1 Themes in Phonology

Paul de Lacy
Rutgers University

1.1 Introduction[*]

This chapter has two aims. One is to provide a brief outline of the structure of this book, and this is the focus of §1.1.1. The other – outlined in §1.1.2 – is to identify several of the major themes that run throughout.

1.1.1 Structure

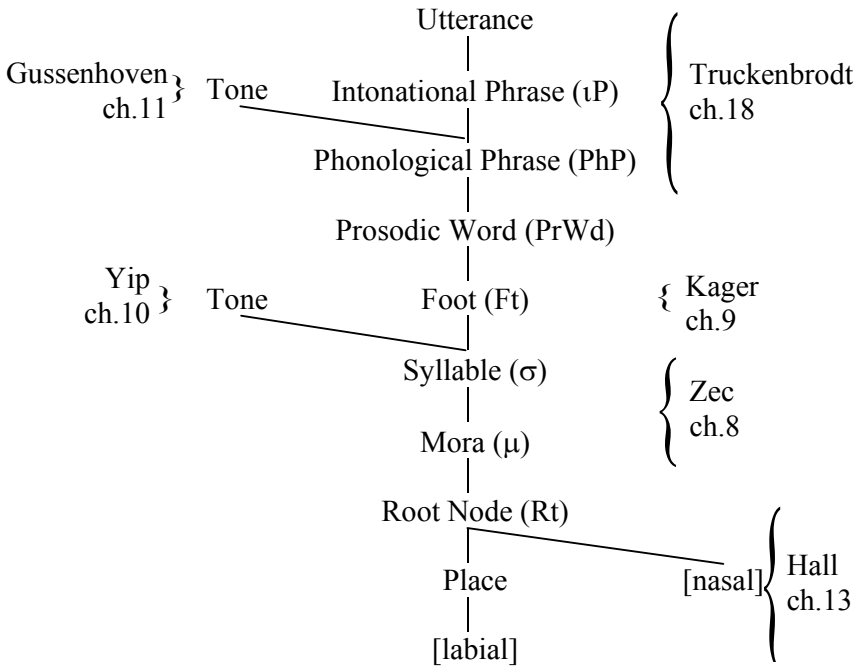
Several different factors have influenced the contents and structure of this Handbook. The topics addressed reflect theoretical concerns that have endured in phonology, but they were also chosen for pedagogical reasons (i.e. many advanced phonology courses cover many of the topics here). There were also 'traditional' reasons for some aspects of organization. While these concerns converge in the main, there are some points of disagreement. For example, there is a traditional distinction between the phonology of lexical tone and intonation, hence the separate chapters by Yip [ch.10] and Gussenhoven [ch.11]. However, Gussenhoven [§11.7] comments that theoretically such a division may be artificial.

Consequently, it is not possible to identify a single unifying theoretical theme that accounts for the structure of this book. Nevertheless, the topics were not chosen at random; they broadly reflect many of the current concerns of the field. In a broad sense, these concerns can be considered in terms of representation, derivation, and the trade-off between the two. 'Representation' refers to the formal structure of the objects that phonological component manipulates. 'Derivation' refers to the relations between those objects.

Concern with representation can be seen throughout the following chapters. Chomsky & Halle (1968) (*SPE*) conceived of phonological representation as a string of segments, which are unordered bundles of features. Since then, representation has become more elaborate. Below the segment, it is widely accepted that features are hierarchically organized (see discussion and references in Hall [ch.13]). Above the segment, several layers of constituents are now commonly recognized, called the 'prosodic hierarchy' (Selkirk 1984). Diagram (1) gives a portion of an output form's representation; it categorizes the chapters of this book in terms of their representational concerns. There is a great deal of controversy over almost every aspect of the representation given below – diagram (1) should be considered a rough expositional device here, not a theoretical assertion; the chapters cited should be consulted for details.

[*] My thanks to all those who commented on this chapter in its various incarnations: José Elías-Ulloa, Kate Ketner, John McCarthy, Nazarré Merchant, Michael O'Keefe, and Alan Prince.

(1)

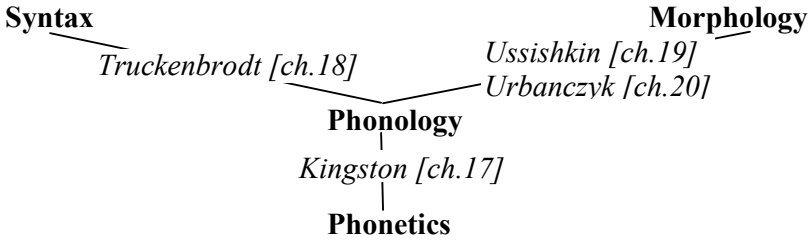


Harris [ch.6] should be added to the chapters cited in (1); Harris' chapter is concerned with broader principles behind representation, including the notion of constituency, whether certain subconstituents are phonological prominent (i.e. headedness), and hierarchical relations.

Not represented in (1) is the interaction between constituents. For example, de Lacy [ch.12] is about the interaction of tone, the foot, and segmental properties. Similarly, a part of Kager [ch.9] is about the relation between the foot and its subconstituents. At the segmental level, three chapters are concerned with the interaction of segments and parts of segments: Baković [ch.14], Archangeli & Pulleyblank [ch.15], and Alderete & Frisch [ch.16]. For example, Baković's chapter discusses the pressure for segments to have identical values for some feature (particularly Place of Articulation).

Diagram (2) identifies the chapters that are concerned with discussing the interaction of different representations. For example, Truckenbrodt [ch.18] discusses the relation of syntactic phrases to phonological phrases. Ussishkin [ch.19] and Urbanczyk [ch.20] do the same for the relation of morphological and phonological structure. Kingston [ch.17] discusses the relation of phonological to phonetic structures.

(2)



There is also a 'derivational' theme that runs through the book chapters. McCarthy [ch.5] focuses on evidence that there are relations between forms, and the nature of those relations. Discussion of derivation has traditionally focused on the relation

between input and output forms, and between members of morphological paradigms. However, the traditional conception of derivation has been challenged in Optimality Theory by McCarthy & Prince's (1995, 1997) Correspondence Theory – the same relations that hold between separate forms (i.e. input~output, paradigmatic base~derivative) also hold in the same output form between reduplicants and their bases; thus Urbanczyk's [ch.20] discussed of reduplication can be seen as primarily about derivation.

Of course, no chapter is entirely about the representation of constituents; all discuss derivation of those constituents. In serialist terms, 'derivation of constituents' means the rules by which those constituents are constructed. In parallelist (e.g. Optimality Theoretic) terms, it in effect refers to the constraints and mechanisms that evaluate competing representations.

There are a set of chapters whose primary concerns relate to both representation and derivation: Gordon [ch.3], Rice [ch.4], and Steriade [ch.7] discuss topics that are in effect meta-theories of representation and derivation. Gordon [ch.3] examines functionalism – a name for a set of theories that directly relate or derive phonological representations (and potentially derivations) to phonetic concerns. Rice [ch.4] discusses markedness, which is effectively a theory of possible phonological representations and derivations. Steriade [ch.7] discusses the idea of phonological contrast, and how it influences representation and derivation.

Rice's discussion of markedness makes the current tension between representation- and derivation- based explanations particularly clear. Broadly speaking, there have been two approaches to generalizations like "an epenthetic consonant is often [ʔ]". One assigns [ʔ] a representation that is different (often less elaborate) than other segments; the favouring of epenthetic [ʔ] over other segments is then argued to follow from general derivational principles of structural simplification. The other is to appeal to derivational principles such as (a) constraints that favour [ʔ] over every other segment and (b) no constraint that favours those other segments over [ʔ]; [ʔ] need not be representationally simple (or otherwise remarkable) in this approach. These two approaches illustrate how the source of explanation – i.e. derivation and representation – is still contentious. The same issue is currently true of subsegmental structure – elaborated derivation mechanisms may allow simpler representational structures (Yip 2004).

