Continuum Companion to Phonology

Continuum Companions

Edited by
Nancy C. Kula, Bert Botma and Kuniya Nasukawa
Contents

Notes on Contributors vii
Editors’ Introduction xi

Part I: Research Methodology 3
1 Methodology in Field Phonology Mary Pearce 16
2 Methodology in Phonological Acquisition Tania S. Zamuner and Elizabeth K. Johnson

Part II: Research Issues 33
3 Features Bert Botma, Nancy C. Kula and Kuniya Nasukawa 64
4 Syllables Péter Szigetvári 95
5 Stress Diana Apoussidou 135
6 Derivations Patrik Bye 174
7 Constraint-Based Phonology Christian Uffmann 202
8 Phonetics-Phonology Interface Silke Hamann 225
9 Syntax-Phonology Interface Anthi Revithiadou and Vassilios Spyropoulos
Notes on Contributors

Heidi Altmann is a lecturer of English and General Linguistics at the University of Konstanz, where she has been teaching phonology, second language acquisition and sociolinguistics. Her research interests comprise of second language prosody, phonology-morphology interface, and language and gender.

Diana Apoussidou works as a postdoctoral researcher at Utrecht University. Her main research interest is the learnability of phonological structure and underlying representations, with a focus on prosody and metrical properties of language. She is currently working on a model of acquiring a mental lexicon as part of a 5-year research project on phonotactic constraints in speech segmentation funded by the Netherlands Organisation for Scientific Research (NWO).

Bert Botma is a lecturer of English language and linguistics at the University of Leiden, where he teaches courses in phonology, phonetics, morphology and historical linguistics. His main research interest is in segmental phonology. He is currently conducting research on the phonological and phonetic properties of sonorants, under the auspices of the Netherlands Organization for Scientific Research.

Patrik Bye is a researcher at the University of Tromsø and Center for Advanced Study in Theoretical Linguistics (CASTL) in Northern Norway. Bye's research has focused on syllable structure and quantity in Finnic and Saami, accent in North Germanic, phonological opacity, and the interaction of phonology with morphology, lexicon and syntax. He is currently involved in a project investigating the intonation of the Saami languages.

Taehong Cho is associate professor in the Department of English Language and Literature, and the director of the Phonetics and Psycholinguistics Lab at Hanyang University, Seoul, Korea. He earned his Ph.D. degree in phonetics at the University of California at Los Angeles (UCLA) and worked at the Max Planck Institute for Psycholinguistics. His main research interest is in the
1. Introduction

It is an ambitious undertaking to write about the history of a field in a chapter length article. This chapter thus only provides an overview that concentrates on a small number of central issues and salient steps in the evolution of generative phonology. With respect to the time scale covered in the chapter, the decision to begin with SPE (The Sound Pattern of English) reflects space limitations rather than a judgement on the important developments in the 1950s and 1960s. There is however a growing body of specialized and well-documented literature on the history of (generative) linguistics in general and on phonology in particular that inquires on individual pieces of the overall picture and sometimes covers larger periods.  

The following pages thus offer a selection based on personal choice, rather than historiography. Nonetheless, the ambition of the chapter follows the traditional motivation of historians: to learn from the past in order to understand the present, and to shape the future. The kind of history that is told is thus not neutral or impartial: it is goal-oriented and functional in the sense that it aims
at isolating important issues, key questions, central ideas, circular movements and real scientific progress.

In this respect, the following topics are discussed: overgeneration, the abstractness debate, (auto-segmental) representations and their balance with computation, the nature of computation (rule-based versus constraint-based, serialism versus parallelism) and restrictions on computation (cyclicity). These issues and the theories from which they emerged will be located with respect to the different evolutionary strands, and with respect to each other.

Finally, the chapter is based on published (or unpublished, but written) scholarly work while the sociology of the field, which may well contribute to our understanding of the overall picture, is not considered.

2. Normal Science: SPE

2.1 The Generative Architecture: Modularity and Translation

As indicated by the section title, the pages below follow Kuhn’s (1962) analytic tool for the evolution of scientific theories first generation generative phonology was established during the 1960s and incarnated as SPE in 1968. It was then practiced for a decade or so with the status of what Kuhn calls normal science, that is, following the canonical rules established without calling them into question even though irritating or frankly conflicting evidence was available and well known. In the case of SPE, its Achilles’ heel was made explicit by Chomsky and Halle themselves in the famous ninth chapter of SPE, to which we will come back below.

Since normal science was practiced and largely spread before the actual publication of SPE (individual chapters were circulated in the 1960s), both the revolutionary and the revisionist movement that it aroused could be launched in 1968, or shortly thereafter. The revolutionary attack that set out to overthrow the system as such was led by the Natural Phonologies (which originate in David Stampe’s 1972 Ph.D.), while the revisionist perspective was opened by Paul Kiparsky’s (1968 [1973]) manuscript How Abstract is Phonology?

The communist and socialist enterprises are the subject matter of Section 3. Before we turn to them, let us look at some key properties of SPE, and namely at those which have set the generative standards and are still in place today against all odds and subsequent evolutions.

SPE makes explicit the architecture of generative grammar that was hailed in Aspects (Chomsky, 1965). This architecture is a direct expression of the frame that the then emerging Cognitive Science defines for cognitive activity (the cognitive revolution of the 1950s, see Gardner, 1985 for an overview): language as such is a module (in the sense of Fodor, 1983, that is, a specialized and autonomous unit in the cognitive system that carries out computation), but its internal structure is modular as well. The following quote documents that generative phonology is conceived of as a cognitive module, that is, a computational unit that modifies an input according to a pre-defined set of instructions that applies 'mechanically', that is, without taking into account any external factors (computation is autistic or, in CogSci terminology, encapsulated):

The rules of the grammar operate in a mechanical fashion; one may think of them as instructions that might be given to a mindless robot, incapable of exercising any judgment or imagination in their application. Any ambiguity or inexplicitness in the statement of rules must in principle be eliminated, since the receiver of the instructions is assumed to be incapable of using intelligence to fill in gaps or to correct errors. (Chomsky and Halle, 1968: 60)

The modular architecture of the generative grammar is embodied in the so-called inverted T model that is depicted in Figure 16.1. While the inverted T was first introduced in Aspects (Chomsky, 1965: 15ff.), the modalities of communication between morpho-syntax and phonology were defined in a book on phonology, SPE, rather than in work on syntax. This phonological bias also runs through further evolution: theories of the interface between morpho-syntax and phonology are phonological theories made by phonologists on the grounds of phonological data (Lexical Phonology, Prosodic Phonology, co-phonologies, indexed constraints etc, with Distributed Morphology being a semi-exception).

![Figure 16.1 The inverted T model: All concatenation before all interpretation](image)

In the perspective of the inverted T, (morpho-)syntax is the central concatenative system which has the privilege of concatenation: its output is sent to two interpretational modules, phonology (PF) and semantics (LF), which assign a form and a meaning to the morpho-syntactic structure. Morpho-syntax and the two interpretational modules are thus procedurally ordered so that words and sentences are pieced together before being shipped off to interpretation at PF and LF. That is, the function of phonology in the (production-oriented) generative perspective is to translate morpho-syntactic structure into a code that may be used by the sensory-motor system. In perception, the movement is reversed: phonology helps to interpret the phonetic signal, that is, to identify morphemes.
‘Phon’ in the word phonology is thus an (understandable) misunderstanding: natural language is modality-independent. That is, linguistic structure may run through any channel as long as humans are able to produce and perceive contrast. Two modalities are known to be used by humans in order to ship messages in natural language: vocal and signed. The former is the default, but the latter is used if the former is not available (there are documented cases where deaf children develop a sign language ex nihilo, that is, in absence of any linguistic stimulus, for example, Senghas et al., 2004). A more accurate definition of phonology and its function is thus the translation back and forth between a physical signal and morpho-syntactic structure: phonology is a translational device.

SPE also adds an extra proviso (which does not follow from the inverted T structure): all concatenation must be done before any interpretation. While the inverted T model remains predominant up to the present day, the full completion of concatenative activity before any interpretation can begin will be a bone of contention in further development. The competing view is so-called interactionism according to which concatenation and interpretation are interspersed. This is the key idea of Lexical Phonology (Kiparsky, 1982a and b; Hargus and Kaisse, 1993 provide an overview), and today also the backbone of minimalist syntax (derivation by phase, Uriagereka, 1999; Chomsky, 2000: passim).

A direct consequence of the modular architecture is the necessity for translation: every module works on its own proprietary vocabulary (this is how modules are defined in CogSci: domain-specificity, for example, Hirschfeld and Gelman, 1994). Morpho-syntax, for example, reacts on things like number, gender, animacy and so forth, which is what the morpho-syntactic part of lexical entries is made of. By contrast, labiality, occlusion and the like are entirely transparent to morpho-syntax. On the other hand, phonology works on these, but is unable to parse number, gender and the like.

Therefore, SPE and all subsequent generative theories establish a translational mechanism that transforms morpho-syntactic structure into objects that can be read by the phonological computation. In SPE, morpho-syntactic structure is translated into hash-marks # by a mapping algorithm (all major categories as well as projections thereof are preceded and followed by #), later on into prosodic constituency (Prosodic Phonology, Selkirk, 1981b: passim). Even though on other grounds and following a different reasoning, structuralist practice was also modular in the sense that so-called Level Independence enforced translation between phonology and morpho-phonology (see Scheer, forthcoming a). The only thing that changes over time is thus the output of the transnational process, which is defined in terms of the phonological units that were used in the respective theories: juncture phanemes in structuralist times, hash marks which were defined as a special kind of segment in SPE (#s were [-segment] segments, while ‘real’ segments were [+segment]), and finally autosegmental constituency in the early 1980s when all areas of phonology were autosegmentalized.

On the other hand, it is also true that SPE violates the translation requirement since apart from hash marks, it also makes brackets and labels available in the phonology, that is cyclic structure (on which more shortly) and genuine morpho-syntactic vocabulary such as verb, noun, adjective. For example, a word such as thead
cality appears in phonology as [[[thea[t]]], i c + all, i + ty], and phonological computation can freely refer to the information that this or that portion of the string is a noun, a verb or an adjective. Hence, typical SPE-style rules such as ‘k palatalizes before e, but only if this e is the dative singular’. We will see below that reference to so-called morphological diacritics is one of the reasons why SPE came under fire in the 1970s. The problem was only (partially) solved in the mid-1980s when Prosodic Phonology established the so-called Indirect Reference Principle (e.g. Nespor and Vogel, 1986) which prohibits reference to unsegmented morpho-syntactic information in phonological processes.

2.2 Cyclic Derivation

Another key property of generative thinking, cyclic derivation, is directly related to the modular architecture. The idea that phonological (and semantic) interpretation follows morpho-syntactic structure, that is, proceeds from inside-out (from the most to the least embedded constituent) bears the stamp of a generative copyright. That is, a string with the morpho-syntactic structure [(A B) C] is computed in such a way that first A is interpreted alone: the result is then assessed together with B (i.e. AB), and finally the product of this computation is interpreted together with C, that is, ABC. It is important to note that the computational system, that is, the set of instructions that cause modifications, is invariable: whatever the string, it is interpreted by the same phonology. Also, there were no requirements regarding so-called strict cyclicity (a notion that we return to below) in SPE, and Chomsky and Halle (1968) are explicit on the fact that cyclic derivation concerns morphological as much as syntactic structure: it applies ‘to all surface structure whether internal or external to the word’ (Chomsky and Halle, 1968: 27).

Cyclic derivation represents a genuinely new idea that was absent from linguistic thinking in Antiquity and the Middle Ages, and from the neogrammarian and structuralist paradigms in modern times. It first appeared in Chomsky et al. (1956: 75), but was codified only in SPE, where it is called the Transformational Cycle. In the 1970s, it was also known as the Phonological Cycle (Mascaró, 1976), and today cyclic derivation (or cyclic spell-out) is known as derivation by phase, which is the spine of current minimalist syntax (Chomsky, 2000: passim).
2.3 Anderson’s Prism: Structure Versus Process

Finally, let us consider something that SPE does not have, and which will become central in the further evolution of the field: representations. The major insight of structuralism was the existence of the system and hence of systemic pressure; a sound is not a linguistic object per se—it acquires this quality only by virtue of its relationship with other sounds in a system. That is, the same physical object may have different phonological behaviour according to the system in which it is couched. This central piece of structuralist thinking was eclipsed in SPE (without argument).

Granting no role at all to structuralist structure, SPE certainly ranges close to the computational end of a scale whose extremes are computation and structure. It is true that SPE is not only made of computation, though: computation necessarily computes something—objects—which are thus different from computation. This is the dualistic philosophy of Cognitive Science: modules are input-output devices that take objects as input, transform them and return their altered versions.

What are the objects that are manipulated by computation in SPE? Feature-matrices that are made of binary features. While serialism (i.e., extrinsically ordered rules) is closely associated with generative phonology (albeit not its invention, see Bloomfield, 1939 and the discussion aroused by the (non-)reception of this article in generative quarters, as reported by Goldsmith, 2008), Chomsky and Halle have never claimed the paternity of binary features: these were introduced by Roman Jakobson (Jakobson, 1939a; Jakobson et al., 1952) in a structuralist environment. SPE has merely replaced their original acoustic definition by articulatory values. In this perspective, thus, extrinsically ordered rules change the values of articulatorily defined binary features, which are the only structure that SPE recognizes. Or rather, to be precise, SPE recognizes yet another type of item beyond binary features: segments, which are an aggregation of features. A segment is defined as a feature matrix whereby all segments are made of the same features and the same number of features. Segmental variation, then, is exclusively a matter of varying feature values.

The non-computational part of SPE, its structure, is thus made of features and feature-matrices. An interesting question that is not usually asked in the (contemporary or later) literature is whether it makes sense to talk of this kind of structure in terms of representations. As we will see below, the advent of autosegmental representations has also introduced the notion of well-formedness, which was central for phonological thinking in the 1980s. It is therefore reasonable, in any case useful, to establish a strong bond between representations and well-formedness: only representations can be well- or ill-formed, and something can only be a representation if it is well-formed. This, however, is clearly not the case with features and feature-matrices: there are some combinations that either cannot or do not occur, but never because of an intrinsic impossibility that is defined by grammar.

For example, a vowel cannot be [−high] and [−low] at the same time—but this is for physiological reasons. And some language may well not possess this or that combination of feature values of the universal set of features—actually most of the astronomical number of logically possible combinations do not occur. But this again is for reasons that have got nothing to do with grammar: inventory selection and overgeneration will drive SPE into trouble as we will see below. Finally, core properties of features (such as their binary nature: they have either a positive or a negative value, no third choice is available) and of feature matrices (e.g., no feature may occur twice in a given matrix) are set in stone before the derivation begins: no lexical item can contravene, and no computational action can create monster features without value or monster matrices. By contrast, autosegmental representations may become ill-formed in the course of a derivation. They therefore introduce a new quality into phonological thinking, which in retrospect turns out to be a milestone in the history of phonology.

