

## Abbreviations

---

BLS	<i>Proceedings of the Annual Meeting, Berkeley Linguistics Society</i>
CLS	<i>Papers from the Annual Regional Meeting, Chicago Linguistic Society</i>
IJAL	<i>International Journal of American Linguistics</i>
JASA	<i>Journal of the Acoustical Society of America</i>
JL	<i>Journal of Linguistics</i>
JPhon	<i>Journal of Phonetics</i>
Lg	<i>Language</i>
LI	<i>Linguistic Inquiry</i>
NELS	<i>Papers from the Annual Meeting of the North East Linguistic Society</i>
NLLT	<i>Natural Language &amp; Linguistic Theory</i>
WCCFL	<i>Proceedings of the West Coast Conference on Formal Linguistics</i>

## 1 Introduction: the phonetic bases of phonological Markedness

---

*Bruce Hayes and Donca Steriade*

If phonological systems were seen as adaptations to universal performance constraints on speaking, listening and learning to speak, what would they be like?  
Lindblom (1990: 102)

### 1 Introduction

Our starting point is a hypothesis central to contemporary phonology: that the markedness laws characterising the typology of sound systems play a role, as grammatical constraints, in the linguistic competence of individual speakers. From this assumption, a basic question follows: How are grammars structured, if markedness laws actively function within them as elements of linguistic competence? We find the answer offered by Optimality Theory (Prince and Smolensky 1993) worth investigating: the grammatical counterparts of markedness laws are ranked and violable constraints and the latter form 'the very substance from which grammars are built: a set of highly general constraints which, through ranking, interact to produce the elaborate particularity of individual languages' (Prince and Smolensky 1993: 217). With qualifications, this view is adopted by many of the contributions to this volume.

The focus of our book is on a different, complementary question: Where do markedness laws come from? Why are sound systems governed by these laws and not by some conceivable others? What is the source of the individual's knowledge of markedness-based constraints? The hypothesis shared by many writers in this volume is that phonological constraints can be rooted in *phonetic knowledge* (Kingston and Diehl 1994), the speakers' partial understanding of the physical conditions under which speech is produced and perceived. The source of markedness constraints as components of grammar is this knowledge. The effect phonetic knowledge has on the typology of the world's sound systems stems from the fact that certain basic conditions governing speech perception and production are necessarily shared by all languages, experienced by all speakers, and implicitly known by all. This shared knowledge leads learners to postulate independently similar constraints. The activity of similar constraints is

a source of systematic similarities among grammars and generates a structured phonological typology.

In this introduction, we explain why it is useful to explore the hypothesis that knowledge of markedness derives from phonetic knowledge: how one's view of markedness changes under this hypothesis and what empirical results come from this change of perspective. We outline first how research on phonetically based markedness can be beneficially explored in the framework of Optimality Theory (section 2); and how the OT search for the right constraint set can be speeded up on the view that markedness is phonetically based (sections 3 and 4). We then discuss a specific example of a phonetically based Markedness constraint that illustrates several options in mapping the facts of phonetic difficulty to the elements of grammar (section 5). In the remaining sections, we relate the general discussion of markedness to the specific contents of the book, noting that despite differences of analytical strategy or general theoretical outlook, the diverse phenomena analysed by our contributors can be viewed in a unified fashion.

## 2 Phonetically based Markedness and Optimality Theory

The idea that phonological Markedness has phonetic roots has particular antecedents in *The sound pattern of English* (Chomsky and Halle 1968), in the theory of Natural Phonology (Stampe 1973), and in the more recent work on Grounded Phonology by Archangeli and Pulleyblank (1994). Optimality Theory makes it worth returning to these issues, since it provides tools with which the questions can be addressed in novel ways. OT takes on a difficulty that held back earlier approaches to naturalness: the *what* is phonetically difficult is not the same as the *how to fix it*. In a rule-based framework, one must provide the theory with multiple fixes, all of which address the same phonetic difficulty. OT separates the problem (embodied in the Markedness constraints) from the solution; the latter is the general procedure at the core of OT, namely creation of a large candidate set by GEN, with the choice from among them determined by the relative ranking of the Markedness constraints with respect to Faithfulness and each other. As a result, OT allows the phonetic principles that drive the system to be expressed directly (Myers 1997): a constraint can embody a particular form of phonetic difficulty, with the issue of how and whether the difficulty is avoided relegated to other parts of the grammar. For a clear case of this sort, see the discussion of postnasal voicing in Pater (1999) and Kager (1999).

The separation of Markedness and Faithfulness also provides a cogent response to an ancient canard: If phonetic optimality is important, why don't sound systems contain nothing but the Jakobsonian optimal [ba]? The answer

is that not all the constraints can be satisfied at once. Faithfulness and Markedness constraints conflict; and moreover, there are conflicts between different types of Markedness constraints (notably, those grounded in production vs those grounded in perception). There is no reason to expect the resolution of these conflicts to be uniform across languages. The postnasal voicing example just mentioned is a plausible case of multiple resolutions of the same difficulty.

The more direct argument for OT is that phonetically based constraints discussed here are frequently both active and violated, yielding Emergence of the Unmarked effects (McCarthy and Prince 1994) which require explicit ranking. Kirchner's, Kaun's, and Crosswhite's chapters provide extensive evidence of this type, as does a voicing example discussed below.

## 3 Markedness

The term *markedness* is ambiguous. It can be used in a strictly typological sense, to identify structures that are infrequently attested or systematically missing, as in *Active use of [-ATR] is marked* (Archangeli and Pulleyblank 1994: 165 and passim). The term can also refer to an element of a formal linguistic theory, as in OT, where the term *markedness* characterises a constraint type, often disambiguated by capitalisation: Markedness constraints penalise particular structures in surface forms, whereas Faithfulness constraints evaluate dimensions of similarity between specified pairs of lexically related structures, such as the underlying and surface representations.

The definition of markedness in OT is also sometimes related to the hypothesis that Markedness constraints are universal and innate. This claim is logically independent of the central tenets of OT about constraint interaction.<sup>1</sup> Accordingly we are free to assume that a constraint need not be universal or innate to qualify as a Markedness constraint; rather, we use the term in the purely technical sense of a constraint whose violations are evaluated solely on surface forms. We use the term *markedness law* to denote patterns found in typological data, which Markedness constraints are often meant to explain. We may add that the correspondence conditions themselves are formulated with the intention of deriving key aspects of phonological typology.<sup>2</sup>

The terms thus clarified, we turn now to the options available to phonologists who study markedness in either of these two senses.

## 4 Inductive and deductive approaches to the study of Markedness

Lindblom (1990: 46)<sup>3</sup> observes that the study of distinctive features can proceed in two ways: inductively and deductively. The inductive approach in the

study of features is to introduce a new feature whenever the descriptive need arises. The deductive approach, for example Stevens' Quantal Theory (1989) or Lindblom's Dispersion Theory (1986), proceeds not from a question of description ('What are the features used in language?') but from a principled expectation: 'What features should we expect to find given certain assumptions about the conditions [under which] speech sounds are likely to develop?' (Lindblom and Engstrand 1989: 107). The deductive approach can thus hope to provide not only an empirically verifiable feature theory, in the form of principles from which feature sets derive, but may also yield answers to further questions, such as 'Why are the mental representations of speech sounds feature-based (and likewise segment-, syllable-, foot-based)?' These questions simply do not arise under approaches that take for granted the existence of such units and merely aim to discover in the data a basis for their classification.

The distinction between inductive and deductive approaches applies equally to research on markedness. Most attempts to discover markedness principles in phonology have proceeded, until recently, in inductive fashion: phonologists accumulate factual observations about languages and, in due course, a cluster of such observations coheres into a law. The law may be absolute ('There are no initial or final systems in which all obstruent combinations are heterogeneous with regard to voicing'; Greenberg 1978: 252), or implicational ('The presence of syllabic [h] implies the presence of syllabic fricatives'; Bell 1978: 183), or only a trend ('If a nasal vowel system is smaller than the corresponding basic vowel system, it is most often a mid vowel that is missing from the nasal vowels'; Crothers 1978: 136). But in most cases the laws originate as generalisations over known languages, not as principles explaining why these laws should be expected to hold. A set of such laws, when they survive peer review, forms a proposed theory of markedness.

The markedness questions asked in earlier typological work seem to have been those for which evidence happened to be available. We cannot exclude the possibility that a priori principles have guided the search for typological generalisations, as reported in the classic work of Trubetzkoy 1938, Jakobson 1941, Hockett 1955, and Greenberg 1978, but these guiding principles were not spelled out and cannot be reconstructed. One may ask, for instance, why the search for clustering universals (Greenberg 1978) proceeds by asking some questions (*Is there an implicational relation between initial [ln] and initial [lt]?*) but not others (*Is there an implicational relation between initial [ht] and initial [th]?*).

There is an issue of research strategy here. The number of conceivable typological observations is so vast that our results will be haphazard if we examine the data in arbitrary order. Without a general conception of what makes a possible markedness principle, there is no more reason to look into the markedness

patterns of, say, initial retroflex apicals (a useful subject, as it turns out; see section 6.1) than into those of prenasal high tones (a topic whose interest remains unproven). The researcher has to take a stab in the dark. In light of this, it seems a sensible research strategy to hypothesise general principles concerning why the constraints are as they are, and let these principles determine a structured search for markedness patterns. We also see below that pursuing the deductive strategy can yield a completely different picture of markedness in several empirical domains.

The work reported in this volume proceeds deductively – as advocated by Lindblom (1990) and Ohala (1983, and much later work) – by asking at the outset variants of the following question: Are there general properties distinguishing marked from unmarked phonological structures, and, if so, what are they? Earlier work in phonetics<sup>4</sup> and phonology<sup>5</sup> suggests that a connection can be found between constraints governing the production and perception of speech and markedness patterns. Certain processes (cluster simplification, place assimilation, lenition, vowel reduction, tonal neutralisation) appear to be triggered by demands of articulatory simplification, while the specific contexts targeted by simplification (e.g. the direction of place assimilation, the segment types it tends to target) are frequently attributable to perceptual factors.

