A Short History of Phonology in America:  
Plus C’est la Même Chose, Plus Ça Change  

Stephen R. Anderson  
Dept. of Linguistics, Yale University  

Abstract  

Although awareness of a difference between the study of the sound patterns of particular languages and the study of the language-independent capacity of humans to produce and perceive sound existed in European and American thought in the early years of the twentieth century, a clear enunciation of the distinction between phonology and phonetics is due to Otto Jespersen in 1924. As phonology became established as a coherent object of inquiry, two themes can be identified in theorizing about it: first, the question of whether phonological structure is in the mind, as an aspect of human cognition, or only a set of facts about the external data of language; and second, the question of whether there are valid universals of phonological structure. These issues are traced across the past century in the work of American linguists. An additional factor identifiable in historical shifts in theoretical perspective is somewhat less principled: as discussion of fundamental issues becomes more technical and relevant data harder to identify, students and scholars looking for productive research topics tend to abandon previous frameworks for others in a search for lower hanging fruit without necessarily having resolved the earlier questions.

Keywords: cognitive science, constraints, generative phonology, phonemics, rules, universals

Although the serious study of language goes back millennia, “Phonology” as a distinct subdiscipline of linguistics is an idea with less than a century of history.1 We can begin with the distinction underlying the emergence of this sub-field from the more general study of sound in natural languages.

Phonetics, on the one hand, I take to be the study of sound related phenomena as these play a role in natural language, but from a point of view that is independent of the properties of any particular language. It is the study of the range of properties of utterances that distinguish one possible human utterance from another, independent of any particular language. Phonetics is thus concerned with the range of human articulatory capabilities, the acoustic consequences of articulatory gestures, and the way acoustic events affect the auditory system. In contrast, phonology is the study of the way sound properties and distinctions function within the systems of individual languages: the study of their sound patterns as an aspect of their particular grammars.

1 I note at the outset that after some initial comments, I will focus below primarily on the history of phonology in recent American linguistics, not because other traditions have nothing to say about the matters under discussion, but rather because they have too much, and a more general treatment would risk losing the thread of a particular interesting story. Also, although phonological notions are applicable to signed as well as spoken languages, I limit myself here to the latter as constituting the center of the developments described.
Section 1 below briefly surveys the understanding of “phonology” in the years prior to the rise of generative views in the late 1950s. Section 2 then discusses the emergence of generative phonology and the associations between that point of view and two important issues: the notion of linguistic structure as an aspect of the mind, and the existence of universals of phonological structure. Section 3 then surveys some of the reactions to Chomsky & Halle (1968) and their parallels to developments in the philosophy of mathematics, and the eventual replacement of the rule-based views of that work with theories based on simultaneously applicable constraints. Over many years, the replacement of one theoretical emphasis by another is seen as motivated not exclusively by the search for solutions to continuing problems, but also by the needs of emerging scholars for immediately significant results. Section 4 concludes with the suggestion that this motivation may still be at play today.

1 The Early History

The word “phonology” itself shows up in literature well before what we think of as the period of modern linguistics: the first citation in the OED is from 1798. In general, though, it does not have a precise meaning beyond something to do with the study of sound in language. Confusingly, de Saussure (1916) uses the word (or its French equivalent), but not in the modern sense: he uses phonologie as essentially what we call phonetics, and phonétique for the historical study of sounds: he had no term for the systematic study of language-particular synchronic sound systems.

It is in the writings of Otto Jespersen that we first find a proposal for a clear delineation of the sort we assume today:

It would, perhaps, be advisable to restrict the word “phonetics” to universal or general phonetics and to use the word phonology of the phenomena peculiar to a particular language (e.g. “English Phonology”). (Jespersen 1924: 35)

Regardless of the words used to describe this difference, however, it was not unfamiliar to serious students of language, and there was quite a lot of interesting work in the European tradition. This might be seen as beginning with Jan Baudouin de Courtenay (1895 [1972]), whose distinction between anthropophonics — essentially physical phonetics — and psychophonetics comes quite close to our contemporary understanding. But this tradition was little known and quite without influence in American linguistics at the beginning of the 20th century. Ferdinand de Saussure is generally thought of as more central to the development of the field, but he too was not at all well known in America in the 1920s, and in any event his ideas of phonological structure are somewhat harder to discern than is sometimes supposed (Anderson 1985: 33–55).

In America, Franz Boas represents a major strand in distinctively American thought about language in these early years, but as far as sound was concerned, he was mostly concerned with accurate phonetic transcription, rather than with things like sound patterns (Anderson 1985: 204–216). Efforts have been made to find support in Boas’ practice for a notion of phonological representation distinct from surface phonetics (Postal 1964), but this interpretation is hard to sustain. Indeed, it is interesting to note that when his principal Kwakw’ala consultant George
Hunt sent him textual material transcribed in a way that abstracted away from certain phonetic differences predictable within the language, Boas ‘corrected’ this so as to restore surface phonetic accuracy (Anderson 1985: 206). Boas’ descriptive practice does reveal a certain amount of interest in potentially universal properties of sound structure, a topic which will occupy us below in section 2.2, but these ideas were quite inexplicit.

