The Cambridge Handbook of Phonology

Edited by **Paul de Lacy**

CAMBRIDGE
UNIVERSITY PRESS

7

Contrast

Donca Steriade

7.1 Introduction: basic notions and outline

Phonological representations are composed of discrete building blocks, drawn from a finite, universal set. The building blocks are feature values and segments. In the representation of any utterance, feature values are linked to each other by relations such as precedence and constituency, and form *phonemes*, or combinations of substantially overlapping features. The same relations group phonemes into larger syntagmatic units, such as syllables. Phonemes *contrast* with each other: a difference between a phoneme pair, embedded in otherwise identical contexts, normally has the potential to convey a meaning difference. This potential for contrast is not actualized in every context: when pairs of phonemes systematically fail to contrast in some position, their contrast has been *neutralized*. A phoneme has contextual variants – *allophones* – which differ from each other in feature composition. Being contextually predictable, differences between allophones cannot convey meaning and thus are *non-contrastive*. Necessarily then, features that differentiate only allophones, not phonemes, are non-contrastive. Based on the universal set, each grammar defines its inventory of phonemes and the contrastive features from which its phonemes are built. Together, the phonemes and contrastive features can be thought of as language-specific alphabets of phonological categories. Universal constraints place limits on the composition of such alphabets.

The contents of the preceding paragraph are widely assumed in pre-generative, structuralist phonology (Sapir 1933, Trubetzkoy 1939, Hockett 1955). A subset of these ideas plays a role in early generative phonology (Chomsky and Halle 1968); virtually all have informed Lexical Phonology (Kiparsky 1982a, 1985; Mohanan 1982) and continue to influence current phonological thinking. Insofar as they can be conceived of as empirical hypotheses, these ideas are increasingly under debate. In particular, the

following have been called into question: the notion of a small and universal set of features and of segment-sized feature combinations; the reality of segments in mental representations; the very existence of a clear cut between contrastive and non-contrastive categories – or of categories *tout court* – in individual grammars. Not all these debates can be discussed in this space. Lindblom (1990b), Pierrehumbert (2003a), Johnson (2004), and Port and Leary (2005) provide some important perspectives on these issues that differ substantially from the assumptions of most working phonologists. Only three broad questions are discussed below, selected for the role they play in the current analytical literature: by what formal mechanisms and at what juncture in the mapping from UR to SR are phonemic alphabets defined (Sections 7.2 and 7.3); what inventories does the grammar define: sounds, features, contrasts (Sections 7.4, 7.5, 7.6); and what factors condition neutralization (Sections 7.3, 7.4, and 7.5, passim).

7.2 Contrast beyond segments

The notions of phoneme and allophone refer to segment-sized units, but the issue of contrast arises in similar terms with larger domains and with non-featural properties. Precedence can differentiate phoneme strings – e.g. /task/ vs. /taks/ – and occasionally pairs of single phonemes (e.g. pre-nasalized /^{n}d/ vs. post-nasalized /d^{n}/). Like feature-sized properties, precedence can be contrastive, as above, or can be thought of as non-contrastive (cf. McCarthy 1989 on V-C precedence).

The relation of temporal overlap (Browman and Goldstein 1992; Sagey 1988) can also be viewed in terms of contrast and neutralization. The extent and consistency of overlap is what distinguishes a bundle of features forming a single segment from a cluster (Byrd 1996;): /k͡p/ vs. /kp/. A few languages contrast unit phonemes C^{j}, C^{w}, C^{h}, C' with corresponding Cj, Cw, Ch, C? clusters: Indo-European, for instance, is reconstructed as having both /k^{w}/ and /kw/ (Ernout and Meillet 1967:200); Yokuts contrasts /t?/ with the ejective /t'/ (Newman 1944). More frequently, however, this kind of contrast is neutralized: thus, it's not obvious what unit Latin or English <qu> spells – C^{w} or Cw – but it is clear that it doesn't contrast in either language with *the other* thing. In Takelma, heteromorphemic combinations of C+h merge with C^{h} (Sapir 1922:43); the same compression of sequenced articulations into segment-sized units functions on a much larger scale in Mazateco (Steriade 1994). It appears that small differences in degree of overlap between otherwise identical articulations are insufficiently distinct to signal a contrast like /k^{w}/ vs. /kw/ (Wright 1996). Thus whether a phoneme class is included in a phoneme inventory depends frequently on whether the language permits a certain cluster, and conversely, whether a language permits a certain cluster depends on whether the inventory contains the relevant phoneme. (More likely, both issues depend on the