Deductive research on phonological markedness starts from the assumption that markedness laws obtain across languages not because they reflect structural properties of the language faculty, irreducible to non-linguistic factors, but rather because they stem from speakers' shared knowledge of the factors that affect speech communication by impeding articulation, perception, or lexical access. Consider the case discussed below, that of the cross-linguistic dispreference for voiced geminates. The deductive strategy starts from the assumption that this dispreference cannot reflect an innate constraint that specifically and arbitrarily bans [b: d: g:], but must be based on knowledge accessible to individual speakers of the factors that might interfere with the production and perception of voicing. This knowledge and its connection to the grammar have then to be spelled out.

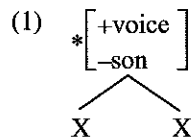
Is the deductive strategy reductionist? Clearly so, but in specific respects. The research presented here bears only on the possibility of systematically deducing the contents of phonological constraints from knowledge of grammar-external factors. This is not the same as deducing the grammar itself: on the contrary, structural properties of the grammar may well filter phonetic knowledge and limit the ways it is mapped onto grammatical statements, as suggested by Gordon (chapter 9) and summarised below (section 5.7). Further, none of the contributions addresses systematically the nature of phonological representations or deduces *their* properties from extra-grammatical factors or discusses whether such reduction is feasible (Gafos 1999). The same

goes for the nature of constraint interaction. On the issue of external grounding for all of these components, see Pierrehumbert's overview (2000), and the discussion of representations and constraint interactions by Flemming (2001).

## 5 Markedness from phonetics: a constraint and its phonetic basis

We now examine a specific example of the deductive strategy. This section introduces a markedness scale and points out its sources in the aerodynamics of speech.

In the phonological analysis of a number of languages, a constraint is needed that penalises voiced obstruent geminates; (1) is a first approximation.



Variants of (1) are active in Ancient Greek (Lupas 1972), Ossetic (Abaev 1964), Nubian (Bell 1971), Lebanese Neo-Syrian (Ohala 1983), Tamil (Rajaram 1972), Yakut (Krueger 1962), Limbu (van Driem 1987), Seleyarese and Buginese (Podesva 2000), and Japanese (Ito and Mester 1995). No language known to us bans just the *voiceless* geminates.<sup>6</sup> The constraint in (1) thus has a typological counterpart, the implicational law in (2):

- (2) The presence of a voiced obstruent geminate in a given language implies, in any context, that of the corresponding voiceless geminate.<sup>7</sup>

If a Markedness constraint like (1) reflects, directly or not, an implicational law like (2), then we must consider the possibility that the constraint is universal, in the sense of being potentially active in any grammar. In the next section we explore the hypothesis that some version of (1) is universal in the sense of being inferable from generally available phonetic knowledge.

### 5.1 From phonetics to grammar

As indicated earlier, we assume that constraints may be universal without being innate (cf. Lindblom 1990; Donegan 1993; Boersma 1998; Hayes 1999). We view Universal Grammar (UG) primarily as a set of abstract analytical predispositions that allow learners to induce grammars from the raw facts of speech, and not as a – dauntingly large – collection of a priori constraints. The project then is to understand how constraints like (1) are induced from evidence

about the conditions under which voicing is perceived and produced and what form they take if they are so induced. It is useful here to make the four-way distinction shown below:

- (3) a. Facts of phonetic difficulty  
 b. Speakers' implicit knowledge of the facts in (a)  
 c. Grammatical constraints induced from the knowledge in (b)  
 d. Sound patterns reflecting the activity of the constraints in (c)

Facts about phonetic difficulty (3a) and sound patterns (3d) are, in principle, accessible; they are obtainable from experiment, vocal tract modelling, and descriptive phonological work. But the precise contents of (3b) and (3c) have to be guessed at. We see no alternative to drawing these distinctions and making some inferences.

With Prince and Smolensky (1993), we assume that constraint organisation, (3c), reflects transparently the structure of markedness scales, (3b).<sup>8</sup> We also assume that the correspondence between the facts of phonetic difficulty (3a) and the markedness scales (3b) is necessarily indirect: the crucial question is how indirect.

The markedness scales phonologists have mainly relied on so far do not, in their current formulations, explicitly relate to scales of articulatory or perceptual difficulty. Examples are: (a) the nucleus goodness scale in Prince and Smolensky 1993; (b) a place optimality scale like ( $\{Labial, Dorsal\} < Coronal < Pharyngeal$ ), where  $<$  denotes 'worse than'; Lombardi (in press); and (c) syllabic markedness scales like  $CVCC, CCVC < CVC < CV$ . This may reflect the fact that there is no connection between Markedness constraints and phonetic scales or that the exact ways in which phonetic scales map onto phonological markedness has no consequences for the functioning of the phonology. However, the research reported in this book as well as in earlier work indicates that there is often evidence for a much closer connection.

In the next subsections we summarise the articulatory difficulties involved in sustaining vocal cord vibration in different obstruents and consider ways in which speakers can encode knowledge of these difficulties in markedness scales. Our point will be that among several types of mapping (3a) onto (3b)–(3c), a more direct one yields more predictive and more successful models of grammar.

### 5.2 Aerodynamics of voicing

Phonetic studies (Ohala and Riordan 1979; Westbury 1979; Westbury and Keating 1986) have located the rationale for the markedness law in (2) in the aerodynamics of voicing production:

- (4) a. Voicing requires airflow across the glottis.  
 b. In obstruents, the supraglottal airflow is not freely vented to the outside world.

For these reasons, active oral tract expansion (for example, by tongue root advancement or larynx lowering) is necessary to maintain airflow in an obstruent. These manoeuvres cannot be continued indefinitely or controlled tightly. It is therefore more difficult to sustain production of voicing in long obstruents. The difficulty is directly witnessed in languages like Ossetic, whose speakers attempt to maintain a voicing distinction in long obstruents but nonetheless lose 'part or all of the voiced quality' (Abaev 1964: 9) in [b: d: g:]. No comparable difficulty exists in sustaining voicelessness in [p: t: k:] or voicing in long sonorants, while the problem of maintaining voicing in singleton stops is necessarily one of shorter duration. So far the discussion motivates a simple voicing difficulty scale of the form  $D_i: < D_j$  where  $D_i$  is a geminate voiced obstruent, and  $D_j$  is the corresponding singleton.

Consider now a second factor that influences phonetic difficulty in obstruents, namely place of articulation. As Ohala and Riordan (1979) observe, the size of the cavity behind the oral constriction affects the aerodynamics of voicing. The time interval from the onset of stop closure to the point where passive devoicing will set in varies with the site of the oral constriction: in one experiment, voicing was observed to continue in [b] for 82 ms, but for only 63 and 52 ms respectively in [d] and [g]. This is because the larger cavity behind the lips offers more compliant tissue, which allows the cavity to continue for a longer time to expand passively in response to airflow. A consequence of this is the known asymmetry (Gamkrelidze 1978) between voicing markedness in singleton bilabials as against alveolars and velars: [g] implies [d] which implies [b].<sup>9</sup> This asymmetry holds, according to Ohala (1983), among voiced geminates as well: a geminate [b:]'s duration will certainly exceed 82 ms, and thus some active expansion of the oral tract must be taking place, just as for [d:] and [g:]. But a difference in ease of voicing maintenance persists among the voiced geminates, because there are more options for expansion available in front than in back articulations.

### 5.3 From aerodynamics to markedness to constraints

There are then at least two sources of articulatory (and indirectly perceptual) difficulty in maintaining voicing: the duration of oral closure and the size of the cavity behind the oral constriction. Phonologically, these are completely different, yet at the level of phonetic difficulty, they are essentially the same thing: in both [g] (a singleton with a small cavity behind the constriction) and [b:] (a geminate with a large cavity) there is difficulty in maintaining voicing

past the point where passive devoicing normally sets in. Thus at the phonetic level we can posit a single scale of difficulty that includes both singletons and geminates.

- (5) \* [+voice]: {g: < d: < b: < g < d < b}

The scales we formulate henceforth distinguish a shared target property – [+voice] in (5) – and the set of contexts in which this property is realised with greater or lesser difficulty: (5) states that the [+voice] feature is hardest to realise in [g:], next hardest in [d:], and so on, and easiest to realise in [b].

The scale in (5) identifies [b:], the best voiced geminate, as harder to voice than short [g], the worst singleton. The difference between a singleton and a geminate consonant is typically much more than the 30 ms that separate the onset of passive devoicing in [b] vs [g] (Lehiste 1970; Smith 1992). Thus the difficulty involved in sustaining voicing should be far more extreme for any geminate obstruent than it would be for any voiced singleton: (5) reflects this point.

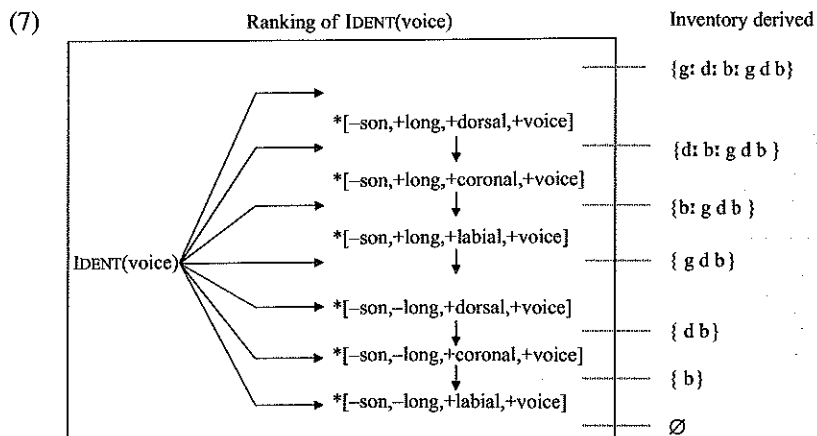
If knowledge about the difficulty of sustaining voicing in obstruents resembles the scale in (5), then its grammatical counterpart cannot be a single constraint; nor can the constraints against voiced geminates remain unrelated to those against voicing in singletons. This is because the voicing difficulty in [g: d: b:] is of the same type – if not of the same magnitude – as that involved in [g d b]. We need a constraint set that reflects the whole scale in (5), not just its upper region. The more general point is that knowledge of markedness, when viewed as phonetic knowledge, generates constraint families and rankings whose structure reflects a broader map of phonetic difficulty, as the learner understands it, rather than isolated points and relations on this map.