The first significant piece of work in America dealing with phonology as we understand it was Sapir’s (1925) paper “Sound Patterns in Language,” which appeared in the first issue of Language. Sapir argued not only that such structure is quite apart from the concrete properties of speech sounds, but that it is a fundamentally psychological notion, in the sense we would today call ‘cognitive’. Elements have their place in such a structure on the basis of complex patterns of associations and not on the basis of external, objective properties. We can see in this paper the beginning of the notion that the study of language, and phonology in particular, is a part of Cognitive Science (Anderson 2001).

Reaction to that conception was not long in coming, however: the very next paper in the same inaugural issue of Language was Albert P. Weiss’s (1925) article “Linguistics and Psychology,” which resolutely rejects the significance of a notion of ‘mind’ and insists that all ‘psychological’ study of language must be grounded in (and limited to) externally observable properties and phenomena. Weiss’s call was enthusiastically taken up by his new colleague at Ohio State, Leonard Bloomfield, and in the scientific atmosphere of the time, Sapir’s conception was displaced for the next several decades by a very different (and non-cognitive) notion of phonemics — a notion that was at the heart of nearly all American theorizing about language during this period, which tended to take phonemics as its model for the study of other components of linguistic structure.

In theorizing about sound structure on a path following Sapir’s early ideas, we see that the Weiss/Bloomfield conception overwhelmingly held sway in the development of phonemics. Bloomfield’s (1933) Language represents the classical period of structuralist thinking about sound structure. This is a picture focused on surface phonetic form, with the abstract notion of the phoneme best thought of not as a cognitive reality in itself, as Sapir had urged, but rather as a set of purely phonetic segments, united by statements of their respective distributions. This way of thinking predominated up until around 1960, when it was fairly abruptly replaced by quite a different point of view.

2 The Rise of Generative Phonology

A new way of approaching sound structure which we can see as that of “phonology” in today’s sense was ushered in by the publication of Morris Halle’s (1959) Sound Pattern of Russian and “Phonology in Generative Grammar” (Halle 1962). These works challenged the fundamental assumptions of structuralist phonology, and the approach they advocated quickly took over the field. Following a heated exchange between Fred Householder (1965) and Chomsky & Halle (1965), the ascendancy of the generative point of view was essentially complete by the time its codification in Chomsky & Halle 1968 appeared.
What provoked the fundamental re-orientation of phonological thinking associated with the replacement of structuralist phonemics by generative phonology and its descendants? The usual picture is that in the late 1950s, Halle (1957) discussed some interesting facts about voicing in Russian obstruents. In this language, voicing is distinctive for some obstruents but not for others, but voicing assimilation is quite general across all obstruents in clusters. Halle argued that a description interposing a phonemic representation between a morphophonemic one and the surface phonetic form necessarily loses the unitary nature of the generalization about voicing assimilation. Presented with that evidence, the story goes, the field rapidly converged on the view that structuralist phonemic representations were a bad idea, and replaced them with the more abstract representations we have since come to know and (mostly) love.

This story leaves out some important history, though (Anderson 2000). In fact, examples just like Halle’s had been discussed before, and taken to lead to very different conclusions. Bloch (1941), for example, discussed essentially similar facts in the distribution of vowel length in some dialects of American English, but instead of concluding that phonemic representations were a bad idea, used the facts to argue that the rigor of phonemic theory could save the linguist from being misled by a seductive apparent regularity.

The force of Halle’s argument, in fact, lay not in the nature of the facts, but in the use Halle made of them. He argued essentially that a description of a language’s phonology had to do more than simply get the forms right — it also had to capture the regularities that govern the distribution of those forms: the rules as well as the representations. Of course, this line of reasoning only makes sense if we assume that the rules are really part of the language, and that in turn only makes sense if we assume that our description has to encompass the knowledge speakers have. And that, of course, is just what phonemic theory was opposed to, based as it was on essentially behaviorist notions according to which only external observables, and not ineffable notions like “the mind” were the province of genuine science.

A key component in the striking success of the generative point of view on phonology, then, was its appeal to linguistics as what would later be called Cognitive Science, and in the resurgence of talk of language as something in the mind, not just out there in the products of speech. But this was not the whole story.