Part 5 of this book contains a diverse array of phonological phenomena which do not fit easily into the themes of representational and derivational concerns. Instead, their unifying theme is that they are all areas which have been the focus of a great deal of recent attention and have provided significant insight into phonological issues; this point is made explicitly by Fikkert [ch.23] for language acquisition, but also applies to the other areas: diachronic phonology (Bermúdez-Otero [ch.21]), free variation (Anttila [ch.22]), learnability (Tesar [ch.24]), and phonological disorders (Bernhardt & Stemberger [ch.25]). There are many points of interconnection between these chapters and the others, such as the evidence that phonological disorders and language acquisition provide for markedness.

In summary, no single theoretical issue accounts for the choice of topics and their organization in this book. However, many themes run throughout the chapters; the rest of this chapter identifies some of the more prominent ones.

1.1.2 Summary of Themes

One of the clearest themes seen in this book is the influence of Optimality Theory (OT), proposed by Prince & Smolensky (1993/2004).[1] The majority of chapters discuss OT, reflecting the fact that the majority of recent research publications employ this theory and a good portion of the remainder critique or otherwise discuss it.[2] However, one of the sub-themes found in the chapters is that there are many different conceptions and sub-theories of OT, although certain core principles are commonly adopted. For example, some theories employ just two levels (the input and output), while others employ more (e.g. Stratal OT – McCarthy [§5.4]). Some employ a strict and totally ordered constraint ranking, while others allow constraints to be unranked or overlap (see Anttila [§22.3.3] and Tesar [§24.4] for discussion). Theories of constraints differ significantly among authors, as do conceptions of representation (see esp. Harris [ch.6]).

Another theme that links many of the chapters is the significance of representation and how it contributes to explanation. The late 1970s and 1980s moved towards limiting the form of phonological rules and elaborating the representation by devices such as autosegmental association, planar segregation, lack of specification, and feature privativity. In contrast, Harris [§6.1] observes that the last decade has seen increased reliance on constraint form and interaction as sources of explanation. Constraint interaction as an explanatory device appears in many of the chapters. Section 1.3 summarizes the main points.

Section 1.4 discusses the increasing influence of Functionalism in phonology, a theme that is examined in detail by Prince [ch.2] and Gordon [ch.3]. Reference to articulatory, perceptual, and parsing considerations as a source of phonological explanation is a major change from the Formalist orientation of *SPE* and its successors. This issue recurs in a number of chapters, some explicitly (e.g Harris [§6.2.2], Steriade [§7.5]), and in others as an implicit basis for evaluating the adequacy of constraints.

Of course, the following chapters identify many other significant themes in current phonological theory; this chapter focuses solely on the ones given above because they recur in the majority of chapters and are presented as some of the field's central concerns.

1.2 The Influence of Optimality Theory

Optimality Theory is explicitly discussed or assumed in many chapters in this volume, just as it is in a great deal of current phonological research ('current' here refers to the time of writing – the middle of 2005). This section starts by reviewing OT's architecture and core properties. The following sections identify particular aspects that prove significant in the following chapters, such as the notion of faithfulness and its role in derivation in §1.2.1, some basic results of constraint interaction in §1.2.2, and its influence on conceptions of the lexicon in §1.2.3. The sections identify some of the challenges facing OT as well as its successes and areas which still excite controversy. The relation of OT to other theories is discussed in §1.2.4.

---

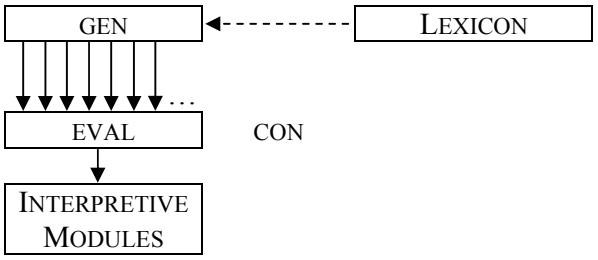
[1] A version of Prince & Smolensky's manuscript is available online for free at the Rutgers Optimality Archive (ROA): http://roa.rutgers.edu/, number 537.

[2] From inspecting several major journals from 1998 to 2004, around three quarters of the articles assumed an OT framework, and many of the others compared their theories with an OT approach.

• *OT Architecture*

OT is a model of grammar – i.e. both syntax and phonology (and morphology, if it is considered a separate component); the following discussion will focus exclusively on the phonological aspect and refer to the model in (3).

(3) *OT architecture*



For phonology, the GEN(erator) module takes its input either directly from the lexicon or from the output of a separate syntax module. GEN creates a possibly infinite set of candidate output forms; the ability to elaborate on the input without arbitrary restraint is called 'freedom of analysis'. In Prince & Smolensky's original formulation, every output candidate literally contained the input; to account for deletion, pieces of the input could remain unparsed (i.e. not incorporated into the representation) which meant they would not be phonetically interpreted. Since McCarthy & Prince (1995a/1999), the dominant view is that output candidates do not contain the input, but are related to it by a formal relation called 'correspondence'; see section 1.2.1 for details (cf. Goldrick 2000).

One significant restriction on GEN is that it cannot alter the morphological affiliation of segments ('consistency of exponence' – McCarthy & Prince 1993b). In practice it is common to also assume that GEN requires every output segment to be fully specified for subsegmental features, bans floating (or 'unparsed') features (except for tone – Yip [§10.2.2], Gussenhoven [§11.5.1]), and imposes restrictions on the form of prosodic and subsegmental structure (though in some work they are considered violable – e.g. Selkirk 1995, Crowhurst 1996, cf. Hyde 2002).

The EVAL(uator) module determines the 'winner' by referring to the constraints listed in CON (the universal constraint repository) and their language-specific ranking. Constraints are universal; the only variation across languages is (a) the constraints' ranking, and (b) the content of the lexicon. The winner is sent to the relevant interpretive component (the 'phonetic component' for phonology – Kingston [ch.17]).

There are two general types of constraint: Markedness and Faithfulness. Markedness constraints evaluate the structure of the output form, while Faithfulness constraints evaluate its relationship to other forms (canonically, the input – see McCarthy [ch.5]).[3] As an example, the Markedness constraint ONSET is violated once for every syllable in a candidate that lacks an onset (i.e. every syllable that does not start with a consonant – Zec [§8.3.2]). [ap.ki] violates ONSET once, while [a.i.o] violates it three times. The Faithfulness constraint I(nput)O(utput)-MAX is violated

[3] There is currently an ambiguity in the term 'markedness'. In OT, 'markedness' refers to a type of constraint. 'Markedness' also refers to a concept of implicational or asymmetric relations between phonological segments and structures (see §1.3 and Rice [ch.5]).

once for every input segment that does not have an output correspondent: e.g. /apki/ → [pi] violates IO-MAX twice (see §1.2.1 for details).

In each grammar the constraints are totally ranked (although evidence for their exact ranking may not be obtainable in particular languages). Constraints are violable; the winner may – and almost certainly will – violate constraints. However, the winner violates the constraints 'minimally' in the sense that for each losing candidate L (a) there is some constraint K that favours the winner over L and (b) K outranks all constraints that favour L over the winner (a constraint 'favours' *x* over *y* if *x* incurs fewer violations of it than *y*).