As a specific proposal to this end, consider the set of Markedness constraints in (6). These constraints are assumed to be ranked a priori, according to the phonetic difficulty of the segments that they ban (but see fn. 8 above on the issue of fixed rankings).

- (6) a. \*[-son, +long, +dorsal, +voice] 'no voiced long dorsal obstruents' >>  
 b. \*[-son, +long, +coronal, +voice] 'no voiced long coronal obstruents' >>  
 c. \*[-son, +long, +labial, +voice] 'no voiced long labial obstruents' >>  
 d. \*[-son, -long, +dorsal, +voice] 'no voiced short dorsal obstruents' >>  
 e. \*[-son, -long, +coronal, +voice] 'no voiced short coronal obstruents' >>  
 f. \*[-son, -long, +labial, +voice] 'no voiced short labial obstruents'

If the rankings in (6) are fixed, then the relative ranking of this constraint family with respect to the Faithfulness constraint IDENT(voice) determines the

inventory of voiced obstruents, as shown in (7):



An interesting aspect of the constraint set in (6) is that it uses very fine categories, each embodying information about both place and length. Phonologists characteristically judge that constraints are based on rather broader categories. One thus could imagine a more modular characterisation of voicing markedness, as in (8):

- (8) a. \*[-son, +dorsal, +voice] 'no voiced dorsal obstruents' >>  
 \*[-son, +coronal, +voice] 'no voiced coronal obstruents' >>  
 \*[-son, +labial, +voice] 'no voiced labial obstruents'  
 b. \*[-son, +long, +voice] 'no long voiced obstruents' >>  
 \*[-son, -long, +voice] 'no short voiced obstruents'

The constraints in (8) are simpler than those of (6), and involve separate chains of a priori rankings for the dimensions of place and length. As a result, this constraint set is silent on how closure duration and cavity size interact – that is, on the [b:] vs [g] comparison – and thus makes rather different predictions. Notably, we find that in ranking IDENT(voice) amid the chains of (8) (interleaving the chains freely), we cannot derive the inventories for two of the crucial cutoff points in (5): {b: g d b} (forbidding \*[d:] and harder) and {d: b: g d b} (forbidding just \*[g:]).<sup>10</sup>

5.4 *From scales to sound patterns: some language data*

The special possibilities implied by (6) (i.e., the constraint set that embodies a unitary scale of voicing difficulty) are confirmed by examples from real

languages. The chart in (9) illustrates patterns of selective voicing neutralisation, on a scale like (5), defined by length and place categories: shaded cells in the chart indicate that the voiced obstruent in the column header does not occur. As we compare the three scales introduced earlier with the chart in (9), we observe first that there exist languages that draw a cutoff on all seven possible points of (5):

(9) Place and length constraints on voicing contrasts

	b	d	g	b:	d:	g:
a. Delaware (Maddieson 1984)						
b. Dakota (Maddieson 1984)						
c. Khasi (Maddieson 1984)						
d. Various (citations under (1) above)						
e. Kadugli (Abdalla 1973), Sudan Nubian (derived environments; Bell 1971)						
f. Cochin Malayalam (Nair 1979), Udaiyar Tamil (Williams & Jayapaul 1977), Sudan Nubian (root-internal only; Bell 1971)						
g. Fula (Maddieson 1984)						

The cases of greatest interest here are (9e) and (9f), which show languages that allow all of the voiced singletons but only some of the voiced geminates. These cases are crucial to the comparison at hand (they are allowed by (6) but not (8)), so we discuss them in greater detail.

A dialect of Sudanese Nubian (Nilo-Saharan; Bell 1971), first discussed in this connection by Ohala (1983), disallows [d<sub>3</sub>:] and [g:] root-internally but does allow [b: d:]. Derived geminates pattern differently: derived [b:] but not [d:] is preserved as such, with only occasional devoicing of [b:], as seen below in (10).

(10)

Stem	Stem + /go:n/ 'and'	Gloss
[fag]	[fak:o:n]	'and goat'
[kad <sub>3</sub> ]	[katʃ:o:n]	'and donkey'
[kid]	[kit:o:n]	'and rock'
[fab]	[fab:o:n], occasional [fap:o:n]	'and father'

As (10) shows, suffixes like /-go:n/ cause gemination of a preceding non-continuant. Gemination entails obligatory devoicing for non-labial stops. There is a difference, then, between the obligatory devoicing of derived [d:] (cf. [kit:o:n] from /kid-go:n/, expected \*[kid:o:n]) and the preservation of root-internal [d:] (e.g. [ed:i] 'hand'). The devoicing of [d:] in derived environments

can be interpreted as an emergence of the unmarked effect (McCarthy and Prince 1994, McCarthy, in press):<sup>11</sup> hence the markedness ranking [d:] < [t:]. The fact that derived [b:] normally surfaces intact suggests a markedness difference relative to derived [d:], which must devoice: this supports the further scale fragment [d:] < [b:]. Moreover, since non-derived [b:] and [d:] are preserved, while [g:] is impossible across the board, a further scale section is established: [g:] < [d:] < [b:]. Finally, singletons are not subject to even optional devoicing, unlike [b:]. We can infer from this that [b:] < [g, d, b]. The Nubian data thus supports a voicing markedness scale that distinguishes at least four intervals: [g:] < [d:] < [b:] < [g d b].

The Nubian pattern of selective voicing neutralisation in geminates is not isolated. A closely related system appears in Kadugli (Niger-Congo; Abdalla 1973): here all voiced singletons are permitted, as well as [b:] and the implosives [ɓ: ɗ:]. No other voiced geminate obstruents occur. Voiceless geminates are found at all points of articulation, including [p: t: k:], but voiced counterparts of the non-labials [d: ɗ: g:] are impossible. Note the \*[d:] vs [ɗ:] difference: larynx lowering in [ɗ:] sustains voicing. Moreover, as seen in (9), some languages exclude just geminate [g:], allowing [b:], [d:], and all singleton voiced C's.

Of related interest to the discussion of voicing markedness is the fact that Nubian lacks [p], a gap related to aerodynamic factors reviewed by Ohala (1983). A short [p] must be actively devoiced, unlike stops at other points of articulation. But [p:] and [p] differ, because the longer duration of [p:] allows it to reach unassisted the point of passive devoicing. In Nubian, this explains why [p] is absent, while [p:] is allowed to arise. We return to this point in section 5.7.

The patterns reviewed in this section and the overall picture in (9) exceed the predictive powers of the most modular statement of voicing difficulty examined, the duo of scales in (8). This is because (8), by hypothesis, limits markedness comparisons to very simple, minimally different pairs of abstract phonological categories: geminates vs singletons and labials vs coronals vs dorsals. This argues that the mapping from voicing difficulty to markedness scales must be more direct and consequently that the scales, and thus the grammars, reflect in greater detail the complexity of phonetic difficulty. The same conclusion is echoed in this volume in the chapters by Kirchner and Zhang.

### 5.5 *Markedness scales and language-specific phonetics*

In comparing (6) and (8), we found that (6), an approach that sacrifices some degree of formal simplicity in order to reflect more closely the asymmetries of production and perception, achieves better descriptive coverage, notably of asymmetrical systems like Nubian. Yet even (6) is not a purely phonetically

based system: it uses standard phonological categories, and refers to only two of the many factors that can influence obstruent voicing. A more thoroughgoing option would be to state that any factor whatsoever that influences difficulty of voicing can be reflected in the constraints and their ranking. This is outlined in the phonetic scale of (11):

- (11) [+voice] { $x < y$ }, where  $x, y$  is any pair of voiced segments or voiced sequences, such that, without active oral tract expansion, the ratio of voiced closure to total closure duration is less in  $x$  than in  $y$ .

This is not a fixed list of sounds but a schema for generating phonetic difficulty scales based on knowledge about the phonetic factors that contribute to voicing maintenance. Such a schema would be expected to respond to fine-grained differences in how particular phonological categories are realised phonetically in individual languages.

Suppose, for instance, that in some particular language [d] is a brief flap-like constriction and [b] is a full stop. In such a case, (11) may predict, depending on the specifics of the durational difference, that [+voice] {[b] < [d]}, contrary to (6) and (8). There are in fact languages that allow [d] but not [b] (Maddieson 1984); but the comparative duration of these [d] relative to other voiced stops is not known to us.

There is some evidence that languages indeed deploy phonological constraints based on the conditions set up by language-specific phonetic factors. Zhang's chapter provides an interesting case, which we review here. In Standard Thai, CVR syllables (V = short vowel, R = sonorant consonant) have richer tone-bearing possibilities than CV:O (V: = long vowel, O = obstruent). In particular, CV:O in Thai cannot host LH or M tones, whereas CVR can host any of the five phonemic tones of the language. The Navajo pattern is close to being the opposite: CV:O can host any phonemic tone (H, L, HL, LH), but CVR cannot host HL or LH.

To explain this type of language-specific difference, Zhang proposes that what licenses contour tones is a combination of length and sonority: vowels make better contour hosts than consonantal sonorants, but, at equal sonority levels, the longer sonorous rhyme is the better carrier. In Zhang's Navajo data, CVR and the V: portion of CV:O are very close in duration. Thus, the sonority difference of R in CVR versus the second half of the long vowel in CV:O implies that it should be CV:O that is the better tone bearer, and the phonology bears this out: CV:O can host more contours.

In contrast, for Thai, it is CVR that is tonally free and CV:O that is restricted. The source of this reversal vis-à-vis Navajo is evidently a pattern of allophony present only in Thai: long vowels are dramatically shorter in closed syllables. As a result, Thai CV:O has considerably less sonorous rhyme duration than CVR, and the difference is plausibly enough to compensate for CVR's inferior

sonority profile. The upshot is that a language-specific difference of allophonic detail – degree of shortening in closed syllables – is apparently the source of a major phonological difference, namely in the tone-bearing ability of different syllable types.