inter-gestural timing relations prevalent in the language – Browman and Goldstein 1992.) Then not only does the notion of contrast extend beyond the segmental domain but also segment-internal contrasts – e.g. /k/ vs. /k^w/ – can't be separately analyzed from non-segmental contrast – e.g. /kw/ vs. /k^w/ – since some clusters give rise to segments. This diminishes the prospects for a separate statement of a language's phonemic alphabet that's somehow analytically prior to the statement of sequence phonotactics. A point we return to below is the role of the factor of *sufficient distinctness* in predicting the existence of contrasts, whether segment internal or not.

7.3 Laws constraining phonemic sets

The composition of phonemic sets is lawful and obeys universal constraints. The best understood are laws of asymmetric implication, or *implicational universals*: if certain segments are selected, then certain other segments also are. Thus if front rounded /y/ is present, then so is front unrounded /i/; but /i/ does not symmetrically imply /y/ (Jakobson 1941/1968; Maddieson 1984). Similarly, if nasalized vowels are present then nasal consonants also are. It is widely assumed, and explicitly so in OT work, that such typological observations have counterparts in the competence of individual speakers: so, not only is there no hypothetical language with a vowel inventory of {y, u, ø, o, a} – as against {i, u, e, o, a} – but, this assumption goes, such systematic gaps arise from a property present in each speaker's grammar and thus are independent of the segmental alphabet the grammar generates. This property could be a ranking between constraints on features (Prince and Smolensky 2004) or a ranking between constraints on contrasts (Flemming 2002, 2004); or a set of filters (i.e. inviolable constraints) activated only in a specific order (Chomsky and Halle 1968:410; Calabrese 1995).[1] Some of these options are discussed below.

There are less well-understood but more general laws which underlie individual implicational constraints in the formation of segmental alphabets. These involve the notions of dispersion and feature economy. *Dispersion* (Lindblom 1990b; Flemming 2004; Gordon 3.3.4) is a relation between pairs of sounds: the general idea here is that contrasting pairs separated by small distances in auditory space (e.g. /ɛ/ and /e/) imply the existence of other contrasting pairs, separated by a larger distance (e.g. /a/ and /i/). *Feature economy* (Clements 2003) is the tendency to minimize the ratio of features over segments in an alphabet. Thus the alphabet in (1) makes a more economic use of features (here [labial], [coronal], [dorsal], [–continuant], [±voice], [±nasal]) than the one in (2), which generates the same number of segments by combining more features, or than (3), which combines the same features as (1), but yields fewer segments[2].

(1) {p t k b d g m n ŋ}
(2) {p t^h k' β r g w n̪ ŋ}
(3) {b t ŋ}

These examples (adapted from Lindblom 1990b) suggest that dispersion and feature economy are, to an extent, conflicting forces: more economic alphabets have less well separated members, because fewer features distinguish them. The more surprising aspect of the comparison between (1), (2) and (3) is that feature economy and dispersion are insufficient to characterize the typology of segmental alphabets. That's because (2) and (3), which achieve vastly better dispersion at the cost of some decrease in economy, are unattested and probably impossible alphabets; in contrast, (1) is widely attested. If an unconstrained tug-of-war between dispersion and economy had been sufficient to characterize the notion of possible alphabet, it would be difficult to exclude (2) and (3). This is a point we will return to.

7.4 Underlying and derived alphabets

7.4.1 Early generative grammar

The interest in modeling the grammatical process that selects phoneme sets is recent. Structuralist and early generative analyses postulate without comment an underlying segment inventory for each language. The assumptions of lexical minimality (minimizing lexically stored information: Chomsky and Halle 1968:381; Steriade 1995:114) and feature economy (as defined above) play an implicit role in these cases. To illustrate the role of feature economy, Kenstowicz and Kisseberth (1979) argue that the underlying vowel set of Yawelmani Yokuts is (4a), as against (4b), which is closer to surface structures. The reason is, in part, that (4a) is "more symmetrical" (1979:206). Cast in feature economy terms, the point is that (4a) makes maximal use of [±long], [±high] and [±round] and eliminates the superfluous use of [±low] implicit in (4b).