• *Tableaux*

The mapping from an underlying form like /oap/ to an output [a] is represented by a tableau, as in (4). The faithful form [.o.ap] violates ONSET twice because it has two syllables without an initial consonant, as does [.o.a]. Candidates (c) and (d) both violate ONSET once, but this constraint can help no further in determining the winner because it fails to distinguish between the two candidates. In contrast, NOCODA, which bans syllable-final consonants, is violated by (c) but not (d), crucially favoring (d). It does not matter that (d) violates IO-MAX more than the other candidates because they all violate higher-ranked constraints more. The dotted line between ONSET and NOCODA indicates that their relative ranking cannot be determined using the available evidence; so, both the ranking || ONSET » NOCODA » IO-MAX || and || NOCODA » ONSET » IO-MAX || are equally consistent with the data.

(4) A 'classical' tableau

/oap/	ONSET	NOCODA	IO-MAX
(a) .o.ap	* *!	*	
(b) .o.a	* *!		*
(c) .ap	*	*!	*
☞ (d) .a	*		* *

To complete the analysis, a candidate in which everything deletes (∅) would have to be ruled out by a constraint ranked higher than ONSET. Similarly, [ʔoʔap], with epenthesis, must be eliminated (e.g. by IO-DEP), and [.o] by some other constraint (e.g. *NUC/e,o – Zec [§8.5.1.1]). There are many other candidates, though a large proportion can be ruled out as perpetual losers as they incur a proper superset of the violations of other candidates (Samek-Lodovici & Prince 1999).

The tableau form in (4) was introduced by Prince & Smolensky (1993/2004) and is the most widely used way of representing candidate competition. Another method is proposed by Prince (2002a), called the 'Comparative Tableau'; it is used in this book by Prince [ch.2], Baković [ch.14], and Tesar [ch.24]. The Comparative Tableau's advantage is that it represents competition directly, rather than indirectly by numbers of violations. The leftmost column lists the desired winner followed by a competitor. A 'W' indicates that the constraint prefers the desired winner, a blank cell indicates that the constraint makes no preference, and an L indicates that the candidate favors the loser. In visual terms, every row should have at least one W before any L.

(5) A comparative tableau

/oap/	ONSET	NOCODA	IO-MAX
.a ~ .o.ap	W	W	L
.a ~ .o.a	W		L
.a ~ .ap		W	L

Bernhardt & Stemberger (1998) propose another way of representing tableaux that is similar to the classic form; see chapter 25 for details.

• *Core principles*

Prince & Smolensky (1993/2004) identify core OT principles for computing input→output mappings, including freedom of analysis, parallelism, constraint violability, and ranking. As they observe, many theories of CON and representation are compatible with these principles. Consequently, a great deal of work in OT has focused on developing a theory of constraints; for proposals regarding other principles, see §1.2.4.

The dominant theories before OT – *SPE* and its successors – employed rules and a 'serial' derivation. For them, the input to the phonological component underwent a series of functions ('rules') that took the previous output and produced the input to the next until no more rules could apply. For example, /oap/ would undergo the rule $C \rightarrow \varnothing / \_]_{\sigma}$ to produce [oa] which would then serve as the input to the rule $V \rightarrow \varnothing /_{\sigma}[\_$ to produce [a]. The rule-based derivation is described in detail in McCarthy [ch.5]. In contrast, the winner in OT is determined by referring to the entire constraint hierarchy and by comparison with (in principle) the entire candidate set (McCarthy & Prince 1993b:ch.1§1).

Certainly, other theories had and have since proposed such concepts as constraints and two- or three-level grammars (e.g. Theory of Constraints and Repair Strategies – Paradis 1988; Harmonic Phonology – Goldsmith 1993a, Two-level Phonology – Koskenniemi 1983, Karttunen 1993; Declarative Phonology – Scobbie 1992, Coleman 1995, Scobbie, Coleman, and Bird 1996). However, OT's combination of these ideas and the key notions of constraint universality, ranking, and violability proved to have almost immediate appeal.

The following sections discuss aspects of the theory that recur or are assumed in many of the following chapters. Section 1.2.1 discusses derivation, correspondence, and faithfulness. Section 1.2.2 discusses the form of the constraint component CON and some important constraint interactions while §1.2.3 examines OT's influence on the concept of the lexicon. Section 1.2.4 discusses the several different versions of OT that currently exist and its relation to other extant phonological theories.

1.2.1 Derivation and Faithfulness

A concept that recurs throughout the following chapters is 'faithfulness' – it is discussed explicitly by McCarthy [ch.5] and faithfulness constraints are used in many of the discussions of empirical phenomena.

In *SPE* and the theories that adopted its core aspects of rules and rule-ordering, there is no mechanism that requires preservation of input material. If input /abc/

surfaces as output [abc], the similarity is merely an epiphenomenon of rule non-application: either all rules fail to apply to /abc/, or the rules that apply do so in such a way to inadvertently produce the same output form as the input.

McCarthy & Prince (1995a, 1999) propose a reconceptualization of identity relations. Segments in different forms can stand in a relation of 'correspondence'. For example, the segments in an input /$k_1æ_2t_3$/ and winning faithful output [$k_1æ_2t_3$] are in correspondence with one another, where subscript numerals mark these relations. Equally, the segments in an unfaithful pair, /$k_1æ_2t_3$/ → [$d_1ɔ_3g_2$], still correspond with one another, even though in this case two segments have metathesized and all have undergone drastic featural change. In keeping with 'freedom of analysis', correspondence relations can vary freely among candidates. For example, input /$k_1æ_2t_3$/ has the outputs [$k_1æ_2t_3$], [$k_1æ_2$] (deletion of /t/), [$k_1æ_2t_3i$] (epenthesis of *i*), [$k_1t_3æ_2$] (metathesis of /æt/), [$k_1\underset{\circ}{æ}_{2,3}$] (coalescence of /æ/ and /t/ to form [$\underset{\circ}{æ}$]), and combinations such as [$t_3æ_2$] (metathesis of /æt/ and deletion of /k/) and forms that are harmonically bounded (i.e. can never win) such as [$k_3æ_3t_2$], and so on.

Constraints on faithfulness regulate the presence, featural identity, and linear order of segments. The ones proposed in McCarthy & Prince (1995a) that appear in this book are given in (6).

(6) *Faithfulness constraint summary (from McCarthy & Prince 1995a)*

(a) Faithfulness constraints on segmental presence (e.g. Zec [§8.3.2])

MAX "Incur a violation for each input segment *x* such that *x* has no output correspondent" (Don't delete)

DEP "Incur a violation for each output segment *x* s.t. *x* has no input correspondent" (Don't epenthesize)

(b) Faithfulness constraints on featural identity (e.g. Steriade [§7.4.3])

IDENT[F] "Incur a violation for each input segment *x* s.t. *x* is [αF] and *x*'s ouput correspondent is [−αF]" (Don't change F values).

(c) Faithfulness constraints on linear order (e.g. de Lacy [§12.6])

LINEARITY "For every pair of input segments *x,y* and their output correspondents *x',y'*, if *x* precedes *y* incur a violation if it is not the case that *y'* precedes *x*'." (No metathesis)

(d) Faithfulness constraints on one-to-many relationships (e.g. Yip [§10.3.3])

UNIFORMITY "Incur a violation for each output segment that corresponds to more than one input segment" (No coalescence)

McCarthy & Prince (1995a, 1999) argue that correspondence relations can also hold within candidate outputs, specifically between reduplicative morphemes and their bases. Consequently, the candidate [$\underline{p_1a_2}p_1a_2t_3a_4$], where the underlined portion is the reduplicant, indicates that the reduplicant's [p] corresponds to the base's [p], and the reduplicant's [a] to the base's. This proposal draws a direct link between the identity effects seen in input→output mappings and those in base-reduplicant relations. Other elaborations of faithfulness are discussed in §1.2.2.

