Chapter 12

Current Developments

The purpose of this textbook has been to explain the foundational logic of phonology, which is a cognitive, symbolic system that maps the output of morphology onto phonetics. The fundamental need for such a system is the fact that the mind cannot already have stored the infinity of utterances constituting a language. Then where does this system come from? Obviously, "it" is learned, but on what basis? And what, exactly, is the "it" that is learned?

The basic logic of language acquisition requires a fixed basis, what we call "Universal Grammar", the architecture of the human language faculty. UG is whatever primitive building-blocks exist that allow construction of a mental grammar, be it a set of rules as presented here, or a finite-state transducer, or whatever other device one wants to posit. With this tool, a baby is born with an already-existing cognitive ability to learn any language just given exposure. They apply some kind of logic to their observations of speech – the words and sentences which they hear – and induce a particular grammatical analysis for the language. This book models the logic needed for a child to arrive at such an analysis of the facts which relate abstracted phenomena to concrete cognitive rules, including the logic of discovering the underlying representations needed to correctly generate language data. This model of data-to-rule analysis is not particularly complicated, it reduces to the art of identifying reasonable and obvious hypotheses, and comparing hypotheses to see which is the best hypothesis. When a child hears the word that refers to "egg" and hears that it is pronounced [ɛg], the simplest hypothesis is that the input to the grammar for that word is /ɛg/. Perhaps a less-obvious hypothesis is correct, perhaps the underlying form is /eg/, but without some positive evidence supporting a more abstract alternative hypothesis, the child would never contemplate the possibility that the underlying form is instead /og/, /øf/ or /livʊjʊ/. This is an important and uncontroversial but under-appreciated aspect of the theory of language, that the nature of the grammar of a given language is only partially influenced by properties of UG, there also has to be a perceptible fact of the language which can lead the child to a non-obvious rule or representation. After this experience in problem-solving, you will have a better idea what non-obvious phonological rules and representations are – and how it is *not* hard to reason to non-obvious analyses if there are facts pointing to that analysis.

A theory of grammar is a mandatory aspect of the logic of phonology, because it defines a limit on what things the child has to choose between. Linguists are skilled at coming up with myriad models of how utterances are generated. We often disagree on whether the goal is to develop abstract mental models of how languages are generated, or is it to develop abstract models of the sets of string that constitute individual languages. The theory of generative grammar posits that there *is* a specific cognitive architecture behind language production and comprehension, but some people consider it premature to talk about causal mechanism when we do not fully understand what are all of the sets of data "defining" languages. Despite such differences in ontological perspective, we can still

sensibly compare models of language, as long as we agree that there is a difference between "possible languages" and "impossible languages". A theory of grammar will then say what a possible solution to a phonology problem is. If a theory of grammar allows numeric operations such as summation or multiplying by 3, or if the theory of grammar admits continuous integer or fractional values for features, the theory requires one to consider certain kinds of solutions that do not exist for the theory assumed here. If a theory of grammar takes individual segments like [æ, n, t^h] to be unanalyzable atoms, rather than being conjunctions of individual properties, that theory forbids the kinds of rules that we have encountered here, and forces the analyst to devise alternative machinery to refer to segment classes such as "voiced stop".

What exactly goes into the correct theory of grammar is an empirical question, answered by asking whether phonological systems have or lack certain properties, as predicted by a given metatheory of phonology. We may ask, "Do grammars have rules that require computation of cosines, square roots, multiplication or sums", and if we find no such evidence (indeed we have found no such evidence), we do not include those tools in the arsenal of phonological computation. The ability to answer such questions depends on two special abilities: the ability to analyze phonological processes, and the ability to frame that analysis in different theoretical frameworks, and therefore to empirically compare competing theories.

The primary emphasis of this textbook is developing analytic, descriptive skills: the ability to see generalizations and untangle interactions between multiple generalizations in a language. These skills of analysis are stable across generations of linguists, and are not invalidated with every tweak of the overarching metatheory. They define how we are *able* to tweak theories, by giving us a stable basis for saying "this new theory still gets all of the old generalizations, and it also explains these puzzles which the competitors cannot explain". A secondary interest here has been appreciating how recurrent factual observations end up being encoded in the theory of grammar, for example how a factual puzzle about tone led to a modification of the theory, indeed a total revolution in phonology.

There are very many aspects of formal theory which were largely or entirely ignored in this work. This textbook does not purport to be an exhaustive or even adequate compendium of phonological theories (I have already said that the primary goal is learning about phonological *analysis*). Little to nothing was said here about markedness, morpheme structure conditions, rule iteration, the cycle, derived environment effects, disjunctive ordering and the Elsewhere Condition, simultaneous rule application, the theory of exceptions, diacritics and morphosyntactic features in phonology, rule strata and level ordering, templates. Perhaps your instructor has mentioned a few of these topics. The theory of non-linear representations brought with it a huge set of controversies as to how things should be represented. Further study is necessary to understand all of the ways of theorizing the facts of language.