(4) *Yawelmani Yokuts vowel inventories (Kenstowicz & Kisseberth 1979)*

(a) Underlying: /a i o u; aː iː oː uː/
(b) Surface: [a e i o u; aː eː oː]

Similar arguments are given by Chomsky and Halle (1968:203) for deriving surface [ʌ] from /u/ in English, by Mohanan and Mohanan (1984) for deriving Malayalam [r] from /t/, among many others. In all these cases, the feature economy arguments are supported by evidence from alternations. How speakers organize their phonemic alphabet when the evidence from feature economy and alternations fails to converge is unknown.

In early generative models, the assumption of lexical minimality has the effect of reducing the underlying alphabet to the minimal sound set needed to express surface differences between distinct morphemes. This requires

then the elimination of allophonic variants from lexical entries: the American English allophone set {[tʰ], [t̚], [ʔt], [ɾ]} for instance, would have to be reduced in the lexicon to one sound. What should the features of this sound be? Here too lexical minimality works to dictate that only those features minimally necessary to distinguish lexical items should be used. So, if [±spread glottis] does not distinguish lexical items, the lexical /t/ sound will be entered as *underspecified* for aspiration: it will bear no value for that feature. How (and whether) learners proceed to eliminate predictable feature values from lexical entries is an unresolved question: so, given that [±round] and [±back] are mutually predictable in the glides /w/ and /j/, do learners represent /w/ and /j/ as [+round] and [−round], respectively, or as [+back] and [−back]? See Dresher, Piggott and Rice (1994) for some proposals. Some doubt that lexical minimality is a useful guideline in constructing lexical entries, noting that the empirical evidence for underspecification is limited and open to a variety of interpretations (Mohanan 1991, Steriade 1995).

In Chomsky and Halle's model, the set of surface speech sounds is the result of the rules of grammar applying in sequence to representations composed, initially, of underlying segments. No regularities characterize the surface inventory. The possibility that a distinct alphabet might be defined at any derivational stage other than the underlying form – e.g. at a "systematic phonemic level" – is explicitly rejected (1968:11) for lack of empirical support. As they note, "the issue is whether the rules of grammar must be so constrained as to provide, at a certain stage of generation, a system of representation meeting various proposed conditions."

7.4.2 Lexical Phonology and Structure Preservation

It is the recognition of just such an intermediate stage of generation that distinguishes the theory of Lexical Phonology (LP; Kiparsky 1982a, 1985; Mohanan 1982) from early generative phonology and from parallelist versions of Optimality Theory. This intermediate level is the output of the lexical component (for whose attempted definitions see Kaisse and Shaw 1985; contributions to Hargus and Kaisse 1993). Here is a clear statement of this position, from Mohanan and Mohanan (1984:575): "Lexical Phonology incorporates three levels of phonological representation: underlying, lexical and phonetic. The lexical 'alphabet' consisting of the 'lexical phonemes' need not be identical to the underlying alphabet consisting of the underlying phonemes." The argument for recognizing the lexical alphabet as distinct from the underlying and phonetic ones is that speakers' judgments of identity and distinctness are rendered at the lexical level: "listeners perceive speech sounds in terms of the grid provided by the lexical alphabet of the languages they speak" (Mohanan and Mohanan 1984:596). For instance, Malayalam dental [n̪], a non-underlying segment, is generated lexically in stem-initial position.

(5) /n/ → n̪ / [stem_

Dental [n̪] is said to be non-underlying because it is in complementary distribution with alveolar [n] stem internally. The rule generating it is said to be lexical because it is conditioned by a morphological factor, the stem boundary. The Malayalam pair [n]-[n̪] is reportedly perceived as clearly distinct by Malayalam speakers. By contrast, the English [n]-[n̪] segments – with [n̪] as in [tɛn̪θ] – are not perceived as distinct by English speakers: the rule generating [n̪] is said to be postlexical and that excludes [n̪], on this view, from the lexical grid of English, explaining the distinctness judgments[3].

