roles of Prior Bias & Experience." *Proceedings of the Annual Meetings on Phonology*.

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Edited by G. Larry Bretthorst. Cambridge University Press.

Jesney, Karen. 2007. "The Locus of Variation in Weighted Constraint Grammars." Paper presented at Workshop on Variation, Gradience and Frequency in Phonology.

Jesney, Karen, and Anne-Michelle Tessier. 2011. "Biases in Harmonic Grammar: The Road to Restrictive Learning." *Natural Language & Linguistic Theory* 29 (1): 251–90.

Jo, Jinyoung. 2024a. "Individual Differences in Phonology: The Realization of Stem-Final Coronal Obstruents in Korean." Ph.D. dissertation, UCLA.

Jo, Jinyoung. 2024b. "Korean Vowel Harmony Has Weak Phonotactic Support and Has Limited Productivity." *Phonology* 40 (1–2): 65–100.

Johnson, Mark. 2002. "Optimality-Theoretic Lexical Functional Grammar." In *The Lexical Basis of Setnence Processing: Formal, Computational and Experimental Issues*, edited by Paola Merlo and Suzanne Stevenson, vol. 3. Natural Language Processing 4. John Benjamins Publishing Company.

Jun, Jongho. 2004. "Place Assimilation." In *Phonetically Based Phonology*, edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge University Press.

Jun, Jongho. 2010. "Stem-Final Obstruent Variation in Korean." *Journal of East Asian Linguistics* 19: 137–79.

Jun, Jongho. 2015. "Korean N-Insertion: A Mismatch between Data and Learning." *Phonology* 32 (3): 417–58.

Jun, Jongho, and Adam Albright. 2017. "Speakers' Knowledge of Alternations Is Asymmetrical: Evidence from Seoul Korean Verb Paradigms." *Journal of Linguistics* 53 (3): 567–611.

Jun, Jongho, Hanyoung Byun, Seon Park, and Yoona Yee. 2025. "How Tight Is the Link between Alternations and Phonotactics?" *Phonology* 42: e3.

Jurafsky, Daniel, and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed.

Jurafsky, Daniel, and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed. Online manuscript released January 6.

Jurgec, Peter, and Jessamyn Schertz. 2020. "Postalveolar Co-Occurrence Restrictions in Slovenian." *Natural Language & Linguistic Theory* 38 (2): 499–537.

Kager, René. 1989. *A Metrical Theory of Stress and Destressing in English and Dutch*. Foris.

Kager, René. 1999. *Optimality Theory*. Cambridge University Press.

Kahn, Daniel. 1976. *Syllable-Based Generalizations in English Phonology*. Vol. 156. Indiana University Linguistics Club Bloomington.

Kahng, Jimin, and Karthik Durvasula. 2023. "Can You Judge What You Don't Hear? Perception as a Source of Gradient Wordlikeness Judgements." *Glossa: A Journal of General Linguistics* 8 (1).

Kang, Yoonjung. 2003. "Sound Changes Affecting Noun-Final Coronal Obstruents in Korean." *Japanese/Korean Linguistics* 12: 128–39.

Kaplan, Aaron. 2018. "Positional Licensing, Asymmetric Trade-Offs and Gradient Constraints in Harmonic Grammar." *Phonology* 35 (2): 247–86.

Kaplan, Aaron. 2021. "Categorical and Gradient Ungrammaticality in Optional Processes." *Language* 97 (4): 703–31.

Kaplan, Aaron. 2024. "Coordinated and Local Optionality in Serial Noisy Harmonic Grammar." *Natural Language & Linguistic Theory* 42 (4): 1415–60.

Kaplan, Aaron. 2025. "Trial Cancellation in NHG." In *Proceedings of the 41st West Coast Conference on Formal Linguistics*, edited by Nikolas Webster, Yağmut Kiper, Richard Wang, and Sichen Larry Lyu. Cascadilla Proceedings Project.

Karttunen, Lauri. 2006. "The Insufficiency of Paper-and-Pencil Linguistics: The Case of Finnish Prosody." In *Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan*, edited by Miriam Butt, Mary Elizabeth Dalrymple, and Tracy Holloway King. CSLI Publications.

Katsuda, Hironori. 2025. "A Probabilistic Model of Loanword Accentuation in Japanese." *Phonology* 42 (9): 1–26.

Katzir, Roni. 2023. "Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi." *Biolinguistics* 17: 1–12.

Kaun, Abigail Rhodes. 2004. "The Typology of Rounding Harmony : An Optimality Theoretic Approach." In *Phonetically Based Phonology*, edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge University Press.

Kawahara, Shigeto. 2006. "A Faithfulness Ranking Projected from a Perceptibility Scale: The Case of [+voice] in Japanese." *Language* 82 (3): 536–74.

Kawahara, Shigeto. 2020. "A Wug-Shaped Curve in Sound Symbolism: The Case of Japanese Pokémon Names." *Phonology* 37 (3): 383–418.

Kawahara, Shigeto. 2021a. "Phonetic Bases of Sound Symbolism: A Review. Https://Psyarxiv.Com/Fzvsu."

Kawahara, Shigeto. 2021b. "Testing MaxEnt with Sound Symbolism: A Stripy Wug-Shaped Curve in Japanese Pokémon Names." *Language* 97 (4): e341–59.

Keller, Frank. 2000. "Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality." Ph.D. thesis, University of Edinburgh.

Keller, Frank. 2006. "Linear Optimality Theory as a Model of Gradience in Grammar." In *Gradience in Grammar: Generative Perspectives*, edited by Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky, and Ralf Vogel. Oxford University Press.