Another contributor to this development was surely the trend in science observed by Max Planck (1949: 33f.): “A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.” That is, scientific progress generally occurs not horizontally, by transmission from one established figure to another, but rather vertically, by the influence of an innovator on students. And this was definitely a factor in the rise of generative views. As Searle (1972) put it, “All of this facilitated a rapid dissemination of new ideas and a dramatic and visible clash of conflicting views. Chomsky did not convince the established leaders of the field but he did something more important, he convinced their graduate students.” Searle refers here to Chomsky’s influence in the emergence of new views on syntax, but the point is also applicable to Halle’s role in the development of phonology.

Planck’s comment on scientific progress is generally taken to require the actual death of proponents of an earlier view: it is commonly paraphrased as “Science advances one funeral at a
time.” Generational change of the implied sort can be precipitated without this, though. In the specific case of phonemic theory, we can observe that by the 1950s, the range of available research questions had become quite constrained. The overall theory of the phoneme was considered a settled matter, and the theoretical thrust of descriptive studies restricted at best. The limitations of the research horizon in phonology were somewhat exacerbated by the methodological rigor and strict procedural bias of the main structuralist phonologists (Harris 1951), together with their insistence that phonemic analyses must be arrived at without reference to any other aspect of language. Significant scholars who argued otherwise, such as Kenneth Pike (1947a,b) with his insistence that all areas of language were to be investigated in their mutual relation, were essentially marginalized in discourse about phonemic theory.

Some remaining controversies, such as the correct analysis of suprasegmental properties such as levels of stress in English, seemed quite intractable, and the structuralist approach to these phenomena in terms of phonemes was seriously undermined by work in what would become the new theoretical framework (Chomsky et al. 1956). As a result of all of these factors, graduate students searching for dissertation topics were more than ready to have opened before them a quite distinct view that, by discarding the specific limitations of classical phonemics, made available a rich array of new research questions. Some of the success of generative phonology, then, must undoubtedly be attributed to a degree of stagnation in structuralist theorizing.

2.1 Phonological Structure as an Aspect of the Mind

In intellectual terms, we can identify two “big picture” issues important to the rise of generative phonology on which positions have shifted over the years. One of these is the question of whether phonological structure and regularity resides in the mind or in the external productions of speakers.²

Sapir and his students, together with most adherents of generative phonology and its various descendants, would maintain that the description of a phonological system is the description of a fundamentally cognitive reality. In contrast, following the lines of behaviorists, American structuralists sought such structure in the sounds themselves. And indeed some recent views, becoming more prominent even as I write, represent in some ways a return to such a concern with the external manifestations of language, and a rejection (or at least considerable weakening) of the view that phonological structure is an aspect of the mind and the biological language faculty.

When we look at the history of notions in the field of psychology that are relevant to language, we can date the emergence of an important perspective to the publication of John B. Watson’s (1924) book *Behaviorism*. This point of view, which we saw already in Weiss’s (1925) article in Language, became the dominant paradigm in psychology for quite a number of years (especially as later taken up by B. F. Skinner), almost completely eclipsing Sapir’s approach among linguists.

² For a vastly more comprehensive study of the role of notions of the mind and cognition in the fields of philosophy, psychology and linguistics, including the relations among these fields in the years leading up to the emergence of generative phonology and syntax, see Goldsmith & Laks 2019.
Matters changed significantly with Chomsky’s (1959) very influential review of Skinner’s (1957) book *Verbal Behavior*, which pointed out various ways in which the behaviorist perspective was unable to deal with the observed facts of how language is acquired and used. This review had an effect not only on thinking about language but more broadly, and it led to the rapid decline of “behaviorist” psychology more generally and the rise of “cognitivist” views. That in turn prepared the way for the revival of mentalist notions specifically in the study of language, as represented by the emergence of generative phonology.

Actually, the decline of behaviorism in psychology was in some ways more apparent than real: a remarkable number of psychologists today, even if they reject being labeled as “behaviorists”, nonetheless see the primary object of inquiry in psychology as an understanding of behavior, something to which talk of “minds” may contribute, but which is primarily about understanding external manifestations. But an analysis of these matters falls outside the scope of the present chapter.

### 2.2 The Place of Universals

A second intellectual theme, characterized by a similar polarity, is the matter of universals: whether these are real or just epiphenomena resulting from the limited range of languages we have to explore; and if they are real, what their status is. Do the regularities we uncover across languages follow from deeply grounded principles of Universal Grammar, or are they simply the natural outcome of more external events?

While Boas is historically associated with the notion that every language must be approached on its own terms, and thus with a rejection of universals, this is probably an over-simplification. Actually Boas had a rather strong, if implicit, view of universal constraints determining the content of phonological systems (Anderson 1985: 202ff.), but that was never really expressed as a theoretical principle, and at any rate had no particular theoretical status in the years when phonology was first coming to be seen as distinct from phonetics.