Even if we grant that judgments of distinctness identify a level of representation intermediate between UR and SR, there may be other ways to look at the specific data cited. For the comparison of Malayalam and English [n̪], what may be relevant is people's tendency to compensate for the effect of context, in speech and other forms of sensory perception: the dental articulation of Malayalam stem initial [n̪] cannot be attributed to a neighboring dental and that's perhaps why [n̪]'s dentality is accurately perceived. By contrast, English [n̪] arises only next to the overtly dental [θ] and thus its dentality can be parsed out of [n̪]'s percept (cf. Gow 2001, for experimental evidence on related points). This scenario also explains why English [ŋ] is perceived as distinct from [n] (Harnsberger 1999). The assimilation of [ŋ] from /n/ is the same process that creates [n̪] before [θ]. But in [ŋ]'s case, the conditioning /g/ disappears word finally, in most English dialects (/lɔng/ → [lɔŋ]): in /g/'s absence, the velarity of [ŋ] becomes salient, because it can't be attributed to the context. On this interpretation, the lexical–postlexical distinction need not be invoked in comparing English and Malayalam [n̪]. It is, in any case, difficult to invoke it: [ŋ] and [n̪] result from the same English assimilation process, but give rise to different judgments of distinctness relative to [n].

Very little empirical work addresses LP's intuition that distinctness judgements tap the lexical – as against the underlying or surface – level: see Whalen et al. (1997) and Jones (2002) on English subjects' ability to judge the contextual appropriateness of allophonic aspiration in English voiceless stops; Paradis and La Charité (2005) and Kenstowicz (2003) on whether L1 allophonic distinctions affect loan adaptation; and some of the contributions to Daniels and Bright (1996) on the derivational level tapped by writing systems. Sapir's (1933) anecdotes about his native informants' spelling preferences are frequently cited as proof of 'phonemic' as against 'phonetic' perception, but their evidence is limited and not clearly about the lexical as against the underlying level.

There are further noteworthy aspects of LP that concern contrast and allophony. First, work in LP (e.g. Kiparsky 1985, Borowsky 1989) introduces filters that jointly characterize an underlying phoneme set. The lexical inventory is defined as the set of sounds obtained by subtracting the feature combinations prohibited by lexical filters from all phoneme-sized combinations otherwise sanctioned by feature theory. To expand on an earlier example, classical LP can characterize the phonemic inventory of English in terms of conditions like (6), which prohibit /ŋ/, /n̪/, /m̪/, /n̪/ in lexical entries.

(6) (a) */ŋ/: *[+nasal, dorsal]; (c) */ɱ/: *[+nasal, labiodental]
(b) */ɲ/: *[+nasal, coronal,–anterior] (d) */n̪/: *[+nasal, +distributed]

Being barred by (6) from lexical entries, [ŋ], [ɲ], [ɱ], [n̪] surface only where the rules of English grammar derive them from other segments. This explains their predictable distribution: [ŋ] surfaces only before [k], [g] – where it is traceable to /ng/, /nk/ – or where the grammar could have eliminated an underlying /g/; [ɲ], [ɱ], [n̪] appear, optionally, only before homorganic consonants, where place assimilation might have generated them.

Unique to LP is the idea that a subset of the lexical filters constrains the effect of rule application in the lexical component. This is the hypothesis of Structure Preservation (Kiparsky 1985). English assimilates /n/ to any following stop but this process applies differently depending on whether the output is [m] as against [ŋ], [ɲ], [ɱ], [n̪]. Word-internal applications yielding [m] are unrestricted and obligatory – cf. *i[np]ermissible, *e[nb]ed. That's as predicted by Structure Preservation: /m/ is a member of the lexical inventory in English, no lexical filter prohibits it, so if place assimilation is to apply at all it will generate at least [m]. But applications yielding [ɲ], [ɱ], [n̪] are optional and absent from slow, careful speech (cf. well-formed *i[nf]allible, e[nf]old*) as are, in certain cases, those yielding [ŋ] (see Borowsky 1989 for the details). Much of this picture is also exactly as predicted by Structure Preservation, based on the blocking effect of the filters in (6) on lexical rule applications. The LP claim is that these sounds could arise only post-lexically, where place assimilation is optional and rate-dependent.