• *Parallelism*

Faithfulness relates to the concept of parallelism: there is essentially a 'flat derivation' with the input related directly to output forms. As the chapters show, a lot of the

success and controversy over parallelism arises in 'local' and 'non-local' interactions. One success is in its resolution of ordering paradoxes found in rule-based approaches. For example, in Ulithian reduplication of /xas/ surfaces as [kakkasi] (Sohn & Bender 1973:45). Coda consonants assimilate to the following consonant, preventing the output from being *[xasxasi]. However, the form does not become the expected *[xaxxasi] because [xx] is banned. Instead, the resulting output is [kakkasi] – this form avoids [xx], satisfies the conditions on codas, and at the same time ensures that the reduplicant is as similar to the base as possible by altering the base's consonant from /x/ to [k].

The ordering paradox can be illustrated by a serialist rule-based analysis in (7). For the reduplicant to copy the base's [k] in [kakkasi], copying would have to be ordered *after* gemination and consequent fortition; however, reduplication *creates* the environment for gemination and fortition.

(7) *A serialist approach to Ulithian reduplication*

Input:	/RED+xasi/
(a) Reduplication:	[xas.xa.si]
(b) Gemination:	[xax.xa.si]
(c) [xx] fortition:	*[xak.ka.si]

In contrast, Correspondence Theory (CT) provides an explanation by positing an identity relationship between the base and reduplicant. In tableau (8), CODACOND requires a coda consonant to agree with the features of the following consonant (after Ito 1986). *[xx] bans geminate fricatives. To force the input /x/ to become [k], both CODACOND and *[xx] must outrank IO-IDENT[continuant], a constraint that requires input-output specifications for continuancy to be preserved. Together, CODACOND and *[xx] favour the candidates with a [kk] – i.e. the winner [kak-kasi] and loser *[xak-kasi]. The crucial distinction between these two is that [kak-kasi]'s reduplicant copies its base's continuancy better than *[xak-kasi]'s. In short, the reason that [kak-kasi] wins is due to a direct requirement of identity between base and reduplicant (cf. discussion in Urbanczyk [§20.2.6]).

(8) *Ulithian reduplication in OT*

/RED-xasi/	CODA COND	*[xx]	BR-IDENT [continuant]	IO-IDENT [continuant]
(a) kak-kasi~xas-xasi	W		L	L
(b) kak-kasi~xax-xasi		W	L	L
(c) kak-kasi~xak-kasi			W	

• *Global Conditions*

Other aspects of faithfulness and parallelism have resulted in a great deal of controversy. One involves 'locality of interaction': a rule/constraint seems to apply at several places in the derivation (globality) or only once (opacity).

'Global rules' or 'global conditions' are discussed in detail by Anderson (1974): global conditions recur throughout a serial derivation. An example that I am familiar with is found in Rarotongan epenthesis (Kitto & de Lacy 1998). There is a ban on [ɾi] sequences, and this ban recurs throughout the derivation. So, while the usual epenthetic vowel is [i] (e.g. [kara:ti] 'carrot', [meneti] 'minute', [naeroni] 'nylon'), to avoid a [ɾi] sequence the epenthetic vowel after [ɾ] is a copy of the

preceding one: e.g. [peːɾe] ‘bail’, [ʔaːmaɾa] ‘hammer’, [poːɾo] ‘ball’, [vuːɾu] ‘wool’. In serialist terms, the condition on [i] epenthesis seems straightforward: ∅ → [i]/$C^{[-rhotic]}$__#, followed by a rule ∅ → V_i /V_iɾ__#. The problem is that the ban on [ɾi] ‘recurs’ in the context […iɾ]: if copying the vowel would result in a [ɾi] sequence, [a] is epenthesized as a last resort (e.g. [piɾa] ‘bill’, *[piɾi]). Consequently, the second rule needs to be reformulated as ∅ → V_i /[¬i]ɾ__#, followed by ∅ → [a] elsewhere. These rules miss the point entirely: there is a constraint on [ɾi] sequences that continually guides epenthesis throughout the derivation.

In OT, global conditions are expressed straightforwardly. A constraint on [ɾi] sequences outranks the constraints that would permit [ɾi]. The constraint M(¬i) is a shorthand for the constraints that favour [i] over all other vowels; AGREE(V) requires vowels to harmonize (Baković [ch.14], Archangeli & Pulleyblank [ch.15]). In tableau (9), *[ɾi] is irrelevant because there is no [ɾ]; so the constraint M(¬i) favours [i] as the epenthetic vowel. In tableau (10), *[ɾi] blocks the epenthesis of [i], so the ‘next best’ option is taken – vowel harmony; this is one of the situations in which *[ɾi] blocks epenthesis. Tableau (11) illustrates the other: when harmony would produce an [ɾi] sequence, it is blocked and [a] is epenthesized instead.

(9) *Epenthesize [i] after non-[ɾ]*

/menet/	*ɾi	M(¬i)	AGREE(V)
(a) meneti~menete		W	L
(b) meneti~meneta		W	

(10) *... except when [i] epenthesis would result in [ɾi], then copy*

/vuːɾ/	*ɾi	M(¬i)	AGREE(V)
(a) vuːɾu ~vuːɾi	W	L	W
(b) vuːɾu ~vuːɾa			W

(11) *… unless copying would create [ɾi], in which case epenthesize [a]*

/piːɾ/	*ɾi	M(¬i)	AGREE(V)	M(¬a)
(a) piːɾa ~piːɾi	W	L	L	W
(b) piːɾa ~piːɾe				W

• *Opacity*

However, OT’s success in dealing with global rules raises a problem. In a sense, the opposite of a global rule is one that applies in only one place in the derivation but not elsewhere, even when its structural description is met. Such cases are called ‘opaque’ and can be broadly characterized as cases where output conditions are not surface true. For example, an opaque version of Rarotongan epenthesis would have *ɾi apply only to block default [i]-epenthesis after [ɾ]; it would not block harmony, so allowing /piːɾ/→[piːɾi]. As McCarthy [§5.4] discusses opacity in detail, little will be said about the details here (also see Bermúdez-Otero [§21.3.2] for an example). Suffice to say that it is perhaps the major derivational issue that has faced OT over the past several years and continues to attract a great deal of attention. It has motivated a number of

theories within OT, listed in McCarthy [ch.5], and a number of critiques (e.g. Idsardi 1998, 2000). It is only fair to add that while opacity is seen as a significant challenge for OT, it also poses difficulties for a number of serialist theories: McCarthy (1999, 2003c) argues that serialist theories allow for unattested types of opaque derivation, where the input undergoes a number of rules that alter its form only for the output to end up identical to the input (i.e. 'Duke of York' derivations).

In summary, McCarthy & Prince's (1995a, 1999) theory that there is a direct requirement of identity between different derivational forms and even within forms has resulted in many theoretical developments and helped identify previously unrecognized phonological regularities. The opacity issue remains a challenge for OT, just as ordering paradoxes and global conditions pose problems for serialist rule-based frameworks.

1.2.2 Constraints and their Interaction

Like many of the chapters in this book, a great deal of recent phonological research has been devoted to developing a theory of constraints. This section discusses the basic constraint interactions and subtypes of faithfulness constraint that appear in the following chapters. The form of markedness constraints is intimately tied to issues of representation and Formalist/Functionalist outlook; these are discussed in §1.3 and §1.4 respectively.

• *Faithfulness*

Many of the chapters employ faithfulness constraints that are elaborations of those in (6), both in terms of their dimension of application and environment-specificity.