Let us now place this discussion in a larger context. The balance between structure and process is the prism that Stephen Anderson (Anderson, 1985) uses to understand the history of phonology in the twentieth century. Anderson has detected a regular see-saw movement between theories that stand far on one side of the spectrum, and others that approach the opposite extreme. Also, both properties are in a reverse proportional relationship: when one goes up, the other goes down.

One idea is that the balance between representation and computation is a valid instrument for understanding what the constant changes in terminology, theories, concepts and schools is all about. Writing at the representational peak of the 1980s, Anderson extrapolates that phonology stands at the dawn of a new computational, hence anti-representational era. Here is the last sentence of his book:

If current attention to the possibilities of novel sorts of representations leads to a climate in which the importance of explicit formulation of rule-governed regularities disappears from view, the depth of our knowledge of phonology will in all likelihood be poorer for it. We hope that this book has demonstrated that neither a theory of rules nor a theory of representations constitutes a theory of phonology by itself. (Anderson, 1985: 350)

Little did he know how right he was, that is, how far on the extreme computational end OT would take phonology a couple of years later. Given its predictive success, the structure versus process prism will be used below as a measure of
the evolution of the field (van der Hulst, 2004 also looks at the history of phonology through this lens). Anticipating the facts to be reported, we may note that the generative micro-cosmos (as compared to Anderson's larger span) has achieved a complete loop from an almost exclusively computational theory over the heavily representation-oriented autosegmental 1980s back to the exclusively computational environment of OT (where literally no structure is left since even structure is supposed to ‘emerge’, that is, to be the result of computation). A reasonable comment on this situation is certainly that reinventing the wheel every now and then, or going in circles, does not witness the maturity of a field.

3. Revolution and Revision

3.1 What is Wrong with SPE: Creating a Revolutionary Situation

3.1.1 Overgeneration
The fundamental problem of SPE is that it can describe all phonological processes that exist, as well as those that do not. For a theory that sets out to build a system (a grammar) that is able to generate all structures that are attested (or well formed), and none that are not attested (or ill formed), this situation amounts to a declaration of bankruptcy. Chomsky and Halle (1968) had lucidly spotted the problem in the ninth chapter of SPE, but the remedy that they suggested, a theory of markedness, never really emerged.

Overgeneration concerns both structure and computation, and is regularly pointed out in the post-SPE literature (e.g. Goyvaerts, 1981; Kaye, 1989: 58ff.). It was already mentioned that given free combinability of positive and negative values, the fixed (and universal) set of binary features (say, 20) creates a bewildering number of logically possible segments, of which only a vanishing fraction exists (either in any given language or cross-linguistically, even when physiological impossibilities are excluded).

On the computational side, the universal rule format \( A \to B / K \) does not impose any restriction on what kind of object can instantiate \( A \), \( B \) and \( K \). Anything may be turned into any other thing (no restriction on the relationship between \( A \) and \( B \): \( n \to n / \_ k \) is just as plausible as \( n \to f / \_ k \) or \( r \to f / \_ k \). Also, anything can provoke any change (no restriction on the relationship between \( A \to B \) and \( K \)): \( n \to n / \_ k \) is just as good as \( n \to n / \_ p \).

It is obvious for most phonologists' that natural language does not work like that. Certain processes are recurrent (e.g. dental nasals become velar before a velar obstruent, dental nasals become labial before a labial obstruent), others are rare, and some do not occur (e.g. a [s] that is turned into [b], see Ewen and van der Hulst, 2001: 3ff. on this). It is therefore wrong to leave the variables \( A \) and \( B \) without any restriction.

It is also self-evident for most phonologists that any context may not provoke any effect: there are precise causalities, and it is not reasonable to assume that a given object may trigger any process and its reverse. Hence, the relationship between \( K \) and \( A \) is not free but obeys precise causal patterns (e.g. dental nasals may become velar before a velar, but not before a labial stop).

Finally, wild overgeneration is also promoted by so-called morpho-phonology: as has already been mentioned phonology could freely refer to morpho-syntactic labels (e.g. velars palatalize before front vowels, but only if these represent the dative singular') since SPE granted free access to labelled brackets that were inherited from morpho-syntax.

3.1.2 Related Items Must Have a Common Underlying Form
Another area where overgeneration produced strange bounds is the relationship between etymologically, paradigmatically or semantically related items. In SPE and early generative phonology, it was admitted without discussion that items for which the linguist can determine a relatedness of this kind must have a common underlying form from which the surface variation is derived by phonological rules. Otherwise, it was argued, a generalization is missed.

For the sake of illustration, it is useful to choose the wildest case on record: the work by Theodore Lightner is certainly not representative of the average post-SPE phonologist, but the fact that the theory allowed Lightner to posit such derivations without hesitation is an indication that something was wrong. A (phonological) theory must somehow define what a possible derivational relationship is, and what is not.

Lightner (1979: 18f., 1981, 1985) holds that the following pairs are derived from a common underlying form: eye and ocular, thunder and detonation, dental and tooth, rebel and bellicose, cardiac and heart, three and third, gynaecology and queen, sweet and hedonism and so on. Since the alternations h-k (heart - cardiac), d-0 (third - fourth) and s-h (sweet - hedonism) suppose Grimm's Law, Verner's Law and the Ancient Greek s > h shift, Lightner concludes that these processes are performed by the grammar of present-day English speakers.

What these examples show is that the more distant (etymologically, paradigmatically, semantically) two items that are supposed to be related by a phonological derivation are, the more demanding they are for underlying representations and for phonological computation. That is, the underlying representation that needs to be set up for sweet and hedonism is highly abstract, that is, fairly distant from the surface form (of at least one of the items). In addition, the phonological rules needed to carry out the transformation are numerous and make reference to diachronic processes that occurred several hundred, sometimes several thousand years ago. Therefore, a typical feature of SPE-style analyses is to propose synchronically underlying forms that are close to or identical with diachronically distant stages of the language (in the case of
SPE Old English, sometimes Common Germanic), and to set up rules that mimic diachronic evolution. It is not quite plausible, though, that present-day natives have acquired Old English lexical forms and the diachronic rules that transformed them over centuries.

The tacit rule that SPE and post-SPE phonology obeyed was thus as follows: a bulky (phonological) computational component associated to a small number of lexical entries make a better world than the reverse situation, or anything in between for that matter. The origin, development and (non-)justification of this view is discussed, for example, in Foley (1977), Kenstowicz and Kisseberth (1977, 1979); Anderson (1985: 331f.) provides an informed overview.

The three alternatives to a phonological derivation that is based on a common underlying form are known and have always been practiced: related items such as electr[i]k and electr[i]s-ity can either represent two distinct lexical entries (electricity is stored as a single lexical entry, which means that its pronunciation does not involve any morphological concatenative or phonological activity). There may also be a case of allomorphy (there are two stems recorded in the lexicon, electr[i]k and electr[i]s, which are selected by a morphological, rather than a phonological context: the morpheme -ity selects the latter). Finally, the third possibility is suppletion: good and better are certainly two distinct lexical entries; as is the case for allomorphy, they are in complementary distribution, but the selection is made according to a grammatical, rather than a morphological context.

On this backdrop, then, the two issues discussed — overgeneration and derivational relatedness — defined the research agenda in the 1970s. To cut a long story short, everybody tried to reduce the expressive power of the grammar — its abstractness. By what means exactly this was to be done was the dominant question. In any event, all phonological theories that emerged in the early and mid-1980s have to some extent learned the lesson that many alternations which early generativists believed were produced by online phonological computation do not represent a synchronically active process. The set of alternations that phonological theory is called to account for, then, is far smaller than what SPE thought it was: phonology has to shrink in order to be viable. This insight had a revolutionary and a revisionist offspring.

3.2 Revolution: Natural (Generative) Phonology

The major challenger of SPE in the 1970s were the two Natural Phonologies, one generative, the other not. The initial spark of the latter is David Stampe’s Ph.D (Stampe, 1972b), which evolved into Natural Phonology (NP) (Donegan and Stampe, 1978, 1979; Dressler, 1974, 1984; Hurch and Rhodes, 1996; Dżubalska-Kołaczky, 2002). The former, Natural Generative Phonology (NGP), was founded by Theo Vennemann (1971, 1976 [1971]) and developed namely with his student Joan Hooper (today Bybee: Vennemann 1972a and b, 1974a and b; Hooper, 1975, 1976).

Both Natural Phonologies share a basic set of assumptions and principles, but as indicated by their name one is more revolutionary than the other: while NGP accepts the basic generative architecture, NP has left generative grounds, namely on the count of functionalism, which is endorsed.

The Natural Phonologies promote a radical means of reducing the expressive power of the grammar: they cut down the set of alternations that represent phonological activity by some 80% or 90% in comparison to SPE (my estimate). In order to do that, a structuralist notion is revived that was abandoned by SPE: the morpho-phonological level. That is, N(G)P does not deny that an alternation such as electr[i]k - electr[i]s-ity is due to online computation – it represents morpho-phonological, rather than phonological computation. In other words, we face allomorphy.

In NGP, the two criteria that divide alternations into one or the other type are the True Generalization Condition (Hooper, 1976: 13ff.) and the No-Ordering Condition (Hooper, 1976: 18ss.). According to the former, only phonetically accessible information can be used in the formulation of phonological rules, while the latter prohibits rule ordering. On this backdrop, alternations that do not suffer any exception in the entire language and exclusively appeal to phonetically retrievable information are called natural and granted phonological status (P-rules). By contrast, alternations that are not surface-true and/or whose statement requires non-phonetic, that is, morpho-syntactic information, are rejected into the pool of morpho-phonemic rules (MP-rules). This contrast is the NGP-adaptation of Stampe’s (1972b) original distinction between processes (P-rules) and rules (MP-rules).

On Stampe’s view, the former are natural, innate, productive, unsuppressible and effect minimal structural changes (hence they apply in loanword adaptation and interfere when non-native languages are spoken), while the latter are conventional, learned, dispensable (they do not interfere in loanwords and when foreign languages are spoken) and may be responsible for structural changes that involve more phonetic distance. In this environment, Stampe’s perspective on language acquisition is that it involves the suppression of those innate processes that are not effective in the language at hand. As was mentioned, the split between processes/P-rules and rules/MP-rules reinstallation morpho-phonological and hence Level Independence in its right.

This restrictive definition of phonology attacks two of the three sources of overgeneration in SPE: reference to morpho-syntactic information (which is prohibited for P-rules) and rule ordering (which is also ruled out in NGP). The third kind of overgeneration is the arbitrary relationship between the context of a rule and the change that it causes. Here it is hoped that the
restriction to phonetically accessible information will automatically introduce sound causality.

This setup was accompanied (or even partly provoked) by a diachronic reasoning regarding the life cycle of alternations. That is, what we see today in alternations such as *electri[k]* - *electri[s]*-ity are vestiges of a once-phonologically controlled alternation that has aged (e.g. Vennemann, 1972a). Alternations are born as phonetic regularities before moving into grammar where they are first phonological but at some point start to be riddled with morphological conditions, followed by lexical factors, and finally are levelled out or eliminated from the language by some other means. Therefore, asking the question how much of what we see is controlled by phonology is if not identical, at least consonant with the question how much diachronics is in synchronic sound patterns.

In sum, then, NP is a radical answer to the overgeneration problem, and the label revolutionary is certainly adequate, be it only in recognition of the fact that extrinsic rule ordering, a prime candidate for the generative identity, is outlawed in NGP, and that NP is based on functional considerations. NP is thus anti-abstract in the sense that it restricts derivational depth (i.e. the distance between underlying and surface forms) to a minimum. As a consequence, phonology looks quite phonetic in its mirror: only phonetically defined items can be taken into account by phonological rules (P-rules). It needs to be noted that the three ingredients mentioned – prohibition of morpho-syntactic information in phonology, anti-abstractness and the reduction of phonology to phonetic and surface-true statements – are logically independent. Gussmann (2007), for example, endorses the former, but rejects the latter two.

### 3.3 Revision: Kiparsky

#### 3.3.1 How Abstract is Phonology?

The revisionist research programme was laid out in Kiparsky (1968 [1973]) and realized through the 1970s mostly by Paul Kiparsky himself. It attacks only one source of overgeneration, the one that is mentioned in the title of Kiparsky’s seminal article: abstractness, that is, the ‘derivational distance’ between underlying and surface forms. The two other sources of overgeneration, reference to morpho-syntactic information and the arbitrary relationship between the context of a rule and the change that is caused, are left untouched.

Abstractness is attacked along two lines: Kiparsky restricts possible underlying forms, and he marshals the complexity of the derivation by two means: the restriction of computation to so-called derived environments and the requirement for rules to use material that is freshly introduced on the last cycle. The two restrictions on computation are united under the roof of Kiparsky’s version of the Strict Cycle Condition (SCC).

The Alternation Condition (Kiparsky, 1968 [1973], 1973b) defines what a possible underlying representation is: in case a morpheme shows no alternation on the surface, it must not be any different in its underlying form. This results in a ban against so-called absolute neutralization, that is, items in underlying forms that never appear on the surface (on which more shortly). On the other hand, Kiparsky (1973b) restricts the application of rules in such a way that a certain rule class may only target derived environments. This was called the Revised Alternation Condition: obligatory neutralization rules apply only in derived environments. An environment is derived if it is created by the concatenation of two morphemes, or by the application of a phonological rule. This embodies the idea that phonological processes do not apply to monomorphic strings, that is, when the trigger and the target belong to the same morpheme.

Trisyllabic Shortening (or Laxing) may illustrate the ban on absolute neutralization and the quest for derived environments. The process produces alternations whereby a long vowel or diphthong of bisyllabic items appears as a short vowel when a suffix is added: *dio[najne] - dio[nunity], apo[e]lque - op[æ]lcity* etc. Monomorphic items such as *najlightingale, nightingale* and *najtory* (or *tory*), however, systematically resist the process, even though they satisfy the trisyllabic condition.

SPE reacts in a non-systematic way that misses the obvious morphological generalization and sets up abstract underlying forms: the application of the rule is eluded simply by destroying either the target or the triggering context of each individual lexical item. Instead of *aj/nightingale* is said to have an underlying /i/, that is, *nixtVngael* (Chomsky and Halle, 1968: 234); and instead of /i/, the last segment of *tory* is made a glide, that is, /ivori/ (Chomsky and Halle, 1968:181). Independent rules that are ordered after Trisyllabic Shortening then take /ix/ and change it to /aj/ (via /i/), and vocalize the final glide of *tory*.

The Revised Alternation Condition kills two birds with one stone: (i) it dispenses with the absolute neutralization of /i/ in /nixtVngael/ and /i/ in /ivori/ (which never appear on the surface) because it (ii) offers a different reason for the non-application of the rule, that is, the request for the triggering environment to be morphologically complex.