This example is striking evidence for the view that at least some of the markedness scales relevant to phonology must be built on representations that contain language-specific phonetic detail: there is, as Zhang argues at length, a cross-linguistically unified theory of optimal contour carriers, based on a single scale of sonorous rhyme duration. But specific rhymes can be ranked on this scale only when their (non-contrastive, language-specific) durations are specified, not by comparing more schematic representations like CVR to CV:O.

Similar conclusions on the nature of markedness scales follow from Gordon's work on weight (chapter 9), which demonstrates that the typology of optimal stress-bearing syllables is generated by scales of total perceptual energy (integration of acoustic energy over time within the rhyme domain). Gordon shows that language-specific facts about coda selection explain why some languages (e.g. Finnish) count VC and VV rhymes as equally heavy, while others (like Mongolian) rank VV as heavier. Relevant in the present context is that Gordon's results, like Zhang's, do not support universal scales composed of fixed linguistic units (say fixed rhyme types like  $V:C_0 > VCC_0 > V$ ) but rather schemas for generating, on the basis of language-specific information, scales of weight or stressability. The advantage of this approach in Gordon's case is that it reveals the basis on which specific languages choose to count specific rhyme types as heavy or light, a choice long believed to be arbitrary.

### 5.6 *The stabilisation problem*

If phonetic factors that are allophonic matter to phonological patterning, we must consider the fact that a great deal of allophonic variation is optional and gradient. If such variation bears on phonology, we would expect to see a number of phonological effects that seem to be missing. For example, we are not aware of any sound system in which slowed-down speech, or phrase-level lengthening, causes categorical obstruent devoicing, for either geminates or singletons. Conversely we know of no case in which fast speech allows voicing distinctions to emerge that are absent at normal rates.

These are instances of what we call the *stabilisation problem*: maintaining a (relatively) stable phonology in the face of extensive variation in the phonetic factors that govern the phonological constraints. The stabilisation problem arises in all markedness domains that one might plausibly link to perception and production factors: most types of articulatory and perceptual difficulties are exacerbated by either excessive or insufficient duration, yet variation in speech rate is seldom associated with phonological neutralisation.

The stabilisation problem can be addressed in a number of ways. One possibility, suggested by Steriade (1999), is to suppose that the computation of optimal candidates is carried out relative to a standard speaking rate and style; stabilisation arises when outputs at other rates and styles are bound to the standard outputs by correspondence constraints. Another approach, suggested by Hayes (1999), posits that phonological learning involves testing candidate constraints against aggregated phonetic experience, stored in a kind of map; those phonological constraints are adopted that achieve a relatively good match to aggregated phonetic experience; thus all speaking rates and style contribute together to constraint creation. For further discussion of stabilisation, see Boersma 1998, Kirchner 1998, Flemming 2001, Pierrehumbert 2001, and Zhang's chapter.

We have compared so far the predictions of three different ways of encoding voicing markedness, making the assumption that the set of Markedness constraints reflects directly properties of phonetic difficulty scales. We have seen that simple statements of markedness like (8), which break down continua of phonetic difficulty into multiple unrelated scales, are unable to reflect cross-class markedness relations such as  $[d:] < [b:] < [g]$  or  $[d:] < [d']$ . For the voicing example considered, the evidence suggests that adherence to a tight-fisted criterion of formal simplicity is therefore untenable. Moreover, we have seen evidence that phonetically based constraints cannot be stated with a priori phonological categories, as in (6), because the phonetic details of how phonological categories are implemented in particular languages turn out to matter to the choice of constraints and their ranking.

### 5.7 *The tension between formal symmetry and phonetic effectiveness*

Cases like the Nubian voicing phenomena are perhaps eye-opening to many phonologists. Nubian appears to pursue the goal of a good phonetic fit despite the phonological asymmetry that is involved: the set of voiced stops that is allowed in the derived contexts of Nubian is the unnatural class  $[b d g b:]$ . Such cases lead one to wonder whether adherence to phonetic factors can give rise to phonological asymmetry on an unlimited basis.

In addressing this question, we should remember that the complexity seen in Nubian only scratches the surface. There are other factors besides gemination and place of articulation that influence voicing, notably whether an obstruent follows another obstruent, or whether it is postnasal or not. Since these factors all impinge on the crucial physical parameter of transglottal airflow, they trade off with one another, just as place and gemination do. Each factor geometrically increases the space of logical possibilities that must be considered in formulating constraints.



Evidence from vocal tract modelling (Hayes 1999), which permits phonetic difficulty to be estimated quantitatively, indicates that pursuing the imperative of good phonetic fit can give rise to hypothetical phonological patterns considerably more complex than Nubian. Consider, for instance, a hypothetical language in which the conditions of (12) hold true

- (12) a. [b] is illegal only after obstruents;  
 b. [d] is illegal after obstruents and initially; and  
 c. [g] is illegal anywhere other than postnasal position

Modelling evidence indicates that this is a system that has a very close fit to the patterns of phonetic difficulty. However, a pattern with this level of complexity has not been documented.

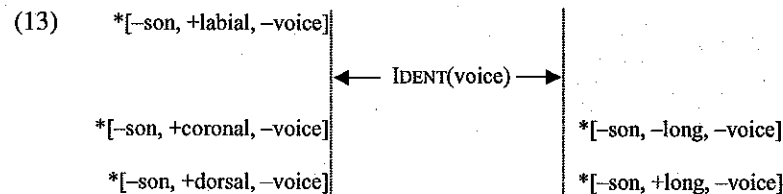
The question of whether there is an upper complexity limit for phonological constraints has also been explored by Gordon (chapter 9), who fitted a large set of logically possible phonological criteria to amplitude and duration measurements made on a variety of languages. Gordon's goal was to assess how well these criteria can classify the syllables of individual languages into groups whose rhymes maximally contrast for total acoustic energy, which appears to be the primary phonetic basis of syllable weight. Gordon finds that the best-distinguished classification often can be achieved by employing a formally very complex phonological distinction – which is never the distinction actually used by the languages in question. Instead, languages evidently adopt whichever of the formally simpler distinctions best matches the patterns of total rhyme energy seen in their syllables. Gordon's conclusion is that formal simplicity places a limiting role on how closely phonetic effectiveness can define phonological constraints.

A puzzle arises here. On the one hand, Gordon found a rather strict limit on the complexity of weight criteria (essentially, two phonological predicates). On the other hand, in the area of segment inventories, languages seem to tolerate complex and asymmetrical systems like Nubian (see scale in (6), which employs minimally four predicates per constraint). Why is the drive for formal simplicity stronger in weight computation? We conjecture that this has to do with the relatively greater difficulty in learning syllable weight categories as compared to segmental categories. Syllables are not actually *heard* as heavy or light; they are *categorised* as such, and this knowledge can only come from an understanding of the prosodic phenomena of the language that depend on weight. Moreover, the primary system reflecting weight, namely stress, is often itself rather complex and difficult to learn. Therefore, any hypothesis about syllable weight is itself dependent for its verification on another complex system, that of stress. Things are different in the case of segmental inventories;

if the grammar under consideration predicts that particular segments should or should not exist, this can be verified fairly directly. Perhaps for this reason, simplicity in computation is not at a premium for inventories and alternations.

Does formal symmetry nevertheless sometimes play a role in determining segment inventories? A possibly relevant case again involves obstruent voicing. We noted in section 5.4 above that the conditions permitting voicelessness in obstruents are essentially the opposite of those for voiced obstruents: [p] is the most difficult obstruent to keep voiceless (particularly in voicing-prone environments, such as intervocalic position); it is followed in order by [t k p: t: k:]. In light of this it is puzzling that Arabic bans geminate [p:], but allows [t k], thus permitting the more difficult sounds and disallowing the easier.

We can interpret this pattern along lines parallel to (8), with [–voice] substituted for [+voice]. There are two families of constraints for [–voice], one based on place, the other on length, with each ranked a priori according to phonetic difficulty. IDENT(voice) is ranked with respect to them as shown in (13); this derives the voiceless inventory [t k t: k:].<sup>12</sup>



Thus, it is possible that languages can vary according to whether the constraints that regulate any particular phenomenon are detailed and closely tailored to phonetics, as in (6), or more general and related to phonetics more abstractly, as in (8) or (13). At present, it appears that both hypotheses like (6) and hypotheses like (8/13) undergenerate, suggesting we cannot account for all the facts unless both are allowed.

## 6 Markedness scales beyond voicing

The voicing example has outlined some of the issues that arise when we pursue systematically the hypothesis that knowledge of Markedness constraints stems from knowledge of phonetic difficulty. We now connect these issues to the contents of the book, outlining the empirical domains covered by the other chapters and pointing out formal parallels to the voicing case.

6.1 *Scales of perceptibility*

A central ingredient in the analyses of segmental phonology are the scales of perceptibility. Certain featural distinctions are more reliably perceived in some contexts than in others. Rounding is better perceived in high, back, and long vowels than in non-high, front, and short vowels (Kaun, chapter 4). Place distinctions in consonants are better perceived in fricatives than in stops; in prevocalic or at least in audibly released consonants than in unreleased ones; in preconsonantal position, a consonant's major place features are better perceived if followed by an alveolar than by a velar or labial (Wright, chapter 2; Jun, chapter 3). All vocalic distinctions are better perceived among longer or stressed vowels, than in short stressless ones (Crosswhite, chapter 7).

Relative lack of perceptibility triggers two kinds of changes: the perceptually fragile contrast is either *enhanced* (Stevens and Keyser 1989) – by extending its temporal span or increasing the distance in perceptual space between contrast members – or it is *neutralised*. Kaun's chapter explores enhancement. She argues that rounding harmony is a contrast enhancement strategy: a vowel whose rounding is relatively harder to identify extends it to neighbouring syllables. In this way, what the feature lacks in inherent perceptibility in its original position it gains, through harmony, in exposure time. The key argument for harmony as a strategy of contrast enhancement – and thus for linking the phonology of rounding to the phonetics of perceptibility – comes from observing systems in which only the harder-to-perceive rounded vowels act as triggers. Thus in some languages only the short vowels trigger harmony, in others just the non-high vowels, or just the front vowels, or just the non-high front vowels. More generally, when specific conditions favour certain harmony triggers, these conditions pick out that subset of vowels whose rounding is expected a priori to be less perceptible compared to the rounding of non-triggers. It is these generalisations on triggers that support the idea of harmony as perceptual enhancement.

