Some Thoughts on Optionality and Phonological Knowledge

Philosophical Priors

1. Probabilities should model ignorance, not actual behavior

> *"...the reification of probabilities as cognitively represented quantities invites a problematic conception of randomness and probability, in which randomness plays a shallow role in the generation of behavior. This shallow stochasticity is at odds with the role of randomness in physical systems that are the analogues of phonological cognition in the MaxEnt analogy. The states of physical systems are not governed by probabilities. Probabilities are merely statistical descriptions that are useful when we are mostly ignorant about the detailed dynamics of a system and its surroundings."* - Tilsen (2023, p. 265)

2. Grammars should model knowledge of language, not use of language

> *"...it is perhaps worth while to reiterate that a generative grammar is not a model for a speaker or a hearer. It attempts to characterize in the most neutral possible terms the knowledge of the language that provides the basis for actual use of language by a speaker-hearer."* - Chomsky (1965, p. 9)

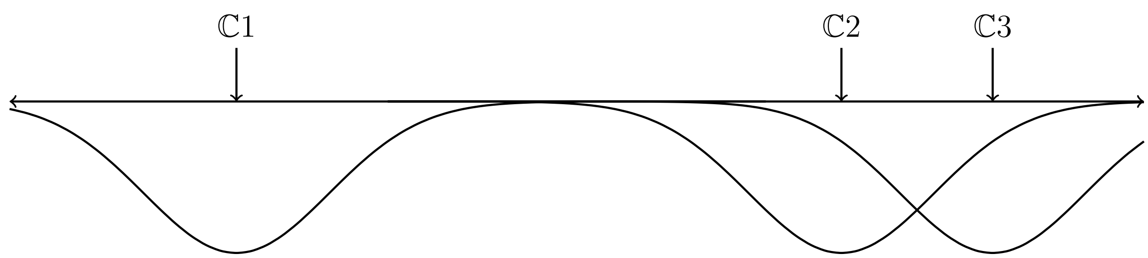
3. Phonological knowledge is only one of many factors governing behavior

> *"Practitioners of phonology often distinguish between internal evidence, which consists of data from distribution and alternation, and external evidence, which consists of data from language production, language comprehension, language acquisition, psycholinguistic experiments of various kinds, sound patterning in versification, language games, etc. [...] The terms "internal" and "external" evidence indicate a bias under which most phonological research is being pursued, namely, the belief that the behaviour of speakers in making acceptability judgments is somehow a more direct reflection of their linguistic knowledge than their behaviour in producing language, understanding language, etc. This bias appears to be related to the fact that linguistic knowledge is only one of the inputs to language production, language comprehension, and other forms of language performance. What accounts for the facts of performance is a conjunct of a theory of linguistic knowledge ('What is the nature of the representation of linguistic knowledge?') and a theory of language performance ('How is this knowledge put to use?')."* - Mohanan (1986, p. 183; emphasis original)

Recent Approaches to Optionality

There are three general approaches with optimization-based grammars. Stochastic OT and Noisy Harmonic Grammar assume that Harmonic Evaluation is a *noisy* procedure which can lead to local re-ranking if the constraint weights are close enough together (Boersma, 1998; Boersma and Hayes, 2001).[1] In stochastic modeling, noise is often represented using a normal distribution (bell curve) centered around 0. This "noise" is then added to a deterministic value to make the entire system non-deterministic or *stochastic*. A probability distribution emerges indirectly in the long term.

For example, the figure below shows potential constraint rankings and their corresponding "noise" distributions. Since $\mathbb{C}1$ is ranked sufficiently higher than the other constraints it will always remain the highest ranked. On the other hand $\mathbb{C}2$ and $\mathbb{C}3$ are close enough to where their local ranking can flip on different evaluations.



Another approach to optionality/variation uses MaxEnt grammars (Hayes and Wilson, 2008; Hayes, 2022). MaxEnt does not have a noisy evaluation procedure. Instead, a normalization procedure is performed over the output candidate weights.[2] This is a three step process where first a harmony score is calculated for each candidate, this is used to determine an eHarmony score, these second scores are the basis for normalization. Below shows how the probability distribution (rightmost column) for some syllable types are determined based of a ONSET constraint with weight 3 and NOCODA constraint with weight 2.