In the next two sections, we will have a brief introduction to two theoretical revolutions which are still with us. The first is Optimality Theory, which totally changes *how* input-output mappings are computed, dispensing with rules in the sense that we have used

them. The second is Substance-Free phonology, which deliberately aims to make formal theory be simpler, even at the expense of making the theory less "explanatory".

12.1 Optimality Theory

One of the leading theoretical desiderata in phonology has been finding a precise answer to the question "what is a possible language?". The first-generation answer in SPE theory was a fairly precise and powerful mathematical system which, for the most part, could express any known input-output relationship. This is certainly better than a theory which cannot state well-known facts, the problem is that SPE theory also allows grammars to be formulated which express completely unattested mappings. A rule could easily be formulated to reverse all feature specifications in a class of segments in a context; a rule can be written to raise a final vowel if the number of syllables in the word is a multiple of 7; a rule can turn a word into its mirror image. The theory is not unlimited: there are no fractional feature values, and the theory cannot compute whether an integer is prime. There was a major concern at the time that the theory was "too powerful", meaning that it was able to do things that do not happen in human languages. Attempts were made to limit what rules could do, by imposing "constraints" on rules of grammar. This led to a widespread change in research goals as "characterizing 'phonological rule' so as to include all and only the phonological rules that the phenomena of a natural language could demand" (a position advocated by the famous linguist James McCawley).

Especially with the advent of autosegmental phonology, the entire SPE concept of "abbreviatory rule" with expressions like X_0 (any number of instances of X), or W(A(B(C))D)Y (a complex condition on rule application "if B is present A must precede…") was abandoned. Instead, rules were reduced to simple single operations of insertion or deletion of single objects, but the success of that approach very much depended on a greatly-enhanced theory of the "repairs" mandated by UG, exploiting the main accepted mechanism for limiting what rules do. This required imposing various universal constraints on rules. We have considered some examples of that approach in the chapter on autosegmental tone. Where this approach fails is that there isn't just one resolution to a particular "defect" such as having detached a vowel from its skeletal position. There are typically many imaginable repairs to a formal defect, so how do we say which of the growing set of supposedly universal constraints and repair strategies is actually followed in a given language?

One response to this problem was to try to reduce rules of language to a set of universal "parameters", an approach which we observed in the metrical analysis of stress. Rather than have a specific rule constructing feet for stress, one would simply say "construct feet!", and specify a set of pre-determined choices, such as directionality, labeling (sw vs. ws), quantity sensitivity and so on. A related problem was the problem of Prosodic Morphology, explored in McCarthy and Prince (1986) which was part of the foundation of Autosegmental Phonology – how does one account for processes of word formation requiring words to have a particular "shape", especially as found in Semitic languages? How do we account for reduplication, a morphological process which copies various parts of a stem in order to create a new inflectional or derivational form of the word? How exactly do you say "copy the first two syllables", or "copy all syllables"? The

answer to many of these questions typically involved representational conditions, to the effect that you "copy as many segments as are needed to create a well-formed foot".

Prince & Smolensky (1993) and McCarthy & Prince (1993) propose a very different theory of grammatical computation, replacing specific learned rules that say "this is how the form changes" with a system of universal constraints which say "this is what is wrong with that form". The idea of a grammatical dispreference stems from the SPE concept of markedness, a set of universal statements intended to distinguish unusual and unnatural phonological processes from common processes. Intervocalic voicing is fairly common in human languages, intervocalic devoicing is distinctly rare, yet simple grammatical theories do not recognize that fact, even though linguists recognise it as a truth. Markedness theory imputes to UG a formalization of various phonetic tendencies, the consequence of which in SPE theory being that it is formally more complex to state a rule of intervocalic devoicing than it is to state intervocalic voicing. This difference in simplicity indirectly results in the rareness fact. In SPE theory, these markedness statements oversee what rules do, always correcting rule outputs to conform to those ubiquitous constraints, unless the rule is expressly stated to override the markedness constraint (which makes the rule more complex, imperiling the acquisition of that statement of the rule).