The evidence for Structure Preservation highlights a fundamental drawback of SPE's views on phonemic alphabets: if constraints characterizing possible phonemes hold exclusively of UR's, then why should rules be blocked from generating, *in derived representations*, sounds absent from the underlying set? Concretely: what is the connection between the absence from English URs of /ɱ/, /n̪/, /ɲ/ – or more neutrally put, their predictable distribution – and the fact that word-internal place assimilation avoids creating these sounds? The same type of question arises in relation to vowel harmony (Kiparsky 1985), metaphony (Calabrese 1995), lenition (Everett 2003), epenthesis (Steriade 1995), and consonant mutation (Lieber 1984) to name only a few processes. There is substantial evidence that alphabet constraints like (6) *can* restrict derived representations. Sometimes they block rules from applying: the constraint against *[ɨ] blocks Finnish harmony from spreading [+back] onto /i/ (Kiparsky 1985). Sometimes they trigger repair processes: the prohibition against high lax vowels in Salentino is expressed when metaphony raises an underlying /ɛ/ not to *[ɪ], as expected, but to [iɛ] (Calabrese 1995). (By contrast /e/ is allowed to raise to [i], without diphthongization to [ie], because tense high vowels are permitted.) It is then inaccurate to say that constraints on alphabets – like (6), or Finnish *ɨ and Salentino *ɪ – only apply to define the underlying inventory:

some of these constraints are *persistent* (Myers 1991) and prohibit the same feature combinations throughout much or all of the derivation.

Structure Preservation also raises a question for LP: how early in the derivation does allophonic differentiation take place? Recall that the assumption of lexical minimality, which LP shares, causes the underlying phoneme inventory to be stripped of contextual variants and segments to be lexically represented as underspecified for predictable features. LP enforces both of these policies through the use of lexical constraints and Structure Preservation, an idea whose benefits were seen above. The problem arises when distinct allophones are generated by processes that have lexical characteristics. The diagnostic tests of lexical status have undergone constant revision but interaction with cyclic morphology has always been on this list (cf. most recently Itô and Mester 2003). It is just such an interaction that we observe in the processes generating nasalized allophones in (7) and (8):

(7) *Sundanese (Cohn 1989, after Robins 1957)*

(a) Underlying	(b) After nasal harmony	(c) Infixed, surface	
/miasih/	mĩãsih	m-ãr-ĩãsih	'love-pl.'

(8) *Madurese (Stevens 1968)*

(a) Underlying	(b) After nasal harmony	(c) Reduplicated, surface	
/nejat/	nɛ̃jãt	jãt-nɛ̃jãt	'intentions'

The nasalized vowels of these languages arise through nasal harmony, which spreads nasality from nasals onto contiguous strings of vowels (and glides, in Madurese). Since nasalized vowels are contextually predictable, lexical minimality excludes them from the lexicon. A lexical filter like *[+nasal, +continuant] used for this purpose will prohibit the underlying contrast between, say /a/ and /ã/. In turn, Structure Preservation will prevent sounds prohibited by this filter from arising in the lexical component. But this is problematic, because the phonology of words created through infixation (in (7)) and reduplication (in (8)) must be computed based on the forms that have undergone nasal harmony (cf. (7b), (8b)). This *cyclic* interaction between morphology and phonology is viewed as diagnosing lexical processes. It follows then that the nasalized allophones are derived by lexical processes.

There are other instances of allophonic processes whose outputs are cyclically transmitted to derived words (Borowsky 1993, Benua 1997, Steriade 2000) and this entire body of evidence suggests the need to revise aspects of LP, such as the idea of a boundary separating lexical from postlexical phonology. However, Structure Preservation cannot be abandoned altogether, because an aspect of it is needed to explain the effect of filters like (6) on derived structures.

7.4.3 Contrast and allophony in OT

The most radical modification of the idea of lexical filters is Optimality Theory's (Prince and Smolensky 2004) move to take Structure Preservation and stand it on its head. While LP views filters like (6) as constraining URs and then rule applications, up to an ill-defined derivational juncture, OT proposes that the function of filters is to directly constrain surface representations, with only an indirect effect on URs. The surface orientation of filters immediately explains why place assimilation has difficulty generating surface [ɱ], [ɲ], [n̪] in English words and why harmony and metaphony can't generate Finnish [i] or Salentino [ɪ]: it is filters on surface structure that prohibit these sounds. The grammar's job is not to first define possible well-formed inputs and then proceed to deform them through rules: it is to characterize the class of well-formed outputs, regardless of their underlying source.