Another case where the underlying form of invariant morphemes could be distinct from their surface form are so-called free rides. In our example, these concern the converse surface situation, that is, cases where the third but last vowel of a monomorphic item is short. Hence Kiparsky (1982b: 148) points out that on the SPE analysis, the underlying forms of *[æ][il]bi, c[a][er]ma* or *Plæi[mel]a* cannot be determined: the third but last vowel could either be faithfully short and hence appear as such on the surface, or it could be underlyingly long, that is, *[æ][il]bi, *[æ][emer], *[æ][i][mel]a*; in this case, the /aj/ will be turned into [æ] by Trisyllabic Shortening – a free ride, that is, one without consequences.
The ban on absolute neutralization also does away with this unwarranted indeterminacy.

Finally, the restriction of rule application to derived environments allows phonology to do away with some of the incriminated reference to morphosyntactic information. For example, vowel shortening applies to *mean* [mi:n] – *meant* [mɛnt] but not to *paint* [peint], *pint* [pɛnt], *moot* [mɔːt] because the latter are morphologically simplex, and their long vowel is specified as such in the lexicon (rather than derived by rule): /meɪnt/ versus /peɪnt/. Shortening, then, applies only to derived environments (Kiparsky, 1985: 87).

On the other hand, Kiparsky’s attempt to define a reasonable line of division between computation and the lexicon also means that certain pieces of morphosyntactically conditioned phonological processes need to be defended against the concretist ambition. By the late 1970s, the abstract SPE-mainstream was opposed by so-called concrete analyses that followed the principle ‘the more concrete an analysis (i.e. the less distance between underlying and surface forms) the better it is’ (e.g. Leben and Robinson, 1977; Tranel, 1981).

This line of thought struggled with the fact that all attempts at devising a formal measure of different degrees of abstractness failed: a so-called evaluation measure (or evaluation metric) could not be set up (Kiparsky, 1974; Campbell, 1981; Goyvaerts, 1981). In this context, abstract analyses with no other limitation than the learnability of rules (cf. Skousen, 1981) persisted (e.g. Dinnissen, 1980; Gussmann, 1980; Dresher, 1981). Kenstowicz and Kisseberth (1977: 1–62) offer extensive discussion of the issue. They argue in Kenstowicz and Kisseberth (1979: 204ff.) that in some cases, there is no alternative to the abstract option.

This is also the position of Kiparsky, who favours grammar-internal principles such as the Revised Alternation Condition and the SCC (to be discussed next) and is not prepared to do anything just in order to pay tribute to concreteness. Lexical Phonology, the theory that emerged from the abstractness debate in the early 1980s, may be viewed as an attempt at maintaining as much morphophonology as possible in the computational device of phonology while cutting away the wildest outgrowths of unrestricted SPI (see Scheer, forthcoming a).

Finally, Kiparsky merged the condition on derived environments with Chomsky’s (1973) Strict Cycle Condition that was originally devised for syntax and later on applied to phonology (Kean, 1974; Mascaro, 1976). Chomsky’s (1973: 243) original formulation is as follows: ‘[n]o rule can apply to a domain dominated by a cyclic node A in such a way as to affect solely a proper subdomain of A dominated by a node B which is also a cyclic node.’ The effect is the blocking of rules whose structural description is met by a string which is made exclusively of material that belongs to previous cycles. That is, given \([AB, Cl]\), a rule that is triggered by AB can apply at cycle i, but not at cycle j. Or, in other words, rules must use material that is introduced on the latest cycle, a restriction that prohibits multiple application of a rule.

Chomsky’s (and Kean’s and Mascaro’s) condition on the applicability of rules is entirely irrelevant for derived environment effects: it will not prevent rules from applying to monomorphic strings since these have necessarily been introduced on the latest (the only) cycle. Thus, Trisyllabic Shortening *[æ]n-ir-a* will happily apply to *n[æ]lingale* and *[æ]bory* under Chomsky’s SCC. Nonetheless, Kiparsky (1982a and b) introduces his version of the SCC (1) as if it were just a restatement of Mascaro’s:14

1) **Strict Cycle Condition (SCC):**
   a. Cyclic rules apply only to derived representations
   b. Def.: A representation \(\varphi\) is derived w.r.t. rule R in cycle j iff \(\varphi\) meets the structural analysis of R by virtue of a combination of morphemes introduced in cycle j or the application of a phonological rule in cycle j.

Kiparsky (1982b: 153f.)

Kiparsky’s attempt to kill two birds (‘use new material!’ and derived environment effects) with one stone (his scrambled SCC) was considered an important achievement in the 1980s, but then turned out to lead to a dead end: the derived environment condition being riddled with counterexamples, Kiparsky (1993) himself declares the bankruptcy of his version of the SCC 10 years later.

This does not mean, however, that the original Chomskyan baby in Kiparsky’s blended SCC ought to be thrown out with the bathwater. The idea that ‘old’ strings which have already experienced interpretation are unavailable for further computation appears in the guise of the Free Element Condition regarding stress assignment (Prince 1985, Poeser 1986, Steriade 1988) and structure preservation regarding syllabification (e.g. Steriade 1982:84ff, Harris 1993). Kaye (1992, 1995) makes it a general condition on computation, and today it is the heart of syntactic Phase Theory, where it is known as the Phase Impenetrability Condition (PIC) (Chomsky, 2000: passim).15

3.3.2 The Evaluation Measure and Common Underlying Forms

Following the footsteps of Kiparsky’s revisionist programme which was only marginally concerned with the question of common underlying forms, a substantial body of literature tried to outlaw Lightner-type excesses without however jumping to the NGP conclusion that phonologically driven alternations must be surface-true and may not make reference to any morpho-syntactic information.

Objective and measurable criteria were sought that could be applied to any given alternation in order to decide whether it is the result of phonological computation on the basis of a common underlying form or not. Candidate criteria that were discussed tried to measure what is natural, simple, elegant, phoetically
plausible, psychologically real or typologically invariant (e.g., Hellberg, 1978; Koutsoudas, 1980; Dinnse, 1980), but this was quite inconclusive.

To date, the question remains open whether an item (such as *electric-ity*) that looks morphologically complex is really considered as such by the grammatical system. All modern theories have somehow swung into a midfield position: it is not good or bad per se to have a big or a small lexicon, or to do little or a lot of phonological computation. Arguments must be made on a case-by-case basis; while some alternating items represent two independent lexical entries for sure, the online computation of others is beyond doubt; the swampy midfield, however, is large enough for still much debate to come. Carvalho (2002a: 134ff.) provides extensive discussion of this question (and Carvalho, 2004, especially considers the role of analogy as a serious player).

The discussion regarding common underlying forms also runs under the header of (anti-)lexicalism (treating *electric-ity* as a single lexical entry is the lexicalist position). The parallel between the oscillating evolution in phonology and syntax in this area is quite remarkable: after a decidedly lexicalist period in the second half of the 1970s and all through the 1980s (triggered in syntax by Chomsky’s 1970 remarks on nominalization), the minimalist environment in syntax, and OT in phonology, today have moved back to the anti-lexicalist heydays of the 1960s (e.g., Newmeyer, 2005; Williams, 2007).

4. (Unintended) Counterrevolution: Autosegmental Representations

4.1 A New Player, Ill-Formedness, Opens New Horizons and Buries the Revolution

In the second half of the 1970s, the SPE-mainstream was drifting in unfriendly waters without real research programme and without any perspective of a significant evolution. At the same time, the revolution initiated by Natural (Generative) Phonology gained ground. Kiparsky’s revisionist work offered some relief, but could not produce more than some kind of SPE-Light. It is then that two movements appeared out of the blue, that is, without being prepared, solicited or envisioned, which acted as a fountain of youth that made generative phonology take a new start on the grounds of renewed premises (see also van der Hulst, 2004 on this episode).

One was Lexical Phonology whose initial spark was Siegel’s (1974) discovery of affix ordering, that is, a non-phonological phenomenon, whose interleaving with associated phonological phenomena allowed for the establishment of a new architecture of grammar (i.e. a new positioning of phonology with respect to morphology and syntax), and for the unification with Kiparsky’s revisionist programme.18

The other, to be discussed in more detail below were autosegmental representations (also called non-linear, multi-linear). These brought an entirely new dimension into phonological theory and certainly deserve to be called, also retrospectively, the most significant scientific advancement that generative phonology has made since its inception. A good indication of this status is the fact that autosegmental representations are the least common denominator of all phonological theories, present and past (including OT, at least in word, on which there is more discussion below).

Autosegmental representations are groundbreaking because they introduce a new player that was absent in neogrammarian as much as in structuralist thinking: ill-formedness. It was mentioned in Section 2.3 that SPE has structure (the material on which computation works, that is, features and feature-matrices), but no representations: a representation is something that can be ill-formed for *grammar-internal* reasons. That is, simple concatenation of morphemes or phonological computation can cause an autosegmental representation to be ill-formed – while nothing can be wrong with features and feature-matrices when they are pieced together or computed.

The potential of ill-formedness as an overgeneration-killer that restricts the expressive power of grammar was rapidly identified, and phonologists put much hope into this new perspective after the straining experience of the abstractness debate. Also, the overgeneration-avoidance virtue of autosegmental representations is *intrinsic* since their mere existence, associated with the statement of the (universal or language-specific) conditions of well-formedness, dramatically cuts down possible results. Note that this does not mean that representations are a filter only on outputs: they restrict possible phonological expressions wherever they occur in the lexicon, in intermediate or in surface forms. A violation of well-formedness may either lead to the crash of the derivation, or cause phonology to act in order to repair the offending property (Goldsmith’s 1976a: 27 initial tone-based conception only built on repair strategies; this line of thought is embodied, for example, in Paradis, 1988 and subsequent work; van der Hulst, 2004 provides an overview of the constraint-and-repair approach).

Representations are certainly a more powerful overgeneration-inhibitor than Kiparsky’s revisionist programme, where a phonological computation could crash for computation-internal reasons. It could now also crash because it either works on or produces an ill-formed structure. This gives new impetus to Stephen Anderson’s perspective on the equilibrium between structure and process: autosegmental representations take the generative cursor away from the computational extreme of SPE into some midfield position.

The notion of ill-formedness is of course not the only virtue of autosegmental representations (more discussion on this below), and it goes without saying that autosegmental representations had not solved the overgeneration problem
as such (e.g. they had no effect on the ‘static’ overgeneration inherent in feature-matrices, see Section 3.1.1). They had, however, brought home the promise of overgeneration-inhibition, at least partly (Lowenstamm, 2006 provides a good overview of the credits).

The new autosegmental possibilities opened up new horizons, and after an initial period of timid proliferation in the late 1970s the entire field was turned upside down when everybody participated in the autostructuralization of all areas of phonology. This emulation rapidly and completely eclipsed the revolutionary vigour of NGP, which all of a sudden appeared toothless and simply petered out (see Laks, 2006 on this decline). By contrast, natural non-generative phonology continued its development outside of the generative paradigm.

4.2 Genesis and Properties of Autosegmental Representations

It is reasonable to daresay that Kiparsky’s revisionist programme associated to Lexical Phonology would not have been able to capture the victory over the revolution alone. It is only autosegmental representations that could drive revolutionary concerns out of the field. Interestingly, this was completely unintended: autosegmental representations came into being as a problem-solving mechanism whose counterrevolutionary virtue and impact on overgeneration was only discovered as things unfolded.

Autosegmental representations quite independently emerged in the analysis of a number of different empirical problems: tone (Leben, 1973; Goldsmith, 1974; Williams, 1976), syllable structure (Fudge, 1969; Kahn, 1976), stress and rhythm (Liberman, 1975; Liberman and Prince, 1977), Semitic non-concatenative (templactic) morphology (McCarty, 1979a) and vowel harmony (Clements, 1977) have played an important role. The emerging non-linear atmosphere was condensed into a general autosegmental theory by John Goldsmith’s (1976a) Ph.D., which is usually credited as the initial spark of autostructuralism (see Anderson, 1985: 347ff.; Lowenstamm, 2006 for a survey). This is correct in the sense that the decisive property of autosegmental representations beyond their graphic and non-linear aspect is introduced in Goldsmith’s Ph.D: the idea that a structure may be ill-formed (Goldsmith, 1976a and b defines well-formedness conditions).

In careful historiographic work, John Goldsmith himself (see also Clements, 2000a) has also identified pre-generative inspiration for non-linear thinking. This is most obvious for the syllable, which is known at least since the neogrammarian school of the nineteenth century and widely acknowledged in structuralist work (e.g. Pike and Pike, 1947; Hockett, 1955: 51ff.; Pulgram, 1970 offers an overview) as well as in NGP (e.g. Hooper, 1972). In addition to this, Goldsmith (1976a) for example, (see also Goldsmith, 1976b; Goldsmith and Laks, forthcoming) mentions Harris’ (1944) long components, Bloch (1948) and the prosodic analysis of the Firthian school of the 1940s. While all of these precursors (in hindsight) heralded the non-linear idea by promoting this or that aspect of it, none of them held all pieces in hand, not to mention their constitution in terms of a formal system. This condensation into a frame that also provides a graphic identity to multi-linear structure (which was absent in the structuralist and the NGP conception of the syllable: in the latter syllable boundaries ‘$’ were merely inserted into the linear string) has only occurred in the second half of the 1970s in generative phonology (Falle and Vergnaud, 1980 provide an early overview of the then recent history).

\[ \begin{array}{cccccccc}
\sigma & R & \sigma & R \\
O & N & O & N & C \\
\hline
x & x & x & x & x & x & \text{skeletal tier} \\
\alpha & \beta & \gamma & \delta & \varepsilon & \zeta & \text{melodic tier} \\
\end{array} \]

Figure 16.2 Shows a typical autosegmental representation

The idea of autosegmental representations is that two phonological objects (features in SPE) may be related in such a way that no linear order can be determined. This is why SPE is called a linear model (all objects, that is, features and feature-matrices, obey an unambiguous precedence relation with all other objects), against autosegmental representations, which are called non-linear. Under Figure 16.2, for example, there is no precedence relation between \( \beta \) and the two \( x \)-slots that it is attached to: \( \beta \) ‘belongs’ to one as much as to the other. Phonological objects may thus occur on different levels, which are called autosegmental tiers (or lines): in Figure 16.2 Greek letters belong to the melodic tier, \( x \)-slots to the skeletal tier, and higher items to the (or several) syllabic tier(s) (more on the specific identity of supra-skeletal tiers shortly). On every tier, items – which are called autosegments – obey a strict linear order. Non-linear effects are produced when items of different tiers are not associated one-to-one: association may be either one-to-many (e.g. \( \beta \)) or many-to-one (e.g. \( \gamma \) and \( \delta \)).

This is the baseline that defines the minimal formal properties of autosegmental representations. Two more properties are agreed upon by all phonologists as far as I can see, although they do not follow from formal requirements. For one thing, autosegmental tiers are disposed in three-dimensional space, rather than only in two dimensions as in Figure 16.2. That is, the melodic tier in Figure 16.2 could well entertain associations with both the skeletal and the
syllabic tier. This specific configuration does not occur in practice, though, because of the second broad agreement, that is, the division of the autosegmental space into two basic areas, one accommodating melodic representations, the other constituent structure of various kinds: syllables, stress, interface with morpho-syntax (it is true that tone, which a priori is a piece of melody, commonly appears on the constituent side).