Optimality Theory (OT) holds that there *are* no rules, just a set of constraints. In previous theories of constraints, all constraints are put together in an unstructured group, absolutely preventing violations. In OT, every constraint can be violated in a linguistic output. Departing from prior theories, constraints in OT have a specific order in a language, so the entire content of the grammar of a language reduces to a statement of the order (ranking) of the constraints, which are themselves universal. For example, there is a constraint Max requiring all input (underlying) segments to be present in the output – deletion is forbidden. There is a constraint Dep requiring all output segments to also be in the input – insertion is forbidden. There is also a constraint NoCoda that forbids coda consonants. Finally there is a constraint Ident which forbids changing the feature content of a segment from its underlying form. These multiple constraints create a paradox for an underlying form like /pɪl/. The output [pɪl] violates NoCoda, [pɪlɪ] violates Dep, [pɪ] violates Max, and [pɪɪ] violates Ident by changing /l/ into [ɪ]. All other imaginable outputs from /pɪl/ violate at least one of these constraints, plus some others. Yet each of these outcomes is what actually happens, in some language.

By judicious ordering of the constraints, we can easily describe the repair strategies of specific languages, by lowering the significance of a particular constraint violation. In the language where /pɪl/ becomes [pɪl], the only constraint violated (in the set under discussion) is NoCoda, so since the correct outcome is [pɪl], we learn that violation of NoCoda is penalized less than violation of Dex, Max or Ident. For the language where /pɪl/ becomes [pɪ], the only constraint violated is Max, meaning that Max has a lower ranking than the other constraints.

The specific computation requires three related sub-components for generating outputs. One is CON, which is the set of constraints (given by UG) and their ordering in the specific language (which is what *defines* the specific grammar of the language). The second is GEN, which lists the set of possible outputs. GEN is simple the set of all

possible linguistic representations (possible in *any* language, i.e. is linguistically well-defined). The third is EVAL, the algorithm for deciding which output from GEN is the "best", given the constraint ranking in CON. The computation is conventionally presented in the form of a tableau like (1). The underlying form is in the top left corner, underneath it are all of the candidates being considered (obviously, not all candidates can be listed, one lists just the most relevant ones). To the left of the vertical double lines are the relevant constraint names and a star for every violation for that constraint within the given candidate.

(1)

pɪl	Dep	Max	NoCoda	Ident
pɪlɪ	*			
pɪ		*		
pɪl			*	
pɪɪ$_1$				*
pɪlɪl	**		*	
pɪɪ$_2$	*	*		

The candidate [pɪlɪ] has only one inserted segment, so only one violation of Dep; [pɪ] has only one deletion, so only one violation of Max; and so on. The second to last candidate [pɪlɪl] not only violates Dep twice, it still violates NoCoda. Compared to the four candidates above it, [pɪlɪl] cannot be the winning candidate (unless there is some other constraint that precludes picking one of those four. There are also two phonetic forms [pɪɪ], labeled with subscripts. [pɪɪ$_1$] simply changes /l/ into [ɪ], violating Ident. [pɪɪ$_2$] however both deletes /l/ and gratuitously adds another [ɪ].

In (1), we have not committed to a specific constraint ranking since the point was simply to give an example tableau with constraint violations, showing which candidates violate which constraint. When two constraints are mutually unranked, the columns are separated with a dashed line. Now let us assume that the correct output in the language we are interested in is [pɪɪ], therefore we are looking at tableau (2). That would mean that changing /l/ into a vowel is least-consequential, i.e. ranked lowest. If Ident is the lowest-ranked constraint, then violation of even one of Dep, Max and NoCoda removes a candidate from consideration and violation of Ident is least consequential – this is marked with an exclamation mark. Therefore [pɪlɪ, pɪ, pɪl] are knocked out of the candidate pool – as are [pɪlɪl] and [pɪɪ$_2$]. The only surviving candidate, [pɪɪ$_1$], is thus the winning candidate (marked with a finger), meaning that this is what the grammar of this language produces for this input.

(2)

pɪl	Dep	Max	NoCoda	Ident
pɪlɪ	*!			
pɪ		*!		
pɪl			*!	
☞pɪɪ$_1$				*
pɪlɪl	*!*		*	
pɪɪ$_2$	*!	*		

Of course, this is not a complete evaluation of all possible candidates. We should at least be descriptively more precise about what generally happens in the language. In any theory, if you know a single instance – underlying /pɪl/ becomes [pɪɪ], we still need more data to determine what the general phonological pattern is (regardless of whether you are doing OT or rule-based theory). Is it that all final consonants become [ɪ]; or does the final consonant delete with lengthening of the preceding vowel; or do particular consonants change to specific vowels so /l/ becomes [ɪ] but /m/ becomes [u] and /k/ becomes [ɨ]? In order to arrive at the correct constraint ranking (thus, the grammar of the language), you have to know what the factual generalizations are. At this point in the course, this should be so self-evident that it hardly needs to be mentioned, unless one somehow loses sight of the fact that you only have the correct analysis when you have ruled out all of the alternatives.