Here are the main lines of an OT analysis for each of the processes reviewed thus far. First, the discussion of nasalized vowel allophones (as in (7), (8)) illustrates the fact that certain sound qualities must be allowed to occur but must not contrast: for instance Sundanese [ã] is permitted, but not in contexts where [a] is. The surface occurrence of [ã] is made possible by ranking *[+nasal, +continuant] below the constraint that triggers nasal harmony: *[+nasal][−cons, −nasal]. The lack of contrast between [ã] and [a] is due, in part, to the effect of *[+nasal, +continuant]: this penalizes all nasal vowels, wherever they occur and whatever their source. (9) illustrates both aspects of the analysis:

(9) *Sundanese: markedness constraints result in neutralization*

/ana/	*[+nasal][−cons, −nasal]	*[+nasal, +continuant]
(a) ana	*!	
☞ (b) anã		*
(c) ãnã		**!

The basic ranking is justified by the comparison of (9a) and (9b). Candidate (9c) illustrates in part how the allophonic distribution of [a] and [ã] is analyzed: for each [ã] that does not follow a nasal segment – e.g. the initial [ã] in (9c) – a better candidate exists that replaces this [ã] by oral [a]. The modified candidate – (9b) – better satisfies *[+nasal, +continuant] without violating *[+nasal][−cons, −nasal]. The idea then is to make it impossible for [ã] to surface *anywhere except where mandated by the higher ranked phonotactic*. To repeat: higher ranked *[+nasal][−cons, −nasal] ensures that no [a] surfaces after nasals and the lower ranked *[+nasal, +continuant] ensures that no [ã] surfaces anywhere else. Under these conditions, contrast between [a] and [ã] is impossible. The opposite ranking yields a different kind of complementary distribution between [a] and [ã], namely systems that exclude nasalized vowels, like English.

The next aspect of the analysis of allophony in OT concerns the role of faithfulness. Phonotactics alone cannot describe the difference between

French, where nasal and oral vowels contrast (Cohn 1989), or Acehnese, where they contrast under stress (Durie 1985), vs. English or Sundanese, where their distribution is predictable. For French and Acehnese, the nasal vowels surface because they are protected by faithfulness to inputs like /ãʲ/. The relevant constraint is *Ident [±nasal] in V's Input-to-Output* (IO): pairs of UR-SR vowels standing in correspondence have identical values for nasality (McCarthy & Prince 1995a). Ident [±nasal] in V's (IO) must be ranked above *[+nasal, +continuant] to allow the French nasal vowels to surface (10):

(10) *French: Preservation of underlying nasality results in a surface contrast*

/ãʲ/ 'year'	Ident[±nasal] in V's (IO)	*[+nasal, +continuant]
☞ (a) ã		*
(b) a	*!	

For Sundanese or English, Ident [±nasal] in V's (IO) is *inactive*: it is outranked by *[+nasal, +continuant] and therefore any underlying nasal vowel will either be oralized, to satisfy *[+nasal, +continuant], or will accidentally preserve its nasality, when required by a phonotactic constraint (like the one triggering nasal harmony), rather than because of faithfulness to the underlying form.

A factorial typology of allophony involving vowel nasality is given in (11).

(11) *Factorial typology of nasal vowel allophony*

(a) Nasal and oral vowels contrast in all environments
IDENT[±nas] in V's (IO) » *[+nas, +cont], *[+nas][–cont, –nas]

(b) Nasal vowels neutralize to oral in all environments
*[+nas, +cont] » IDENT[±nas] in V's (IO), *[+nas][–cont, –nas]

(c) Vowels must be nasal after nasal consonants; elsewhere they neutralize to oral
*[+nas][–cont, –nas] » *[+nas, +cont] » IDENT[±nas] in V's (IO)

(d) Vowels must be nasal after nasal consonants; elsewhere they contrast
*[+nas][–cont, –nas] » IDENT[±nasal] in V's (IO) » [nas, cont]

(e) Some systems predicted to be impossible:
(i) Oral vowels neutralize to nasal vowels in all environments
(ii) Vowels contrast for nasality after a nasal consonant, but neutralize elsewhere

This result can be generalized in interesting ways. Kirchner (1997) shows that the contrastive status of any feature F – that is, F's ability to differentiate lexical items in surface forms – is determined in an OT grammar by the position of the Ident F IO constraint in the constraint hierarchy (see also Itô, Mester and Padgett 1995). Inactive Ident F IO yields a strictly allophonic distribution for the feature. A grammar in which Ident F IO outranks some conflicting phonotactic constraints, but not all, describes the case in which F is contrastive for some segments – or some segments in some contexts – but not across the board.