The red line that separates the melodic and the constituent area is the skeleton, which has a special status among autosegmental tiers. Skeletal slots are timing units that define the chronological progression of the overall structure: they are the instrument that defines the linear order of the non-linear structure. Items on other tiers may be related in various ways among themselves and with skeletal slots. When it comes to transform the structure into a linear phonetic signal, though, the skeleton is parsed slot by slot (left to right), and every autosegment of whatever tier that is attached to a given slot will be pronounced on that particular slot. This also means that items that are not attached to any skeletal slot (such as so-called floating segments, on which more is discussed below) are not pronounced.

Indeed, an important ground rule that is shared by all versions of autosegmental representations is that a result on the phonetic side supposes three things: (i) a piece of melody, (ii) a constituent that is itself integrated into the constituency above the skeleton and (iii) an association line that relates both items. In the absence of either, no audible trace is produced (e.g., R6, 1986). This is illustrated in Figure 16.3, where the skeleton is used for the sake of exposition (in moraic theory, other constituents take over its role).

<table>
<thead>
<tr>
<th>(1) autosegmental conditions on phonetic action</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. phonetic result: [α]</td>
</tr>
<tr>
<td>b. no phonetic result</td>
</tr>
<tr>
<td>c. no phonetic result</td>
</tr>
<tr>
<td>d. no phonetic result</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>α</td>
</tr>
<tr>
<td>α</td>
</tr>
</tbody>
</table>

Figure 16.3 Autosegmental conditions on phonetic action

There are two autosegmental theories that in one way or another do not fit the picture presented thus far. Moraic theory (Hyman, 1985; Hayes, 1989a) is based on the claim that the skeleton does not exist. Phonologically relevant units are based on weight (mora), rather than on chronological precedence. Moraic theory thus challenges standard syllable structure, which is x-slot-based. It assigns mora to segments that contribute syllabic weight, that is, to vowels (and syllabic consonants) in all languages, parametrically to coda consonants in some, but never to onset consonants.

The other theory to be mentioned in this context is based on what is called the metrical grid (Prince, 1983; Hayes, 1984; Iverson, 1992; see Halle, 1998 for an overview). Metrical grids are erected over segments and indicate their relative prosodic prominence, that is, regarding stress and rhythm. Prominence is marked by piling up a number of diacritic grid marks (which are graphically represented as ‘*’). Metrical grids thus compete with regular autosegmental arborescence (feet) for the representation of stress and, depending on the particular implementation, also with other autosegmental structure (e.g., the prosodic hierarchy that is responsible for the interface with morpho-syntax in Selkirk, 1984b). Metrical grids are clearly a child of autosegmentalism and the prevailing non-linear atmosphere of the 1980s; however, they are not autosegmental themselves because they are in no way non-linear: every segment has its own pile of grid marks, which may shift from one segment to another through computation, but which do not interrupt the linear sequence of segments. That is, a given grid mark is never related to more than one segment, and a given segment never associates to a grid mark that dominates a neighbouring segment. I leave the autosegmental status of metrical grids an open question here.

Let us now consider ill-formedness, the inherent overgeneration-inhibiting virtue of autosegmental representations. A very early and intuitive condition on well-formedness that remains in place to date is the ban on line-crossing (Goldsmith, 1976a and b): two items of the same tier (not of different tiers, though, recall that the autosegmental space is three-dimensional) may not entertain associations with another tier if they result in the crossing of association lines. In Figure 16.3, for example, /t/ could not be associated to the x-slot of the codan, and /s/ simultaneously to the x-slot of the nucleus. The autosegmental literature has produced a number of well-formedness conditions, which have been debated over the years. Three prominent cases in point are the Obligatory Contour Principle (OCP), the Strict Layer Hypothesis (SLH) and the ban on two empty nuclei in a row. The OCP (Goldsmith, 1980; McCarthy, 1986; Yip, 1988) prohibits two identical (or similar) autosegments in a row on the same tier. It has played the role of an output filter for derivations, and also of a trigger for various repair strategies. The SLH concerns the constituency of the Prosodic Hierarchy (Selkirk, 1981b; Selkirk, 1984b: 26f.). It expresses the idea that a prosodic constituent of a given layer only dominates constituents of the immediately lower level, and is exhaustively contained in a constituent of the immediately higher level. Hence, there can be no nested constituents, and no association line can bypass a layer.

Finally, a ground rule in Government Phonology (Kaye et al., 1990; Harris, 1994) is that a structure which accommodates two empty nuclei in a row (in the same domain) is ill-formed.3 Hence, in French il faut y revenir ‘one needs come back to this’ where the two e’s of revenir are schwae, either can be left out (‘revenir is ok as much as rev’nir’), but not both (‘r’o’nir is agrammatical). On the analysis
of Government Phonology, this is because the non-pronunciation of a vowel only means that the association between the relevant piece of melody and the nucleus is interrupted. The nucleus itself is still present and thus empty, and two empty nuclei in a row as in *r’o’nir are ill-formed.

4.3 Autosegmentalization of All Areas of Phonology, Blooming Theoretical Diversity

The potential of autosegmental representations, regarding overgeneration but also a host of analytic advances, was well understood in the early 1980s. Therefore, early spartan representations were continuously enriched throughout the 1980s, leading to rather complex structures. This evolution was parallel to the expansion of arboreal structure in [GB] syntax during the same period.

In contrast to the theoretical monoculture of the 1990s and 2000s (the large array of OT-internal variation notwithstanding), a blooming variety of new theories emerged in the early and mid-1980s on the autosegmental ticket. As far as I can see, this is the first time in the history of generative phonology that the mainstream divided into clearly distinguished, partly or completely competing theories that were defined by original assumptions and carried a cleaving name. The only clear precedent is Natural (Generative) Phonology – otherwise phonologists gathered around general tendencies or issues such as abstractness/concreteness.

A good witness of this extraordinary vitality are the first two volumes of the then newly established journal Phonology Yearbook (edited by Colin Ewen and John Anderson, today Phonology), which were published in 1984/1985 and became the voice of the autosegmental project (the two volumes edited by van der Hulst and Smith, 1982, were also a focal point). The Phonology Yearbook 1/2 contains a number of contributions that give Lexical Phonology a more precise shape after Kiparsky’s (1982a and b) founding papers (Paul Kiparsky, Jerzy Rubach and Geert Booij, Jerzy Rubach, K.P. Mohanan, Patricia Shaw, including also the first overview by Ellen Kaise and Patricia Shaw), and they herald the new thematic focus that dominated the 1980s: the internal structure of segments.

The founding statements of the new autosegmental theories of segmental structure also appeared in the two volumes of Phonology Yearbook. On the one hand, Feature Geometry (Clements, 1985) proposed an autosegmental arrangement of a set of articulatorily defined binary features that is more or less the same as in SPE. Each feature resides on an autosegmental tier of its own, and features are bundled under labelled nodes that define natural classes of segments, a new and important notion (that was condensed a year later in Elizabeth Sagey’s 1986 Ph.D).

The alternative system of melodic representation rejects binary primes and instead favours so-called privative (or monovalent, holistic) melodic primes that (i) define items which are bigger than a single feature (e.g. the prime /i/ represents the high front tongue body position) and (ii) produce contrast by being either present or absent (rather than by being always present but having two distinct values). The privative idea was not new: it was first aired by Anderson and Jones (1974), but now received three distinct implementations in Dependency Phonology (Anderson et al., 1985, condensed in Anderson and Ewen, 1987), Government Phonology (Kaye et al., 1985) and Particle Phonology (Schane, 1984) (see Botma et al., this volume, on features).

It is to be noted that the privative option has rapidly gained ground: it entered Feature Geometry under the header of underspecification (Archangeli, 1988; Pulleyblank, 1988). The idea is that plus-minus specifications of some (but not all) features are absent from lexical representations and come into being through spreading from the context, or by feature-filling default rules that apply at the end of the derivation. Today the question is not so much whether there should be underspecification, but rather which features should be underspecified (see Hall, 2007 for an overview).

This implementation of the privative idea is still quite different from the aforementioned theories that are direct heirs of Anderson and Jones (1974), though: in underspecification approaches, features are not just temporarily unspecified for a value: rather, they are missing altogether. Melodic primes are different in nature (bigger than a classical feature), and they are independently pronounceable, which is not the case of (underspecified) features. Finally, underspecified features often afford a three-way opposition (plus, minus, unspecified), while truly privative systems only oppose the presence to the absence of a prime.

Another important issue was the kind of relationship that melodic primes entertain: while everybody agreed that relations between linguistic objects, in phonology and elsewhere, are asymmetric and hierarchical, opinions diverged as to how this insight should exactly be implemented. The most obvious take is the one of Feature Geometry, which simply extends autosegmental trees to the infrasegmental area. The alternative is the heart of Dependency Phonology, that is, a dependency relation: the idea is that one item is more important than the other(s) and contributes more to the result of the union (cf. also Liberman and Prince’s 1977 labelled trees where sisters are either weak or strong). Government Phonology follows this line of thought (here the distinction is between a head and (an) operator(s)), while Particle Phonology expresses asymmetry by the number of copies of the same prime that contributes to the definition of a segment: [i] consists of the prime /i/, [e] of /i+a/, [æ] of /i+a+a+æ/ (the presence of several copies of the same prime does not have any
impact on the result in other theories). Van der Hulst and Ritter (1999, 2003) provide an informed overview of the various implementations of the dependency programme.

Melodic representation of course was not the only sub-field that was autosegmentalised: all areas of phonology were recast in autosegmental terms. Figure 16.2 shows syllable structure above the skeleton, but this is only an arbitrarily chosen example. The representation of stress and the interface with morpho-syntactic information was also autosegmentalized; while the former is expressed in terms of feet since Libermann and Prince (1977) (with the aforementioned competing metrical grid). Prosodic Phonology has replaced linear carriers of morpho-syntactic information in phonology, that is, SPE-type boundaries \( \# \), with a multi-layered areal pattern, the Prosodic Hierarchy (Selkirk, 1981b, 1984b; Nespor and Vogel, 1986; see Scheer, 2008, forthcoming a on this move). Prosodic Phonology also proposed a global unification of all types of constituent structure above the skeleton: the Prosodic Hierarchy encompasses (with some variation regarding, for example, morae and the clitic group) the following units of growing size in a single hierarchical structure (which is marshalled by the Strict Layer Hypothesis): mora < syllable \( (c) \) < prosodic (phonological) word \( (\omega, \text{about the size of a word}) \), prosodic (phonological) phrase \( (p, \text{about the size of an NP or a VP}) \), intonational phrase \( (IP, \text{an intonational unit}) \), phonological utterance \( (U, \text{about the size of a sentence}) \).

Finally, it must be noted that in parallel to the autosegmentalization of phonology, the new research programme for generative linguistics that was laid out in Chomsky's (1981) Pisa lectures and gave rise to the Government and Binding framework in syntax very rapidly seeped through into phonology. Grammar was divided into principles and parameters: the ban on line-crossing, for example, was a (universal) principle, while there was a fair amount of parametric variation in the application of the OCP. The import of the new syntactic model into phonology as a major change in perspective is especially (and explicitly) recognized not only in Government Phonology (Kaye and Lowenstamm, 1984: 123, Kaye et al., 1985: 305, Lowenstamm and Kaye, 1986: 97ff., 124ff.), as well as in Declarative Phonology (Scobbie, 1991: 7).

4.4 What about Computation?

In this rush for representational expansion, a fair question to ask is what happened to computation. At first sight, the answer is: nothing at all. All through the 1980s, computation was carried out exactly as before, only that the set of ordered rules now applied to representations, rather than to linear strings of segments and feature bundles. Computation was called to do labour that it did not do before, though, since the most widespread idea regarding representations above the skeleton was that they are absent from the lexicon and hence need to be created by rule. This means that underlying representations were not really different from what they were in SPE (and from what they have returned to today in OT): a linear sequence of segments (except of course that the internal structure of segments was now autosegmental).

Syllable structure, for example, was erected over unsyllabified lexical entries by a syllabification algorithm on the grounds of major category information provided by the segments. In further evolution, bits and pieces of constituent structure were stored in the lexicon as well. Rubach (1986), for example, carries out the autosegmentalization of vowel-zero alternations in Slavic languages (where alternating vowels are called yers), whose major advance was to distinguish alternating (Polish pies - ps- a 'dog, Nsg. Gsg') from stable (bries - bies-a 'devil, Nsg. Gsg') vowels that are phonetically and phonemically identical by associating the latter, but not the former, to the skeleton in the lexicon. Look on this analysis, thus, alternating vowels are floating, while stable vowels are lexically associated to the skeleton. The very notion of floating segment, which has rapidly gained ground, actually supposes the existence of some constituent structure in the lexicon: it only makes sense to talk about something that floats if there are other items that do not float, that is, are associated.

Another typical case of floating segments that require the presence of constituent structure in the lexicon is French liaison, where word-final consonants are pronounced or not according to whether the following word begins with a consonant (les [le] café 'the coffee') or a vowel (les [lez] enfants 'the children'). The standard autosegmental analysis of liaison is in terms of floating consonants (Encrevé, 1983, 1988). Government Phonology (Kaye et al., 1990) has gone farthest in this direction by assuming fully syllabified lexical entries.

Finally, the nature of computation was significantly changed by the fact that, operating over representational material, all it could do was to link or to delink autosegmental items. Indeed, the typical computational operation in an autosegmental environment is to spread some item, that is, to create an association line between it and some other unit. The palatalization of velars before front vowels, for example, is understood as the spreading of some palatal prime from the trigger into the target. This contrasts with pre-autosegmental rule application where the structural change occurred before the modification of a feature value. Therefore, Harris (1994: 111), for example, holds that the class of possible phonological processes is restricted to operations of delinking or spreading.

We will see in the next section that computation, which was by and large ignored in the autosegmental tradition, will come back like a boomerang in the 1990s in order to set back the cursor on Anderson's structure versus process-scale right to where it stood in the 1960s (or even farther towards the computational extreme). Recall from Section 2.3 that Anderson (1985) predicted at the representational peak of the 1980s or rather, on the way uphill, that
computation will take its revenge: his outlook was that phonology stood at the
dawn of a new computational, that is, anti-representational round.

5. Second Revolution: Anti-Derivationalism

5.1 Anti-Derivationalism without Argument

The second revolution rages against derivationalism in general and ordered
rules in particular. It is far more serious than the first revolution for two rea-
sons: (i) it attacks one of the deepest layers of generative thinking and
more generally of Cognitive Science and; (ii) it combines an internal concern,
that is, the growing discomfort with ordered rules, with an external solution,
Parallel Distributed Processing (PDP) (i.e. parallel computation), which is the
spine of the connectionism. Connectionism, however, is the neo-empiricist
(neo-behaviourist) alternative that started to challenge the rationalist standard
theory of Cognitive Science in the second half of the 1980s (Rumelhart et al.,
1986: passim). Section 5.5 discusses the connectionist import at greater length.