According to Crosswhite (chapter 7), enhancement and neutralisation of perceptually difficult contrasts are not incompatible strategies. Certain types of vowel reduction display both. Crosswhite notes that the lowering of stressless mid vowels (as in Belarussian) creates a stressless vowel inventory [a i u] whose elements are maximally distinct acoustically. The lowering of [e o] to [a] neutralises the mid-low contrast, but contrast enhancement is also needed to explain why the non-high vowels fail to shift to [ə] (an option exercised by a different reduction type), but rather lower to [a].

Better documented are cases in which the less perceptible features are eliminated altogether. The class of phenomena discussed in Jun's chapter (see also Jun 1995; Myers 1997; Boersma 1998: ch. 11) involve perceptibility scales for consonantal place. Jun argues that place assimilation is just one more

consequence of the general conflict between effort avoidance – whose effect is to eliminate or reduce any gesture – and perceptibility sensitive preservation. The latter corresponds, in Jun's analysis, to a set of constraint families whose lower-ranked members identify less perceptible gestures as more likely to disappear. Thus corresponding to the scales in (14), Jun proposes the families of correspondence constraints in (15):

- (14) a. perceptibility of C-place: {(strident) fricative > stop > nasal}  
 b. perceptibility of C-place: {velar > labial > coronal}  
 c. perceptibility of C-place: {before V > before coronal C > before non-coronal C}
- (15) a.  $\text{PRES}(\text{pl}(\frac{\text{C}}{[+\text{cont}]})) \gg \text{PRES}(\text{pl}(\frac{\text{C}}{[\text{stop}]})) \gg \text{PRES}(\text{pl}(\frac{\text{C}}{[\text{nasal}]}))$   
 b.  $\text{PRES}(\text{pl}(\text{dorsal})) \gg \text{PRES}(\text{pl}(\text{labial})) \gg \text{PRES}(\text{pl}(\text{coronal}))$   
 c.  $\text{PRES}(\text{pl}(-\text{V})) \gg \text{PRES}(\text{pl}(-[\frac{\text{C}}{+\text{coronal}}])) \gg \text{PRES}(\text{pl}(-[\frac{\text{C}}{-\text{coronal}}]))$

Unlike the voicing scales discussed above, the three scales in (14) represent independent dimensions of perceptibility, hence do not seem to be reducible to a single scale: the scales in (14b, c) reflect the effect of the external context (duration of vocalic transitions; masking effect of following segment) on the perceptibility of C-place, while (14a) ranks the effectiveness of place cues internal to the segment. Correspondingly, Jun observes variation in the typology of place assimilation, suggesting that the manner of the target consonant, the place of the target, and the context of assimilation do not interact and are not mutually predictable. This is what one might expect given the option of intersecting at different points the distinct constraint hierarchies in (15).

The phonological relevance of the perceptibility scales is strengthened by the broader correlation between perceptibility and neutralisation (Steriade 1999). Normally, place distinctions are better identified in pre- than post-V position (Fujimura, Macchi and Streeter 1978; Ohala 1990). However, one-place contrast (that between apico-alveolars like [t] and retroflexes like [ɖ]) concentrates essential place cues in the V-to-C transitions and thus is best perceived if the apicals are postvocalic. Indeed, confusion rates among apicals – but not other C-places – rise steeply in contexts where V-to-C transitions are absent (Ahmed and Agrawal 1969; Anderson 1997). The phonology of place neutralisation is sensitive to this difference in the contextual perceptibility of different place contrasts. In a VC<sub>1</sub>C<sub>2</sub>V sequence, assimilation for major place features (dorsal, coronal, labial) targets C<sub>1</sub>. This follows, as Jun notes, from the fact that, in VC<sub>1</sub>C<sub>2</sub>V, C<sub>1</sub> occupies a lower rank in the place

perceptibility scale relative to the  $C_2$ . But this only follows for major place and not for the apical place contrast [t] vs [t̥]: apicals in  $C_2$  position of  $VC_1C_2V$ , should be less perceptible, hence more likely to neutralise, than postvocalic apicals in  $C_1$  position. This is indeed what happens: non-assimilatory neutralisation always targets  $C_2$  apicals in  $VC_1C_2V$  strings (Hamilton 1996; Steriade 1995); and moreover place assimilation in apical clusters is predominantly progressive (Steriade 2001): we find mostly  $/VtV/ \rightarrow [Vt̥V]$  and  $/Vt̥V/ \rightarrow [VtV]$  assimilations.<sup>13</sup> As before, this observation suggests that phonological constraints track the phonetic difficulty map rather faithfully: we do not observe the adoption of any general-purpose context of place licensing, employed for all contrasts, regardless of differences in their contextual perceptibility.

One of the many questions left open by the study of perceptibility on segmental processes relates to the choice between the strategies of place enhancement and place neutralisation. Thus Jun's study of C-place neutralisation, when read in the light of Kaun's results on V-place enhancement, invites the speculation that there exists a parallel typology of C-place enhancement that affects preferentially those Cs whose place specifications are either inherently or contextually weaker. Thus, if every perceptually weak segment is equally likely to be subject to either place enhancement (say via V-epenthesis) or to place neutralisation, then the preferential targets of C-place assimilation identified by Jun should correspond, in other systems, to preferential triggers of epenthesis. We are unaware of cases that fit exactly this description; however, Wright (1996) and Chitoran, Goldstein and Byrd (2002) have documented timing differences among CC clusters tied to differences in perceptibility: the generalisation emerging from these studies is that  $C_1C_2$  clusters containing a less perceptible oral constriction in  $C_1$  typically tolerate less overlap. Further research is needed to determine whether the polar strategies of enhancement and neutralisation are equally attested across all contrast types.

## 6.2 Scales of effort

One option we did not explore in the earlier discussion of voicing scales like (5) ( $\{g: < d: < b: < g < d < b\}$ ) was to identify more directly the difficulty posed by voicing maintenance with biomechanical articulatory effort. This is the strategy pursued by Kirchner (chapter 10) in analysing consonant lenition. Kirchner draws several comparisons, some of which are outlined below, and which suggest a global connection between patterns of lenition and effort avoidance.

### (16) Effort avoidance and lenition patterns: three comparisons

(a) Vertical displacement of articulators active in C constriction	Greater displacement stop	Lesser displacement approximant
(b) Rate of change	Faster displacement V-stop-V (fast rate)	Slower displacement V-stop-V (slow rate)
(c) Jaw displacement in C constriction relative to neighbouring V	Greater displacement $\left[ \begin{array}{c} v \\ +low \end{array} \right] //stop$	Lesser displacement $\left[ \begin{array}{c} v \\ +high \end{array} \right] //stop$
(d) Number of jaw displacement gestures	Two gestures: C-to-V and V-to-C VCV	One gesture C-to-V or V-to-C (C)CV, VC(C)

Lenition typically turns stops into approximants and, as the comparison in (16a) suggests, this substitutes a less extreme displacement for a more extreme one. Lenition is also more likely at faster rates, a point Kirchner exemplifies with evidence from Tuscan Italian: (16b) suggests that at faster rates the articulators active in Cs have to accelerate in order to cover the same distance to the constriction site in less time. Thus the faster rate makes it more urgent that a less effortful approximant constriction be substituted for the more effortful stop. In Tuscan (and elsewhere: cf. Kirchner 1998) lenition is more likely when one or both flanking vowels are low or at least non-high; less likely if both vowels are high. (16c) suggests that the lower jaw position of low vowels adds to the distance that the articulators must cover in order to generate a stop constriction. Again, the additional effort required here makes it more likely that the active consonantal articulators will fall short of the target, and thus more likely that an approximant will be substituted for the stop. Finally, lenition in one-sided V contexts (pre- or post-V) implies lenition in double sided  $V\_V$ . This can be tied, as (16d) suggests, to the larger number of jaw displacement gestures required in  $V_1CV_2$  (jaw raising from  $V_1C$  and lowering from  $C$  to  $V_2$ ) relative to (C)CV or VC(C).

Rather than recognise as many isolated scales of articulatory difficulty as there are comparisons like (16) – a safe but less interesting move – Kirchner

opts for a single scale of biomechanical effort, which underlies all of them. This scale generates a single constraint family – LAZY – whose members penalise articulations in proportion to the degree of effort exertion they entail. This makes it possible to compare disparate gestures, not just oral constrictions, as realised in diverse contexts: the common grounds for the comparison between them being the level of effort expenditure required of each. (Faithfulness constraints cut down on the range of possible articulatory substitutions.) The clear benefit is that, when an independent method for identifying effort costs is found, a precise and elaborate system of predictions will be generated about the circumstances under which one articulation replaces another.

### 6.3 *Scales combining effort and perceptibility*

Zhang's study of contour tone licensing (chapter 6) offers an additional possibility: instead of constraint families based exclusively on articulatory or perceptual difficulty, there may be constraints that simultaneously reflect both factors. The formal apparatus Zhang develops relies ultimately on a quantitative measure one could call *steepness*: of two otherwise identical contour tones  $x$  and  $y$ ,  $x$  is steeper than  $y$  if  $x$ 's duration is shorter than  $y$ 's, or else if the pitch range covered in  $x$  is greater than in  $y$ . Thus, for example, HL on [a] is steeper than HL on [a:], as well as ML on [a].

The steepness comparisons among contour tones are similar to those drawn by Kirchner between sequences more or less likely to undergo lenition: thus the same articulatory trajectory from the low jaw/dorsum position in [a] to the high position needed for [k] is steeper if it has to be completed in less time, for instance at a faster speech rate. Responses of the system to excessive steepness are likewise similar: tonal contour flattening and stop lenition (as well as vowel reduction; see Crosswhite and Flemming's chapters) all reduce steepness.

However, Zhang's point is that, at least for contour tones, steepness is not simply a measure of articulatory difficulty: adequate duration is not only needed for the speaker to complete an articulatory trajectory but also for the listener to identify what contour tone has been articulated. Thus the steepness measure for contour tones should be neutral between articulation and perception. It remains to be seen if Zhang's effort/perceptibility scales are appropriate strictly for contour tones (and diphthongs; Zhang 2001) or whether they extend to facts now analysed by reference to scales that refer to effort or to perceptibility alone.