As Kirchner (1997) and Goldsmith (1995b:9–13) note, contrastiveness never functions as an on/off switch in grammars (cf. also Kager 2003). Rather, "phonological systems exert varying amounts of force on the specification of the feature F" (Goldsmith 1995b:12) resulting in a cline from the standard contrastive status, to intermediate states of "modest asymmetry, not yet integrated semi-contrasts, just barely contrastive [features]," all the way to standard allophony and complementary distribution. The variable position of faithfulness constraints like IDENT F relative to conflicting phonotactics is well suited to formalize Goldsmith's gradual cline from contrast to allophony.

The distribution of Acehnese nasality (Durie 1985) is a particularly good example of this cline. Here, nasal vowels contrast under stress with oral vowels: ['bɛ̃h] 'a calf's cry', [ca'hɛ̃ʔt] 'sever with a knife attached to the end of a pole', ['pĩʔp] 'to suck' (Durie 1985:15–27). After nasal segments, stressed and stressless vowels and glides are nasalized by a rule comparable to that of Sundanese: [mã'ɲɛ̃t] 'corpse'; [mã'w̃ʌ̃] 'rose'; [paŋli'mã] 'army leader'; [p-ɯn-ã'ɟoh] 'food delicacies', cf. [pa'ɟoh] 'to eat'. Stressless vowels can be nasal only through nasal harmony and are predictably oral elsewhere: no forms like *[pã'ɟoh] occur. The analysis of this system involves a *positional* faithfulness constraint (Casali 1996, Beckman 1998): the constraint requires identity for the value of the feature [±nasal] between pairs of correspondent vowels, in which the surface vowel is stressed. This is abbreviated as IDENT [±nasal] in 'V (IO). (12) illustrates how stressed vowels preserve their nasality.

(12) *Acehnese: Positional faithfulness allows contrast in stressed syllables*

/cahɛ̃ʔt/	IDENT[±nasal] in 'V (IO)	*[+nasal, +continuant]
(a) ca'hɛʔt	*!	
☞ (b) ca'hɛ̃ʔt		*

IDENT [±nasal] in 'V (IO) must be outranked by *[+nasal][−cons, −nasal] to allow stressed vowels to undergo nasal harmony. To demonstrate this, we use [mãɲɛ̃t] 'corpse', whose original form (in a borrowing from Arabic) was [ma'jit]. (13) models one step in the mapping of [ma'jit] to [mã'j̃ɛ̃t], from which present-day [mãɲɛ̃t] must have resulted.

(13)

/majit/	*[+nasal][−cons, −nasal]	IDENT[±nasal]/in 'V (IO)
(a) mã'j̃ɛt	*!	
☞ (b) mã'j̃ɛ̃t		*

The interest of Acehnese is that its vocalic nasality is contrastive, but not unrestrictedly contrastive: it does not contrast outside of the stressed syllable. Even under stress, a vowel is not protected by its contrastive orality from undergoing harmony. An analytical system (such as those reviewed in Steriade 1995; cf. also Dresher, Piggott and Rice 1994) in which contrastive status for F results in full specification for both values of F and where full specification blocks rules like harmony will have some difficulty with this case. Its OT analysis is simple:

(14)

*[+nasal][–cons, –nasal] » IDENT[±nasal] in 'V(IO) » *[+nasal, +continuant] » IDENT [±nasal]

The further ascent of IDENT [±nasal] in 'V (IO) above *[+nasal][–cons, –nasal] describes Guaraní (Kiparsky 1985 and references there), where stressed vowels are both distinctively nasal *and* protected from undergoing nasal harmony. This entire range of attested options seems compatible only with the constraint system whose factorial typology was shown in (11).

Section 7.2 noted that segmental features are not the only contrastive properties in a phonological system: relations like precedence and temporal overlap, and in addition non-segmental features like tone, and global properties like stress or relative prominence, signaled by a variety of phonetic means, all represent potential sources of lexical distinctions. The difference between contrastive and allophonic status for all these properties can be formalized in the same terms as above, given the necessary faithfulness and conflicting phonotactic constraints.