Sapir, however, already in his little book *Language* (Sapir 1921), rejected the notion that there are real, substantive universals of language. Sapir draws this in part from the Boas tradition of emphasizing the study of languages on their own terms (while neglecting Boas’ rather strong views about typological frameworks for linguistic description, which imply a strong set of substantive universals), and in part from the liberal politics of the time, which emphasized ‘nurture’ at the expense of ‘nature’ in accounting for the properties of individuals. Unlike his views on language as an aspect of the mind, Sapir was more in tune with the way the field would develop in rejecting language universals and attributing observed cross-linguistic regularities to factors of culture and historical transmission. This would be true at least until the 1960s, when the re-emergence of nativist views of innate structure resulted in the ascendance of strong theories of linguistic universals.

In the rise of generative phonology, we can see a resurgence of the idea that there are strong universals governing linguistic structure (and phonology in particular), principles that are grounded in the biological nature of the human language faculty. The embedding of facts of
particular languages in a framework of universals of linguistic structure is certainly the approach associated with almost all strains of generative phonology, up to and including Optimality Theory.

More recently, though, we see some scholars staking out positions that are more skeptical about the status of apparent universals: for example, Juliette Blevins’s (2004) arguments that these are just the consequences of independent principles of linguistic change, whose outcomes show regularities that are not due to any innate faculty of UG or the like. Similarly, Jeff Mielke (2008) has argued that even the basic descriptive framework of the feature system is not grounded in biologically determined universals, but just follows from the range of classes that arise in phonological regularities as a consequence of change and related factors; see also the papers in Clements & Ridouane 2011. These and other writers assert that an adequate understanding of observed cross-linguistic regularities can be obtained on the basis of statistical analyses of surface phenomena without an appeal to a rich human faculty of language whose characteristics determine and constrain the substantive content of linguistic systems, in phonology as in other areas of grammar.

In contrast, the most recent major innovation in phonological theory, Optimality Theory (as discussed below in section 3.3) takes a strong position on universals. The collection of markedness and faithfulness constraints that constitute the core of a grammar in this system are asserted to be universal in nature, with only their relative ranking subject to language-particular specification. The effectiveness of this claim is open to question, but the commitment to it on the part of theoreticians is clear, indicating that the issue of phonological universals is far from a settled matter.

3 Phonological Theory After Chomsky & Halle 1968

The appearance in print of Chomsky & Halle (1968) after many years of circulation in samizdat’ form was a major landmark in the development of phonological theory. By no means simply a description of the “sound pattern of English”, the work presented a more or less definitive formulation of the theory of generative phonology as this had emerged over the preceding decade or so. As such, it offered a clearer target for criticism of the underlying assumptions of that theory than had previously been available, and the field took up the challenge with enthusiasm.

3.1 The Logicism of SPE and Reaction to it

Chomsky and Halle’s work presented a theory of phonological structure in the form of a notation for the expression of phonological regularities, together with an architecture (based on sequential application of rules) within which the relation between underlying phonological form and surface phonetics could be computed. Following the classic discussion in Chomsky (1964), the theory aspired to success at three levels.

---

3 For a fuller presentation of the points discussed in this section, see Anderson (1980) and Anderson (1985: 328–347).
Minimally, the level of observational adequacy could be reached to the extent descriptions formulated within the theory could accurately reproduce the empirical data on which they were based. Secondly, these descriptions could be characterized as descriptively adequate to the extent they could be shown to reflect accurately not only the data from which they were constructed, but the underlying knowledge on the part of speakers that gave rise to those data. This could best be shown by demonstrating that the grammar correctly predicts novel data not present in its original basis, but which are confirmed by the intuitions of speakers.

There was little controversy over the claim that descriptively (and a fortiori, observationally) adequate grammars could be constructed within the SPE framework. Much less clear, however, was the claim that the theory itself could be shown to be explanatorily adequate, in the sense of providing a mechanical procedure by which to determine which of multiple candidate descriptions of a given set of facts, each observationally adequate, was in fact descriptively adequate.

Explanatory adequacy was to be achieved by incorporating into the theory an evaluation measure: an algorithm for computing, for each candidate grammar, its overall complexity, with the claim that the least complex observationally adequate grammar was in fact the descriptively adequate one. This evaluation was to consist in a counting of the feature specifications in each candidate grammar, with fewer features corresponding to less complexity. In order to make the description’s complexity correspond to its likelihood of being “correct” in the relevant sense, a variety of notational conventions were proposed, each supposedly corresponding to a type of linguistically significant generalization to be favored, and allowing the description to be shortened to reflect this. These notational conventions, far from being purely matters of aesthetics, were in fact a crucial, central component of the theory in its goal of attaining explanatory adequacy, each constituting an empirical claim about what sorts of generalization were in fact “linguistically significant”.