Kerkhoff, Annemarie. 2007. "Acquisition of Morpho-Phonology: The Dutch Voicing Alternation." University of Utrecht.

Kilbourn-Ceron, Oriana, Meghan Clayards, and Michael Wagner. 2020. "Predictability Modulates Pronounciation Variants through Speech Planning Effects: A Case Study on Coronal Stop Realizations." *Laboratory Phonology* 11 (1): 5.

Kilbourn-Ceron, Oriana, and Matthew Goldrick. 2021. "Variable Pronounciations Reveal Dynamic Intra-Speaker Variation in Speech Planning." *Psychonomic Bulletin and Review* 28: 1365–80.

Kiparsky, Paul. 1971. "Historical Linguistics." In *A Survey of Linguistic Science*, edited by William Orr Dingwall. Linguistics Program, University of Maryland.

Kiparsky, Paul. 1973. *Three Dimensions of Linguistic Theory*. Edited by Osamu Fujimura. With Donald L. Smith, S. I. Harada, and Julie B. Lovins. TEC.

Kiparsky, Paul. 1982. "Lexical Phonology and Morphology." In *Linguistics in the Morning Calm*. Hanshin Publishing.

Kiparsky, Paul. 1993. "An OT Perspective on Phonological Variation."

Kiparsky, Paul. 2006a. "A Modular Metrics for Folk Verse." In *Formal Approaches to Poetry: Recent Developments in Metrics*, edited by Bezalel Elan Dresher and Nila Friedberg, vol. 11. Phonology and Phonetics. De Gruyter Mouton.

Kiparsky, Paul. 2006b. "Iambic Inversion in Finnish." In *A Man of Measure: Festschrift in Honour of Fred Karlsson on His 60th Birthday*, edited by Mickael Suominen, Antti Arppe, Anu Airola, et al. Linguistics Association of Finland.

Kirchner, Robert Martin. 2001. *An Effort Based Approach to Consonant Lenition*. 1st ed. Oustanding Dissertations in Linguistics. Routledge.

Kisseberth, Charles W. 1970. "On the Functional Unity of Phonological Rules." *Linguistic Inquiry* 1 (3): 291–306.

Kodner, Jordan, Sarah Payne, and Jeffrey Heinz. 2023. *Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi*. arXiv Preprint arXiv:2308.03228.

Koo, Hahn, and Young-il Oh. 2007. "Onset-to-Onset Probability and Gradient Acceptability in Korean." *Language Research* 43 (2): 289–309.

Korkmaz, Yunus, and Aytuğ Boyacı. 2018. "Classification of Turkish Vowels Based on Formant Frequencies." *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 1–4.

Korzybski, Alfred. 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. The International Non-Aristotelian Library Pub. Co.

Kroch, Anthony S. 1989. "Reflexes of Grammar in Patterns of Language Change." *Language Variation and Change* 1 (3): 199–244.

Kubozono, Haruo. 2006. "Where Does Loanword Prosody Come from? A Case Study of Japanese Loanword Accent." *Lingua* 116: 1140–70.

Kumagai, Gakuji, and Shigeto Kawahara. 2018. "Stochastic Phonological Knowledge and Word Formation in Japanese." *Gengo Kenkyu (Journal of the Linguistic Society of Japan)* 153: 57–83.

Kuo, Jennifer. 2023a. "Evidence for Prosodic Correspondence in the Vowel Alternations of Tgdaya Seediq." *Phonological Data and Analysis* 5 (3): 1–31.

Kuo, Jennifer. 2023b. "Phonological Markedness Effects in Paradigm Reanalysis." Ph.D. dissertation, UCLA.

Kuo, Jennifer. 2024a. "Hiatus Avoidance and the Development of Maori Passive Allomorphy." *Proceedings of NELS 54* 2: 31–40.

Kuo, Jennifer. 2024b. "Phonetic Naturalness in the Reanalysis of Samoan Thematic Consonant Alternations." *Journal of Phonetics* 107: 101355.

Kuo, Jennifer. 2024c. "Phonological Reanalysis Is Guided by Markedness: The Case of Malagasy Weak Stems." *Phonology* 41 (e3): 1–35.

Kuo, Jennifer. 2024d. "Phonological Reanalysis Is Guided by Markedness: The Case of Malagasy Weak Stems." *Phonology* 41: e3.

Labov, William. 1969. "Contraction, Deletion, and Inherent Variability of the English Copula." *Language* 45 (4): 715–62.

Labov, William. 1973. *Sociolinguistic Patterns*. 1st ed. University of Pennsylvania Press.

Labov, William. 1989. "The Child as Linguistic Historian." *Language Variation and Change* 1: 85–97.

Labov, William. 1994. *Principles of Linguistic Change, Volume 1: Internal Factors*. Blackwell.

Laplace, Pierre-Simon. 1814. "Essai Philosophique Sur Les Probabilités." In *Œuvres Complètes de Laplace*, tome VII. Gauthier-Villars.

Lasdon, Leon S., Richard L. Fox, and Margery W. Ratner. 1974. "Nonlinear Optimization Using the Generalized Reduced Gradient Method." *Revue Française d'automatique, Informatique, Recherche Opérationnelle* 8 (3): 73–104.

Lau, Jey Han, Alexander Clark, and Shalom Lappin. 2016. "Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge." *Cognitive Science* 41 (5): 1202–41.

Lavoie, Lisa M. 2001. *Consonant Strength: Phonological Patterns and Phonetic Manifestations*. Garland Pub.

Leben, Will. 1973. "Suprasegmental Phonology." Ph.D. dissertation, MIT.