## 7 *How the picture changes*

In a reply to *Natural Phonology* (Donegan and Stampe 1979) and phonetic determinism (Ohala 1979), Anderson writes: 'the reason [to look for phonetic explanations] is to determine what facts the linguistic system proper is *not* responsible for: to isolate the core of features whose arbitrariness from other

points of view makes them a secure basis for assessing properties of the language faculty itself' (1981: 497). Any scholar's interest in the phonetic components of phonological markedness could in principle grow out of an Andersonian belief that we will gain a better understanding of phonology proper once we learn to extract the phonetics out of it. But the project of extracting the phonetics out of phonology can take unexpected turns: in trying to discover those aspects of phonological markedness that are 'arbitrary from other points of view', our views of phonological organisation have changed. Here we outline two changes of this nature that relate to the contents of this volume.

### 7.1 *Segment licensing: syllables vs perceptibility*

An important role of syllable structure in contemporary phonology is to deliver compact statements of permissible segment sequences. The hope is that an explicit description of minimal syllabic domains, like onsets and rhymes, should suffice to predict the phonotactic properties of larger domains, like the phonological word. Syllables look like good candidates for Anderson's 'core of features whose arbitrariness from other points of view makes them a secure basis for assessing properties of the language faculty itself', because the choice between different syllable structures seems to be simultaneously central to phonology and unrelated to any extra-grammatical consideration: what phonetic or processing factors could determine the choice between parses like [ab.ra] and [a.bra]? Syllables are also invoked as predicates in the statement of segmental constraints. Thus the fact that final or pre-C consonants are more likely to neutralise place and laryngeal contrasts is attributed (Ito 1986; Goldsmith 1990) to the idea that codas license fewer features than onsets do. Thus contexts like 'in the onset' or 'in the coda' come to play a critical rule in constraints and rules alike. The licensing ability of onsets is of interest to us precisely because it is 'arbitrary from other points of view': nothing about perception, articulation or processing leads us to expect any licensing asymmetry among syllable positions.

As shown in Wright's chapter, the content of the onset-licensing theory can be reconstructed on a phonetic basis. Steriade (1999, 2001) argues that languages tend to license segmental contrasts where they are maximally perceptible. For segments of low sonority this is harder to do, because the perceptibility of a low-sonority segment depends not on its own internal acoustic properties (e.g., all stops sound alike during closure), but on the *external cues* present on neighbouring high-sonority segments, which are created by coarticulation. Thus, there is strong pressure for low-sonority segments to occur adjacent to high-sonority segments. Moreover, not all forms of adjacency are equal: for the psychoacoustic reasons outlined in Wright's chapter, external cues are more salient at CV transitions than at VC transitions.

When incorporated into phonetically based constraints, these principles largely recapitulate the traditional syllable-based typology: branching onsets and codas, which are assumed to be marked, always include consonants that are suboptimally cued:  $\underline{CCV}$ ,  $\underline{VCC}$ . Moreover, the preference for cues residing in the CV transitions takes over the burden of the traditional arbitrary postulate that onsets are better licensors than codas. Thus, in the following cases,  $C_1$  normally is better cued than  $C_2$ :  $\#C_1VC_2\#$ ,  $VC_1VC_2$ ,  $CV$ .

A cue-based theory not only recapitulates the syllabic theory in non-arbitrary form, but outperforms the syllabic theory when we move beyond the broad outlines to the specific details. Thus, for instance, a preconsonantal nasal in onset position (as in many Bantu languages) is very unlikely to have place of articulation distinct from the following consonant; nor are onset obstruents that precede other obstruents (as in Polish [ptak] 'bird') likely to take advantage of their putatively privileged onset position to take on phonologically independent voicing values (as in '[btak]'). Both cases fall out straightforwardly from the cue-based theory. Wright's chapter further notes that sibilant-stop initials should be preferred to other obstruent clusters, on the grounds that sibilants, unlike stops, are recoverable from the frication noise alone. In terms of sonority sequencing, sequences like [spa] are as bad or worse than [tpa], but in terms of perceptual recovery of individual oral constrictions, there is a clear difference that favours [spa]. The typology of word initial clusters (Morelli 1999) clearly supports Wright's approach.<sup>14</sup>

Jun's survey of place neutralisation (chapter 3) also bears on the issue of onset vs coda licensing, by showing that not all codas are equally likely to neutralise: recall from (14) that nasals assimilate more than stops and stops more than fricatives, even when all three C-types are codas. What does distinguish the codas that are more likely targets of assimilation from less likely ones are the scales of perceptibility discussed earlier. Importantly, these factors explicate the entire typology of place neutralisation, with no coverage left for onset licensing. Recall further that C-place neutralisation targets onsets, not codas, whenever the C-place contrast is cued primarily by V-to-C transitions (Steriade 1999 and above): assimilation is strictly progressive in combinations of apicals and retroflexes, because these sounds are more confusable in post-C than post-V position. In this respect too a syllable-based theory of C-neutralisation simply cannot generate the right predictions.

## 7.2 Contrast and contrast-based constraints

Flemming's chapter shows that the deductive approach to markedness leads to a fundamental rethinking of the ways in which constraints operate. The issue is whether constraints evaluate individual structures – sounds or sequences – or systemic properties, such as the co-existence of certain sequences in the same

language. Flemming starts from the simple observation that perceptibility conditions cannot be evaluated by considering single sounds or single sequences: when we say, for instance, that [ĩ] and [ẽ] are more confusable than [i] and [e], we mean that [ĩ] and [ẽ] are more confusable with each other, not that they are confusable with unspecified other sounds or with silence. It matters, then, what exactly is the set of mutually confusable sounds that we are talking about.

From this it follows that if there exist phonological constraints that evaluate perceptibility, the candidates considered by those constraints consist of sets of contrasting sequences, not of individual sequences. This implies a quite radical conclusion, that OT grammars must evaluate abstract phonotactic schemata, rather than candidates for particular underlying forms, since no one individual form specifies the other entities with which it is in contrast.

Since the implications of this conclusion are daunting, it is important to determine if Flemming's proposal is empirically warranted. For instance, does it make a difference in terms of sound patterns predicted whether we say that nasalised vowels are avoided *because* they are mutually confusable (a perceptibility constraint that requires evaluating whole nasal vowel sets) or whether we say that nasalised vowels are just marked, with no rationale supplied?

Flemming's fundamental argument is that traditional OT constraints, based on Markedness and Faithfulness, simply misgenerate when applied to areas where the effect of contrast is crucial. For instance, the languages that maintain a backness distinction among high vowels could be analysed with a constraint banning central vowels (\*[i]), letting only [i] and [u] survive to the surface. The seemingly sensible \*[i] constraint becomes a great liability, however, when we consider vertical vowel systems, which maintain no backness contrasts. It is a liability because it predicts the existence of vertical systems in which the only vowel is [i] or [u]; such cases are systematically missing. Evidently, it is the perceptually salient contrast (maximal F2 difference) of [i] and [u], and not any inherent advantage of either of these two vowels alone, that causes them to be selected in those languages that maintain a backness contrast. Thus, it is the entire system of contrasts (at least in this particular domain) that the grammar must select, not the individual sounds. The constraints of conventional OT, which reward or penalise individual segments, cannot do this. Parallel results can be obtained, as Flemming shows, in the study of contrastive and non-contrastive voicing and nasality and (Padgett 2001; Lubowicz 2003) in other phonological domains as well.

## 8 Other areas

### 8.1 The role of speech processing

Frisch's chapter makes the important point that what we have been calling 'phonetic difficulty' characterises only the periphery of the human sound processing

apparatus; that is, the physical production of sound by the articulators and the initial levels of processing within the auditory system. The deeper levels of the system, such as those that plan the execution of the utterance, or that access the lexicon in production or perception, are just as likely to yield understanding of how phonology works. Frisch covers a number of areas where we might expect to find such effects, focusing in particular on how the widely attested OCP-Place effects might reflect a principle of phonological design that helps avoid 'blending of perceptual traces', and thus avoid misperception.

## 8.2 *The diachronic view of phonetics in phonology*

Blevins and Garrett's chapter takes a sharply and intriguingly different approach to the role of phonetics in phonology. Their view<sup>15</sup> is that articulatory ease and perceptual recoverability channel historical sound changes in certain directions, but lack counterparts in the synchronic grammar. Whatever the constraints may be that learners actually internalise, they are believed not to impose articulatory ease or perceptual recoverability on phonological structure.

The core of Blevins and Garrett's approach is the phenomenon of 'innocent misapprehension' (Ohala 1981, 1990). First, phonetic factors determine a pattern of low-level variation. Then, language learners assign to the forms that they are mishearing a novel structural interpretation, differing from that assigned by the previous generation; at this point, phonological change has occurred. To call this process 'innocent misapprehension' emphasises its lack of teleology: phonology is phonetically effective, not because grammars tend to be designed that way, but because innocent misapprehension allows only phonetically effective phonologies to survive.

Various other authors in our volume (Kaun, Frisch) also take the view that diachrony helps explain some aspects of phonological naturalness, and we believe there is clear empirical support for this possibility. But the heart of the controversy, and what makes it interesting to us, lies with Blevins and Garrett's view that the diachronic account suffices entirely, and that we can adopt a theory of phonology (whatever that ends up being) that is entirely blind to phonetically based markedness principles; or perhaps to any markedness principles at all.

Large differences of viewpoint are scientifically useful because they encourage participants on both sides to find justification for their opinions. In this spirit, to further the debate, we offer the following attempted justification of our own position.

First, the study of child phonology shows us many phonological phenomena that could not originate as innocent misapprehensions. Child phonology is characteristically endogenous (Menn 1983): the child inflicts her own spontaneous changes on the adult forms, which in general have been heard accurately (Smith 1973). Child-originated phonological changes often constitute solutions

to specific phonetic difficulties, and include phenomena such as cluster simplification, sibilant harmony, and [f]-for-[θ] substitution. Child-originated changes are often adopted by other children and carried over into the adult language (Wells 1982: 96).<sup>16</sup> If children can deploy phonetically natural constraints on their own, it becomes a puzzle that this very useful capacity is not employed in acquiring the adult language.