The resulting program intended to allow all problems associated with the discovery of a descriptively adequate account of sound structure in a given language to be reduced to the mechanical manipulation of expressions in a fully explicit notational system. It was not claimed that SPE completely accomplished this goal, but that was nonetheless the program of the theory. As a result, the research issues that it presented for study centered on matters of the correct choice of notational conventions and the elaboration of the architecture within which the rules were to be applied.

The goals of SPE were in fact strikingly similar to those of another fundamental work of 20th century thought, Whitehead & Russell’s (1910–1913) Principia Mathematica. That work also took as its project the reduction of all of the intellectual content of a field (mathematics) to the formal manipulation of expressions in a logistic system by means of fully explicit mechanical operations. Such an attempt to express all of the content of a field in terms subject to purely formal manipulation by explicit rules came to be known as logicism in the philosophy of mathematics, a term that can also be appropriately applied to the program of SPE.

As is well known, the initial successes of Whitehead & Russell’s work soon gave rise to discontent, as it became clear that there were fundamental challenges to it. These included for instance the fact that it apparently gave rise to a number of unresolvable paradoxes. Russell
attempted to resolve these through the addition of a theory of “types”, but this in turn led to counterintuitive and overly restrictive emendations of the system, and the logicist program for mathematics gradually came to be abandoned.

As summarily described by Kleene (1952: ch. 3), the problems posed by logicism gave rise to alternative views. Simplifying grossly, David Hilbert and others pursued on the one hand what is called formalism, attempting to pursue as much mathematical content as possible by logical manipulation without claiming that all of mathematics has this character. On the other hand, L. E. J. Brouwer attempted a resolution under the name of intuitionism by rejecting all expressions purporting to refer to objects that cannot in fact be fully constructed (such as infinite sets). This has the consequence that the paradoxes arising in Whitehead & Russell’s system are avoided, since the problematic classes cannot be constructed within the limits of an intuitionist logic. It also has the consequence, however, that a great deal of mathematics must fall outside the scope of the program. In the end, both formalism and intuitionism have generally been seen as inadequate bases for an understanding of mathematics.

Interestingly, just as the logicist program for mathematics gave rise to apparent problems and reactions, the same was true for Chomsky & Halle’s program as expressed in SPE. One of these was already recognized in the concluding chapter of that work: the fact that in principle, the theory excludes all matters of concrete phonetic content in favor of purely formal representation. This they attempted to remedy by the addition of a theory of markedness intended to play a role in reducing the substantive content of phonological representations and rules to matters of pure form. Just as the theory of types failed to find acceptance as an adequate resolution of the problems that arise within Whitehead & Russell’s system, however, the theory of markedness advanced in SPE failed to attract significant adherents.

A second problematic aspect of the SPE theory was noted by Paul Kiparsky in papers circulated in the 1960s, later published in fuller form as Kiparsky 1973. The issue here was the fact that descriptions in classical generative phonology can be constructed that are unrealistically abstract, in that they posit representations and sequences of rules that are probably inaccessible to native speakers. Again, a variety of restrictive additions to the theory of SPE were proposed to exclude such analyses, but no fully satisfactory resolution was achieved.

One attempt to avoid the problem of excessive abstractness was made by Theo Vennemann and his students at UCLA in the early 1970s, and most fully presented in the work of Joan [Bybee] Hooper (1976) under the label of Natural Generative Phonology. This theory took as its basic premise the requirement that phonological statements should be confined to ones that are literally true of surface phonetic forms. Such a principle is strikingly reminiscent of the intuitionist requirement that mathematical statements should not be allowed to refer to objects that cannot be explicitly constructed; like intuitionism in mathematics, Natural Generative Phonology allows for an account of a limited subset of its intended domain (here phonology, as opposed to mathematics), but in both cases the limitations are sufficiently severe that neither achieved broad acceptance.

Another effort to address the issue of phonetic content in phonological expressions can be found in the work of David Stampe (1973) under the label Natural Phonology. This view completely rejects the purely formal nature of phonological regularities as expressed by rules in
SPE. It posits a rich inventory of substantive natural processes, all presumed to be innate and active at the outset of language acquisition. On this view, acquiring the phonology of a particular language consists in learning to suppress the activity of some processes and to subordinate others, while also adding a limited set of arbitrary and language particular rules. The resulting system has some resonance with important aspects of Optimality Theory (section 3.3 below), but in its basic form it attracted only limited attention, due in part to the fact that a much more important role must apparently be ascribed to the rules that fall outside the set of natural processes than can easily be accommodated in the system.

Within a few years after the publication of SPE, then, there was enough discontent with its program to encourage students to look elsewhere. On the one hand, the program itself offered few obvious topics for significant advances: the problems raised by matters of notational conventions and rule ordering were highly technical, and difficult to address in terms of readily accessible empirical evidence — difficulties similar to those presented by phonemic theory a decade or so earlier. On the other hand, the attempts to deal with limitations of the SPE program within its general Weltanschauung all seemed unappealing for reasons of their own. The climate was ripe for a more serious re-orientation of theoretical attention.