McCarthy & Prince (1995a,1999) proposed that faithfulness relations held both on the input-output (IO) dimension and between bases and their reduplicants (BR) (see Urbanczyk [ch.20] for more on BR faithfulness). In its fundamentals, McCarthy & Prince's original conception of faithfulness relations have remained unchanged: i.e. the core ideas of regulating segmental presence, order, and identity are still at the core of faithfulness. However, the dimensions over which faithfulness has been proposed to apply have increased. Correspondence relations within paradigms have been proposed by McCarthy (1995/2000c, 2005) and Benua (1997) (see McCarthy [§5.5]), from inputs to reduplicants by Spaelti (1997), Struijke (2000a/2002) and others cited in Urbanczyk [§20.2.6], and correspondence relations within morphemes have been explored by Kitto & de Lacy (1999), Hansson (2001b), and Rose & Walker (2004) (see Archangeli & Pulleyblank [§15.3]).

Others have proposed that there are environment-specific faithfulness constraints. For example, Beckman's (1997, 1998) 'positional faithfulness' theory proposes that constraints can preserve segments specifically in stressed syllables, root-initial syllables, onsets, and roots (also see Lombardi 1999, Casali 1997). For example, ONSET-IDENT[voice] is violated if an onset segment fails to preserve its underlying [voice] value, as in /aba/ → [a.pa] (but not /ab/→ [ap]) (see e.g. Steriade [§7.4.3], Baković [§14.4.3]). There is currently controversy over whether positional faithfulness constraints are necessary, or whether their role can be taken over by environment-specific markedness constraints (Zoll 1998). For further elaborations on the form of faithfulness constraints in terms of environment, see Jun (1995, 2004), Steriade (2001b), and references cited therein.

In addition, some work seeks to eliminate particular faithfulness constraints, such as Keer (1999) for UNIFORMITY and Bernhardt & Stemberger (1998) and [§25.3.4] for DEP.

A significant controversy relates to segment- and feature-based faithfulness. In McCarthy & Prince's (1995a) proposal, only segments could stand in correspondence with each other; a constraint like IDENT[F] then regulates featural identity as a property of a segment. In contrast, Lombardi (1999) and others have proposed that features can stand directly in correspondence – a constraint like MAX[F] requires that every feature have an output correspondent. The difference is that the MAX[F] approach allows features to have a life of their own outside of their segmental sponsors. Consequently, the mapping /pa/→[a] does not violate IDENT[labial], but does violate MAX[labial]. For tone, MAX[Tone] constraints seem to be necessary (Yip [§10.3], Myers 1997), but for segments, it is common to use IDENT[Feature]. For critical discussion, see Keer (1999:ch.2), Struijke (2000a/2002:ch.4), de Lacy (2002§6.4.2), Howe & Pulleyblank (2004).

• *Interactions of markedness and faithfulness*

The source of much phonological explanation in OT derives from constraint interaction. At its most basic, the interaction of faithfulness and markedness determines whether input segments survive intact in the output (e.g. FAITH(α) » *α) or are eliminated (*α » FAITH(α)). In constraint terms, this is putting it fairly crudely: there are subtleties of constraint interaction that can prevent elimination of underlying segments in different contexts. For example, Steriade [§7.4.3] shows how the general ranking *βγ » *α » IDENT[α] prevents an otherwise general /α/→[β] mapping before γ (i.e. 'allophony').

One theme that the chapters here lack is explicit discussion of constraints on inputs. This is because interactions of faithfulness and markedness constraints preclude the need for restrictions on the input ('richness of the base' – Prince & Smolensky 1993§9.3). For example, there is no need to require that inputs in English never contain a bilabial click /⊙/; the general ranking || *⊙ » FAITH[⊙] || will eliminate clicks in all output environments.

Turning to more subtle consequences of constraint interaction, a number of the following chapters employ a consequence of OT: the decoupling of rule antecedents and consequents. A rule like α→β describes both the 'problem' – i.e. α, and the 'solution' – i.e. β. In contrast, a constraint like *α identifies the problem without committing itself to any particular solution. *α could be satisfied by deleting α or altering α to β, for example. The proposal that the same constraint can have multiple solutions – both cross-linguistically and even in the same language – is called 'heterogeneity of process, homogeneity of target' (HoP-HoT – McCarthy 2002c§1.3.2).[4] Examples are found in various chapters: Baković [§14.3] discusses the many ways that AGREE[F] can be satisfied, including assimilation, deletion, and epenthesis, with some languages employing more than one in different environments, Yip [§10.3.3] shows how the OCP – a constraint on adjacent identical tones – can variously force tone deletion, movement, and coalescence in different languages, and de Lacy [§12.6] shows how constraints that relate prosodic heads to sonority and tone can motivate metathesis, deletion, epenthesis, neutralization, and stress shift.

[4] The opposite is identified and exemplified by Ketner (2005) as 'heterogeneity of target, homogeneity of process', where the same process is used to satisfy a number of different conditions.

While HoP-HoT has clearly desirable consequences in a number of cases, one current challenge is to account for situations where it over-predicts. For example, Lombardi (2001) argues that a ban on voiced coda obstruents can never result in deletion or epenthesis, only neutralization (e.g. such a ban can force /ab/ to become [ap] but never [a] or [a.bi]). This situation of 'too many solutions' is currently an area of increasing debate in OT (Lombardi 2001, Wilson 2000, 2001, Steriade 2001b, Pater 2003, de Lacy 2003b, Blumenfeld 2005).

Another consequence of constraint interaction is the Emergence of the Unmarked (TETU): a markedness constraint may make its presence felt in limited morphological or phonological environments (see e.g. Rice [§4.5.1,§4.5.2], Urbanczyk [§20.2.4]). For example, a number of languages have only plain stops (e.g. Māori – Bauer 1993), so constraints against features like aspiration (*^{h}) must exist and in Māori outrank IDENT[^{h}]. In other languages where aspiration can appear fairly freely, IDENT[^{h}] outranks *^{h}. In contrast, in Cuzco Quechua *^{h} has an 'emergent' effect – while aspirated stops appear in roots, they do not appear at all in affixes. Beckman (1997§4.2.3) shows that this pattern can be accounted for by the ranking | Root-IDENT[^{h}] » *^{h} » IDENT[^{h}] |, where Root-IDENT[^{h}] is a positional faithfulness constraint that preserves aspiration in root segments only.

TETU has provided insight into many areas of phonology. However, there are some challenging issues related to it. One is that in some languages, TETU results in a segment that is otherwise banned. For example, Dutch has an epenthetic [ʔ] in onsets, even though [ʔ] is otherwise banned in the language. For discussion, see Łubowicz (2003:ch.5).

1.2.3 The Lexicon

The chapters make both explicit and implicit assumptions about the form of the lexicon and the sort of information it provides in OT. The lexicon has been traditionally seen as the repository of 'unpredictable information' – it contains morphemes (or words) and their unpredictable properties, such as their morphological and syntactic categories, their phonological content, and their semantic content. Two ongoing issues with the lexicon are (a) where to store unpredictable information and (b) how much predictable information to store in the lexicon. In post-*SPE* phonology, the dominant view was to put all unpredictable information into the lexicon and to try to minimize predictable information. From the other point of view, Anderson (1992) proposed that lexical items could effectively be expressed as rules.

Ussishkin [ch.19] adopts a popular middle ground in OT, with some unpredictable aspects of morphemes implemented as constraints. For example, McCarthy & Prince (1993a) propose constraints such as ALIGN-L(*um*, L), which requires the left edge of the morph of the Tagalog morpheme *um* to align with the left edge of a stem (i.e. be a prefix); this approach is discussed in detail by Ussishkin [§19.3.2]. So, whether a morpheme is prefixing or suffixing is not expressed in the lexicon as a diacritic that triggers a general concatenative rule (e.g. Sproat 1984), but as a morpheme-specific constraint.