But let us start by looking at what anti-derivationalism actually is, and what
its motivation was. Rooted in the properties of the universal Turing
machine (Turing, 1936–1937), derivationalism lies at the heart of the standard theory
of Cognitive Science that emerged in the 1950s, and whose application to linguistics
produced generative grammar. It is the idea that computation in the mind
involves a set of instructions that act on the input in such a way that it experi-
ences step-by-step modifications which occur in a chronological and logical
order where the output of step n-1 is the input to step n. Serialism (perhaps
more appropriate a word than derivationalism), then, boils down to the exis-
tence of a set of extrinsically ordered instructions that produce a chronological
and logical order of events (and hence of the action of computation).

In generative grammar, serialism incarnates as extrinsically ordered rules in
phonology, and in early syntax as extrinsically ordered transformations. These
were abandoned in the early 1980s when GB replaced them by so-called move α,
a system where movement (computation) is free in itself, but marshalled by
constraints on representations (e.g. Newmeyer, 1986: 163ff.). Move α represents
an important turn in syntactic theory away from restrictions on computation
itself (Chomsky’s 1973 original Strict Cycle Condition, extrinsically ordered
rules) in favour of a central role of well-formedness constraints on represen-
tations such as barriers, the Empty Category Principle (ECP), case checking and
so forth. The autosegmental evolution in phonology that was described in
Section 4.2 follows the same track: representations are marshalled by well-
formedness conditions such as the OCP. On the first page of their book,
Prince and Smolensky (2004 [1993]) explicitly draw on the evolution in syntax
and declare that their new theory extrapolates the timid phonological precedent
into a formal system:

[...our goal is to develop and explore a theory of the way that
representational well-formedness determines the assignment of
grammatical structure. [...] The basic idea we will explore is that
Universal Grammar consists largely of a set of constraints on
representational well-formedness’ Prince and Smolensky. (2004 [1993]: 16.)

There is thus a causal chain beginning with the emergence of autosegmental
representations, which produced the need for well-formedness conditions on
outputs, which in turn were generalized to constraint-based computation.9 In
practice, Prosodic Morphology (McCarthy and Prince, 1986), which fully
explores the autosegmental tool, was instrumental as a precursor of constraint-
based computation. It developed at least two central devices of OT: correspond-
dence theory and alignment of prosodic and morphological constituents (the
foreword to the 2001 edition of the manuscript, McCarthy and Prince, 2001,
explains this evolution in greater detail). We will see below that for reasons to
be determined, the implementation of the constraint-based programme in OT
has initiated the dissolution of representations as an autonomous actor in
grammar.

For the time being, it is enough to take stock of the fact that while generative
syntax abandoned serialism since 1981, the representational blossoming of the
early 1980s left serialism entirely untouched in phonology.26 In the second half
of the 1980s, though, a diffuse and inscrutable discomfort with ordered rules
arose, which quickly turned into a vigorous and lasting antipathy. I have read
through the literature of that period in search for indications why serialism is
supposed to be wrong, or why it aroused scepticism – without success. I have
also asked phonologists who have lived through that period: all confirm that
there was indeed a deeply rooted antipathy against ordered rules, and that this
feeling was shared by about everybody across theories, but that it somehow
remained below the surface. The broad reference to the evolution in syntax set
aside, nobody could name, and I was unable to hunt down, sources in print that
would have explained why serialism is wrong, and why phonology needed to
engage in a major cultural break.

I happen to be aware of the reason why Government Phonology participated
in the anti-serialist movement: because Jonathan Kaye and Jean Lowenstamm
considered extrinsic rule ordering empirically vacuous. According to them,
examples where serial ordering of instructions is alleged to be critical are either
based on erroneous data, involve misanalysis or concern processes whose prop-
erties disqualify them as instances of phonological computation (no plausible
relationship between trigger and effect, exceptions, appeal to morphological
information). An example for erroneous data is Martin Joos’ famous dialect B of Canadian English for which there is no evidence (Kaye, 1990c, 2008), but which was used by Bromberger and Halle (1989) as the litmus test for rule ordering (see note 21). Examples for processes that are not phonological in nature are Trisyllabic Shortening (or other traces of the great vowel shift), and velar softening (see the aforementioned electri[k] - electri[s]-ity). The trouble is that apart from Kaye (1990c) which is only a short notice about the non-existence of dialect B, there is no trace of this programme in print (Lowenstamm and Kaye, 1986: 97 mentions that the model of the authors is referred to as the ‘no-rule approach’, but does not say why).

In contrast to the non-overt sources of anti-serialism, the origin of parallel computation is evident: connectionism. Since Rumelhart et al., (1986), the central argument in favour of parallel computation was clearly made and pedagogically repeated (e.g. Rumelhart, 1989: 134ff.): the implementation of serialism in a neuron-based environment appears to be unrealistic given the computational complexity that would be required and the time that it would take to carry out all intermediate steps one by one. We know that neurons are not serially ordered in the brain, but rather multiply interconnected. Hence, in order to get to grips with a realistic implementation in the brain, several things must be done simultaneously, just like many neurons fire at the same time and thus transmit information simultaneously. This is why we need ‘brain-style computation’ (Rumelhart, 1989).

We thus have an argument that calls for the modification of grammar, which is a model of competence, under the pressure of performance, that is, neural implementation. It is interesting to observe that Prince and Smolensky (2004 [1993]: 215ff.) categorically reject this reasoning (‘[t]he model is not incumbent upon grammar to compute’), which they consider a category mistake. The rejection of performance-based reasoning is their general-purpose shield against the numerous objections that were raised against the astronomical inflation of computational complexity in OT (see Section 5.4).

It is idle to further speculate where the general antipathy against serialism came from, how it spread in absence of written record, and how an entire field could throw overboard a fundamental piece of its identity without any discussion of the reasons. In his article In defense of serialism, Clements (2008b: 193ff.) makes the obvious point that the rise of connectionism on the Cognitive Science scene has played an important role in the development of anti-derivationalism in phonology. It remains to be seen, though, whether the development in Cognitive Science really seeped through into phonological quarters at this early stage: it was only unfolding when anti-serialism was already widespread in phonology.

In any event, the fact is that the defenders of serialism – of which Morris Halle is the most prominent figure whose position has not varied since the 1950s – reacted on the anti-derivational atmosphere by exposing arguments in favour of ordered rules. The article by Sylvain Bromberger and Morris Halle published in 1989 (Bromberger and Halle, 1989) discusses the question whether the abandoning of ordered instructions by GB syntax and its replacement by the Principles and Parameters approach should lead phonology to follow the same track. The authors reject this perspective because, as they try to show, the subject matter of phonology and syntax is intrinsically different. Bromberger and Halle thus defend serialism against an extra-phonological trend, and also mention the fact that extrinsically ordered rules have come under fire in phonology. Significantly, though, they can come up with only two anti-derivational references: Lowenstamm and Kaye (1986) and a 1987 LSA (Linguistic Society of America) presentation by B. Majdi and David Michaels. This is good indication of the fact that the anti-derivational atmosphere was by and large absent from print: Bromberger and Halle were fighting against an invisible enemy.

5.2 Consequences of the Latent Anti-Derivationalism: New Constraint-Based Theories in the Early 1990s

The latent anti-derivationalism of the late 1980s was the driving force of the events in the early 1990s. Anti-derivationalism was now made explicit as the driving force behind the emergence of new theories, but there was still no discussion of comparative merits with serialism, or of the reasons why serialism is a bad thing to have. That is, parallel computation is typically introduced as an alternative to ordered rules, but authors leave it at this juxtaposition (e.g. Prince and Smolensky, 2004 [1993]).

Three theories that have emerged in the early 1990s (or late 1980s) are based on the anti-derivational mantra: Optimality Theory (Prince and Smolensky, 2004 [1993]), Declarative Phonology (Scobbie, 1991; Scobbie et al., 1996) and Government Phonology (Kaye et al., 1993, which was prepared by work quoted earlier since the mid-1880s). The computation in all three cases is based on constraints, which however do not have the same status: while they are ranked and violable in OT, they are absolute (i.e. non-violable) in Declarative Phonology (see discussion of constraints in Uffmann, this volume). Computation in Government Phonology was not explicitly regulated for some time (the only indication that could be found in print was Kaye’s 1992: 141 statement according to which processes ‘apply whenever the conditions that trigger them are satisfied’), but its constraint-based character is obvious since the mid-1990s (Licensing Constraints, Charette and Gökser, 1994, 1996; Kaye, 2001; Guussmann’s 2007 recent book on Polish is an application to an entire language).

Constraints in Government Phonology thus apply whenever a form may be modified by them, with no extrinsic ranking or ordering, and without being
able to be violated: the set of constraints (the $\varphi$-function in Kaye's 1995 terms) is (simultaneously and) iteratively applied to the string that is submitted to interpretation, and computation ends when no further modification can be made (this is an obvious parallel with Harmonic Serialism, on which more discussion below). To use serial vocabulary, this system is thus able to handle a feeding relationship (the input for the application of a constraint is created by the modification of the input string by another constraint), but no other (i.e., bleeding, counter-feeding, counter-bleeding). A difference must therefore be made between serial computation as in GP and serialism as such. In the former, computation is serial in the sense that constraints may apply to the same string several times, and that intermediate steps may exist, whereas in the latter there is no extrinsic or logical ordering of instructions, that is, classical extrinsic rule ordering. There is also no ranking or prominence relationship among constraints: all instructions are equally important, and there is no selective application (all items of the $\varphi$-function apply when a string is computed). However, Gussmann's (2007) book on Polish (as well as Polgárdi, 1999) indicates that constraints may also be ranked, that is, that in case of a conflict between two constraints one will be given priority (see Scheer, 2010b).

Declarative Phonology is directly affiliated with HPSG (Head-Driven Phrase Structure Grammars) on the syntactic side and therefore does not stand on generative grounds: under the pressure of overgeneration, HPSG has taken the radical step to eliminate computation altogether. The result are so-called monoserial representations, which are fully informed with morpho-syntactic, semantic and phonological information that is available at any point in the derivation (talking about a derivation is actually improper because monoserialism rejects the existence of distinct underlying and surface representations). The major issue that HPSG has with the generative approach is thus about modularity: there are simply no modules in the HPSG landscape, which is a fully scrambled everything-is-one environment.

Regarding OT, it is to be noted that the juxtaposition of serial and parallel computation was really a matter of indecision in the early days: Prince and Smolensky (2004 [1993]) consider so-called Harmonic Serialism all through their manuscript. Harmonic Serialism works like regular OT, only that the candidate set produced by GEN is much more local, and the winner of the strictly parallel evaluation procedure is fed back into GEN. This procedure is repeated until no harmonic improvement can be achieved anymore. Hence, Prince and Smolensky (2004 [1993]: 6) write that

\[\text{[d]efinitive adjudication between parallel and serial conceptions, not to mention hybrids of various kinds, is a challenge of considerable subtlety, as\}]

indeed the debate over the necessity of serial Move-$\alpha$ illustrates plentifully [...], and the matter can be sensibly addressed only after much well-founded analytical work and theoretical exploration.

A more direct application of connectionism to phonology. John Goldsmith's Harmonic Phonology (Goldsmith, 1992, 1993; Larson, 1992), also follows this track: the successive application of rules progressively increases harmony, and the well-formedness of representations is measured in gradient, rather than in categorical terms. Finally, Harmonic Grammar (Legendre et al., 1990b; Smolensky and Legendre, 2006) is also a direct application of connectionism, in fact one which is most closely related to the interests of Cognitive Science: it is based on so-called weighted constraints, which like Goldsmith's gradual well-formedness are a direct transcription of the central connectionist notion of connection weight (and the activation level of neurons which defines their output). In this perspective, the relationship between constraints is one of lesser or greater prominence, rather than of strict dominance: less important constraints can leagie together and outrank a more important constraint on account of their cumulated weight.

Serialism, which stood unchallenged and without alternative since the 1950s, has thus first fallen into disgrace for reasons that are not clear and have not been made explicit, and was then replaced by parallel computation when this option was available as an import from connectionism (more on this import in the following section).

Parallel computation has an important corollary: it can only be done on the basis of constraints. We have seen in Section 5.1 that autosegmental representations and well-formedness conditions were the initial spark of the development of constraint-based computation. Evicting rules as the basic carrier of computational instructions in favour of constraints was thus the result of the conjoint action of parallel computation (which in turn was the instrument of anti-derivationalism) and the increasing importance of well-formedness conditions.

Another factor, whose impact on the movement towards constraints is difficult to evaluate, though, is what was known as global conditions and conspiracy in the 1970s: the latter notion is known at least since Kisseberth (1970a). It builds on the observation that a number of seemingly independent processes may conspire to produce the same result (such as the avoidance of closed syllables). In rule-based computation, each process requires its own rule, and the obvious generalization regarding the restriction on the output is missed. The relevant generalization may be captured by a single constraint on the output, and this is why current parallel computation is often called output-oriented.
5.3 How Computation Works: Rules Versus Constraints, Eventual Hybrid Solutions

Rules and constraints have different properties. Rules are made of a structural change (the part on the left-hand side of the slash in A → B / C...D) and a structural description (the part on the right-hand side). The latter defines a string, CAD, that is in need of modification. Constraints do the same thing (if in different vocabulary, see below): they issue general requirements or prohibitions, in our case, for example, *CAD. However, they do not specify what should be done in order to satisfy the requirement or prohibition at hand. This is another reason why constraint-based computation is said to be output-oriented: constraints specify how things should or must not look like, but do not give any indication how the desired or prohibited state of affairs should be achieved. That is, constraints divorce the structural description and the structural change of rewrite rules. Goldsmith and Laks (forthcoming) point out that this split was already suggested by Sommerstein (1974).

Hale and Reiss (2008: 195ff.) discuss the formal difference between rules and constraints at length. They describe rules as a function that maps an input representation to an output representation (i.e. something is modified), while a constraint maps an input representation to the binary set ‘violation’ or ‘no violation’, that is, without modifying anything. The equivalent of the modification that is operated by rules is achieved by comparing the input form to candidates, and by determining which candidate incurs the least harmful set of violations. OT-style parallel computation the fact falls into three independent steps which are serially ordered: (i) GEN: the candidate set is generated on the basis of a given input; (ii) EVAL: candidates are evaluated for violations of the constraint set. This is the time-saving (in the connectionist sense) and truly parallel piece of the computation since the evaluation of a candidate by different constraints can be done simultaneously: a given computational action does not need to ‘wait’ for another computational action since constraints assess candidates independently. (iii) Finally, the optimal candidate is determined once the results of the computational action of all constraints are known.

Hale and Reiss (1998: 196) build on this difference between rules and constraints in order to make a point against the latter from the logical and cognitive point of view: a grammar ought not to contain explicit statements against monsters (they use the NoBANANA example in order to show that there is no point in explicitly excluding real bananas from UG by an explicit statement therein). Constraints, however, only inform some other part of the grammar that a given representation is ill-formed, and there is an infinite number of ways in which a representation can be ill-formed. Hence, explicit statements (constraints of the NoBanana kind) are needed in order to rule out the monsters.