Therefore to get the final grammar, you first have to know what the valid language generalization is: change the consonant into a fixed vowel, or delete and lengthen the surviving vowel, or some other option. Assuming that the generalization in the language is that all coda consonants become [ɪ], a more complete analysis would then say why the replacement vowel is [ɪ] and not [i, e, a…]. Amongst the constraints provided by UG are markedness constraints penalizing the presence of specific vowels, such as *[i], *[e], *[a]… In order to guarantee that a consonant becomes [ɪ], the constraint prohibiting [ɪ] would be ranked lower than the other constraints prohibiting [i, e, a].

(3)

pɪl	Ident	*i	*e	*a	*ɪ
pɪi	*	*!			
pɪe	*		*!		
pɪa	*			*!	
☞pɪɪ	*				*

Observe that in this selection of candidates and constraints, violation of Ident has no effect because all candidates being considered violate that constraint equally, so the choice is left to the individual vowel prohibitions. The only crucial ranking in this part of the tableau is that *ɪ must be ranked below the other anti-vowel constraints, so that its effect is nullified.

OT (at least the classic version) is considered to be "non-derivational" in the sense that there is a two-level mapping between input and output, whereas in standard rule-based theory there can be any number of "levels", i.e. intermediate products, where one must compute the outcome of applying Rule *n* to String *j*, before computing the outcome of applying Rule *n+1* to String *j´* (because prior computation by Rule *j* can affect how Rule *n+1* applies). In classic OT, the computation of the number of stars for a candidate only considers the input and proposed output. Heuristically, it may appear that one has to walk column-at-a-time from left to right in order to work out which candidate has the fewest violations weighted so that violation of Dep, Max or NoCoda are penalized more than violation of Ident and violation of *i, *e, *a are penalized more than violation of *ɪ. Yet

the binary numerals 10000101, 01000101 and 00000101 can be sorted instantly to reach the result that 00000101 is the smallest of these numbers, even though it might seem that we have to go through a meticulous column-by-colum comparison of digits.

As articulated by the creators of the theory, OT is a theory of constraint interaction, the above computation of stars and outputs, so there is technically very little left to say in OT unless the basic architecture is changed so that there are multiple strata of computation analogous to the levels of the derivational theory of Lexical Phonology (that is, Stratal OT), or we change to a completely serial version of OT that computes one change at a time, until the output is reached (Candidate Chains). Most research effort in OT has centered around the constraints themselves.

To date, there is no formal meta-language for expressing constraints analogous to the formalism of rule-based theories. Instead, constraints are usually stated in plain English, and part of the art of OT is knowing how to convert constraint names back to their full definitions (e.g. Ons, *Coda, *Complex, Agree). Constraints are subclassified into two basic types: Well-formedness and Correspondence (popularly termed Markedness and Faithfulness). Well-formedness constraints state what internal properties of a candidate are prohibited, so *Voiced means that voicing is prohibited, *NC̥ means that a nasal followed by a voiceless consonant is prohibited, and these are all requirements computed on the properties of individual candidate. Correspondence constraints regulate relations between forms or substrings of forms, coming in under a dozen types (the main constraint name identifying the relation) and about three flavors (a suffix identifying the levels involved). The main flavor of correspondence constraint is IO correspondence which governs the relationship between the output candidate and its input, thus classical “Dep” is more formally Dep-IO because it computes a relationship between the input form and the candidate – has anything been added to the output, given an inspection of the input? A violation of Dep-IO is thus defined so that the output [$p_1ɪ_2l_3ɪ_4l_5$] which contains segments not present in the input /$p_1ɪ_2l_3$/ violates the constraint (twice), and a violation of Max-IO is where the input /$p_1ɪ_2l_3$/ contains a segment which is not present in the output [$p_1ɪ_2$], resulting on one violation. Other correspondence constraints require that two input segments not merge into one segment (Uniformity), they do not split into two (Integrity) and they are not reordered (Linearity). In a language that allows two segments to merge into one, Uniformity will be low-ranked. At this point, you can go back to the analysis of Kamba in 6.1.2 to determine whether Uniformity-IO is ranked relatively high or low.

Beside stating particular Dep / Max / Ident relations between forms, there are also choices of what kind of forms enter into a relationship. Thus we find Base-Reduplicant correspondence constraint, which governs the relationship between the reduplicant portion of an output candidate and the reduplication-base portion of an output candidate (requiring the base and reduplicant forms to be the same in some way). There is Input-Reduplicant correspondence, which requires the reduplicant to be the same in some respect as the underlying form, rather than the output base. Output-Output correspondence requires the computed form to be the same as some other word-form. For example in some dialects of English [æ] cannot precede coda [ɹ], except that truncated names like [hæɹi, bæɹi, læɹi, sæɹə] become [hæɹ, bæɹ, læɹ, sæɹ] – the vowel of the truncated name must be the same as the vowel of the un-truncated name.