Lee, Seung Suk, Joe Pater, and Brandon Prickett. forthcoming. "Representing And Learning Stress: A Maxent Framework for Comparing Learning across Grammatical Theories." Unpublished manuscript.

Lefkowitz, Lee Michael. 2017. "Maxent Harmonic Grammars and Phonetic Duration." Ph.D. dissertation, Department of Linguistics, UCLA.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. "Harmonic Grammar – A Formal Multi-Level Connectionist Theory of Linguistic Well-Formedness: An Application." *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (Mahwah, NJ), 884–91.

Liang, Kevin, Victoria Mateu, and Bruce Hayes. submitted. "An Experimental Study of Catalan Consonant Alternations." Department of Linguistics, UCLA.

Lindblom, B. 1990. "Explaining Phonetic Variation: A Sketch of H and H Theory." In *Speech Production and Speech Modeling*, edited by William J. Hardcastle and Alain Marchal. Kluwer Academic Publishers.

Linzen, Tal, Sofya Kasyanenko, and Maria Gouskova. 2013. "Lexical and Phonological Variation in Russian Prepositions." *Phonology* 30: 453–515.

Lubowicz, Anna. 2002. "Derived Environment Effects in Optimality Theory." *Lingua* 112: 243–80.

Magpale, Teri An Joy. 2026. "Probabilistic Gender Indexing in Filipino Names: A Maximum Entropy Analysis of Phonotactic Structure." *Folia Linguistica*.

Magri, Giorgio. 2012. "Convergence of Error-Driven Ranking Algorithms." *Phonology* 29: 213–69.

Magri, Giorgio. 2023. "Stochastic Harmonic Grammars Do Not Peak on the Mappings Singled out by Categorical Harmonic Grammars." *Proceedings of the Society for Computation in Linguistics*, 269–77.

Magri, Giorgio. 2024. "How Many Maximum Entropy Grammars Are Predicted by a Constraint Set When We Ignore Small Differences among Grammars?" *Proceedings of the Society for Computation in Linguistics*, 190–204.

Magri, Giorgio. 2025. "Constraint Interaction in Probabilistic Phonology: Deducing MaxEnt from Hayes and Zuraw's Shifted Sigmoids Generalization." *Linguistic Inquiry* 56: 1–54.

Magri, Giorgio. 2026. "Constraint Interaction in Probabilistic Phonology: Deducing Maximum Entropy Grammars from Hayes and Zuraw's Shifted Sigmoids Generalization." *Linguistic Inquiry* Early Access.

Magri, Giorgio, and Arto Anttila. 2022. "Paradoxes of MaxEnt Markedness." *Proceedings of the Annual Meeting on Phonology*.

Magri, Giorgio, and Arto Anttila. 2026. "MaxEnt Fails at Reasoning by Transitivity." *Proceedings of the Annual Meeting on Phonology* 2 (1).

Magri, Giorgio, and Edward Flemming. 2024. "Variation." In *The Cambridge Handbook of Phonology*, 2nd ed., edited by Adam Jardine and Paul de Lacy. Cambridge University Press.

Magri, Giorgio, and Benjamin Storme. 2020. "Calibration of Constraint Promotion Does Not Help with Learning Variation in Stochastic Optimality Theory." *Linguistic Inquiry* 51: 97–123.

Manning, Christopher D. 2003. "Probabilistic Appoaches to Syntax." In *Probabilistic Linguistics*, edited by Rens Bod, Jennifer Hay, and Stefanie Jannedy. MIT Press.

Martin, Andrew. 2011. "Grammars Leak: Modeling How Phonotactic Generalizations Interact within the Grammar." *Language* 87 (4): 751–70.

Mascaró, Joan. 2007. "External Allomorphy and Lexical Representation." *Linguistic Inquiry* 38: 715–35.

Mayer, Connor. 2021. "Issues in Uyghur Backness Harmony: Corpus, Experimental, and Computational Studies." Ph.D. dissertation, University of California, Los Angeles.

Mayer, Connor. 2025. "Reconciling Categorical and Gradient Models of Phonotactics." *Proceedings of the Society for Computation in Linguistics*, 144–54.

Mayer, Connor, Arya Kondur, and Megha Sundara. 2025. "The UCI Phonotactic Calculator: An Online Tool for Computing Phonotactic Metrics." *Behavior Research Methods* 57: 218.

Mayer, Connor, Adeline Tan, and Kie Zuraw. 2024. "Introducing Maxent.Ot: An R Package for Maximum Entropy Constraint Grammars." *Phonological Data and Analysis* 6 (4): 1–44.

McCarthy, John J. 1988. "Feature Geometry and Dependency: A Review." *Phonetica* 45 (2): 84–108.

McCarthy, John J. 2002. *A Thematic Guide to Optimality Theory*. Cambridge University Press.

McCarthy, John J. 2003. "OT Constraints Are Categorical." *Phonology* 20: 75–138.

McCarthy, John J. 2008. *Doing Optimality Theory: Applying Theory to Data*. Blackwell Publishers.

McCarthy, John J., and Alan Prince. 1995. "Faithfulness and Reduplicative Identity." In *Papers in Optimality Theory*, edited by Jill N. Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, vol. 1. University of Massachusetts Occasional Papers in Linguistics 18. GLSA, UMASS.

McCawley, James. 1968. *The Phonological Component of a Grammar of Japanese*. Mouton.

McPherson, Laura. 2013. *A Grammar of Tommo So*. De Gruyter.

McPherson, Laura, and Bruce Hayes. 2016. "Relating Application Frequency to Morphological Structure: The Case of Tommo So Vowel Harmony." *Phonology* 33 (1): 125–67.