Another important difference between constraints and rules is the vocabulary in which they are stated: while rules can only refer to the specific vocabulary items that phonology is made of (features or other items of autosegmental representations), constraints are made of prose statements and can express anything that prose can express (including very broad instructions such as ‘be lazy!’), which is the formulation of the constraint Lazy that Kirchner, 1998 believes is the motor for lenition; *Structure is another case in point). This loss of reference to a specific phonological vocabulary is meaningful in terms of Cognitive Science: we will see below that while so-called domain-specificity is a defining property of cognitive modules (which thus operate over a specific and proprietary vocabulary), it is denied by connectionism, where computation is content-free.

The fact that the formulation of constraints is not constrained in any way is a well known and oft-mentioned property of this carrier of computational instruction (see, for example, Carvalho’s 2002b formulation ‘from constraint-based theories to theory-based constraints’), which to date however has failed to produce any effect: as far as I can see, no attempt is made in order to define what an (in-)possible formulation of constraints looks like.33

Obviously, the freedom to express anything that prose can express dramatically increases overgeneration, which takes us back to the post-SPE debate. OT has often been charged with computational irresponsibility: in principle, GEN produces an infinite set of candidates that cannot even be stored, let alone computed, and the number of grammars that a set of, say, two or three hundred universal constraints (which is an extremely conservative estimate) produces given free ranking is astronomical (even when logically impossible rankings and those that produce identical patterns are counted out). Factorial typology is commonly advertised as a trump for modelling dialectal variation, but the hundreds, thousands or more systems that are produced without empirical echo are usually not mentioned. Also, the concern for limiting the number of constraints that was present in early OT has more or less disappeared from the agenda: new constraints are proposed every day, and hardly anybody is able to establish a comprehensive inventory.

From the beginning, the answer of OT to criticisms regarding excessive computational complexity (and overgeneration) was to call on the competence-performance distinction (but recall from Section 5.1 that it is precisely performance-based arguments that are used by connectionism in order to promote parallel, that is, brain-style computation):

It is not incumbent upon a grammar to compute, as Chomsky has emphasized repeatedly over the years. A grammar is a function that assigns structural descriptions to sentences; what matters formally is that the function is well-defined. [. . .] Grammatical theorists are free to
contemplate any kind of formal device in pursuit of these goals; indeed, they must allow themselves to range freely if there is to be any hope of discovering decent theories. Concomitantly, one is not free to impose arbitrary additional metaconstraints (e.g. computational plausibility) which could conflict with the well-defined basic goals of the enterprise. (Prince and Smolensky, 2004 [1993]: 215f., emphasis in original)

Since its inception, generative grammar indeed followed this line of thought: Chomsky has always argued that competence is not about implementation, and that implementational arguments have no bearing on the properties of the model of competence that linguists are supposed to develop. The minimalist programme (Chomsky, 1995, 2000: passim), however, clearly suspends this perspective: grammar must respond to implementational requirements. The whole point of the minimalist approach is to make grammar evolve in response to extra-grammatical factors, that is, interface requirements. Phase Theory, for example, cuts the computation of a full sentence into independent pieces for reasons of computational economy regarding the limited availability of active memory, a costly cognitive resource (e.g. Chomsky, 2000: 101; 2001: 15).

In this environment, OT cannot hide behind competence anymore in order to escape the discussion of computational complexity. The greater generality of constraints with respect to rules and their dissociation from a specifically phonological vocabulary is thus a problem. On the other hand, it has crucially contributed to the success of OT and parallel computation. De Lacy (2007b: 14ff.), for example, recapitulates the advantages of parallel computation over serialism: ordering paradoxes, global conditions, conspiracy. The question, then, is whether the generalizing remedy (candidate computation by prose-based constraints) is not worse than the original disease (time-consuming serial application of rules).

5.4 Tabula Rasa in 1993 and Hybrid Rule-Constraint Computation?

A worrisome property of the OT literature and of the self-understanding of OT is the constant attempt to assess a tabula-rasa interpretation of the year 1993: de Lacy (2007b), for example, is much concerned with showing that OT is different from anything that phonology has produced before 1993 in general, and with making OT antithetical with respect to SPE in particular (de Lacy, 2007b: 13, opposes OT to ‘the dominant theories before OT – SPE and its successors’). The tabula-rasa claim may help to assess a new theory, to raise funds and to gain academic positions, as it did in the 1960s when generative phonology rolled over the structuralist establishment – but it is as wrong now as much as it was then (see Goldsmith, 2008). Even a cursory look through Anderson’s structure-process prism is enough to be convinced: OT takes phonology way back to the 1960s when representations did not exist and computation was king (more on this in Section 5.6). Van der Hulst and Ritter (2000) describe the SPE-heritage of OT in greater detail. Also, it was shown in Sections 5.1 and 5.2 that OT roots in autosegmentalism-promoted well-formedness conditions. Also, Prince and Smolensky’s (2004 [1993]) original manuscript does not really license the picture that present-day OT-based story telling tries to establish.

This is not to say, of course, that there is nothing new in OT: our conception of computation has profoundly changed since OT and other theories laid the focus on constraints. The merit of OT is the application of parallel computation to phonology, as well as the promotion of the specific view whereby constraints are ranked and violable.

In this context, it is interesting to compare the evolution of computation in syntax and phonology. As Hale and Reiss (2008: 202ff.) point out, the Pisa turn syntax and the movement against extrinsically ordered rules in phonology were not only abolishing the same kind of serialism in both areas, and were not only doing away with rules. They also replaced rule-based serialism by free generation-cum-filters. In GB syntax, morphemes could ‘freely’ be concatenated, and Move-α could ‘freely’ apply. Ill-formed results were then filtered out by global constraints (on locality, case etc.). The minimalist version of this conception of computation focuses on the interfaces, PF and LF, which impose conditions that make the derivation either crash or converge. Constraint-based computation in phonology follows the same track: in OT GEN does the free generation, and constraints filter out the optimal candidate. Unlike in syntax where filters have equal rights and are inviolable, however, constraints are ranked and violable in OT.

Given this overall picture, an interesting question is certainly whether the correct solution for computation is only binary, that is, either rule-only or constraint-only. Calabrese (2005), for example, holds the principled position that a sound theory of phonology must have both serially ordered rules and constraints: while the former are instructions to create a given configuration, the latter specify which configurations must be avoided (constraints apply only to Kisseberth’s 1970a conspiracies). This is indeed the naturally grown state of the art of the 1980s (with enlarged competences for well-formedness constraints, which may also concern purely melodic configurations): recall from Sections 4.2 and 5.1 that well-formedness constraints (i.e. output filters) are a direct consequence of (autosegmental) representations and peacefully cohabited with traditional rule-based computation. Hale and Reiss (2008: 209ff.) offer instructive discussion of the rules-cum-constraints option (concluding that both hybrid and constraint-only systems are inaccurate: computation must be purely rule-based in their opinion).
5.5 OT and its Connectionist Endowment

It was briefly mentioned in Section 2.1 that generative grammar is the linguistic outgrowth of the cognitive revolution of the 1950s, which was anti-behaviourist, anti-empiricist and dualistic in nature. Connectionism is the modern version of the reverse intellectual position. Prince and Smolensky (2004 [1993]: 217ff.) provide an informative discussion of the relationship of OT with connectionism. They argue for cherry-picking; while parallel computation is taken over, other major connectionist tenants are rejected. For example, Prince and Smolensky do not accept the connectionist anti-symbolic stance that rejects symbolic representations (such as syntactic trees) because the only relevant level where decisions are made is neuronal: in a connectionist environment, neurons only react on activation levels, hence cannot parse, or distinguish, between symbolic objects (e.g. Fodor and Pylyshyn, 1988; Dinsmore, 1992).

Just like the standard theory of Cognitive Science, OT recognizes a symbolic level of representation. The place for connectionist non-symbolic (‘colourless’) computation, then, is an intermediate level between the symbolic level and the physiologically neural functioning of the brain. This conciliatory position that rejects reductionism (i.e. the denial of the mind as an independent level of analysis) and where the connectionist level mediates between the mind and the brain is defended by Paul Smolensky since his earliest work (Smolensky, 1986, 1987, 1988, 1991) and up to the present day (Smolensky and Legendre, 2006).

Also, Prince and Smolensky reject the neo-behaviourist take of connectionism regarding acquisition according to which ‘knowledge of language can be empirically acquired through statistical induction from training data’ (Prince and Smolensky, 2004 [1993]: 217).

Although Prince and Smolensky (2004 [1993]) do not mention this issue, OT obviously also rejects the connectionist claim that there is no difference between computation and storage. In the connectionist perspective, the ‘experience’ of a neural network – the equivalent notion of memory – is acquired when the patterns of connectivity change: neurons may develop new connections (synapses), may lose old connections or modify the strength (weight) of existing connections (the two former are often viewed as a special case of the latter). The computational units themselves have no variable behaviour that contributes to the properties of the whole, which is exclusively determined by the connective network (see, for example, Stillings et al., 1995: 114ff. on connectionist models of memory).

All linguistic theories since Antiquity of course rely on the assumption that there is a lexicon that exists independently of grammatical activity which transforms lexically stored objects into actual speech. The linguistic mirror of the connectionist non-separation of storage and computation is so-called ‘Cognitive Grammar’, which was founded by Langacker (1987) and is overtly empiricist, (neo-)behaviourist and anti-generative (see Taylor, 2002). Langacker (1987, Vol.1: 42) talks about the ‘rule/list fallacy’. The phonological offspring of this line of thought is represented by exemplar- and usage-based approaches in general, and by Joan Bybee in particular, who writes that

[...]linguistic regularities are not expressed as cognitive entities or operations that are independent of the forms to which they apply, but rather as schemas or organisational patterns that emerge from the way that forms are associated with one another in a vast complex network of phonological, semantic, and sequential relations. (Bybee, 2001: 20ff.)

Finally, another important connectionist headline is the aforementioned PDP, which contrasts with the regular assumption that computation is serial. While on the count of standard Cognitive Science the output of one computation is the input to another, multiple computations take place in parallel in the connectionist perspective. Also, the units that carry out computation – neurons, or clusters thereof – are not specialized for any particular computational task, or for a particular input material (computation is non-symbolic and not specific to any domain). Rather, neurons are all-purpose computational units that are able to perform any computation on the grounds of any type of information submitted. This is why connectionist computation is called distributed.

A corollary of distributed computation is the claim that there are no specialized computational units made of clusters of neurons that can be delineated in the brain: computation is opportunistic and does not need any specialization of its support units, the neurons. The modular theory of the mind (Fodor, 1983), which incarnates as the inverted T model in the generative architecture of grammar, is based on the reverse assumption: there are stable, genetically endowed, content-sensitive (i.e. symbolic) computational units that are devised to a very narrow and specific function, which can only work with a particular type of input vocabulary (domain-specificity), and can do nothing else than what they have been designed for.

In sum, the connectionist perspective may be characterized by the fact that computation is content-free: all other properties follow from this assumption. That is, the mind does not know what it is doing when computation takes place: computation is only general-purpose, that is non-specialized for any task or function; it works without reference to any symbolic code, which would make the operations specific to a particular domain or content since symbols are symbols of something, and may be opposed to symbols of a different kind.

The question is whether the cherry-picking in the densely interrelated network of connectionist assumptions that Prince and Smolensky (2004 [1993]) propose is viable: the genetic code of OT rejects basically all tenets of connectionism save one, parallel computation. Parallel computation is represented by the two Ps in
PDP (Parallel Distributed Processing), but Prince and Smolensky do not address the question of the D, which is anti-modular. We will see in Section 5.7 that the D appears to be a direct consequence of parallel computation: it is constantly working on OT practice (if without explicit discussion) and has induced what I call the scrambling trope, that is, the creeping dissolution of modular contours. The same holds true for content-free computation, which has made representations irrelevant and interchangeable, before dissolving them in computation altogether (representations are ‘emergent’, rather than given). This computation-trope is discussed in the following section.

The conclusion, then, is that parallel computation has probably entered the generative paradigm with some more empiricist luggage, and the question for further development is a theory that maintains the rationalist and anti-empiricist core of generative grammar while also implementing constraint-based and parallel computation.

5.6 Grammar Reduces to Computation: Representations Are Demoted to Decoration

Knowing about the connectionist roots of OT helps to understand the extreme computational orientation that phonology has taken under its lead since 1993. It is sometimes rightly recalled that OT is a theory of constraint interaction, not of constraints. This means that OT does not supply any substance itself: there are genuine vocabulary items in structuralism (phonemes, SPE (segments) and autosegmental theory (autosegmental structure), but there are no OT-specific representational items. OT uses whatever representational material is available, and may well produce the same result with entirely different (and incompatible) vocabulary.67 It is difficult not to establish a direct relationship between the fact that OT is a purely computational theory where representations make no sovereign contribution to the definition of grammaticality (which is decided by constraint interaction alone) and its content-free connectionist prototype.

Paul de Lacy’s (2007a) edited handbook, and de Lacy’s (2007b) introduction to the volume in particular, document the global trend from representation to computation in much detail. It is hard to find a thematic chapter of the book that does not insist on this evolution. A good example is John Harris’s 2007 contribution on representations. The author, who is not exactly known for anti-representational positions, defends the categorical, that is, non-gradient character of representations. In the end, however, the reader learns why it is that representations can be non-gradient at all: because they do not exist. The world, including grammar, is definitely gradient and computational: categorical objects can exist at best as a product of gradient computation (structure is emergent, rather than given: ‘categorical behaviour is emergent rather than an inherent property of these descriptors’, Harris’ conclusion, p. 137). Rather than write about representations, Harris thus ends up talking about ‘descriptors’ since this is the only role that they can play in a world where decisions are only made by constraint interaction: representations ‘still have an important heuristic value as descriptors to be used in the building and experimental testing of models of phonological grammar’ (Harris, 2007: 137). In other words, representations are decoration: they may help the linguist to picture the result of constraint interaction – but they have no impact on grammar at all.

This is indeed the conclusion that follows from the OT-tenet that the only means to determine grammaticality is constraint interaction. Hence, whatever items of the representational furniture of the 1980s are used, they are mere decoration that do not function as sovereign arbitrator, and do not have anything to say regarding grammaticality (a structure with line-crossing, for example, may be the optimal candidate if all other candidates violate higher-ranked constraints).

Given Stephen Anderson’s prism and the constant see-saw movement observed, a fair question to be asked in the face of the new OT-loop back into the 1960s is why theories go down the representational or the computational road. When autosegmental representations were developed, the motivation was clear: gain of insight (tone spreading, the possibility to characterize the coda disjunction [f, C] as a single phonological object etc.) and the promise of an efficient instrument against the plague of overgeneration. De Lacy (2007b) examines the question why OT has progressively replaced representations by computation. The answer is more or less that representations are eliminated because their function can be taken over by computation: ‘OT has allowed the burden of explanation to move from being almost exclusively representation-based to being substantially constraint-based’ (24). The question why there should be such a movement (which, recall, is not at all what Prince and Smolensky, 2004 [1993] had originally envisioned) is left unanswered: we do X not so much because we want to and have good reason to do it, but simply because we can do it. In other words, OT is a theory of computation and promotes what it is competent for without looking left or right. The progressive elimination of representations is thus but a side effect of this computational trope, which roots in the decision that the only thing that determines grammaticality is constraint interaction.