[1] There is a variant of NHG where noise is instead added to each cell rather than to the constraint weight (Goldrick and Daland, 2009). Interestingly, this was proposed specifically to account for speech errors which sometimes yield non-grammatical forms: "Being blind to the nature of constraints, the disruptions caused by random noise will not be constrained to produce only those mappings that can specified by the grammar."

[2] MaxEnt grammars are usually *phonotactic grammars*, but they can be used for input-output mappings as well.

x	*#V	*C#	Score	$P^*(x)$	P(x)
C	0	2	2	0.1353353	0.1019709
V	3	0	3	0.04978707	0.03751299
CV	0	0	0	1	0.7534685
VC	3	2	5	0.006737947	0.00507683
CVC	0	2	2	0.1353353	0.1019709

The final approach is a "competing grammars" approach exemplified by Anttila (1997). He gives the following two explanations:

1. A candidate is predicted by the grammar if it wins in some tableau.
2. If a candidate wins in n tableaux and t is the total number of tableaux, then the candidate's probability of occurrence is n/t.

Given three constraints $\mathbb{C}1$, $\mathbb{C}2$, $\mathbb{C}3$, there are $3!$ ways in which they can be ordered. This gives us six grammars/tableaux. The idea behind this third approach is that the probability of one form over another is tied to the proportion of rankings that choose that candidate as the winner. Suppose we have two candidates x_1 and x_2. x_1 violateds $\mathbb{C}1$ and $\mathbb{C}2$ while x_2 violates $\mathbb{C}3$. The following table shows the six possible rankings and which candidate would win in each case.

Ranking	Winner
$\mathbb{C}1 \gg \mathbb{C}2 \gg \mathbb{C}3$	x_2
$\mathbb{C}1 \gg \mathbb{C}3 \gg \mathbb{C}2$	x_2
$\mathbb{C}2 \gg \mathbb{C}1 \gg \mathbb{C}3$	x_2
$\mathbb{C}2 \gg \mathbb{C}3 \gg \mathbb{C}1$	x_2
$\mathbb{C}3 \gg \mathbb{C}1 \gg \mathbb{C}2$	x_1
$\mathbb{C}3 \gg \mathbb{C}2 \gg \mathbb{C}1$	x_1

This predicts that x_2 should appear about two-thirds of the time while x_1 should appear about one-third of the time. A probability distribution emerges from the types of forms the constraint set predicts.

Issues with these approaches

In the case of Stochastic OT and Noisy Harmonic Grammar, the specific form is chosen almost completely by chance. It is one thing to match the probability distributions of given forms, but it is something completely different to state which factors lead to a specific form being chosen over another. Distributions should emerge based on the factors themselves leading to some type of distribution. The competing grammars approach in its basic form struggles from the same issue: when are we choosing one grammar over another? MaxEnt removes the chance element, but does not have a theory of actuation: given that we have a probability distribution, when is one form chosen over the other.

Alternative Approach 1: optionality as a property of the phonetic implementation

- Under the BLUEPRINT MODEL OF PRODUCTION Nelson and Heinz (in press), a recent model of the phonetics-phonology interface, optionality is explained based on properties of the post-phonological phonetic production.
- The BLUEPRINT MODEL OF PRODUCTION, explicitly structures the phonetic production module such that phonological knowledge (surface form), lexical knowledge (underlying form), and extra-grammatical knowledge all interact to produce a phonetic output.
- Nelson and Heinz (in press) show that access to both the underlying and surface form allows for straightforward explanations of phenomena such as incomplete neutralization and variation in homophone duration while maintaining discrete phonological knowledge.

Under this approach phonological knowledge itself is never optional: it exists as one type of information that is consulted during language use. Optionality therefore shifts from being a property of the grammar to choosing to apply phonological knowledge when computing the phonetic exponent a given lexical item (Nelson, to appear).