Mielke, Jeff. 2005. "Modeling Distinctive Feature Emergence." In *Proceedings of the West Coast Conference on Formal Linguistics*, edited by John D. Alderete, vol. 24. Cascadilla Proceedings Project.

Mikheev, Andrei. 1997. "Automatic Rule Induction for Unknown-Word Guessing." *Computational Linguistics* 23: 405–23.

Moore-Cantwell, Claire. 2016. "The Representation of Probabilistic Phonological Patterns: Neurological, Behavioral, and Computational Evidence from the English Stress System." PhD Thesis, University of Massachusetts Amherst.

Moore-Cantwell, Claire. 2020. "Weight and Final Vowels in the English Stress System." *Phonology* 37 (4): 657–95.

Moore-Cantwell, Claire. 2026. *The Gradual Lexicon and Phonology Learner*. Released.

Moore-Cantwell, Claire, Dana Bosch, Ethan X. Kahn, Christine Kim, and G. Shoemaker. 2024. "Production Priming of Stress in Nonwords." *Laboratory Phonology* 15 (1): 1–31.

Moore-Cantwell, Claire, and Joe Pater. 2016. "Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints." *Catalan Journal of Linguistics* 15: 53–66.

Moreton, Elliott, and Joe Pater. 2012a. "Structure and Substance in Artificial-Phonology Learning: Part I, Structure." *Language and Linguistics Compass* 6: 686–701.

Moreton, Elliott, and Joe Pater. 2012b. "Structure and Substance in Artificial-Phonology Learning: Part II, Substance." *Language and Linguistics Compass* 6: 702–18.

Moreton, Elliott, Joe Pater, and Katya Pertsova. 2017. "Phonological Concept Learning." *Cognitive Science* 41 (1): 4–69.

Moreton, Elliott, and Katya Pertsova. 2024. "Implicit and Explicit Processes in Phonological Concept Learning." *Phonology* 40 (2): 1–53.

Nagahara, Hiroyuki. 1994. "Phonological Phrasing in Japanese." Ph.D. dissertation, UCLA.

Nakisa, Ramin Charles, Kim Plunkett, and Ulrike Hahn. 2001. "A Cross-Linguistic Comparison of Single and Dual-Route Models of Inflectional Morphology." In *Models of Language Acquisition : Inductive and Deductive Approaches*, edited by Peter Broeder and Jaap Murre. MIT Press.

Nazarov, Aleksei. 2016. "Dutch Reduction Domains:Between Syllables and Feet." *Proceedings of the 2014 Annual Meeting on Phonology*.

Newman, Stanley. 1944. *The Yokuts Language of California*. Viking Fund Publications in Anthropology 2. Viking Fund.

Newmeyer, Frederick J. 2003. "Grammar Is Grammar and Usage Is Usage." *Language* 79: 682–707.

Odden, David. 1981. "Problems in Tone Assignment in Shona." Ph.D. dissertation, University of Illinois at Urbana-Champaign.

O'Hara, Charlie. 2022. "MaxEnt Learners Are Biased against Giving Probability to Harmonically Bounded Candidates." *Proceedings of the Society for Computation in Linguistics 202*, 229–34.

Olejarczuk, Paul, and Vsevolod Kapatsinski. 2018. "The Metrical Parse Is Guided by Gradient Phonotactics." *Phonology* 35 (3): 367–405.

Olson, Erin. 2020. "Loanwords and the Perceptual Map : A Perspective from MaxEnt Learning." Ph.D. dissertation, Massachusetts Institute of Technology.

Oostendorp, Marc van. 1995a. "Style Levels in Conflict Resolution." In *Language Variation in Phonological Theory*, edited by Frans Hinskens, Roeland van Hout, and W. Leo Wetzels. John Benjamins.

Oostendorp, Marc van. 1995b. "Vowel Quality and Phonological Projection." Ph.D. dissertation, Tilburg University.

Orr, Carolyn. 1962. "Ecuadorian Quichua Phonology." In *Studies in Ecuadorian Indian Languages I*, 1st ed., edited by Benjamin F. Elson, with Catherine Peeke. Linguistic Series 7. Summer Institute of Linguistics of the University of Oklahoma.

Pater, Joe. 1999. "Austronesian Nasal Substitution and Other NC Effects." In *The Prosody-Morphology Interface*. Cambridge University Press.

Pater, Joe. 2000. "Non-Uniformity in English Secondary Stress: The Role of Ranked and Lexically Specific Constraints." *Phonology* 17 (2): 237–74.

Pater, Joe. 2004. "Austronesian Nasal Substitution and Other NC Effects." In *Optimality Theory in Phonology: A Reader*, 1st ed., edited by John J. McCarthy. Blackwell.

Pater, Joe. 2008. "Gradual Learning and Convergence." *Linguistic Inquiry* 39: 334–45.

Pater, Joe. 2009. "Weighted Constraints in Generative Linguistics." *Cognitive Science* 33 (6): 999–1035.

Pater, Joe. 2010. "Morpheme-Specific Phonology: Constraint Indexation and Inconsistency Resolution." In *Phonological Argumentation: Essays on Evidence and Motivation*, edited by Steve Parker. Advances in Optimality Theory, edited by Ellen Woolford and Armin Mester. University of Toronto Press.

Pater, Joe. 2016. "Universal Grammar with Weighted Constraints." In *Harmonic Grammar and Harmonic Serialism*, edited by John McCarthy and Joe Pater. Equinox.