Our second objection rests on our doubt that innocent misapprehension is capable of driving systematic phonological changes (Steriade 2001). Consider, for instance, the possible roots of regressive place assimilation ( $/V\eta+bV/ \rightarrow [VmbV]$ ) in the misapprehension of the place feature of a preconsonantal nasal. Hura et al. (1992), who have investigated the phenomenon of perceptual assimilation, report that the nasals in stimuli like  $[V\eta bV]$  are indeed misperceived, but not primarily in an assimilatory fashion. They suggest, then, that simple confusion cannot alone explain the typological fact that nasals frequently assimilate in place to a following obstruent. Confusion alone would predict some form of nonassimilatory neutralisation. Thus, unless there is some factor present in real language-change situations that was absent in Hura et al.'s experiments, 'innocent misapprehension' seems to lack the directional stability that would be needed for it to drive diachronic change.

Lastly, we consider the typology of stop-sibilant metathesis (Hume 1997, Steriade 2001, and Blevins and Garrett's chapter) as supporting the teleological approach to phonology assumed in phonetically based OT. The crucial observation is that stop-sibilant metathesis acts to place the stop – which requires external cues more strongly than the sibilant does – in a position where the best external cues will be available. Usually, this means that the stop is placed in prevocalic (or merely released) position; thus  $/VksV/ \rightarrow [VskV]$  is phonetically optimising. The single known exception (Blevins and Garrett, section 3.4) occurs in a strong-stress language, where it is plausible to assume that post-tonic position provides better cues than pre-tonic position; hence  $/VskV/ \rightarrow [VksV]$ . This cross-linguistic bias in metathesis is unexpected if stop-sibilant metathesis is merely random drift frozen in place by innocent misapprehension, but makes sense if it is implemented 'deliberately' in language, as a markedness-reducing operation.

We believe that most of the evidence that could bear on either side's position remains to be gathered or considered, and thus that further attention to this debate could lead to research progress.

## References

- Abaev, Vasilii I. (1964). *A grammatical sketch of Ossetic*. The Hague: Mouton.  
 Abdalla Ibrahim Abdalla (1973). *Kadugli language and language usage*. Institute of African and Asian Studies. University of Khartoum, Sudan.

- Ahmed, R. and S. S. Agrawal (1969). Significant features in the perception of (Hindi) consonants. *JASA* 45: 758–63.
- Anderson, Stephen R. (1981). Why phonology isn't 'natural'. *LJ* 12: 493–539.
- Anderson, Victoria B. (1997). The perception of coronals in Western Arrernte. In G. Kokkinakis et al. (eds.), *Proceedings of the 5th European conference on speech communication and technology*, Vol. 1. University of Patras, Greece, 389–92.
- (2000). *Giving weight to phonetic principles: The case of place of articulation in Western Arrernte*. PhD dissertation, University of California, Los Angeles.
- Archangeli, Diana and Douglas Pulleyblank (1994). *Grounded phonology*. Cambridge, MA: MIT Press.
- Barnes, Jonathan (2002). *The phonetics and phonology of positional neutralization*. PhD dissertation, University of California, Berkeley.
- Baroni, Marco (2001). How do languages get crazy constraints? Phonetically-based phonology and the evolution of the Galeata Romagnolo vowel system. *UCLA Working Papers in Phonology* 5: 152–78. <http://home.sslmit.unibo.it/~baroni/research.html>.
- Beckman, Jill N. (1998). *Positional Faithfulness*. PhD dissertation, University of Massachusetts, Amherst.
- Beddor, Patrice, Rena Krakow, and Stephanie Lindemann (2001). Patterns of perceptual compensation and their phonological consequences. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. San Diego: Academic Press, 55–78.
- Bell, Alan (1978). Syllabic consonants. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.), *Universals of human language*. Stanford: Stanford University Press, vol. 2, 153–201.
- Bell, Herman (1971). The phonology of Nobiin Nubian. *African Language Review* 9: 115–39.
- Boersma, Paul (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.
- Browman, Catherine P. and Louis Goldstein (1990). Tiers in articulatory phonology, with some implications for casual speech. In John Kingston and Mary Beckman (eds.), *Papers in Laboratory Phonology I*. Cambridge: Cambridge University Press, 341–76.
- Calabrese, Andrea (1995). A constraint-based theory of phonological markedness and simplification procedures. *LJ* 26: 373–463.
- Casali, Roderic (1997). Vowel elision in hiatus contexts: Which vowel goes? *Lg* 73: 493–533.
- Chitoran, Ioanna, Louis Goldstein and Dani Byrd (2002). Gestural overlap and recoverability: Articulatory evidence from Georgian. In Carlos Gussenhoven and Natasha Warner (eds.), *Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 419–48.
- Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. New York: Harper and Row.
- Crothers, John (1978). Typology and universals of vowel systems. In Joseph H. Greenberg et al. (eds.), *Universals of human language*. Stanford: Stanford University Press, vol. 2, 93–152.
- De Lacy, Paul (2002). *The formal expression of markedness*. PhD dissertation, University of Massachusetts, Amherst.
- Donegan, Patricia (1993). On the phonetic basis of phonological change. In Charles Jones (ed.), *Historical linguistics: Problems and perspectives*. London: Longman, 98–130.
- Donegan, Patricia J. and David Stampe (1979). The study of natural phonology. In Daniel A. Dinnsen (ed.), *Current approaches to phonological theory*. Bloomington: Indiana University Press.
- Flemming, Edward (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18: 7–46.
- (2002). *Auditory representations in phonology*. New York: Routledge.
- Fujimura, Osamu, Marian J. Macchi, and L. Streeter (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech* 21: 337–46.
- Gafos, Adamantios (1999). *The articulatory basis of locality in phonology*. New York: Garland Publishers.
- Gamkrelidze, T. V. (1978). On the correlation of stops and fricatives in a phonological system. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.), *Universals of human language*. Stanford: Stanford University Press, vol. 2, 9–46.
- Goldsmith, John (1990). *Autosegmental and metrical phonology*. Oxford: Basil Blackwell.
- Grammont, Maurice. (1933). *Traité de phonétique*. Paris: Delagrave.
- Greenberg, Joseph (1978). Some generalizations concerning initial and final consonant clusters. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravcsik (eds.) *Universals of human language*. Stanford: Stanford University Press, vol. 2, 243–80.
- Guion, Susan G. (1996). *Velar palatalization: Coarticulation, perception, and sound change*. PhD dissertation, University of Texas, Austin.
- Halle, Morris and Kenneth Stevens (1971). A note on laryngeal features. In *Quarterly Progress Report 101*. Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology, 198–213.
- Hamilton, Philip (1996). *Phonetic constraints and markedness in the phonotactics of Australian aboriginal languages*. PhD dissertation, University of Toronto.
- Hansson, Gunnar (2001). *Theoretical and typological issues in consonant harmony*. PhD dissertation, University of California, Berkeley.
- Hayes, Bruce (1999). Phonetically-driven phonology: The role of Optimality Theory and inductive grounding. In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly (eds.), *Functionalism and formalism in linguistics, volume I: general papers*. Amsterdam: John Benjamins, 243–85.
- Hockett, Charles F. (1955). *A manual of phonology*. Baltimore: Waverly Press.
- Hume, Elizabeth (1997). The role of perceptibility in consonant/consonant metathesis. *WCCFL* 17: 293–307.
- Hume, Elizabeth and Keith Johnson (2001a). *The role of speech perception in phonology*. San Diego: Academic Press.
- (2001b). A model of the interplay of speech perception and phonology. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. San Diego: Academic Press, 3–26.



- Hura, S., B. Lindblom, and R. Diehl (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech* 35. 1–2: 59–72.
- Hyman, Larry M. (2001). The limits of phonetic determinism in phonology: \*NC revisited. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. San Diego: Academic Press, 141–85.
- Ito, Junko (1986). *Syllable theory in prosodic phonology*. PhD dissertation, University of Massachusetts, Amherst.
- Ito, Junko and Armin Mester (1995). Japanese phonology. In John Goldsmith (ed.), *The handbook of phonological theory*. Oxford: Blackwell, 817–38.
- Jakobson, Roman (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Uppsala: Almqvist & Wiksell.
- Jun, Jongho (1995). *Perceptual and articulatory factors in place assimilation: An optimality theoretic approach*. PhD dissertation, University of California, Los Angeles.
- Kager, René (1999). *Optimality Theory*. Cambridge: Cambridge University Press.
- Kavitskaya, Darya (2002). *Compensatory lengthening: Phonetics, phonology, diachrony*. New York: Routledge.
- Keating, Patricia A. (1985). Universal phonetics and the organization of grammars. In Victoria Fromkin (ed.), *Phonetic linguistics*. Orlando, FL: Academic Press, 115–32.
- Kingston, John and Randy Diehl (1994). Phonetic knowledge. *Lg* 70: 419–53.
- Kirchner, Robert (1997). Contrastiveness and Faithfulness. *Phonology* 14: 83–111.
- (1998). *An effort-based approach to consonant lenition*, PhD dissertation, University of California, Los Angeles. Rutgers Optimality Archive 276, <http://roa.rutgers.edu>.
- Kochetov, Alexei (2002). *Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization*. London: Routledge.
- Krueger, John R. (1962). *Yakut manual*. Bloomington: Indiana University.
- Lehiste, Ilse (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Lindblom, Björn (1986). Phonetic universals in vowel systems. In John J. Ohala and Jeri Jaeger (eds.), *Experimental phonology*. Orlando: Academic Press, 13–44.
- (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech production and speech modelling*. Dordrecht: Kluwer, 403–39.
- Lindblom, Björn and Olle Engstrand (1989). In what sense is speech quantal? *JPhon* 17: 107–21.
- Lombardi, Linda (in press). Coronal epenthesis and markedness. *Phonology* 19: 219–51. Rutgers Optimality Archive 579.
- Lupas, L. (1972). *Phonologie du grec attique*. Mouton: The Hague.
- Lubowicz, Ania (2003). *Contrast preservation in phonological mappings*. PhD dissertation, University of Massachusetts, Amherst.
- Maddieson, Ian (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.
- McCarthy, John (in press). Comparative markedness. *Theoretical Linguistics*.
- McCarthy, John and Alan Prince (1994). The emergence of the unmarked: Optimality in prosodic morphology. *NELS* 24: 333–79.
- Menn, Lise (1983). Development of articulatory, phonetic, and phonological capabilities. In Brian Butterworth (ed.), *Language production*. London: Academic Press, vol. 2, 3–50.
- Morelli, Frida (1999). *The phonotactics and phonology of obstruent clusters in Optimality Theory*. PhD dissertation, University of Maryland, College Park.