This proviso, however, does not follow from OT: OT is a theory of parallel computation that uses ranked and violable constraints. This does not lay any claim on how much of the explanatory pie is computational: the view that 100% has been installed without discussion or comment and today is part and parcel of OT. Yet it is only one possible attitude. Another view is expressed in a small but growing body of literature to which Marc van Oostendorp (van Oostendorp, 2002, 2003, 2005, 2006) has contributed a good deal, and which is condensed in
Continuum Companion to Phonology

Blahe et al., (eds.) (2007); Optimal Domains Theory belongs to the same family (Cole and Kisseberth, 1994; Cassimjee and Kisseberth, 1998). Blahe et al., (eds.) challenges Freedom of Analysis: you ought not to be free to do what you want with representations. In terms of the classical OT grammar, this means that there are restrictions on GEN, which produces only a subset of logically possible candidates.

The idea that OT is a complete theory of grammar has been tacitly entertained since its inception. The bare existence of variations of OT that place restrictions on GEN show that this view is overstated: OT is not a theory of grammar; it is a theory of a piece of grammar, computation. Anderson (1985) says that it takes more than just computation to make a grammar. The least common denominator of OT, then, is parallel computation that uses ranked and violable constraints. All the rest is free and a matter of choice of the analyst, who may or may not be a generativist, may or may not be a functionalist, may or may not assume a modular architecture, may or may not be representationally oriented, may or may not believe in the virtue of serial ordering of phonological (and/or grammatical) events (see below), may use this or that representational system, and so forth (this issue is further discussed in Scheer, 2009).

5.7 The Scrambling Trope

Closely related to the trope for computation is the pervasive reflex of OT to make distinct things indistinct; that is, to put them in the same constraint hierarchy, to intersperse them and to assess them in one go. Of course there are many different degrees of scrambling, but the tendency is clearly towards a single grammatical space where anti-derivationalism is enforced globally. The most visible result of this trope is anti-cyclicality, that is, the rejection of inside-out interpretation because of its serial character (e.g. Kager, 1999: 277). OT has produced a whole anti-cyclicality literature (which has an important intersection with the anti-opacity literature) that proposes alternative, strictly parallel ways of communicating with morpho-syntax: co-phonologies, indexed constraints, OO (Output-Output) faithfulness and so-called interface constraints.

Three more diagnostics for OT's misty (and largely unreflected) relationship with modularity may be mentioned. First, uncontradicted violations of Indirect Reference (i.e. the prohibition to make reference to untranslated morpho-syntactic categories that was established by Prosodic Phonology) are commonplace in OT. That is, ALIGN and WRAP constraints make constant reference to morpho-syntactic structure and labels; interface constraints such as Farm-root and Farm-affix make reference to designated morpho-syntactic categories (even reference to individual morphemes is not a problem, Anttila, 2002, provides an overview) and thereby reincarnate the SPE-practice of supplementing rules with morphological diacritics. Second, mapping (of morpho-syntactic into phonological prosodic categories) is done in the phonological constraint hierarchy (rather than outside of the phonology as was the case in Prosodic Phonology): ALIGN and WRAP are interspersed with purely phonological constraints. Finally, constraints whose formulation combines phonological and morphological instructions are customary (see Yip, 1998, on this issue).

Kingston (2007) discusses the scrambling trope on the example of the abandon of the distinction between phonetic and phonological constraints. He also points out the causal relationship between the move from serial to parallel computation on the one hand and the everything-is-the-same programme on the other:

[r]eplacing serial derivation by parallel evaluation removes the barrier to phonetic constraints being interspersed among and interacting with phonological constraints. [... ] Future research will determine whether phonological and phonetic constraint evaluation are a single, integrated process, as advocated by Steriade and Flemming or instead sequential, as advocated by Zsiga. (Kingston, 2007: 432)

There is thus reason to believe that the commitment to non-serialism is the driving force behind OT's scrambling trope. Despite this in-built tendency, however, adhering to modular-destructive indistinction in grammar is a personal choice of the analyst, not a fatality. Those who argue with anti-derivationism in order to challenge cyclic derivation and to set up a single constraint hierarchy where phonetic, phonological, morphological and even syntactic constraints (sic. Russell, 1999) are interleaved make a category mistake: derivation and computation is not the same thing (see Scheer, forthcoming a). OT is committed to parallel computation, and in generative grammar the unit where computation takes place is the module. Grammar is made of several modules, each with a distinct computation that works on distinct (i.e. domain specific) vocabulary. Hence, nothing withstands a perspective where all linguistic computation is perfectly parallel, but distributed over distinct and serially ordered computational systems (modules). It is only when the non-derivational claim is laid on the entire grammar that the scrambling trope appears.

That one can resist the scrambling trope is shown by the fact that there are derivational versions of OT that allow for serial communication among modules while observing strictly parallel modular-internal computation: Derivational Optimality Theory (DOT) (Rubach, 1997: passim) and Stratalt OT (Kiparsky, 2000; Bermudez-Otero, forthcoming a) are reincarnations of Lexical phonology in the new constraint-based environment.
Beyond the diluting of modular contours in one single constraint chamber, another target of the scrambling trope is the lexicon. Richness of the Base prohibits the introduction of any distinction in the lexicon that goes beyond melodic contrast (this is one of a number of readings, the literature offers various interpretations) because GEN must not be marshalled in any way when creating candidate variation. Hence, nothing beyond segmental contrast can be hard-wired in the lexicon. This is why input forms in OT typically look like underlying forms in SPE: they are made of a linear chain of segments. In practice, this means that the progressive integration of constituent structure into the lexicon that was undertaken in the 1980s (see Section 4.4) is ruled out, and the corresponding labour is transferred to the constraint chamber. But even more generally there is a trend in OT to introduce more and more unpredictable information by constraint. This movement is documented, for example, by de Lacy (2007b: 19f.); it contributes to the computationalization of grammar, and is also a step in the direction of the connectionist indistinction between computation and storage (whose linguistic expression, recall, is Langacker’s 1987 Vol.1: 42 and Bybee’s 2001: 20f. ‘rule/list fallacy’).

Finally, two other types of static information are also expressed by constraints, rather than by some hard-wired (lexical) recording: parameters and inventories. The transformation of the lexical recording of language-specific parameters into a variable constraint ranking (children do not set parameters on-off, but rank relevant constraints) is one of the most basic ambitions of OT, and also one that is generally held to be an area where OT is successful: cross-linguistic variation is described in terms of different rankings of the same set of universal constraints (factorial typology).

Regarding inventories, the idea is to kill two birds with one stone: inventory properties and some of the language-specific processing are derived from the same source, that is, a set of constraints. Steriade (2007) provides a well-informed overview of this project, which is also developed in Government Phonology since the mid-1990s (where the instrument are so-called Licensing Constraints, see Charette and Göksel, 1994, 1996; Kaye, 2001). In my opinion, the computationalization of parameters and of inventory information has a different quality when compared to other aspects of the trend towards scrambling: it does not destroy any distinction that is central to generative thinking (as is the case with modular contours), nor does it undo analytic advance (as is the case with the prohibition of constituent structure in the lexicon); rather, it gives a different expression of something that must be stated somehow anyway (parameters), or provides a new analytic perspective for the (belated) reconciliation with the heart of structuralist thinking (inventories and systemic pressure, which may also be at the origin of phonological processes).

In conclusion, it appears that all letters in PDP (Parallel Distributed Processing) make sense and are interrelated. This casts doubt on the idea that individual connectionist tenets may be cherry-picked: recall that Prince and Smolensky (2004 [1993]) import parallel computation into OT, but do not mention its distributed aspect. It looks like the distributed character of computation, that is, the connectionist everything-is-the-same programme, is a direct (perhaps even ineluctable) consequence of parallel computation.

6. Counterrevolution #2: Neo-Serialism and Representations

6.1 Serialism within Phonological Computation: OT-CC

The latest evolution within OT has initiated the second counterrevolution: John McCarthy was at the forefront of the anti-derivational movement in the 1990s and 2000s. In his 2007 book (McCarthy, 2007), he makes a radical about-turn and now promotes serialism. His OT-CC (Candidate Chains) has an entirely different quality than OT versions of Lexical Phonology (DOT, Stratal OT) where the relationship between modules is serial, but phonological computation itself strictly parallel: OT-CC holds that phonological computation itself is done on the basis of serially ordered instructions. This does not mean, however, that parallel computation is abandoned, or that McCarthy goes back to ordered rules. Rather, OT-CC is a version of Harmonic Serialism (see Section 5.2, and actually is more and more referred to under this name): the overall computation of an input is cut into a series of different evaluations where the output of EVAL is fed back into GEN, which adds harmonically improved candidates on each round. Candidate chains thus reflect the chronology of computational events (like intermediate forms in a serial SPE-derivation). EVAL is strictly parallel and invariant: it evaluates candidate chains, rather than single candidates, and so-called PREC constraints, a new constraint family which is the actual locus of ordered instructions, assign violation marks on the grounds of the comparison of different members of the candidate chain (PREC constraints react on the order of faithfulness violations).

There is an obvious parallel to computation as conceived of in Government Phonology (see Section 5.2) where input strings are also modified by a fixed set of instructions (which however are unordered and unranked) in iterative fashion: the output of computation also loops back in order to become the new input, and the derivation also ends when the string cannot be further modified ("harmonically improved").

OT-CC is the capitulation of anti-serialism in the face of opacity: John McCarthy has tried hard to find a solution for opacity within parallelism (Sympathy Theory, Comparative Markedness), but was not any more successful than the half-dozen of other opacity killers that can be found in the literature (e.g. targeted constraints, enriched inputs, F&F conjunction, M&F conjunction,
Continuum Companion to Phonology

6.2 Representations: You Cannot Compute Nothing

The other strand of the counterrevolution are representations that are not just mere decoration. Meaningful representations contribute a sovereign and unextinguishable arbitration role. That is, their verdict is absolute and not subject to any further evaluation: an ill-formed representation is ill-formed no matter what and independently of constraint interaction. An ill-formed structure where, say, association lines cross is out; it cannot turn out to be the optimal candidate because all other candidates violate some higher-ranked constraint. An ill-formed representation is either repaired, or the derivation crashes. In other words, representations that deserve their name break with the OT-baseline according to which the only way to determine (relative) grammaticality is constraint interaction.

Another property of meaningful representations is that they are not necessarily the result of computation: structure may be given, and neither its genesis nor its role as arbitrator has to be ‘emergent’. Representations of this kind are also meaningful because they are not interchangeable. It was mentioned in Section 5.6 that OT is a theory of computation and therefore has no opinion on representations: it has not developed any and can (actually does) work with any furniture of the 1980s, including competing and incompatible structure (see note 25). Constraint interaction will always squeeze out the right result. The effect is that representations first become decoration, and then completely disappear. Since representations are irrelevant, they have stopped being developed: there is hardly any discussion regarding the comparative merits of competing representational theories (‘I do this analysis using morae, but if you prefer x-slots that will do as well, you will just have to change the constraints accordingly’).

Theories that develop meaningful representations include Government Phonology (Kaye et al., 1990; Charette, 1991; Harris, 1994, 1997; Kaye, 2005; Pichler, 2006), Dependency Phonology (van der Hulst and Ritter, 1999; Botria, 2004; van der Hulst, 2005), Substance-Free Phonology (Hale and Reiss, 2008; Blaho, 2008) and the aforementioned rule-based approaches.

Government Phonology is especially often regarded as a representation-oriented theory. It has continuously worked on the development of representations since the mid-1980s, namely in the area of syllable structure. The genuine idea defended is that constituent structure is lateral, rather than arborescal: in Standard Government Phonology (see the references in the preceding paragraph), whether or not a consonant can be a coda depends on the availability of licensing from a following onset. That is, r in VrTv can be a coda because a licensing onset to its right is available, but word-final consonants cannot be codas since there is no following onset; therefore r in VrV is an onset (followed by an empty nucleus).

So-called CVCV (or strict CV, Lowenstamm, 1996; Scheer, 2004; Szigetvári, 2002; Szigetvári and Scheer, 2005; Cyran, 2003), a development of Standard Government Phonology, takes the lateral idea to its logical end: instead of a hybrid system where arborescal and lateral structure cohabitate, constituent structure boils down to a strict sequence of non-branching onsets and non-branching nuclei (no branching constituents, no rhymes, no codas). In this environment, the coda behaviour of a consonant is a consequence of the fact that it is followed by an empty nucleus which is unable to license its onset because it is empty (consonants that show onset behaviour are followed by filled nuclei that is hence able to dispose licensing).

The flat, that is, non-arborescal constituent structure that is the result of strict CV offers an in-built explanation why there is no recursion in phonology: the absence of arborescal structure witnesses the fact that phonology has no access to Merge (or an equivalent tree-building device); in absence of Merge, however, there can be no recursion (Scheer, 2004: xlix ff., Neellemann and van de Koot, 2006).

Finally, it was already mentioned that there is also a representation-rehabilitating strand within OT that was condensed in Blaho et al., (2007) (see also Optimal Domains Theory, Cassimjee and Kisseberth, 1998). Van Oostendorp and van der Weijer, (2005), for example, make the argument based on vocabulary (recall domain-specificity that is required by the modular architecture): OT needs a universe of discourse. That is, you cannot compute nothing (or interchangeable phantoms). Regarding melodic representation, Hall (2001) and Clements (2001) may certainly be counted as expressions of this line of thought as well.
7. Conclusion

I try to condense in (2) below what I take to be the important issues and the true advances that have been made in the field and that will probably still play a role in 50 years.

(2) Important issues

a. the system
   systemic pressure on phonological events and their implementation into a formal system
b. melodic primes
   binary, privative or a blend thereof?
c. type of relatedness of alternating forms
   1. computational: by a phonological process (common underlying form) or by a non-phonological process (allomorphy, suppletion, analogy)
   2. non-computational: two separate lexical entries
d. phonological computation
   1. type of computational instruction
      rules (specifying the modification of the input and its triggering conditions) or constraints (general requirement or prohibition with no modificational instruction?)
   2. serial, parallel or a blend thereof? If serial, what kind of serialism: extrinsically ordered rules, harmonic serialism? If parallel, what kind of parallelism: hard or violable constraints, if violable, dominance expressed by weight or by rank?

None of these issues is settled, and I doubt that (2c) will ever be. For (2a) is not about whether or not phonological processes are under the spell of systemic pressure: hardly anybody doubts that they are; the question is how exactly this could be expressed in a formal system.

The list under (3) below gives what seem to me are true achievements and advances of generative phonology.

(3) Achievements and advances

a. cyclic derivation
b. modular architecture of grammar
c. interactionism
d. autosegmental representations, that is, which can be ill-formed, act as arbitrator and are not necessarily the result of computation
e. privative melodic primes
f. constraints and parallel computation.