Pater, Joe, and Brandon Prickett. 2022. "Typological Gaps in Iambic Nonfinality Correlate with Learning Difficulty." Paper presented at Annual Meeting on Phonology. *Proceedings of the 2021 Annual Meeting on Phonology*.

Peperkamp, Sharon, and Emmanuel Dupoux. 2003. "Reinterpreting Loanword Adaptations: The Role of Perception." *Proceedings of the International Congress of Phonetic Sciences* 15: 367–70.

Phillips, Betty. 2006. *Word Frequency and Lexical Diffusion*. Palgrave Macmillan.

Piantadosi, Steven T. 2024. "Modern Language Models Refute Chomsky's Approach to Language." In *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, edited by Edward Gibson and Moshe Poliak. Language Science Press.

Pierrehumbert, Janet. 1980. "The Phonology and Phonetics of English Intonation." Massachusetts Institute of Technology.

Pierrehumbert, Janet. 2000. "Exemplar Dynamics: Word Frequency, Lenition and Contrast." In *Frequency Effects and Emergent Grammar*, edited by J. Bybee and P. Hopper. John Benjamins.

Pierrehumbert, Janet. 2006. "The Statistical Basis of an Unnatural Alternation." In *Laboratory Phonology 8: Varieties of Phonological Competence*, edited by Louis Goldstein, D. H. Whalen, and Catherine Best. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet. 2022. "More than Seventy Years of Probabilistic Phonology." In *The Oxford History of Phonology*, edited by B. Elan Dresher and Harry van der Hulst. Oxford University Press.

Pierrehumbert, Janet B. 1993. "Dissimilarity in the Arabic Verbal Roots." *Proceedings of the North East Linguistics Society 23* (University of Ottawa) 2: 367–81.

Piñeiro, Gervasio, Susana Perelman, Juan P. Guerschman, and José M. Paruelo. 2008. "How to Evaluate Models: Observed vs. Predicted or Predicted vs. Observed?" *Ecological Modeling* 216: 316–22.

Pinta, Justin, and Jennifer Smith. 2017. "Spanish Loans and Evidence for Stratification in the Guarani Lexicon." In *Guarani Linguistics in the 21st Century*, edited by Bruno Estigarribia and Justin Pinta.

Podesva, Robert, and Devyani Sharma, eds. 2013. *Research Methods in Linguistics*. Cambridge University Press.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Poplack, Shana, Alicia Pousada, and David Sankoff. 1982. "Competing Influences on Gender Assignment: Variable Process, Stable Outcome." *Lingua* 57 (1): 1–28.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt, and Michael Becker. 2010. "Harmonic Grammar with Linear Programming: From Linear Systems to Linguistic Typology." *Phonology* 27 (1): 77–117.

Prasada, Sandeep, and Steven Pinker. 1993. "Generalization of Regular and Irregular Morphological Patterns." *Language and Cognitive Processes* 8: 1–56.

Prentice, D. J. 1971. *The Murut Languages of Sabah*. Australian National University.

Prickett, Brandon. 2025. *Hidden-Structure-MaxEnt*. Released.

Prickett, Brandon, Kaden Holladay, Shay Hucklebridge, et al. 2019. "Learning Syntactic Parameters without Triggers by Assigning Credit and Blame." *Proceedings of the 55th Annual Meeting of the Chicago Linguistic Society (CLS)* (Chicago, Illinois).

Prince, Alan. 2003a. "Anything Goes." In *A New Century of Phonology and Phonological Theory : A Festschrift Professor Shosuke Haraguchi on the Occasion of His Sixtieth Birthday*, edited by T. Honma, M. Okazaki, T. Tabata, and S. Tanaka. Kaitakusha.

Prince, Alan. 2003b. "Arguing Optimality." *University of Massachusetts Occasional Papers in Linguistics* 26: 268–304.

Prince, Alan S. 1984. "Phonology with Tiers." In *Language Sound Structure: Essays Presented to Morris Halle by His Teacher and Students*, edited by Mark Aronoff and Richard T. Oehrle. MIT Press.

Prince, Alan S., and Bruce Tesar. 2004. "Learning Phonotactic Distributions." In *Constraints in Phonological Acquisition*, edited by René Kager, Joe Pater, and Wim Zonneveld. Cambridge University Press.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory : Constraint Interaction in Generative Grammar*. Blackwell Publishers.

Pullum, Geoffrey K. 2013. "The Central Question in Comparative Syntactic Metatheory." *Mind & Language* 28: 492–521.

Rasin, Ezer, Iddo Berger, Nur Lan, Itamar Shefi, and Roni Katzir. 2021. "Approaching Explanatory Adequacy in Phonology Using Minimum Description Length." *Journal of Language Modelling* 9: 17–66.

Rasin, Ezer, and Roni Katzir. 2016. "On Evaluation Metrics in Optimality Theory." *Linguistic Inquiry* 47: 235–82.

Rebrus, Péter, Péter Szigetvári, and Miklós Törkenczy. 2023. "Morphological Restrictions on Vowel Harmony: The Case of Hungarian." In *The Wiley Blackwell Companion to Morphology*.

Regier, Terrry, and Susanne Gahl. 2004. "Learning the Unlearnable: The Role of Missing Evidence." *Cognition* 93: 147–55.

Reiss, Charles. 2017. "Substance Free Phonology." In *The Routledge Handbook of Phonological Theory*.

Riggle, Jason. 2009. "Generating Contenders." Preprint, Rutgers Optimality Archive.

Riggle, Jason. 2010. "Sampling Rankings." Preprint, Rutgers Optimality Archive.

Ringen, Catherine O., and Orvokki Heinämäki. 1999. "Variation in Finnish Vowel Harmony: An OT Account." *Natural Language & Linguistic Theory* 17 (2): 303–37.