The idea that unpredictable information can be expressed as a rule/constraint is not due to OT, but OT has allowed expression of such information by constraints to be straightforward, and it is now widely assumed (cf. Horwood 2002). It is also debatable how much lexical information should be expressed as a constraint: Golston

(1995) and Russell (1995) argue that even morphemes' phonological material should be introduced by constraint.

In *SPE*, as much predictable information was eliminated from the lexicon as possible and given by rule. For example, if medial nasal consonants always have the same place of articulation as the following consonant, pre-consonantal nasals in lexical entries were not specified for Place of Articulation. This idea was adapted in underspecification theories of the 1980s and 1990s. The explanatory power of *SPE* and its later rule-based successors partly relied on the fact that the input to the phonology was restricted in predictable ways.

In contrast, Prince & Smolensky's (1993/2004) principle of 'Richness of the Base' (RoB) forced this idea to be reconsidered. Because OT eschews constraints on input forms, a language's grammar must be able to account for every conceivable input, so whether underlying lexical forms lack predictable information or not becomes irrelevant with RoB. Consequently, a great deal of work in OT and in the chapters here assumes that lexical entries are fully specified for phonological information (cf. Ito et al. 1995, Inkelas et al. 1997, cf. Artstein 1998). The irrelevance of the specification of predictable information in the lexicon does not indicate any greater level of complexity in OT. In fact, the principle it relies on – the lack of restrictions on inputs – has allowed resolution of some long-standing problems (e.g. the Duplication Problem – McCarthy 2002c§3.1.2.2).

Finally, a large amount of work in OT has re-evaluated the formal expression of morphological relatedness. As McCarthy [§5.5] discusses, Correspondence Theory has been extended to account for phonological similarities among morphologically related words, such as the syllabic nasal in 'lighten' [laɪtn̩] and 'lightening' [laɪtn̩ɪŋ] (cf. 'lightning' [laɪtnɪŋ], *[laɪtn̩ɪŋ] – in the formal register of my dialect of New Zealand English) (e.g. Benua 1997). This issue is discussed more fully by McCarthy [§5.5].

In short, the lexicon in OT is different in significant ways from the lexicon in previous work. Some unpredictable information has been moved out of the lexicon and expressed as constraints, and some predictable information is commonly assumed to remain in the lexicon. The formal expression of 'morphological relatedness' and paradigms has changed fundamentally as part of the development of Correspondence Theory; it is no longer necessary to appeal to a serial derivation to account for phonological similarities between morphologically related words.

1.2.4 OT theories and other theories

One point that emerges from surveying the chapters in this volume is that it is misleading to imply that there is a single unified theory of OT that everyone adheres to. It is more accurate to say that there is an OT framework and many OT sub-theories.

Almost every aspect of OT has been questioned. For example, McCarthy & Prince's (1995a, 1999) theory of GEN with Correspondence is fundamentally different from the Containment model of Prince & Smolensky (1993/2004). There are also fundamentally different approaches to the constraint component CON: some view constraints from a Functionalist perspective and others from a Formalist one (see §1.3). In addition, some approaches see each constraint as independent of every other, while others attempt to identify general schemas that define large classes of

constraints (e.g. McCarthy & Prince's 1993a ALIGN schema (Ussishkin [§19.2.1]), Beckman's 1997 positional faithfulness schema, and markedness schemas in a variety of work). The concept of a totally ordered and invariant ranking has been questioned from several perspectives (see Anttila [ch.22] for details). Wilson (2000) proposes an EVAL that is fundamentally different from Prince & Smolensky's (1993/2004) (cf. McCarthy [§5.4], 2002b). McCarthy (2000b) considers – but does not advocate – a Serialist OT theory (also Rubach 2000). Finally, a number of proposals involving more than two levels have been put forward recently (see McCarthy [§5.4]).

In addition, the core principles of OT are compatible with aspects of other theories. For example, Harris & Gussman (1998) combine representational elements of Government Phonology with OT's theory of derivation, Smith (2000) proposes a melding of Dependency Phonology and OT, and some key features of the rule-based Lexical Phonology have been recast in an OT framework (see McCarthy [§5.5]).

In summary, there are many subtheories of OT, there are mixtures of OT and other theories' devices, and there also are a number of other theories that are the focus of current research (e.g. Government Phonology in Scheer 1998, 2004; Declarative Phonology – Coleman 1998, Bye 2003, and many others). Nevertheless, it is clear from the chapters here that Prince & Smolensky's (1993/2004) framework has had a profound impact on the field and helped to understand and reconceptualize a wide variety of phonological phenomena.

1.3 Representation and Explanation

Harris [§6.1] observes that "recent advances in derivational theory have prompted a rethink of… representational developments." Comparison of the chapters in Goldsmith's (1995a) *Handbook* with the ones here underscores this point: there is less appeal to specific representational devices and more reliance on constraints and their interaction to provide sources of explanation.

To give some background, in Autosegmental Phonology (Goldsmith 1976, 1990) and Metrical Phonology (see Hayes 1995, Kager [ch.9] for references) the aim throughout the 1980s and early 1990s was to place as much of the explanatory burden as possible on representation as possible with very few operations (e.g. relinking and delinking of association lines, clash and lapse avoidance). In contrast, constraint interaction in OT allows ways to analyze phonological phenomena that do not rely on representational devices.

- *Markedness and representation*

An example is found in the concept of Markedness, which has been a central issue in phonological theory since the Prague School's work in the 1930s (Trubetzkoy 1939, Jakobson 1941/1968). It is the focus of Rice's chapter [ch.4] in this handbook, and markedness theory is explicitly discussed in many others (e.g. Zec [§8.5], de Lacy [ch.12], Fikkert [ch.23], Bernhardt & Stemberger [§25.2.1]).

'Markedness' refers to asymmetries in linguistic phenomena. For example, it has often been claimed that epenthesis can produce coronals, but never labials or dorsals (e.g. Paradis & Prunet 1991b and references cited therein). Coronals are therefore less marked than labials and dorsals, and this markedness status recurs in many other processes (e.g. neutralization). In general, phonological phenomena such

as neutralization and epenthesis are taken to produce exclusively unmarked feature values.[5]

SPE's approach to markedness was to define feature values – *u* for unmarked and *m* for marked – which were interpreted by special 'marking conventions' which essentially filled in a phonetically interpretable value of '+' or '–' (*SPE*:ch.9). *SPE*'s approach was therefore essentially representational: markedness follows from the form of feature values. After *SPE*, a more elaborate theory of representation and markedness developed in the Autosegmental Theory of representation (Goldsmith 1976, 1990), and in theories of underspecification (e.g. Kiparsky 1982b, Archangeli 1984) and privativity (e.g. Lombardi 1991) (see Harris [§6.3], Hall [§13.2]). The unmarked feature value was indicated by a lack of that feature; for example, coronals had no Place features at all (articles in Paradis & Prunet 1991b, Avery & Rice 1989, Rice 1996, also see Hall [ch.13]). Coupled with the view that neutralization is feature deletion, the fact that neutralization produces unmarked elements is derived.

While the representational approach to markedness has continued in OT work (for recent work – Causley 1999, Morén 2003), Prince & Smolensky (1993/2004) and Smolensky (1993) opened up an entirely different way of conceiving of the concept (its most direct precursor is in Natural Generative Phonology – Stampe 1972). Instead of relying on representation, constraint ranking and form is central: coronals are not marked because they are representationally deficient, but because all constraints that favour dorsals and labials over coronals are universally lower-ranked than those constraints that favour coronals over other segments: i.e. || *DORSAL »» *LABIAL »» *CORONAL ||, where '»»' indicates a ranking that is invariant from language to language. There is no need to appeal to the idea that coronals lack Place features in this approach: they are the output of neutralization because other options – labials and dorsals – are ruled out by other constraints (for examples of fixed ranking, see Zec [§8.5], Yip [§10.3.2], de Lacy [§12.2.2])

The idea of universally fixed rankings is found in the opening pages of Prince & Smolensky (1993/2004); its success at dealing with markedness hierarchies in the now famous case of Imdlawn Tashlhiyt Berber syllabification is probably part of the reason that OT's influence spread so quickly (see Zec [§8.5.1] for discussion). Recent approaches to markedness in OT have rejected universally fixed rankings; they instead place restrictions on constraint form to establish markedness relations (see de Lacy [ch.12]). However, the principle is the same: markedness relations are established by ranking and constraint form, not by representational devices.