This list may be divided into two groups. On the one hand, cyclic derivation and the modular architecture of grammar are the generative baseline: they were present upon inception of the generative enterprise, and they concern architectural properties of the grammar that lie beyond narrowly phonological concerns. Also, while modularity is shared with structuralism (Level Independence), cyclic derivation is a genuinely generative discovery. It is difficult to see how anything could be called generative that does not endorse these two properties.

It was shown that OT in its present state freely violates modularity, and that this appears to be the application of the D (distributed: the scrambling trope) of PDP, which itself is a direct consequence of the P (parallel computation). It remains to be seen whether Prince and Smolensky's (2004 [1993]) cherry-picking (we are true generativists and hence reject the empiricist freight of connectionism, except for parallel computation) can bear fruit.

The other group of the list under (3) encompasses items that have appeared in the course of the development of generative theory. Two of them, (3e-f) also appear in the list of issues: there is still debate whether they exist at all and some believe they do not, but in any case the existence of this debate alone and the issues raised are valuable progress. While constraints are used by all theories in one form or another (recall that well-formedness conditions of the 1980s such as the OCP are a form of constraints), it is true that strictly rule-based approaches may reject the idea of parallel computation altogether. In any event, the parallel idea is an entirely new perspective on computation that has injected fresh air into the settled waters of the serial model of Standard Cognitive Science in general, and of its linguistic outgrowth in particular. Note that unlike modularity, serialism is not an essential of the rational conception of the mind: it is simply the default of the 1950s that was introduced by the Turing/von Neumann machine. All other things being equal (and this is the question raised by cherry-picking), computation could be parallel as well. Generative syntax has shown that non-serialism is possible within the generative paradigm (recall, however, that the constraints on syntactic well-formedness are hard, rather than ranked and violable).

Finally, the two other items of the second group, (3c-d) are the major scientific advances of generative phonology in my opinion. Again, one concerns an architectural property: interactionism was invented by phonologists in Lexical Phonology but has made a brilliant career in syntax: it is the spine of current minimalist thinking (under the name of derivation by phase). The other item is
the face of modern phonology: autosegmental representations are certainly the least common denominator of all currently practiced theories, even if they are demoted to pure decoration in (some versions of) OT.

This record appears to be quite positive and encouraging, an impression that needs to be tuned down, though, when we consider that generative phonology seems to go in circles with respect to some fundamental issues, and spends quite some time reinventing the wheel. The systemic question was thrown over board in the 1960s without argument, and reappeared on the research agenda only recently. But worse is the situation regarding Anderson's (1985) concern for a reasonable balance between representation and computation. Generative phonology has made a complete loop (which is also described by van der Hulst 2004: computational SPE → super-representational 1980s → super-computational OT) and is maybe engaging into another round (cf. the still timid return of representations within and outside of OT). The parallel with syntax in this area is quite striking: like phonology, syntax was anti-lexical and very computationally oriented in the 1960s, explored the lexicalist and representational track in the 1970s and 1980s, but has gone back to extreme anti-lexicalism and proceduralization since the turn of the minimalist programme.

It would be nice if phonology (and syntax for that matter), like adult science, could be said to follow a linear trajectory from less to more knowledge, in a cumulative movement that builds on and learns from the experience and errors of the past. This has certainly not been the case thus far (regular assertions of the contrary from OT quarters notwithstanding), but maybe the see-saw movement can be broken this time: we do not know where exactly the red line runs, but there is good reason to believe that both representations and computation are needed in order to make a grammar.

8. Notes


2. The revolutionary metaphor is commonplace in linguistic historiography and, as Koerner (2002b) points out, part of generative story-telling. Koerner dwells on the (in)adequacy of this reference at greater length, also in regard to Kuhn's model.

3. A fair debate is about whether we are talking about sub-modules, or just about an aggregation of independent modules whose collaboration happens to produce language. The same issue arises in other areas of the cognitive system; we know today that vision, for example, is made of different sub-systems that compute colour, shape, face recognition and movement independently.

Another aspect of the debate regarding language-internal computational units is what Chomsky (1981) has called the sub-systems of GB modules (case theory, theta...

4. Syntacticians usually do not go beyond statements such as 'and then PF takes over'. In the recent minimalist environment, they actually use PF as a dustbin for things that they want to get rid of in order to have a 'clean' syntax (e.g. deletion of words or even entire phrases at PF). The resulting 'dirty' phonology, then, is not anything that they are much concerned with (see Scheer forthcoming a).

5. Whether syntax and morphology are the same or two distinct computational systems is a debated issue that is orthogonal to the present discussion.

6. Bermúdez-Otero (forthcoming b) provides an informed overview of cyclicity and the role it plays in the development of generative phonology (see also Scheer forthcoming a).

7. Though not for all: today substance-free phonology (Halle and Reiss, 2008; Blaho, 2008) holds that substance, that is, melodic properties of segments, is irrelevant for phonological computation; substance is the subject matter of phonetics (if phonology talks about substance, it is liable for substance abuse). Hence regarding melody, anything and its reverse can be turned into anything and its reverse in any context and its reverse. So-called crazy rules (Bach and Harns, 1972) are a relevant factor in the discussion (Scheer, forthcoming b).

8. This principle has later been made explicit in Government Phonology: 'non-arbitrariness: There is a direct relation between a phonological process and the context in which it occurs' (Kaye et al., 1990: 194).


10. Von Neumann (1976 [1971]) is quoted as a 1971 manuscript by Hooper (1975) and Zwicky (1974). Even though Stampe's Ph.D. is typically referred to as an important source in the NGP literature, it appears that the development of NP and NGP was parallel, rather than based on a common ancestor.

11. See Newmeyer (1998) for an overview of what he takes to be the two basic approaches to linguistics, formalist and functionalist thinking. There is no place to further discuss the differences between the two Natural Phonologies. Laks (2006) offers more material on this issue (see also Hooper, 1975: 544ff.) and discusses their further evolution: while NGP has no modern offspring (Joan Hooper, who today publishes as Joan Bybee, is engaged in a usage-based, that is, empiricist approach, and thus has left generative grounds altogether), NP continues to be actively developed (see the references quoted).

12. Cole (1995) and Bermúdez-Otero (forthcoming a) provide a good overview of this movement, which is also explained by Kiparsky himself in Kiparsky (1982b, 1993), and by Anderson (1981: 530ff.).

13. Today Trisyllabic Shortening is typically not considered a synchronically active process: it faces quite a number of counterexamples such as e.g. [wibbi:] – wibbi: (wibbi:), Hayes (1995) and Green (2007: 172ff.) provide an informed review of its status.

14. Cole (1995: 72) discusses the two distinct origins of Kiparsky's SCC at greater length, and Scheer (forthcoming a, in press a) follows the development of all no look-back devices from Chomsky (1975) until the modern syntactic Phase Impenetrability Condition. Finally, it is to be noted that the marriage of Chomsky's SCC with derived environment effects was first proposed by Halle (1978).
Continuum Companion to Phonology

15. On the syntactic side, the precursor of the PIC was Riemsdijk’s (1978: 160) Head Constraint, which does the same labour as the PIC, but not for the same reasons (see Scheer, forthcoming a).


17. This is only a rough and incomplete formulation of the phonological ECP, which also does not mention the central notion of government; see Uegver and Koe (1993), Scheer (2004) for details.

18. The linear solution was to add extra items to the vowel inventory, i and y, that were then either transformed into [e] (‘lowered’) or deleted, which is subject to absolute neutralization.

19. Constraints such as this, however, are probably as old as linguistic analysis: so-called morpheme structure constraints defined what a possible morpheme is in SPE, and structuralism as much as neogrammarians thinking appealed to restrictions, requirements and prohibitions. Van der Kruyt (2004) provides an overview of the generative record regarding constraints.

20. There was an attempt to do away with extrinsically ordered rules in the 1970s around the work of Andreas Koutsoudas (Koutsoudas et al., 1974; Koutsoudas, 1976a); Koutsoudas claimed that there is no empirical basis for extrinsic rule ordering: its effects can be derived (and predicted) from other factors. The evidence presented was not judged convincing, though, and the proposal did not have any portent. In any event, it was entirely unrelated to the representationally dominated environment of the 1980s in general and the 1980s anti-particulism in particular.

21. Bremmer and Hall argue along the lines of the strongest case, which they claim is the comparison of the well-attested phenomenon of Canadian raising in Canadian English (dialect A) with another Canadian dialect (dialect B) whereby the only difference is the order of application of two rules. Unfortunately, the only source for dialect B is a three-page article by Jorns (1942), whose informants were among his pupils in an Ontario public school. Despite extensive study of Canadian varieties by Canadian dialectologists such as Chambers (1972), though, no trace of dialect B could be found some 30 years later. Despite the arguments of Koe (1990), the Canadian raising case continues to spook though the literature as alleged support for rule ordering.

22. Mohan (2000: 146) argues that they do not because one can always be translated into the other. The point, however, is not whether the ‘correct result’ can be achieved, rather, the way in which this result is achieved matters, and is different. See the informative discussion in Hale and Reiss (2008: 195f.) on this issue.

23. The only exception that I am aware of is van Oostendorp and van de Weijer’s (2005) attempt to define what they call a universe of discourse for the expression of instructions in OT.

24. I use quotation marks in order to refer to this framework because its name falls apparently suggests that it has a copyright on cognitive aspects of grammar, and that anything which is non-Langackerian must be non-cognitive.

25. For example, Lombard (2001a: 3) writes with respect to melodic representation: ‘the tenets of OT, regarding constraint violability and ranking, make no particular claims about phonological representations. We could, for example, do OT with any kind of feature theory; SPE feature bundles or feature geometric representations, privative or binary features, and so on.’

26. An eloquent example is the ‘representation’ of melody that is currently practiced in OT: binary features are about as sophisticated as they were in SPE (i.e. an amorphous, unarticulated set), and constraints that manipulate melody boil down to ‘star [feature]’ (e.g. *dorsal etc.). Clements (2001) and Hall (2001) bemoan this evolution.

Bibliography


Bibliography


Feest, S.V.H. van der and P. Fikkert (under revision). Learning to represent phonological contrasts.


Bibliography


Bibliography


Kabak, B., K. Maniwa and N. Kazanina (to appear) Listeners use vowel harmony and word-final stress to spot nonsense words: A study of Turkish and French. Laboratory Phonology 1.


Legendre, G., Y. Miyata and P. Smolensky (1990b) Harmonic grammar — a formal multi-level connectionist theory of linguistic well-formedness: Theoretical


Padgett, J. (2001a) Constraint conjunction vs. grounded constraint subhierarchies. Ms., University of California, Santa Cruz.


Bibliography


Smith, N. and D. Apousiouds (in prep.) Recessive accent in Ancient Greek revisited: The final coda effect.


Bibliography


Bibliography


Language Index

<table>
<thead>
<tr>
<th>Language/Sign Language</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adamorobe Sign Language</td>
<td>275</td>
</tr>
<tr>
<td>Al-Sayyid Bedouin Sign Language</td>
<td>255, 275</td>
</tr>
<tr>
<td>American Sign Language (ASL)</td>
<td>255, 257–8, 268–9</td>
</tr>
<tr>
<td>Arabic</td>
<td>314</td>
</tr>
<tr>
<td>Cairene Arabic</td>
<td>75, 93</td>
</tr>
<tr>
<td>Classical Arabic</td>
<td>152</td>
</tr>
<tr>
<td>Egyptian Arabic</td>
<td>311</td>
</tr>
<tr>
<td>Jordanian Arabic</td>
<td>311</td>
</tr>
<tr>
<td>Moroccan Arabic</td>
<td>83–4</td>
</tr>
<tr>
<td>Palestinian Arabic</td>
<td>94</td>
</tr>
<tr>
<td>Armenian, New Julfa</td>
<td>150</td>
</tr>
<tr>
<td>Barra Gaelic</td>
<td>378</td>
</tr>
<tr>
<td>Berber (Tashliht)</td>
<td>197</td>
</tr>
<tr>
<td>British Sign Language (BSL)</td>
<td>255, 260–4, 266–7, 273</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>97</td>
</tr>
<tr>
<td>Cama (Ebrìë)</td>
<td>57</td>
</tr>
<tr>
<td>Catalan</td>
<td>305, 388, 390</td>
</tr>
<tr>
<td>Chukwulunu</td>
<td>237</td>
</tr>
<tr>
<td>Chinese</td>
<td>304, 312, 314–15</td>
</tr>
<tr>
<td>Cantonese Chinese</td>
<td>300, 307</td>
</tr>
<tr>
<td>Mandarin Chinese</td>
<td>307</td>
</tr>
<tr>
<td>Shanghai Chinese</td>
<td>238</td>
</tr>
<tr>
<td>Xianmen Chinese</td>
<td>237</td>
</tr>
<tr>
<td>Czech</td>
<td>72, 312, 315</td>
</tr>
<tr>
<td>Dakota</td>
<td>97</td>
</tr>
<tr>
<td>Dutch</td>
<td>23, 89, 93, 98, 218–21, 279, 333–4, 388</td>
</tr>
<tr>
<td>Sign Language</td>
<td>255, 272</td>
</tr>
<tr>
<td>American English</td>
<td>77, 344, 389</td>
</tr>
<tr>
<td>Canadian English</td>
<td>138, 144, 169, 424</td>
</tr>
<tr>
<td>Middle English</td>
<td>45, 389</td>
</tr>
<tr>
<td>Modern English</td>
<td>76</td>
</tr>
<tr>
<td>Old English</td>
<td>406</td>
</tr>
<tr>
<td>Scottish Standard English</td>
<td>77</td>
</tr>
<tr>
<td>Singapore English</td>
<td>151</td>
</tr>
<tr>
<td>Standard British English</td>
<td>77</td>
</tr>
<tr>
<td>Standard Southern British English</td>
<td>77</td>
</tr>
<tr>
<td>English, Tyneside</td>
<td>86, 77</td>
</tr>
<tr>
<td>Filipino</td>
<td>282</td>
</tr>
<tr>
<td>Finnish</td>
<td>97, 99, 161, 313</td>
</tr>
<tr>
<td>French</td>
<td>64, 77–8, 83, 98, 184, 227–8, 284, 301, 312, 314–15, 349–50, 363, 417, 421</td>
</tr>
<tr>
<td>Canadian French</td>
<td>146, 167, 304, 311</td>
</tr>
<tr>
<td>Sign Language</td>
<td>274, 276</td>
</tr>
<tr>
<td>Standard French</td>
<td>149</td>
</tr>
<tr>
<td>Greek</td>
<td>100–1, 233, 245, 249</td>
</tr>
<tr>
<td>Athenian Greek</td>
<td>230</td>
</tr>
<tr>
<td>Ancient Greek</td>
<td>69, 101, 105, 405</td>
</tr>
<tr>
<td>Modern Greek</td>
<td>95, 100–1, 106</td>
</tr>
<tr>
<td>Hausa</td>
<td>240–1, 251</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>188</td>
</tr>
<tr>
<td>Hungarian</td>
<td>75, 83, 311, 313–14</td>
</tr>
</tbody>
</table>