The basic story with the BLUEPRINT MODEL OF PRODUCTION is that you can account for subtle phonetic variation by considering both surface and underlying form. We implement this with a scaling formula for a given phonetic cue c. To account for optionality, I add a parameter $v \in \{0, 1\}$ that chooses whether or not to use the surface form (phonological knowledge) in the computation. When $v = 1$ no phonological knowledge is applied and when $v = 0$ phonological knowledge is applied. The scaling formula is

$$c = c_{UR} \times \textbf{max}(v, i^{\alpha}) + c_{SR} \times (1 - \textbf{max}(v, i^{\alpha}))$$

Global optionality in Waoro is reported by Osborn (1966). Labial voicing is a context free process which voices labial stops. Examples (1–2) provide specific surface forms where there is free variation between the all or nothing cases (a) but the cases where there is a mismatch in application are not observed (b).

(1) *weak*

a. [paroparera] ∼ [**b**aro**b**arera]

b. * [paro**b**arera] ∼ * [**b**aroparera]

(2) *he will put them*

a. [apaupute] ∼ [a**b**au**b**ute]

b. * [apau**b**ute] ∼ * [a**b**aupute]

The figure below shows the abstract characterization of how optionality emerges. The center box with a function labeled A is what implements the scaling formula above. P corresponds with SR, L corresponds with UR, and E corresponds with v.

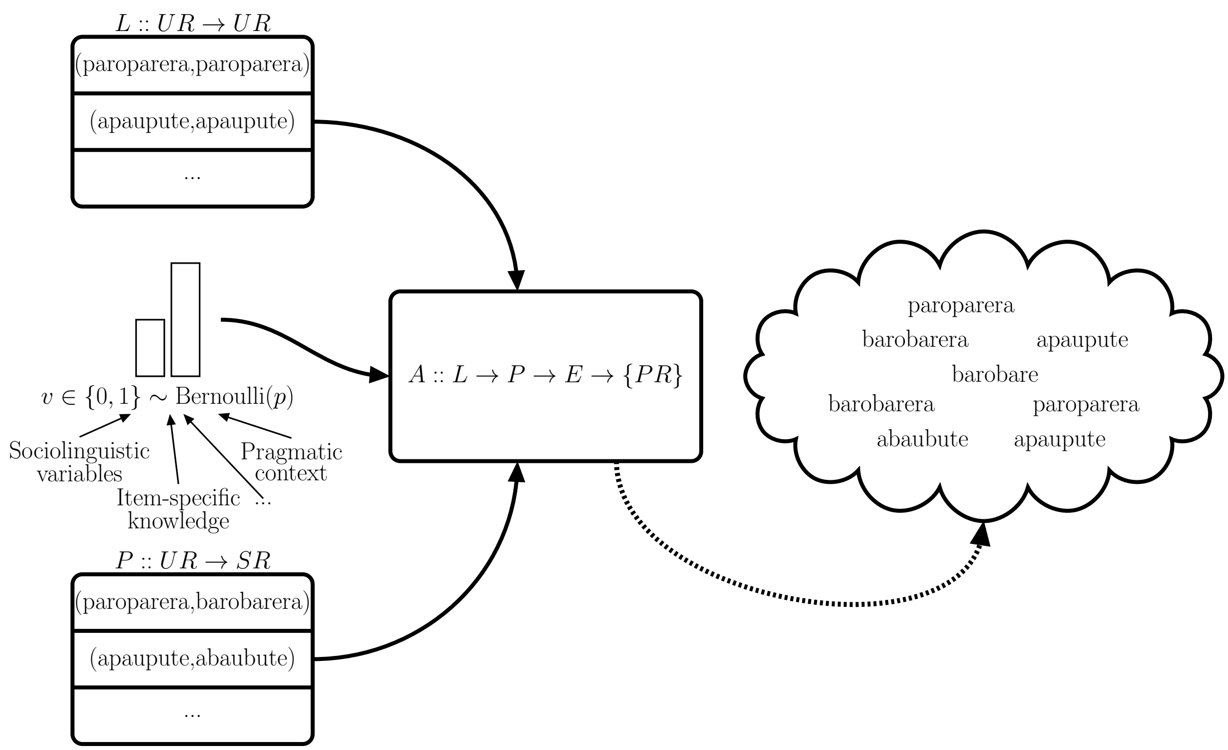


Figure 1: An example of data generation under the Blueprint Model of Production for labial voicing in Waoro. The lexicon and phonology are discrete, but the parameter v is controlled by various external factors which create a probability distribution over its application.