One of the most persistent problems for OT to solve is how to accomplish what rule ordering in rule-based theories does. There is a traditional taxonomy of rule ordering relations where A precedes B, based on the effect of that ordering compared to the opposite ordering. Classically, A *feeds* B if application of A to a string S creates an input to B, and where B would not otherwise apply to S. If A is vowel epenthesis (preventing a coda consonant) and B is intervocalic stop voicing, the derivation /itka/ → *itika* → [idiga] where A precedes B is a feeding order. The opposite order where B precedes A is a counter-feeding order: /itka/ → *itka* (voicing, not applicable) → [itika]. Counterfeeding orders are particularly problematic for OT. Vowel epenthesis simply requires that Dep-IO be ranked relatively low compared to Max-IO and *Coda. Intervocalic voicing requires the constraint *VC̬V ("no intervocalic voiceless consonants") to be ranked above Ident-IO. The feeding relation where the output is [idiga] is easy to derive: Max-IO, *Coda, *VC̬V > Dep-IO, Ident-IO. This predicts that you do not delete in order to avoid a coda, therefore you must insert (a vowel). Because (and this would be an analytic fact that needs to be independently established) /iki/ → [igi] in response to *VC̬V we know that *VC̬V dominates Ident-IO. It follows from this ordering of constraints that the feeding output *[igida] is superior to the counter-feeding output [ikita]. So how can we *prevent* the constraints from producing *[igida]? The only constraint that might block voicing is Ident-IO, but we know that Ident-IO is subordinated to *VC̬V (since /iki/ → [igi]), therefore counterfeeding is an impossibility, in classic OT. Myriad solutions to this problem have been proposed, an entire course could be dedicated to evaluating the problems and their proposed solutions. Most of these solutions focus on expanding what constraints can do, for example by saying that only input-present vowels will trigger a violation of *VC̬V. Another way to put the generalization is that you can violate Ident-IO, and you can violate Dep-IO, but you cannot violate both Ident-IO and Dep-IO in the same substring.

The fundamental question about OT grammatical theory then is somewhat outside the strict purview of the original theory, namely "What is the theory of constraints".

12.1 Substance Free Phonology

Substance-Free Phonology is a framework which denies what has long been axiomatic in generative phonology, that the theory of grammar should contain mechanisms predicting all of the linguistically-significant generalizations about phonological systems. For example, if we conclude, factually, that no language has intervocalic devoicing or word-final voicing, then grammatical theory should contain some mechanism that forbids such rules, even if such a restriction is probabilistic.

Generative phonology originally focused more on formal aspects of the phonological computation, hence we were initially most interested in the syntax of rule-writing and application. Even within SPE, dissatisfaction was expressed with the lack of attention to phonetic underpinnings of rules, leading to a great expansion of phonetic principles in phonological analysis. Chomsky & Halle (1968: 400) themselves say that

> There is nothing in our account of linguistic theory to indicate that the result would be the description of a system that violates certain principles governing

> human languages. To the extent that this is true, we have failed to formulate the principles of linguistic theory, of universal grammar, in a satisfactory manner. In particular, we have not made any use of the fact that the features have intrinsic content.

Prince & Smolensky (1993: 216) urge a greatly-increased role for physical explanations in phonological theory, commenting that

> We urge a re-assessment of this essentially formalist position. If phonology is separated from the principles of well-formedness (the "laws") that drive it, the resulting loss of constraint and theoretical depth will mark a major defeat for the enterprise.

If we take the goal of phonological theory to be precisely distinguishing possible rules and representations, it seems inevitable that there will be some aspect of grammatical theory that has to explain why final devoicing is very rare (leading some people to argue that it is non-existent). These mechanisms are substantive, in the sense that they refer to specific features and values.

Changes in the philosophical foundations of generative grammar have resulted in a questioning of the premise that the theory of grammar needs to give a self-contained explanation for all significant generalizations about languages. The nature of a given language is obviously significantly influenced by the nature of genetic endowment for language and an individual's experience with language (the facts of the particular language that a child is exposed to). A third factor which was not at first properly appreciated was that some aspects of human language computation may be consequences of general cognition, not something specific to language. For example, phonological rules do not compute prime numbers, but this is not just a fact about the language faculty, it is a fact about human cognition. Adult humans can in principle comprehend the concept of prime numbers if they have learned how to divide two numbers and they understand the concept of a remainder, a reasonably clever person will know that all even numbers other than 2 are not prime and a very clever person may be able to compute on the spot whether 373 is prime (if they didn't memorize that fact before). Prime numbers simply do not enter into natural human cognitive processing, even though they are useful in encryption.