Rose, Sharon, and Rachel Walker. 2004. "A Typology of Consonant Agreement as Correspondence." *Language* 80: 475–531.

Rosenblatt, Frank. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386.

Rosenthall, Samuel. 1989. "The Phonology of Nasal-Obstruent Sequences." M.A. Thesis, Department of Linguistics, McGill University.

Ross, John R. 1967. "Constraints on Variables in Syntax." MIT.

Rousseau, Pascale, and David Sankoff. 1978. "Advances in Variable Rule Methodology." In *Linguistic Variation : Models and Methods*. Academic Press.

Rumelhart, David E., James L. McClelland, and PDP Research Group. 1986. *Parallel Distributed Processing*. The MIT Press.

Ryan, Kevin M. 2010. "Variable Affix Order: Grammar and Learning." *Language* 86 (4): 758–91.

Sadeghi, Vahid. 2014. "Optionality and Gradience in Persian Phonology: An Optimality-Treatment." *Iranian Journal of Applied Language Studies* 6 (1): 157–82.

Samek-Lodovici, Vieri, and Alan Prince. 1999. *Optima*. Rutgers Optimality Archive ROA-363.

Sankoff, David. 1988. "Variable Rules." In *Sociolinguistics : An International Handbook of the Science of Language and Society*, 2nd ed., edited by Ulrich Ammon, Norbert Dittmar, and Klaus J. Mattheier, vol. 2. De Gruyter.

Sankoff, David, and William Labov. 1979. "On the Uses of Variable Rules." *Language in Society* 8 (2–3): 189–222.

Schachter, Paul, and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. University of California Press.

Scholes, Robert J. 1966a. *Phonotactic Grammaticality*. Janua Linguarum. Series Minor ; Nr. 50. Mouton.

Scholes, Robert J. 1966b. *Phonotactic Grammaticality*. Mouton & Co.

Schütze, Carson. (1996) 1996. *The Empirical Base of Linguisticcs: Grammaticality Judgments and Linguistic Methodology*. Classics in Linguistics 2. Chicago University Press. Reprint, Language Science Press.

Šegedin, Bruno Ferenc, and Uriel Cohen Priva. 2026. "A Large-Scale Investigation of Vowel Co-Occurrence Patterns in the World's Lexicons." *Glossa: A Journal of General Linguistics* 11 (1).

Selkirk, Elisabeth O. 1978. "On Prosodic Structure and Its Relation to Syntactic Structure." In *Nordic Prosody II*, edited by T. Fretheim. TAPIR.

Sereno, Joan A., and Allard Jongman. 1997. "Processing of English Inflectional Morphology." *Memory and Cognition* 25 (4): 425–37.

Shademan, Shabnam. 2007. "Grammar and Analogy in Phonotactic Well-Formedness Judgements." Phdthesis, UCLA.

Shaw, Jason, and Shigeto Kawahara. 2018. "Predictability and Phonology: Past, Present and Future." Article No. 20180042. *Linguistic Vanguard* 4 (2): 1–11.

Shih, Stephanie S. 2017. "Constraint Conjunction in Probabilistic Weighted Constraint Grammar." *Phonology* 34 (2): 243–68.

Shih, Stephanie S. 2018. "Learning Lexical Classes from Variable Phonology." *ICUWPL* 4: 1–15.

Shih, Stephanie S., Ackerman, Jordan, Noah Hermalin, and Sharon Inkelas. 2018. "Pokémonikers: Study of Sound Symbolism and Pokémon Names." *Proceedings of the Linguistic Society of America* 3 (42): 1–6.

Shih, Stephanie S., Jason Grafmiller, Futrell, Richard, and Bresnan, Joan. 2015. "Rhythm's Role in Predicting Genitive Alternation Choice in Spoken English." In *Rhythm in Cognition and Grammar: A Germanic Perspective,*. De Gruyter Mouton.

Shih, Stephanie S., and Kie Zuraw. 2017. "Phonological Conditions on Variable Adjective and Noun Word Order in Tagalog." *Language* 93: e317–52.

Shinohara, Shigeko. 2000. "Default Accentuation and Foot Structure in Japanese: Evidence from Japanese Adaptations of French Words." *Journal of East Asian Linguistics* 9: 55–96.

Siah, Jian-Leat. in press. "Prosodic End-Weight Effects in Malay Echo Reduplication: The Role of Phonetic Naturalness and Lexical Frequency." *Phonology*.

Simpson, Edward H. 1951. "The Interpretation of Interaction in Contigency Tables." *Journal of the Royal Statistical Society, Series B* 13: 238–41.

Siptár, Péter, and Miklós Törkenczy. 2000. *The Phonology of Hungarian*. Oxford University Press.

Skousen, Royal, Deryle Lonsdale, and Dilworth P. Parkinson, eds. 2002. *Analogical Modeling: An Exemplar-Based Approach to Language*. John Benjamins Publishing Company.

Slootweg, Annemarie. 1988. "Metrical Prominence and Syllable Duration." In *Linguistics in the Netherlands*, edited by P. Coopmans and A. Hulk. Foris.

Smith, Brian W., and Claire Moore-Cantwell. 2017. "Emergent Idiosyncrasy in English Comparatives." In *NELS 47: Proceedings of the 47th Meeting of the North East Linguistic Society*, edited by Andrew Lamont and Katie Tetzloff. Graduate Linguistic Student Association.

Smith, Brian W., and Joe Pater. 2020. "French Schwa and Gradient Cumulativity." *Glossa: A Journal of General Linguistics* 5 (1): 24–33.