As Lakatos (1970) writes, “A ‘model’ is a set of initial conditions (possibly together with some of the observational theories) which one knows is *bound* to be replaced during the further development of the programme, and one even knows, more or less, how.” The Blueprint Model of Production was originally proposed to provide an explanation for how discrete phonological knowledge could still account for continuous phonetic phenomena such as incomplete neutralization and variation in homophone duration. Here I have shown how a slight alteration of some of the original assumptions could lead to a further understanding of optionality and variability as it pertains to rate of application of phonological knowledge. In both cases, being explicit about the structure of the phonetics-phonology interface has provided insights into potential causal mechanisms underlying these phenomena.

Alternative Approach 2: deterministically generating multiple forms with weighted logic

Turns out, I forgot to push my LaTeXcode from my desktop in Illinois for this section. So I just added the pdf of a recent abstract that highlights this approach.

Before that is some data from the Southeast dialect of Finnish (reported by Anttila, 1997) which results in four potential surface forms. This is the result of non-optional hiatus resolution (low vowel deletes before high vowel suffix), optional assibilation (t→s/_i), and optional unstressed /i/ deletion (apocope). For input /lentä-i/ 'fly-PAST' the following four forms can be generated.

	/lentä-i/	/lentä-i/	/lentä-i/	/lentä-i/
Resolution	lenti	lenti	lenti	lenti
Assibilation	–	lensi	–	lensi
Apocope	–	–	lent	lens
	[lenti]	[lensi]	[lent]	[lens]

If time allows, I can discuss how these data may be problematic for my approach.

References

Anttila, A. (1997). Deriving variation from grammar. *Variation, Change, and Phonological Theory*, pages 35–68.

Boersma, P. (1998). *Functional phonology*. Netherlands Graduate School of Linguistics.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, 32(1):45–86.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Goldrick, M. and Daland, R. (2009). Linking speech errors and phonological grammars: Insights from harmonic grammar networks. *Phonology*, 26(1):147–185.

Hayes, B. (2022). Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics*, 8(1):473–494.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Lakatos, I. (1970). *Falsification and the Methodology of Scientific Research Programmes*, volume 4, page 91–196. Cambridge University Press.

Mohanan, K. P. (1986). *The theory of lexical phonology*, volume 6. Springer.

Nelson, S. (to appear). Optionality and the phonetics-phonology interface. In *Proceedings of WCCFL 42*.

Nelson, S. and Heinz, J. (in press). The blueprint model of production. *Phonology*.

Osborn, H. A. (1966). Warao I: phonology and morphophonemics. *International Journal of American Linguistics*, 32(2):108–123.

Tilsen, S. (2023). Probability and randomness in phonology: Deep vs. shallow stochasticity. *Studies in Phonetics, Phonology and Morphology*, 29(2):247–269.

Phonological Knowledge, Weighted Logic, and the Competence/Performance Distinction

Overview There has been much debate over the role that gradience and probability play in the formalization of phonological generalizations, but a continuously growing body of evidence has shown that said generalizations do not always hold over the entire lexicon and are often variable in application even when they do. Many researchers take this to mean that phonological knowledge is itself gradient (Hayes 2022). Here, we take a different approach. Drawing on the competence/performance distinction, we formalize the difference between *phonological knowledge* (the possible discrete mappings from underlying to surface form) and *phonological usage* (the observed gradient probability distributions of applied forms) with weighted logic and semirings.