The importance of extra-grammatical influences on phonology is brought into focus in various works leading up to Hale & Reiss (2008). It is well-known that languages have a "prejudice" against voiced obstruents, and SPE markedness theory holds that [+voice] is "more costly" that [–voice] in obstruents (the opposite is true for sonorants). In OT, there is posited a constraint *Voice which penalizes voiced obstruents, but not a constraint *Voiceless which would penalize voiceless obstruents. The physical explanation for this prejudice against voicing is well understood in phonetics: vocal fold vibration cannot be easily maintained when there is a significant narrowing of the vocal tract (as is the case – by definition – with obstruents), because that narrowing results in pressure build-up above the glottis, reducing the pressure drop across the glottis which causes vocal fold vibration.

Phonetically-driven markedness devices in grammatical theory are thus scientifically redundant. We know from non-grammatical physiology and physics what the problem is

with voiced obstruents, we do not learn anything new by imposing on grammar a principle that voiced obstruents are difficult. There are very many such phonetically-explicable phonological patterns:

Nasals frequently assimilate place of articulation of a following consonant
Vowels often nasalize after a nasal
Intervocalic stops frequently voice or weaken to fricatives
Coda consonants, especially stops, very often devoice
Voiceless coda stops often change to [ʔ]
Consonants tend to delete adjacent to other consonants and vowels tend to delete adjacent to other vowels, but vowels do not have a similar tendency to delete adjacent to consonants and consonants do not likewise delete next to vowels
Velars become alveopalatals most often before front vowels, rather than before back vowels

Hale & Reiss direct our attention to the importance of language acquisition and the difference between synchrony versus diachrony in seeking explanations for such tendencies. It is obvious that individual grammars are learned, and that the properties of the data produced by speakers of the ambient language affect the nature of the grammar that will be learned. Children do not directly inherit grammars from other speakers, they create them on the basis of the outputs enabled by previously-acquired grammars. Grammars are only imperfectly learned, so that a strong statistical tendency in the data from one generation may lead to a categorial rule in the next generation – the child may mistakenly think that there is a two-way choice in degree of constriction, rather than a continuum (as you experienced with velars in Logoori, Chapter 2). Phonetic factors especially influence how forms are articulated and perceived, which affects the primary linguistic data. When phonetic facts predict and explain the nature of the primary linguistic data, thereby influencing the data to be "phonetically natural", we do not need any additional mechanism to say why the child did not learn an arbitrary unnatural phonological rule. The desideratum of explanation is that all facts should be explained, not that all facts should be explained by the theory of phonological computation.

One branch of research in the Substance Free framework has thus been identifying and explaining the historical origins of unnatural phonological processes, some of which are discussed in §8.4, for example *pw→tʃ* in Nguni languages, *mb→mp* in Sotho-Tswana, other examples are *nŋ → mp* in Nyole, consonant strengthening in Campidanese Sardinian, backing and rounding of /i/ after /d/ in Kashaya, open-syllable shortening in Menominee, and voicing of voiceless geminate stops as a realization of "strengthening" in North Saami. Investigation of how these rules came to exist shows that languages do not spontaneously develop arbitrary, phonetically-unnatural rules, they phonologize natural physical tendencies, then with that phonetic origin firmly encoded in the phonological rule system, they become subject to historical reanalysis, because a child only knows the data of the current language, not the entire history of how the language got where it is.

The seemingly unnatural rule where /mb/ → [mp] in Sotho-Tswana (*χʊbɔ́ná* 'to see', *χʊmpɔ́ná* 'to see me') arose from a natural historical reanalysis of ordinary sound changes. Proto-Bantu had a phonemic contrast between nasals, plain voiced, and voiceless stops, in (4a). The columns headed by *mb, b* and *mp, p* are not just phoneme distribution columns, they represent active alternations in root-initial consonants, where a root-initial consonant can be preceded by either by a vowel, Ø, or a nasal. Daughter languages developed stop / fricative differences in (b) due to various sound changes related to voicing, preceding nasals and following high vowels. The change b→ʋ/β (widespread in Bantu) is much earlier than the development of aspiration for voiceless stops and the change of p^h→φ except after a nasal. Our main concern here starts at stage (b), the subgroup which developed into the related Makua and Sotho-Tswana languages. At this stage, the child learning the language was faced with a choice between underlying /b/ versus /ʋ/. Since [b] has the most restricted distribution, the simplest rule is that ʋ → b /[nasal]__ rather than b → ʋ /{#,[+syllabic]}__ , leading to reanalysis of the voiced labial as /ʋ/.