Smith, Jennifer. 2025. "Loanword Phonology." In *The Cambridge Handbook of Phonology 2nd Ed.*, edited by Adam Jardine and Paul de Lacy. Cambridge University Press.

Smith, Jennifer L. 2011. "Category-Specific Effects." In *Companion to Phonology*, edited by Marc van Oostendorp, Colin Ewen, Beth Hume, and Keren Rice. Oxford: Wiley-Blackwell.

Smolensky, Paul. 1986. "Information Processing in Dynamical Systems: Foundations of Harmony Theory." In *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, by David E. Rumelhart, James L. McClelland, and the PDP Research Group, vol. 1. MIT Press.

Smolensky, Paul. 1996. "On the Comprehension/Production Dilemma in Child Languagev." *Linguistic Inquiry* 27: 720–31.

Smolensky, Paul, and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. MIT Press.

Solé, Maria-Josep, Patrice Speeter Beddor, and Manjari Ohala, eds. 2007. *Experimental Approaches to Phonology*. Oxford University Press.

Sorace, Antonella, and Frank Keller. 2005. "Gradience in Linguistic Data." *Lingua* 115: 1497–524.

Sprouse, Jon. 2007. "Continuous Acceptability, Categorical Grammaticality, and Experimental Syntax." *Biolinguistics* 1: 123–34.

Stefanowitsch, Anatol. 2005. "New York, Dayton (Ohio), and the Raw Frequency Fallacy." *Corpus Linguistics and Linguistic Theory* 1 (2): 295–301.

Steriade, Donca. 2000. "Paradigm Uniformity and the Phonetics/Phonology Boundary." In *Papers in Laboratory Phonology*, edited by Janet Pierrehumbert and Michael Broe, vol. 6. Cambridge University Press.

Steriade, Donca. 2008. "The Pseudo-Cyclic Effect in Romanian Morphophonology." In *Inflectional Identity*, edited by Asaf Bachrach and Andrew Nevins. Oxford University Press.

Steriade, Donca. 2026. "The Pseudo-Cycle and Lexical Conservatism." In *The Wiley Blackwell Companion to Phonology, Second Edition*.

Steriade, Donca, and Igor Yanovich. 2015. "Accentual Allomorphs in East Slavic: An Argument for Inflection Dependence." In *Understanding Allomorphy: Perspectives from Optimality Theory*, edited by Eulàlia Bonet, Maria-Rosa Lloret, and Joan Mascaró. University of Toronto Press.

Storme, Benjamin. 2026. "A Method to Evaluate Systemic Constraints in Probabilistic Grammars." *Linguistic Inquiry* 57 (1): 216–28.

Suomi, Kari. 1983. "Palatal Vowel Harmony: A Perceptually Motivated Phenomenon?" *Nordic Journal of Linguistics* 6 (1): 1–35.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, et al. 2017. "Spoken Syntax in a Comparative Perspective: The Dative and Genitive Alternation in Varieties of English." *Glossa* 2 (1): 1–27.

Tarnóczy, Tamás. 1964. *Acoustic Analysis of Hungarian Vowels*. Speech Transmission Laboratory: Quarterly Progress Status Report No. 4. KTH.

Tesar, Bruce. 2004. "Computing Optimal Forms in Optimality Theory: Basic Syllabification." In *Optimality Theory in Phonology: A Reader*, edited by John J. McCarthy. Blackwell Publishing Ltd.

Tesar, Bruce. 2014. *Output-Driven Phonology*. Cambridge University Press.

Tesar, Bruce. 2014. *Output-Driven Phonology: Theory and Learning.* Cambridge University Press.

Tesar, Bruce, and Paul Smolensky. 1993. *The Learnability of Optimality Theory: An Algorithm and Some Basic Complexity Results.* Technical Report CU-CS-678-93. Department of Computer Science, University of Colorado at Boulder.

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

Tessier, Anne-Michelle. 2016. *Phonological Acquisition: Child Language and Constraint-Based Grammar*. Palgrave-Macmillan.

Tilsen, Sam. 2012. "Utterance Preparation and Stress Clash: Planning Prosodic Alternations." In *Speech Production and Perception: Planning and Dynamics*, edited by Susanne Fuchs, Pascal Perrier, Melanie Weirich, and Daniel Pape. Peter Lang Verlag.

Treiman, Rebecca. 1985. "Onsets and Rimes as Units of Spoken Syllables: Evidence from Children." *Journal of Experimental Child Psychology* 39: 161–81.

Truckenbrodt, Hubert. 1999. "On the Relation between Syntactic Phrases and Phonological Phrases." *Linguistic Inquiry* 30: 219–55.

Tufte, Edward. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press.

Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory." *The American Statistician* 34: 23–25.

Turk, Alice, and Stefanie Shattuck-Hufnagel. 2020. *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. Vol. 5. Oxford University Press.

Vago, Robert M. 1976. "Theoretical Implications of Hungarian Vowel Harmony." *Linguistic Inquiry* 7 (2): 243–63.

Velldal, Erik, and Stephan Oepen. 2005. "Maximum Entropy Models for Realization Ranking." In *Proceedings of the 10th Machine Translation Summit*, edited by Jun-ichi Tsuji. Asia-Pacific Association for Machine Translation.

Vendelin, Inga, and Sharon Peperkamp. 2006. "The Influence of Orthography on Loanword Adaptations." *Lingua* 116 (7): 996–1007.

Vitevitch, Michael S., and Paul A. Luce. 2004. "A Web-Based Interface to Calculate Phonotactic Probability for Words and Nonwords in English." *Behavior Research Methods, Instruments, & Computers* 36: 481–87.