Background Model-Theoretic Phonology accounts for phonological knowledge by using mathematical logic to describe properties of finite structures (models). Structures are defined by assigning properties to elements of a domain (e.g. - labels, ordering). The *interpretation* of one structure in terms of another structure provides a way to formalize phonological transformations. Using boolean logic, the interpretation defines a discrete input-output mapping. For example, formulae like $\phi_P(x) \stackrel{\text{def}}{=} Q(x)$ are interpreted as "domain element x has property P in the output structure if it's corresponding element has property Q in the input structure. A phonological process like $a \rightarrow b/c\_\_d$ can be turned into a logical statement directly: $\phi_{\mathbf{b}}(x) \stackrel{\text{def}}{=} [\mathbf{a}(x) \wedge \mathbf{c}(p(x)) \wedge \mathbf{d}(s(x))] \vee \mathbf{b}(x)$. This reads, "domain element x has the property of being a **b** on the output if it was an **a** on the input and it was preceded ($p(x)$) by a **c** and followed by ($s(x)$) a **d** in the input or if it was already a **b**.

Competence and Performance with Weighted Logic First, we declare a function $k : \Sigma^* \rightarrow \mathcal{P}(\Sigma^*)$ representing phonological **k**nowledge. This is a function from strings to sets of strings based on optional application of a phonological process. Furthermore, we make the claim that all phonological processes are "optional" at this stage. In other words, the output set will always contain the fully faithful input as well as the additional strings where the phonological process may have applied. Second, we declare a function $u : \mathcal{P}(\Sigma^*) \rightarrow P(X \in E)$ representing phonological **u**se. This is a function from sets of strings to a discrete probability distribution over that set. The event space $E \in \mathcal{P}(\Sigma^*)$ represents the possible outputs of the application or non-application of phonological knowledge. Suppose for input string w, $k(w) = \{w, t\}$ and $u(\{w, t\})$ outputs a probability distribution where $P(w) = 0$ and $P(t) = 1$. Given this distribution of phonological use, "optionality" disappears. Under this interpretation, phonological knowledge is the knowledge of possible mappings while phonological use refers to how frequently that knowledge is applied.

These functions can be formalized using weighted logic and semirings (Chandlee and Heinz 2017; Mayer 2024). While Boolean logic provides a way to characterize functions of the type $f : \Sigma^* \rightarrow \{\texttt{TRUE}, \texttt{FALSE}\}$, weighted logic allows for more general interpretation of functions of the type $f : \Sigma^* \rightarrow \mathcal{S}$ where $\mathcal{S}$ is any semiring. The important characteristics of a semiring are that it is a set that is closed under two binary operations $\oplus$ (with a corresponding identity element 0) and $\otimes$ (with a corresponding identity element 1). The set $\{\texttt{TRUE}, \texttt{FALSE}\}$ is a semiring where disjunction ($\vee$) acts as $\oplus$ and FALSE is the identity 0, and conjunction ($\wedge$) acts as $\otimes$ and TRUE is the identity 1. Weighted logics also vary in that negation can only be applied to atomic formula and only conjunction and all formula must be written with only conjunction and disjunction. The two semirings that will be used to describe the functions above are given below.

Name	$\mathcal{S}$	$\oplus$	$\otimes$	0	1
Boolean	$\{\texttt{TRUE}, \texttt{FALSE}\}$	$\vee$	$\wedge$	FALSE	TRUE
Finite Language	FIN	$\cup$	$\cdot$	$\emptyset$	$\{\lambda\}$
Probablity	$\mathbb{R}^+$	$+$	$\times$	0	1

The finite language semiring interprets union ($\cup$) as $\oplus$ with $\emptyset$ as its identity 0 and interprets string concatenation ($\cdot$) as $\otimes$ with the set containing the empty string ($\{\lambda\}$) as its identity 1. The set $\mathcal{S}$ in this instance is the set of all finite sets of strings. Additionally, members of $\mathcal{S}$ may appear as atoms in the logical sentences. Using the finite language semiring, the phonological rule mapping **a** to **b** between **c** and **d** is made optional with the following formula. $[\mathbf{a}(x) \wedge \mathbf{c}(p(x)) \wedge \mathbf{d}(s(x)) \wedge \{a, b\}] \vee [\mathbf{b}(x) \wedge \{b\}]$. When predicates evaluate to TRUE, they get replaced with the set containing the empty string $\{\lambda\}$ and when they are false they get replaced with the empty set $\emptyset$. Conjunction and disjunction are replaced with union ($\cup$) and language concatenation ($\cdot$) (which results in the empty set if either set is empty). Domain element 2 of the input string cad would evaluate to TRUE for the first three predicates and FALSE for the last one. The sentence is therefore reinterpreted as $[\{\lambda\} \cdot \{\lambda\} \cdot \{\lambda\} \cdot \{a, b\}] \cup [\emptyset \cdot \{b\}] \equiv \{a, b\}$. This results in domain element 2 being interpreted as both **a**, and **b**, hence optionality. The same formula can be used with the probability semiring replacing sets of finite strings with weights that are multiplied or summed depending on the connective used. By defining conjunctive sets of these types of formulae we can describe the function k with the finite language semiring and the function u with the probability semiring. I assume the probability distribution output by u results from normalization of the output weights (as is the case for MaxEnt (Hayes and Wilson 2008)).