3.2 A Focus on Representations

This found its expression in the re-orientation of phonological research over the next two decades from the study of phonological rule systems to the study of representations. SPE and related theories were built on the assumption that phonological (and phonetic) representations took the form of sequences of segmental units, each composed of values for features taken from a universally available set. The result could be represented as a matrix whose rows correspond to the features and whose columns to the successive segments. A variety of challenges arose to this formally simple mode of representation, each presenting the possibility of new research problems for investigation and new solutions to old problems. In each case the novel vista provided a certain amount of low hanging fruit for exploitation by a generation of graduate students and others; once that had been gathered, the tendency was to look for other novelties, and the field was quite prepared to provide these.

One of the earliest successes of (what would become) the generative approach to phonology was the elegant analysis of English stress provided by Chomsky et al. (1956), an account which evolved into that of Chomsky & Halle (1968). The apparatus necessary to support this picture, however, presented some problems. Since features in the SPE framework were associated with individual columns (segments) in the representation, this resulted in stress values being associated only with a single vowel, rather than with an entire syllable — and indeed, syllables had no status at all in this theory. Furthermore, the feature [Stress], unlike others, was required to take a range of numeric values, rather than a binary choice between ‘+’ and ‘−’; and a rather unwieldy convention of stress reduction had to be posited such that assignment of [1Stress] anywhere within a form resulted in the demotion of all other values from [nStress] to [n+1Stress].

A resolution to these difficulties was provided by Mark Liberman’s (1975) MIT dissertation. There it was proposed that instead of a single homogeneous matrix of features, a phonological
representation should be regarded as a binary branching tree whose terminal elements were prosodic constituents: syllables (whose phonological role was recognized in another MIT thesis, Kahn 1976). The hierarchical representation of the relations among these, then, could be annotated as a relation of strong to weak, and this organization could be interpreted as relative stress, instead of treating stress as a feature like others.

At one shot, this theoretical move (which came to be known as the theory of Metrical Phonology) elegantly resolved the problematic aspects of stress in the purely featural account. It also, however, opened the door to other innovations based on attributing more structure to phonological representations than that of a single uniform matrix of features. Metrical phonology was an immediate success. Analyses of a variety of languages in these terms were produced, and something of a consensus about the range of possible stress systems in the languages of the world emerged in Bruce Hayes’s (1980) dissertation a few years later.

A variant of metrical representations already foreshadowed in its presentation by Liberman & Prince (1977) was the treatment of rhythmic phenomena in terms of a grid, rather than hierarchical constituent structure. This was developed and extended to other properties by Prince (1983), though work in this framework has generally been quite limited.

In contrast, the recognition of syllables and higher metrical constituents such as feet and prosodic words as structurally significant units was taken up widely and quickly became part of the basic vocabulary of phonological description. The theory of these categories was organized in work such as Nespor & Vogel (1986), and the relation of hierarchical prosodic structure to similar structures in syntax was explored by Selkirk (1984).

The new perspective on stress and other prosodic systems provided by the richer notions of structure in Metrical and Prosodic Phonology encouraged students to look for other areas in which comparable moves would provide better accounts of problematic phenomena. Such a domain was the analysis of tonal systems: like stress, tonal features seemed to be associated with phonological content in ways that were not satisfactorily represented by features of individual segments. To accommodate this, John Goldsmith (1979) proposed a view known as Autosegmental Phonology, on which rather than all being present in unitary columns of the representation, individual features were linked to one another by association lines, such that a single specification of one feature could be linked to one or more specification of others (thus allowing for a tonal value to take a number of segments as its unitary domain) or multiple values of a single feature could be linked to a single specification of some other (thus allowing for the description of contour tones as sequences of levels associated with a single vowel).

While initially motivated by phenomena of tone, the apparatus of Autosegmental Phonology was soon pressed into service for a variety of segmental processes. Assimilation became reassociation, for example, as the classical picture of a matrix of rows and columns was replaced by a generalized notion of association among features, such that traditional segmental structure is simply the limiting case where all associations are one-to-one. As work of this sort developed, the question was raised of which features tend to associate together, and which independently. This in turn led to the proposal that the features themselves are organized into a sort of tree structure, such that for instance a node [Place] dominates a number of features specifying place of articulation, and can associate as a unit (in e.g. nasal assimilation from a following obstruent).
The proposal of such organization produced the program of Feature Geometry as introduced by Clements (1985), Sagey (1990) and others. The research agenda set by this view was to uncover a single uniform organization of features into higher-level categories valid across languages. Despite considerable effort, however (e.g. McCarthy 1988), such an organization did not emerge, and the focus on work in this program declined rapidly.