Via a subsequent change (c), voiced stops were devoiced everywhere, which changed the original [ʋ ~ mb] alternation to a [ʋ ~ mp] alternation: this change characterizes the entire Sotho-Tswana subgroup. Again, the child had to make choices about underlying forms, the simplest rule accounting for the alternation being one which changes /ʋ/ into [p] after a nasal, rather than one changing /p/ into [ʋ] everywhere except after a nasal. Once you have the right rule identified, the underlying form follows automatically. The final step from (c) to (d), which results in this synchronically "unnatural" rule, involves three separate changes. The velar approximant descendant of PB *[g], presumed [ɰ], deletes everywhere (all forms, all dialects) – consequently, vowel-initial roots insert [k] after a nasal prefix. The coronal approximant [l] is retained except before high vowels where it becomes [d] or [ɺ], depending on language. The labial approximant [ʋ] becomes the voiced stop throughout the language cluster except in Sepedi where it remains [ʋ].

(4)	a.	mb	b	mp	p	original (Bantu)
	b.	mb	ʋ	mp^h	φ	voiceless spirantization (Pre-Sotho)
	c.	mp	ʋ	mp^h	φ	stop devoicing (Sotho general)
	d.	mp	b	mp^h	φ	approximant occlusivization (dialect specific)

Unconditioned occlusivization of ʋ (or β) may not be one of the most frequent historical sound changes in language and it is unusual in Bantu, but there are analogous historical changes in Germanic, Provençal, Indic and Iroquoian. The key to reconciling the phonetic arbitrariness of phonological rules like /mb/→[mp] and the desideratum of seeking explanations for all facts is to recognize that grammatical theory is only responsible for capturing the relationship between observed data and a system of rules and representations in an individual mental grammar. Phonological theory is reponsible for explaining how rules and representations are induced from phonological data; historical linguistics, the theory of language acquisition, and phonetics are responsible for explaining how the data that form the basis for inducing a grammar are the way they are. The primary desideratum of Substance Free theory is then setting forth a formal theory of phonological representation and computation. A formal theory considers the abstract

symbolic form of a computation or representation, and not the physical substance that the computations and representations refer to. The explanation for the grammatical oddity is that the language facts demand a rule /mb/ → [mp], and the nature of the language facts is a consequence of multiple sound changes over generations that do not follow a long-range grand plan.

Substance Free theory treats rules and representations as substantively arbitrary, meaning that a rule A→B/C_D is valid only with respect to the syntactic properties of the symbols – is replacement of A by B a valid operation, can a rule refer both to what precedes the focus and what follows the focus, irrespective of what entities A, B, C and D refer to? Analogously, in formal logic it is invalid to conclude Q from the premise "If P then Q", regardless of whether P stands for "Apples are fruits" or "All men are mortal", and whether Q stands for "Peaches are fruits" and "All dogs are mortal".

Basic SPE rule theory is substance-free in that the formal definition of rules does not refer to specific features or values, therefore a rule [+nasal] → [+back] / [+lateral]__[–round] is as valid as a rule [+voice] → [+continuant] / [+syllabic]__[+syllabic], even though the later is statistically very common and the former is not observed in any human language. SPE rules were self-contained in that a given rule directly encodes what action the rule performs, only requiring an interpretatation of the notation "→ means 'becomes'" and "X_means X precedes". However, SPE itself did abandon the goal of purely formal self-contain rules, by encoding highly substance dependent rules of "markedness" which provided certain feature values "for free" in rules. Self-contained rules which directly say what they do got replaced with conditions on rules which fleshed out vague rule statements, because it was felt that rules became more explanatory by recognizing those physical causes.

Another example of the rising trend away from self-contained rules is the "Elsewhere Condition". While the ordering A > B generally implies that the output of A can form an input to B, the "Elsewhere Condition" adds an overriding clause that if A is "more general" and B is "more specific" and the changes performed by the rules are incompatible, the output of A does not undergo B. Another generally-accepted rule-external rule condition was the "Revised Alternation Condition" which says that automatic neutralization rules only applied to derived forms. Questions do arise as to what constitutes an "automatic neutralization rule" and what counts as a "derived form", but with these questions answered we still end up with a radically different theory of rules compared to original SPE theory (sans markedness) because the rule itself does not say what the rule does, that can only be determined through an auxiliary computation of conditions on rules.

An early example of an overarching constraint on rules from the autosegmental era is the Obligatory Contour Principle, which in its most general form said that there cannot be two adjacent identical elements in a representation. Originally, this applied to tones and meant that a representation could not contain the sequences HH or LL, but the scope of the constraint and the theory of representations changed to the point that it became a parameter in rule statement, where a rule might be subject to the condition that two adjacent vowels cannot contain H tones, or that two adjacent syllables cannot contain aspirated consonants: or, such conditions might hold only for adjacent segments, or

adjacent words. The available parameters governing rule application may then also refer to phonetic substance, because it turns out that tone rules are never sensitive to whether the tones are on adjacent segments, such conditions – but only for tone rules – are always based on some higher prosodic unit, not segments.