The OT ranking/constraint form approach to markedness has been widely accepted in current work, but the representational theory has not been rejected: the two approaches are often employed together. As discussed in Harris [ch.6], the debate continues as to where the balance lies.

[5] There is a great deal of controversy over the role of markedness in phonology. For example, Blevins (2004) proposes that markedness effects can be ascribed to diachronic change, and Hume (2003) rejects the idea that there are any markedness asymmetries (at least with respect to Place of Articulation). Rice [§4.7] and de Lacy (2005§1.3) re-evaluate the scope of markedness effects, arguing for a strict division between Competence and Performance. There has also been an ongoing re-evaluation of the empirical facts that support markedness. While there is much debate about which markedness asymmetries exist at the moment, it is at least clear that many traditional markedness diagnostics are not valid (e.g. Rice [§4.6]; also Rice 1996 et seq., de Lacy 2002, 2005, Hume & Tserdanelis 2002).

• *Representation in current theory*

The chapters identify and exemplify a number of reasons why there was a shift towards explanation through constraint interaction. One function of representation was to express markedness; as explained above, from the first Prince & Smolensky (1993/2004) showed how to capture markedness effects with constraint interaction. Similarly, much of the theory of representation relied on, or at least employed, serial derivations. For example, assimilation was seen as a three-step process of delinking a feature, adding an association to a nearby feature, then deleting the stray feature (also see Harris [§6.3.3]). With a two-level approach to grammar, the concepts of delinking and reassociation have no clear counterpart (though see Yip [ch.10] and the discussion below).

In many of the chapters here, Correspondence Theory is used instead of representational devices. For example, reduplication was seen in Marantz (1982) and McCarthy & Prince (1986) as a series of associations followed by delinking due to a ban on crossed association lines; Urbanczyk [ch.20] shows how reduplication can be analyzed using correspondence – another type of relation entirely. Representation was also relied on to express dependency relations. For example, if a feature F always assimilates whenever feature G does, then F was assumed to be representationally dependent on G. Harris [§6.3.3] observes that Padgett's (2002) work shows that at least some dependency relations between features and classhood can be expressed through constraint interaction and do not rely on an explicit representational hierarchy of features (also see Yip 2004).

Of course, it is crucial for any theory of phonology to have a well-defined restrictive theory of representation. However, OT has allowed the burden of explanation to move from being almost exclusively representation-based to being substantially constraint-based.

In fact, while most recent work in OT has focused on constraint interaction, a good deal has examined or employed representational devices as a crucial part of explanation. For example, Beckman (1997, 1998) employs an OT version of Autosegmental phonology to deal with assimilation. Cole & Kisseberth (1994) propose Optimal Domains theory, which certainly relies less on representational devices than its predecessors but crucially refers to a representational notion of featural alignment. McCarthy (2004a) proposes a theory of representation that builds on autosegmental concepts. Interestingly, the representation of Tone has been least affected by the move to OT. Very little has changed in representational terms: pre-OT notions such as multiply-linked (i.e. spread and contour) tones, floating tones, and tonal non-specification are commonly used in OT work – see Yip [ch.10] for details.

One reason for the lack of in-depth discussion of representation is that it has become common to focus on constraint interaction and violations in OT work, while there has been less necessity to provide explicit definitions of constraint form. An example is the AGREE[F] constraint (Lombardi 1999, Baković 2000), defined by Baković [§14.1] as "Adjacent output segments have the same value of the feature *x*." The constraint is defined in this way because the definition aims to express the *effect* of the constraint (i.e. how it assigns violations) rather than providing a formal structural description. In formalizing the definition, though, it is necessary to deal with representational issues: what does the term "have the same value" mean? In formal terms, is this phrase necessarily expressed as a multiply-associated feature? Or can it be expressed through correspondence relations? These issues are receiving more attention in recent work.

In summary, much of the burden of explanation has shifted from representational devices to constraint interaction. However, many of the representational devices that were developed in the 1980s remain integral to current phonological analyses, as exemplified by the detailed prosodic structures used by Zec [ch.8], Kager [ch.9], Yip [ch.10], Gussenhoven [ch.11], and Truckenbrodt [ch.18], and the feature structure discussed by Hall [§13.2]. As the authors discuss, justification for the structures remains despite the effects of constraint interaction.

1.4 Functionalism

Gordon [ch.3] observes that "the last decade has witnessed renewed vigour in attempting to integrate functional, especially phonetic, explanations into formal analyses of phonological phenomena." Functionalist principles are discussed in many of the chapters in this book (including Rice [§4.7], Harris [§6.2.2], Zec [§8.6], Steriade [§7.3], Yip [§10.4.2], Hall [§13.2], Baković [§14.4.1], Alderete & Frisch [§16.3], Kingston [§17.3], Bermúdez-Otero [§21.4], Anttila [§22.3.3, §22.4], Fikkert [§23.2]). This section provides some background to both Functionalist and Formalist approaches to phonology (also see McCarthy 2002c§4.4).

Gordon [ch.3] identifies a number of core principles in Functionalist approaches to phonology. A central concept is expressed by Ohala (1972:289): "Universal sound patterns must arise due to the universal constraints or tendencies of the human physiological mechanisms involved in speech production and perception". Certainly, a number of researchers advocated a Functionalist approach (e.g. Stampe 1972, Ohala 1972 et seq., Liljencrants & Lindblom 1972, Archangeli & Pulleyblank 1994, Bybee 2001 and many others), but it is only recently that Functionalist theories employing OT-like frameworks have gained a great deal of popularity, as documented by Gordon [ch.3] (also see the articles in Gussenhoven & Kager 2001, Hume & Johnson 2001, and Hayes et al. 2004; Flemming 1995, Jun 1995, Boersma 1998, Kirchner 1998, 2001, Gordon 1999, 2002b, and many others). Research has focused on issues such as how concepts such as markedness are grounded in concepts of articulatory ease and perceptual distinctiveness, and how to express these influences in constraint form.

The property common to all current functionalist approaches is the idea that phonological effects (especially markedness) are not due to innate constraints or constraint schemas. Instead, one functionalist view (called 'Direct Functionalism' here) holds that constraints are constructed by mechanisms that measure articulatory effort and perceptual distinctiveness (and perhaps also parsing difficulty). Constraints are defined in units that directly record this effort and distinctiveness; consequently, the approaches use finely differentiated units (e.g. real numbers) not used in traditional conceptions of phonology (see Harris [§6.2.2] for discussion, also Anttila [§22.3.3] and Tesar [§24.4]).

Another view combines direct functionalism with the idea that the phonological component is limited in terms of its expressive power. In this view, constraints are constructed with reference to articulation and perception, but they must be expressed in terms of a small set of phonological primitives: i.e. "phonological constraints tend to ban phonetic difficulty in simple, formally symmetrical ways" (Hayes 1999§6.2). The phonological primitives may not be well-adapted to expressing phonetic categories, so there may be various mismatches.