Each of the theoretical innovations just surveyed resulted in an initial burst of enthusiasm and proliferation of research results. As the work became more standardized, however, phonologists sought out new topics, sometimes leaving problems of the previous round of innovation unresolved — just as the move to richer theories of representation had left behind unresolved problems of the SPE theory, such as the relation of notational conventions to special principles of application, the nature and generality of rule ordering, and others. In some instances, innovations left their traces in the general view, as with the acceptance of much of the apparatus of Metrical, Prosodic and Autosegmental theory in forming a broadly accepted view of phonological structure. By the early 1990s, however, the field — which had become accustomed to a rapid turnover of ideas and research topics — was in need of a new infusion of both, and had begun to develop a sense of stagnation. What happened then represented a more extreme change than anything since the appearance of SPE.

3.3 The Rise of Optimality Theory

During the 1980s, a great deal of artificial intelligence research within the computer science community was focused on the development of the architecture of neural networks (Rumelhart et al. 1986), “Connectionist” systems that were claimed to be able to learn, on the basis of exemplars as training data, complex associations between inputs and outputs without explicit instruction in the nature of the relation involved. One application of this work was in the analysis of natural language, and an influential paper in that framework (Rumelhart & McClelland 1986) claimed to document such a system that acquired a significant segment of the morphology of English (past tense forms of verbs) without direct instruction apart from a training set. The assumptions and adequacy of this model were strongly criticized by Pinker & Prince (1988), and linguists and cognitive scientists generally were not impressed with the promise this approach might hold for their fields.

Among those in the computer science community engaged in the exploration of neural network models, Paul Smolensky had especially broad interests in cognition more generally, and in particular in the nature of the representations that could be attributed to these models and ways in which symbolic processes could be modeled. When he as a prominent Connectionist, and Alan Prince as a prominent critic of that approach, met and began to work together, there was actually a substantial amount of common ground for them to explore.

After several years of collaboration, with occasional presentations to other phonologists (e.g. in a workshop at the 1991 Linguistic Institute at UC Santa Cruz), Prince & Smolensky (1993; later

---

4 For a more detailed account than can be provided here of the rise of Optimality Theory and of the theoretical proposals that preceded and anticipated it, see Burzio 1995; Griffiths 2019; van Oostendorp to appear.
published as Prince & Smolensky 2004) appeared as a photocopied manuscript that was widely disseminated to large numbers of phonologists. This set out a radically new approach to the description of phonological systems, dispensing entirely with language-particular rules functioning in a serial derivation.

The new framework was based centrally on generalizations about surface phonetic forms, represented as constraints taken from a universal set. These were of two sorts: *faithfulness* constraints, which require that the output phonetic form resemble the phonological input in various ways, and *markedness* constraints, requiring the output form to conform to various universal conditions of phonetic naturalness. These are typically in conflict with one another, and thus must be ranked with respect to their importance: a grammar then consists precisely of a ranking of the members of the universal set of constraints. A given constraint is allowed to be violated in the surface form precisely to the extent this is required by other higher-ranking constraints. A derivation consists in taking a phonological input form, allowing a component of the system (“GEN”) to generate a potentially unlimited range of possible corresponding outputs, and then evaluating in parallel each of these against the ranked constraint set (“EVAL”). The candidate form whose constraint violations are least serious (the “optimal” candidate) is selected as the output.

It is not necessary for our purposes here to go into more of the details of how such an Optimality Theoretic (“OT”) grammar works. What is most important is the fact that it provides a radically different account of phonological organization from that of the SPE model, even supplemented with all of the representational innovations discussed above. The absence of language particular rules, or any sort of serial derivational structure, combined with the focus on surface-oriented constraints, made this quite unlike anything that had gone before.

It is true that the potential importance of regularities over surface phonetic forms had been brought into discussion previously. Kisseberth (1970) had noted in the early days of work within the SPE model that such regularities sometimes did not find any expression in such a grammar. Multiple rules might for instance “conspire” to have the effect that stress in a given language never falls on a weak syllable (sometimes moving the stress, sometimes lengthening a vowel, sometimes deleting a syllable, etc.), but nowhere in the grammar is that stated in a unitary way, as for instance by a constraint preventing stressed weak syllables. While such “conspiracies” often seem to constitute quite real aspects of a language’s structure, the theory provided no effective way to incorporate that observation into the description, and so it had to go unstated. OT, in contrast, provided a clear status for such generalizations.

It is also important to observe (as Prince & Smolensky do) the resemblance between OT’s positing a universal inventory of constraints — especially markedness constraints — and the system of natural processes in Stampe’s (1973) Natural Phonology. In both cases it is presumed that the intrinsic nature of the system implementing speech has an important role to play in language, and that the effects of this have to be ranked with respect to one another and to the need to maintain distinct signals for distinct content (a primary role of the faithfulness constraints).