The Substance Free approach seeks to eliminate such rule-external conditions, requiring that the only consideration determining rule application is stated directly in the rule: is this a syntactically well-formed rule? This theory does reject the dominant theory of constraints assumed in OT, that a constraint is a statement of dispreferred phonetically-defined states (e.g. *Voice means that segments with the specification [+voice] are penalized, but there is no constraint penalizing voiceless segments). However, this disagreement is not necessarily over the idea that a rules should be replaced with negative statements, it is over the idea that the constraints themselves are pre-defined in UG in terms of phonetic properties, and are justified based on their phonetic efficacy. Blaho (2008) advances a substance-free version of OT with theory of constraints where a child only has access to the form of constraints (e.g. Dep(X), Max(X), Ident(X), for arbitrary values of X, so a formal theory of constraints is logically possible, even if it is not actively pursued.

A significant difference between the Substance Free theory of rules and SPE rule theory is that Substance Free rule theory gets rid of SPE abbreviatory devices such as numeric subscripts, parentheses, braces and value variables, to the greatest degree possible. This is made possible by embracing the fundamental representational insights of autosegmental theory, allowing segments to be unspecified for some features, or a feature can have multiple segments (or no segments) as its domain. We eliminate the need for expressions like "$(C_0\check{V})_0$" in stress rules by accepting higher representational objects "syllable" and "foot" and directly saying in the rule of foot construction that certain kinds of syllables must be the head of the foot.

Another branch of inquiry in the Substance Free approach is the question of what features are. The original SPE theory held that features are exact phonetic descriptions of language sounds, precise to the point that a phonetic observation can be said to be a property of a language and not a consequence of human anatomy. A well-known example of such extreme precision in grammar is the fact that the high vowel [i] does vary across languages – as we noted in Chapter 2, the vowel [i] in Dutch is pronounced higher than it is in Turkish, and [i] in English is between Dutch and Turkish [i]. This is handled in SPE theory by saying that features have numeric values so that two languages can differ in whether their high vowels are [1high] versus [2high]. The limit on numeric values for features is that they are always introduced by language-specific phonological rule.

This approach was mandated in SPE theory because that theory did not include a language-specific component of phonetic interpretation saying how [i] or [t] are physically realized, instead the phonology must say exactly how feature matrices are pronounced in the language. With a language-specific phonetic component, phonology no longer carries the burden of exact continuous phonetic descriptions, and phonology can focus on abstract categories of "sounds" in the non-physical sense. Features can be more vague: we can say that English, Navaho and Thai have "aspirated stops" (e.g. [+spread glottis]) even though the physical realization of those stops differs substantially between

the languages. The Unified Features Theory approach (Clements & Hume 1995) collapses SPE feature distinctions substantially, so that vowel fronting including palatalization and coronality (alveolar, dental, alveopalatal etcs) reduced to one feature, [coronal], just as "labial" and "round" are a single feature [labial] distinguished by what node dominates the feature (C-place versus V-place). What UFT contributes is the idea that structural contrasts (what dominates what) can obviate the need for certain feature distinctions. The Parallel Structures model (Morén 2003) goes further, largely eliminating strict phonetic definitions, instead putting more emphasis on phonological class behavior rather than phonetic property as the basis for feature assignments. At the logical end of this continuum of theorizing we find Radical Substance Free Phonology (Blaho 2008, Odden 2022) which holds that UG only contains the fact that segments are conjunctions of features, and features have no phonetic definitions so it is meaningless to ask whether the features of one language are the same as the features of another language, instead features are learned entirely on the basis of class behavior in rules.

12.3 Synopsis

I will close with a bit of a personal memoir. When I was a student, my dissertation advisor, Charles Kisseberth, summarized his experiences studying phonology at MIT in 1967-8. The first term was spent reading everything that there was to read in generative phonology, because the entire literature was a small collection and could be covered in a quarter. The second term was then cutting edge research projects. I started studying phonology in 1974, and by 1981 I might have read about a quarter of what had been written in generative phonology. Currently, an ambitious student might aspire to read 10% of what has been written since 2000. It is not that we are getting slower at reading, it is that there has been a massive expansion of research in phonology, making it hard to "keep up". At times it may seem that our many conflicting theories impede communication and progress in the field: but the ability to reduce theoretical constructs to the perceptible through the standard logic of phonological analysis means that we still have a hope of judging whether competing theories say "the same thing".