- Myers, Scott (1997). Expressing phonetic naturalness in phonology. In Iggy Roca (ed.), *Derivations and constraints in phonology*. Oxford: Clarendon Press, 125–52.
- Nair, Somasekharan P. (1979). *Cochin dialect of Malayalam*. Trivandrum: Dravidian Linguistic Association.
- Ohala, John J. (1979). Universals of labial velars and de Saussure's chess analogy. In *Proceedings of the Ninth International Congress of Phonetic Sciences*, Copenhagen, vol. 2, 41–7.
- (1981). The listener as a source of sound change. In Carrie S. Masek, Roberta A. Hendrick, and Mary Frances Miller (eds.), *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistic Society, 178–203.
- (1983). The origin of sound patterns in vocal tract constraints. In Peter F. MacNeilage (ed.), *The production of speech*. New York: Springer Verlag, 189–216.
- (1990). The phonetics and phonology of aspects of assimilation. In John Kingston and Mary Beckman (eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press, 258–75.
- Ohala, John J. and Carol Riordan (1979). Passive vocal tract enlargement during voiced stops. In J. J. Wolf and D. H. Klatt (eds.), *Speech communication papers*. New York: Acoustical Society of America, 89–92.
- Padgett, Jaye (2001). Contrast dispersion and Russian palatalization. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. San Diego: Academic Press, 187–218.
- Passy, Paul (1890). *Etude sur les changements phonétiques et leurs caractères généraux*. Paris: Firmin-Didot.
- Pater, Joe (1999). Austronesian nasal substitution and other NC effects. In Harry van der Hulst, René Kager, and Wim Zonneveld (eds.), *The prosody morphology interface*. Cambridge: Cambridge University Press, 310–43.
- Pierrehumbert, Janet (2000). The phonetic grounding of phonology. *Les cahiers de l'ICP*, Bulletin de la Communication Parlée, 5: 7–23.
- (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes* 16: 691–98.
- Podesva, Robert (2000). Constraints on geminates in Buginese and Selayarese. *WCCFL* 19: 343–56.
- Prince, Alan (2001). Invariance under re-ranking. Talk given at WCCFL XX, University of Southern California, 24 February 2001. <http://ling.rutgers.edu/gamma/talks/WCCFL2001.pdf>
- Prince, Alan and Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers Optimality Archive 537, <http://roa.rutgers.edu/>.
- Rajaram, S. (1972). *Tamil phonetic reader*. Mysore: CILL.
- Smith, Caroline (1992). *The temporal organization of vowel and consonant gestures*. PhD dissertation, Yale University.
- Smith, Neilson (1973). *The acquisition of phonology: A case study*. Cambridge: Cambridge University Press.
- Smolensky, Paul (1993). Harmony, markedness, and phonological activity. Rutgers Optimality Archive 537, <http://roa.rutgers.edu/>.
- Stampe, David (1973). *A dissertation on natural phonology*. PhD dissertation, University of Chicago. Distributed 1979 by Indiana University Linguistics Club, Bloomington.
- Steriade, Donca (1995). Positional neutralization. Ms., Department of Linguistics, UCLA. <http://www.linguistics.ucla.edu/people/steriade/papers/PositionalNeutralization.pdf>.



- (1999). Phonetics in phonology: the case of laryngeal neutralization. In Matthew Gordon (ed.), *Papers in Phonology 3*. (UCLA Working Papers in Linguistics 2) Los Angeles: Department of Linguistics, UCLA, 25–146. <http://www.linguistics.ucla.edu/people/steriade/papers/phoneticsinphonology.pdf>.
- (2001). Directional asymmetries in place assimilation: A perceptual account. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. New York: Academic Press, 219–50.
- Stevens, Kenneth N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. B. Denes and E. E. David, Jr (eds.), *Human communication, a unified view*. New York: McGraw-Hill, 51–66.
- (1989). On the quantal nature of speech. *JPhon* 17: 3–46.
- Stevens, Kenneth N. and S. Jay Keyser (1989). Primary features and their enhancement in consonants. *Lg* 65: 81–106.
- Suomi, Kari (1983). Palatal vowel harmony: A perceptually-motivated phenomenon? *Nordic Journal of Linguistics* 6: 1–35.
- Trubetzkoy, Nikolai S. (1938). *Grundzüge der Phonologie*. (Travaux du cercle linguistique de Prague, 7) Prague: Cercle linguistique de Prague.
- van Driem, George (1987). *A grammar of Limbu*. Berlin: Mouton de Gruyter.
- Wells, John (1982). *Accents of English I: An introduction*. Cambridge: Cambridge University Press.
- Werner, Roland (1987). *Grammatik des Nobiin*. Hamburg: Helmut Buske Verlag.
- Westbury, John (1979). *Aspects of the temporal control of voicing in consonant clusters in English*. (Texas Linguistic Forum, 14) Austin, TX: Dept. of Linguistics, University of Texas, Austin.
- Westbury, John and Patricia Keating (1986). On the naturalness of stop consonant voicing. *JL* 22: 145–66.
- Williams, T. Edward and V. Y. Jayapaul (1977). *Udaiyar dialect of Tamil*. Annamalainagar: Annamalai University.
- Wright, Richard (1996). *Consonant clusters and cue preservation in Tsou*. PhD dissertation, University of California, Los Angeles.
- Zhang, Jie (2001). The contrast-specificity of positional prominence – evidence from diphthong distribution. Paper presented at the 75th annual meeting of the Linguistic Society of America, Washington, DC.

### Notes

1. Indeed, the view that all the substantive elements of phonological theory are innate is not unique to OT; cf. Calabrese 1995 or Archangeli and Pulleyblank 1994.
2. See in particular work on ‘positional faithfulness’, such as Jun 1995, Casali 1997, Beckman 1998, Steriade 1995, Steriade 2001.
3. Cf. Lindblom and Engstrand 1989, Lindblom 1990.
4. See Passy 1890, Grammont 1933, Ohala 1983, 1990, Lindblom 1990, Browman and Goldstein 1990, Halle and Stevens 1973, Keating 1985, and Stevens and Keyser 1989.
5. See Chomsky and Halle 1968, Stampe 1973, and Archangeli and Pulleyblank 1994.

6. Maddieson (1984) lists Wolof as such a case; this is evidently an error; cf. forms like *japp* ‘do one’s ablutions’, *wacc* ‘leave behind’ (personal communications from Pamela Munro, Russell Schuh, and Mariam Sy).
7. See discussion of Arabic below for a possible counterexample and ways of analysing it.
8. Formally, the link between markedness scales and Optimality-Theoretic grammar can be achieved in (at least) two ways. Consider a markedness hierarchy  $M(S_1) > M(S_2) > \dots > M(S_n)$ , where  $S_1$ – $S_n$  are phonological structures and  $M(S)$  refers to their relative degrees of markedness. This hierarchy can correspond to a universally fixed ranking in which  $*S_1 \gg *S_2 \gg \dots \gg *S_n$ , as in Prince and Smolensky 1993. Alternatively (Prince 2001), the constraints on  $S_1 \dots S_n$  are formulated so that each one bans all elements on the scale at the same markedness level or higher: thus  $*S_2$  penalises  $S_2$  as well as the more marked  $S_1$  structures, whereas  $*S_1$  penalises just  $S_1$ . In this system, less marked structures like  $S_2$  are penalised by a proper subset of the constraints that ban more marked ones  $S_1$ : no fixed ranking is needed. Empirical arguments favouring the second approach are outlined in Prince 2001 and De Lacy 2002.
9. Maddieson (1984) reports seven languages with a voicing contrast limited to labials; and seventeen where labials and coronals contrast in voicing but velars do not. For discussion, see section 5.5.
10. Moreover, the constraints of (8) derive two inventories that those of (6) cannot derive:  $\{d: b: d\}$  and  $\{b: b\}$ . We return to the question of such unnatural-but-symmetrical inventories in section 5.7 below.
11. Comparable avoidance of derived-only voiced geminates is documented for Egyptian Nubian (Werner 1987) and Buginese (Podesva 2000).
12. An alternative interpretation of the missing  $[p:]$  in Arabic could invoke the fact that a majority of geminates arise through gemination of underlying singletons: if  $[p]$  is prohibited and if IDENT(voice) between correspondent segments is undominated, there will be few occasions for the geminate  $[p:]$ 's to arise. This predicts that there will be few  $[p:]$ 's in this type of system; the fact that there are none does not directly follow.
13.  $/VtV/ \rightarrow [VtV]$  and  $/VtV/ \rightarrow [VtV]$  are limited to cross-word boundary cases, where greater faithfulness plausibly protects  $C_2$ ; cf. Casali 1997.
14. Initial  $[mb]$ ,  $[pt]$ , and  $[sp]$  are sometimes considered not to consist of a single onset; rather, the initial consonant is said to be under an appendix node, attached directly to the prosodic word, or stray. Such theories must add stipulations for why these structural configurations occur where they do, and why they behave differently in licensing richer ( $[st]$ ) or more impoverished ( $*[nb]$ ,  $*[bt]$ ) phonotactic possibilities.
15. Other work along these general lines includes Ohala 1983, 1990, Suomi 1983, Guion 1996, Baroni 2001, Beddor et al. 2001, Hansson 2001, Hume and Johnson 2001b, Hyman 2001, Kavitskaya 2002, Kochetov 2002, and Barnes 2002.
16. Such changes imply the possibility of a theory that is both diachronically based (in agreement with Blevins and Garrett) and phonologically teleological (in disagreement with them).