The universal nature of constraints assumed in OT is not a self-evident property of the theory. In particular, as analyses of individual languages have proliferated within this framework, it has
become increasingly clear that the constraints that need to be posited for individual languages can in fact be quite specific. It is not obvious that the assumption of a universal set of constraints is other than a placeholder for a system by which the constraints active in a given language could be learned — a result that would change the cognitive commitments of the theory in important ways.

OT was also not the first phonological descriptive framework to focus on constraints as a formal method. Paradis & LaCharité (1993) survey three such theories as they existed at the time OT first appeared on the scene, but none of the others caught the attention of the field in the way OT did. For whatever reasons, OT came along just at the right time (when the field was eager for something new), and was aggressively promoted by its originators. Within a remarkably short time, most new work in phonology was being produced in this framework.

Initially, on the basis of the illustrative analyses provided by Prince & Smolensky (1993), OT appeared to be primarily useful for the description of prosodic properties, including syllable structure and related effects. Soon, however, research had pushed the techniques of the framework into essentially all areas of phonological structure, and its victory was to all intents and purposes complete. Some important linguists (including, notably, both Chomsky and Halle) were unconvinced, and argued against OT, but others came on board, and most importantly, students rushed to adopt the new approach in formulating dissertation topics. By the end of the 1990s, rule-based serial derivations were only to be seen in the work of a few outliers, a state of affairs that continued well into the new millennium.

Although differing in many fundamental ways from preceding phonological theories, OT can be seen as essentially developing a basic view of the architecture of grammar very similar to that of various forms of “Generative Phonology.” Strongly contrasting with this conception is that of the Laboratory Phonology movement, an approach originating in the 1980s in work of a number of phoneticians. The essential goal of this program was to remake phonology, deriving all of the regularities of sound structure in natural language from phonetic phenomena observable in the laboratory with little or no appeal to higher-level cognitive principles. Although much work in this tradition proceeds with strongly expressed hostility to traditional generative views, there has in fact been very little actual interaction between the two approaches. Given the tenuous connections between this view and others surveyed above, a serious consideration of the Laboratory Phonology literature and related theories such as that of Articulatory Phonology (Browman & Goldstein 1989 and subsequent publications) is beyond the scope of the present chapter.

4 Conclusion

In the developments reviewed above over nearly a century, we can find a considerable amount of progress in ideas that has, overall, led to a richer and more substantial view of the sound systems of natural languages. By no means all of what has been seen as progress, however, can

---

5 For a review of early work in Laboratory Phonology and an assessment of its relation to the phonological tradition, see Dobrovolsky 1994.
be attributed solely to the superiority of new approaches over old ones in terms of their ideas. In some cases, we see that when the problems arising within a theoretical perspective become sufficiently complex, and the data necessary to resolve them too elusive, the result is a search for a new research agenda that would allow students to achieve results in quite another direction. Science does, undeniably, make progress, but at least some of this progress results as much from a need for novelty as from the resolution of old problems.

As I write, such a change may be taking place in phonology once again, though no landmark work as significant as that of Bloomfield (1933), Harris (1951), Chomsky & Halle (1968) or Prince & Smolensky (2004) has appeared to incarnate it. While still without doubt the most widespread and influential theoretical position among phonologists, Optimality Theory too has begun to lose its force. As was the case for earlier ascendant views, the theory has reached a point where the outstanding matters of controversy are somewhat obscure and hard to resolve. In addition, the important phenomenon of opacity — generalizations that are importantly true, but true in a way not susceptible of formulation in terms of surface form — has not received a satisfactory resolution, despite serious effort and attempts to incorporate aspects of previous theories such as serial derivation (McCarthy 2007).

Partially in reaction to such accumulating problems, phonologists (along with some other linguists) have turned away from traditional methodologies to seek solutions in computational analyses, statistical inferences and the data-mining study of increasingly available large corpora of language materials. It is far too early to say whether this approach will succeed in replacing linguists’ conceptions of phonological structure with something quite different. While it may be possible that long unresolved questions of phonological organization will yield to these methods, the temptation to see in this turn yet another search for low-hanging fruit is hard (for the present author) to resist.

Acknowledgements
This chapter has its origins in a talk I intended to present to the LSA at its Annual Meeting in 2014 commemorating 90 years of the Society. [...].
References

Planck, Max. 1949. *Scientific autobiography, and other papers*. New York: Philosophical Library. Translated from German by Frank Gaynor.
de Saussure, Ferdinand. 1916. Cours de linguistique générale. Lausanne: Librairie Payot. Edited by Charles Bally and Albert Sechehaye, with the collaboration of Albert Riedlinger.