English nasal-place assimilation A specific example is given using English nasal-place assimilation. For ease of exposition, it relies on the predicates $\texttt{velStop}(x)$ and $\texttt{labStop}(x)$ which evaluate to TRUE for any velar or labial stop and the variable ? as a stand in for all non-nasal sounds in the language. The following formula defines function k for the English nasal-place assimilation map: $[\text{ŋ}(x) \wedge \{\text{ŋ}\}] \wedge [\text{n}(x) \wedge \texttt{velStop}(s(x)) \wedge \{\text{ŋ}\}] \wedge [\text{m}(x) \wedge \{\text{m}\}] \wedge [\text{n}(x) \wedge \texttt{labStop}(s(x)) \wedge \{\text{m}\}]] \wedge [\text{n}(x) \wedge \{\text{n}\}] \wedge [?(x) \wedge \{?\}]$ This ensures that domain elements with an **n** followed by a velar stop can be either **n** or **ŋ** and similarly **n** followed by labial stops can be either **n** or **m**. All other domain elements surface faithfully. The following formula defines the pre-normalized weights for function u. In order to match the average probabilities for each form observed by Coetzee (2016), certain weights are dependent on the length of the string ($|w|$). $[\text{ŋ}(x) \wedge 1\}] \wedge [\text{n}(x) \wedge \texttt{velStop}(s(x)) \wedge \frac{|w|}{1.857}] \wedge [\text{m}(x) \wedge 1] \wedge [\text{n}(x) \wedge \texttt{labStop}(s(x)) \wedge \frac{|w|}{1.222}]] \wedge [\text{n}(x) \wedge 1] \wedge [?(x) \wedge 1]$. This ultimately gives an output probability distribution of 35% and 65% for an input sequence **nk** to be realized as **nk** or **ŋk** and a probability distribution for an input sequence **np** of 45% and 55% for an input sequence to be realized as **np** or **mp**. These values are further scaled by Coetzee (2016) to match perceptual inhibition and speech rate which are extra-grammatical factors and, given the approach taken here, are added to the u function. They can be added to the logical formula above with no difficulty.

Conclusion This abstract shows how phonological analyses using logic and model theory can account for gradient phenomena. The analysis is split into two functions: one covering *phonological knowledge* (what can be mapped to what) and one covering *use* (the rate at which forms appear). Composing the two functions suggests that they can be viewed as a single function, only the latter is considered to be affected by extra-grammatical factors and is therefore different in kind from the former. The choice of specific weighted logic echoes this competence/performance distinction.

Chandlee, Jane and Jeffrey Heinz (2017). “Computational phonology.” In: *Oxford Research Encyclopedia of Linguistics*.

Coetzee, Andries W (2016). “A comprehensive model of phonological variation: Grammatical and non-grammatical factors in variable nasal place assimilation.” *Phonology* 33.2, pp. 211–246.

Hayes, Bruce (2022). “Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation.” *Annual Review of Linguistics* 8.1, pp. 473–494.

Hayes, Bruce and Colin Wilson (2008). “A maximum entropy model of phonotactics and phonotactic learning.” *Linguistic inquiry* 39.3, pp. 379–440.

Mayer, Connor (2024). “One (semi)ring to rule them all: Reconciling categorical and gradient models of phonotactics.” LSA Session on Formal Language Theory in Morphology and Phonology.