Distinct from these views is the 'diachronic functionalist' approach (Ohala 1971 et seq., Blevins 2004). Blevins' approach in particular allows the phonological component to generate virtually any sound pattern (Gordon [§3.5]). However, not every sound pattern is learnable, or the conditions under which certain sound patterns can be learned are so restricted that the pattern is incredibly rare. Consequently, markedness effects are due to the process of language learning, and explanation for diachronic change and synchronic processes are treated as one. Diachronic functionalism is discussed by Gordon [§3.5], so will not be examined further here.

1.4.1 The Formalist Approach

A great deal of current phonological work has its roots in Formalist approaches (see Chomsky 1966 for phonology specifically, Chomsky 1965 et seq., and more recently Hale & Reiss 2000b). In OT, the Formalist approach plays out as the assumption that all constraints or constraint schemas are innate.

The Formalist approach does not necessarily rule out functional grounding in constraints. As Chomsky & Lasnik (1977§1.2) discuss, Formalist approaches can assume a 'species-level' functionalism: this is the idea that a particular constraint has been favoured in evolution because it helps with articulation, perception, or parsing. For example, Chomsky & Lasnik suggest that the syntactic constraint *[$_{NP}$NP TENSE VP] is innate, but has been favoured in evolution because it simplifies parsing (p.436).

The implication of the Formalist approach to phonology is that derivation, representation, and constraints can have 'arbitrary' aspects – i.e. they may not directly aid (and could even act against) reduction of articulatory effort and increase in perceptual distinctiveness. However, it is not surprising to find that some (or even many) mechanisms or constraints do serve to aid in articulation, perception, and processing; this functional grounding would be seen as following from 'species-level' adaptations.

With 'species-level functionalism', it may seem that the Formalist and Functionalist approaches would have very similar effects. However, the difference resides in the Formalist possibility for arbitrary phonological structures, hierarchies, and constraints. For example, Zec [§8.5] and de Lacy [§12.2] employ the sonority hierarchy as a central part of their analyses of prosodic structure, yet determining phonetic basis of sonority – and therefore its articulatory and perceptual value – has proven notoriously difficult (Parker 2002 and references cited therein). It seems that the sonority hierarchy is at least partially arbitrary (i.e. without functional motivation), and only partially adapted to aiding articulation and perception; this sort of mismatch is expected in the Formalist approach. Of course, the difficulty in identifying arbitrariness is that we may simply not be looking at the right articulatory or perceptual property.

Also expected in the Formalist view is the idea that there could be arbitrary (and even functionally non-sensical) restrictions on phonological processes. An example is found in tone- and sonority-driven stress, discussed in de Lacy [ch.12]. Longer segments (e.g. long vowels, diphthongs) often attract stress, and there are plausible functional reasons for such attraction (Ahn 2000). In fact, this attraction may (partially) account for the attraction of stress to high sonority vowels like [a] because they typically have a longer inherent duration than low sonority vowels like [i], [u], and [ə]. However, in many languages there is a correlation between tone level and vowel duration: the lower the tone, the longer the vowel (e.g. Thai – Abramson

1962). Thus, low-toned [à] is longer than high-toned [á], and so on. Now, if low tone increases duration, and stress is attracted to longer elements, functional reasoning should lead us to believe that stress will be attracted to low tone over high tone. However, this is *never* the case: stress always prefers high toned vowels to low-toned ones. Of course, there may be some other functional reason for favouring high-toned stressed syllables, but given the fact that languages can vary as to which functional factor they favour (i.e. through ranking), it is surprising that *no* language favours stressed low-toned vowels over high-toned ones (cf. functional approaches to vowel inventories, where articulatory and perceptual factors can conflict, but one can take precedence over the other in particular languages).

To summarize, support for the Formalist view (with 'species-level functionalism') can be sought in phonological arbitrariness and Competence-Performance mismatches.

1.4.2 Challenges

Gordon [§3.1] observes that one reason for the increase in Functionalist popularity is OT's formalism: OT can be easily adapted to expressing gradient phenomena; it also provides a framework for expressing the concept of 'tendency' through constraint ranking. However, it is important to emphasize that OT is not an inherently Functionalist theory, and some Functionalist versions of OT depart significantly from Prince & Smolensky's (1993/2004) proposals (e.g. versions of Stochastic OT – see discussion and references in Anttila [§22.3.3], McCarthy 2002c§4.4).

Another reason may be that the Formalist explanation for sound patterns is seen by some as insufficiently profound. For example, the fact that dorsals are more marked than coronals receives the explanation that *DORSAL universally outranks *CORONAL in a Formalist approach, and this universal ranking is innate. In other words, the constraint ranking is an axiom of the theory. Yet there is clearly a good articulatory reason for this ranking – dorsals require more articulatory effort than coronals (if effort is measured from rest position), and there may be perceptual reasons as well. A Functionalist approach makes a direct connection between the substantive facts and the formalism.

A further reason is skepticism about the ability of species-level functionalism to account for phonological facts. For example, how could the fixed ranking *DORSAL »» *CORONAL evolve? A fixed ranking *DORSAL »» *CORONAL would have to appear through a random mutation (or exaptation), then provide some advantage that a speaker who had to learn their ranking did not have (e.g. faster learning). Identifying the exact advantage (whether survival or sexual) is challenging. There may also be the issue of plausibility, though as Pinker & Bloom (1990) have observed, tiny advantages can have significant influence over time.

The problem that Formalist approaches face is not that they lack explanation, but that it is difficult to provide proof. Little is understood about the biology of phonological evolution, and so evolutionary arguments are hard to make (though see Hauser, Chomsky, and Fitch 2002 for discussion and references). Given the burgeoning popularity of Functionalist approaches, the onus currently seems to be on the Formalist approach to close the 'plausibility gap' and identify testable predictions that differ from Functionalist ones.

There are also challenges for the Functionalist perspective. For the diachronic functionalist view, one challenge is to account for cases where a diachronic change

has no synchronic counterpart, and why there are unattested synchronic grammars which could easily be created by a series of natural diachronic changes (Kiparsky 2004). Mismatches also pose a challenge for the 'direct' functionalist point of view, as do cases of arbitrariness (as in the sonority hierarchy), as all constraints and markedness hierarchies should be tied directly into Performance considerations.

Functionalist approaches have already had a significant impact on phonological theory. There are many works that explicitly advocate functionalist principles (cited in §1.3 above). It is also commonplace in recent publications to see a constraint's validity evaluated by whether it is related to a decrease in articulatory effort or helps in perception or parsing. For example, in a widely-used textbook, Kager (1999a:11) comments that "phonological markedness constraints should be *phonetically grounded* in some property of articulation and perception". Of course, a Formalist perspective does not accept the validity of such reasoning. In the immediate future, it is likely that the functionalist perspective will continue to gain ground, but it is also likely that there will be increasing dialogue between the various Formalist and Functionalist approaches and also an increasing understanding of the implications of Formalist tenets in phonology.

1.5 Conclusions

The preceding sections have attempted to identify some of the major theoretical themes that appear throughout the following chapters. Of course, there are many others in the following chapters that are not covered here (e.g. the role of contrast in phonology – Steriade [ch.13], Rice [§4.6]). There are also a number of 'empirical' themes: Fikkert [§23.1] comments that for language acquisition there has been an increase in research and resources, and Tesar [ch.24] discusses the growing field of learnability. As detailed in the chapters in Part 5, areas of phonology that have traditionally been under-studied or seen as peripheral (e.g. free variation – Anttila [ch.22]), are having a significant influence on central issues in the field.

The chapters in this Handbook show that phonological theory has undergone enormous theoretical changes compared with ten years ago, and it continues to change rapidly. It is probably for this reason that none of the chapters in this book attempt to make predictions about the broad issues that will dominate phonology in the next ten years. Perhaps the only safe bet is that any prediction about the future of phonology will be wildly inaccurate.