Walker, James A. 2010. *Variation in Linguistic Systems*. 1st ed. Routledge.

Wang, Yang, and Bruce Hayes. to appear. "Learning Underlying Representations: The Role of Abstractness." *Linguistic Inquiry*.

Wasserman, Larry. 2004. *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Werner, Raphael, Jürgen Trouvain, and Bernd Möbius. 2022. "Optionlity and Variability of Speech Pauses in Read Speech across Languages and Rates." *Proceedings of Speech Prosody 2022*, 312–16.

Westbury, John R., and Patricia A. Keating. 1986. "On the Naturalness of Stop Consonant Voicing." *Journal of Linguistics* 22 (1): 145–66.

White, James. 2017. "Accounting for the Learnability of Saltation in Phonological Theory: A Maximum Entropy Model with a P-Map Bias." *Language* 93: 1–36.

White, James, and Megha Sundara. 2014. "Biased Generalization of the Newly Learned Phonological Alternations by 12-Month-Old Infants." *Cognition* 133: 85–90.

White, Lydia. 2003. *Second Language Acquisition and Universal Grammar*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Wilson, Colin. 2006. "Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization." *Cognitive Science* 30 (5): 945–82.

Wilson, Colin, and Gillian Gallagher. 2018a. "Accidental Gaps and Surface-Based Phonotactic Learning: A Case Study of South Bolivian Quechua." *Linguistic Inquiry* 49 (3): 610–23.

Wilson, Colin, and Gillian Gallagher. 2018b. "Accidental Gaps and Surface-Based Phonotactic Learning: A Case Study of South Bolivian Quechua." *Linguistic Inquiry* 49: 610–23.

Wilson, Colin, and Marieke Obdeyn. 2009. "Simplifying Subsidiary Theory: Statistical Evidence from Arabic, Muna, Shona and Wargamay." M.S. Thesis, Johns Hopkins University.

Wolfram, Walt. 1993. "Identifying and Interpreting Variables." In *American Dialect Research*, edited by Dennis R. Preston. John Benjamins Publishing Company.

Xu, Lily. 2025. "Non-Concatenative Morphological Domains Constraint Phonotactics: A Case Study of Egyptian Arabic." *Phonology* 42: e16.

Xu, Lily. 2026. "A Model of Iterative Morphophonological Learning: Two Empirical Studies." Ph.D. dissertation, UCLA.

Yang, Charles. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT Press.

Young, Richard, and Robert Bayley. 1996. "Varbrul Analysis for Second Language Acquisition Research." In *Second Language Acquisition and Linguistic Variation.*, edited by Robert Bayley and Dennis R. Preston. John Benjamins.

Zellou, Georgia, Mohamed Afkir, Mohamed Lahrouchi, and Karim Bensoukas. 2025. "Cross-Language Variation in the Acceptability of Vowelless Nonwords." *Frontiers in Communication* 10: 1518754.

Zhang, Jie. 2004. "The Role of Contrast-Specific and Language-Specific Phonetics in Countour Tone Distribution." In *Phonetically Based Phonology*, edited by Bruce Hayes, Robert Kirchner, and Donca Steriade. Cambridge University Press.

Zhang, Jie, Yuwen Lai, and Craig Turnbull-Sailor. 2006. "Wug-Testing the 'Tone Circle' in Taiwanese." In *Proceedings of the 25th West Coast Conference on Formal Linguistics*, edited by Donald Baumer, David Montero, and Michael Scanlon. Cascadilla Press.

Zhang, Jie, Yuwen Lai, and Craig Turnbull-Sailor. 2009. "Opacity, Phonetics, and Frequency in Taiwanese Tone Sandhi." In *Current Issues in Unity and Diversity of Languages: Collection of Papers Selected from the 18th International Congress of Linguists*. Linguistic Society of Korea.

Zheng, Shuang, and Youngah Do. 2025. "Substantive Bias in Artificial Phonology Learning." *Language and Linguistics Compass* 19: e70005.

Zimmer, Karl E. 1969. "Psychological Correlates of Some Turkish Morpheme Structure Conditions." *Language* 45 (2): 309–21.

Zimmermann, Richard. 2017. "Formal and Quantitative Approaches to the Study of Syntactic Change: Three Case Studies from the History of English." Ph.D. thesis, University of Geneva.

Zuraw, Kie. 2002. "Aggressive Reduplication." *Phonology* 19 (3): 395–439.

Zuraw, Kie. 2007. "The Role of Phonetic Knowledge in Phonological Patterning: Corpus and Survey Evidence from Tagalog." *Language* 83: 277–316.

Zuraw, Kie. 2010. "A Model of Lexical Variation and the Grammar with Application to Tagalog Nasal Substitution." *Natural Language & Linguistic Theory* 28 (2): 417–72.

Zuraw, Kie. 2013. "*MAP Constraints."

Zuraw, Kie, and Bruce Hayes. 2017. "Intersecting Constraint Families: An Argument for Harmonic Grammar." *Language* 93 (3): 497–548.

Zuraw, Kie, Isabelle Lin, Meng Yang, and Sharon Peperkamp. 2021. "Competition between Whole-Word and Decomposed Representations of English Prefixed Words." *Morphology* 31: 201–37.

Zuraw, Kie Ross. 2000. "Patterned Exceptions in Phonology." University of California, Los Angeles.

Zymet, Jesse. 2018. "Learning a Frequency-Matching Grammar Together with Lexical Isiosyncrasy: MaxEnt vs. Heirarchical Regression." *Proceedings of the Annual Meeting on Phonology*, ahead of print.

N.d.