The notion of *derived contrast* (Harris 1990) is also definable in terms of phonotactics-faithfulness rankings. A phonological property +P that is predictable by reference to both morphosyntactic and phonological information, may appear to contrast with –P, if the contribution of the morphosyntax is ignored. Thus Malayalam [a-n̪a], with predictably dental stem-initial [n̪], may appear to contrast – if we overlook the silent stem boundary – with [ana], with stem medial alveolar [n]. Similarly, the nasal [ĩ] of the infixed Sundanese form [m-ăr-ĩãsih] 'love-pl' ((7) above) contrasts with the oral [i] of monomorphemic [mãrios]. In a rule-based phonology, derived contrasts are created by letting allophonic rules apply cyclically: a later cycle inherits the allophones generated by the immediately preceding one. So the infixation cycle /m-ar-ĩãsih/ inherits the effects of nasal harmony from the previous cycle /mĩãsih/. As seen above, this move is problematic in LP if cyclicity is restricted to the lexical component and if allophony is excluded from it. The OT approach to cyclic effects is described in McCarthy (Ch.5) and makes use of *output-to-output (OO) faithfulness* constraints, as against the *IO faithfulness* constraints, whose interactions with phonotactics define non-derived contrast. What is strictly relevant to derived contrasts is that in a system where IDENT F IO is inactive, IDENT F OO may be active,

by outranking a critical phonotactic constraint. Forms like Sundanese [m-ãr-ĩãsih] are generated by letting IDENT [±nasal] OO » *[+nas, +cont]; in the same system, the ranking *[+nas, +cont] » IDENT [±nas] IO describes the predictable status of vocalic nasality stem internally (Benua 1997). There is no contradiction here: the two types of faithfulness constraints are distinct, because they relate distinct pairs of representations, and thus can occupy different places in the constraint hierarchy.

7.4.4 Richness of the Base and Lexicon Optimization

Two distinct ideas underlie the notion of Richness of the Base, summarized by Prince and Smolensky (2004:191) as "for the purposes of deducing the possible outputs of a grammar, [. . .] all inputs are possible." One of these is the distinction between the Lexicon *sensu stricto*, containing the actual entries a subject happens to know, vs. the full set of potential lexical entries. The example in (13) – showing how Arabic [majit] surfaces as [mājẽt] in Acehnese - illustrates this distinction. An L2 word like [majit] is necessarily absent from the L1 lexicon: but the grammar must still characterize its realization as a well-formed L1 word, if this form is borrowed. If the surface L1 pattern displays a certain regularity – e.g. "every contiguous string of vowels and glides following a nasal is nasalized; and no stressless nasalized vowels and glides occur elsewhere" – the right grammar will guarantee this output pattern, no matter what the input is. This idea is not specific to OT: any generative grammar is responsible for mapping to surface not only entries in the Lexicon in the narrow sense, but those from the unrestricted list of potential inputs.

The second component of Richness of the Base, as currently understood, is the strictly parallelist idea of a one-step mapping from any potential lexical entry, *sensu lato*, to the surface form. This hypothesis rejects conditions holding specifically of lexical entries, because they amount to a two-step filtering of potential inputs: one step eliminates impossible UR's, while the subsequent step, the derivation proper, maps the residue to well-formed SR's. The theory of Stratal OT (Kiparsky 2006) and variants of it (Itô and Mester 2003) which distinguish lexical from postlexical constraint hierarchies have roughly this multi-step property, dictated by empirical considerations, such as the analysis of opacity (see McCarthy 5.4). In principle, then, any form that is an input to the grammar may be underspecified for some or all features, or might contain all manner of redundant phonological information, or a mix of redundant and underspecified material. To require either systematic underspecification of features in lexical entries, or systematic full specification amounts to a condition on inputs; such requirements are rejected by the second component of Richness of the Base on the parallelist grounds outlined above and, one may add, because the necessity for any such conditions on lexical entries is yet to be proven.

In practice, however, the related hypothesis of Lexicon Optimization (Prince and Smolensky 2004:ch.9) does have an effect on how certain inputs are lexically represented. The idea is that non-alternating phonological properties – say the aspiration of initial [k^h] in [k^hæt] – are always present in the lexical entry, for the following reason: the right grammar will guarantee the surface occurrence of [k^h] in [k^hæt] no matter whether [k^h] or [k] is present underlyingly, but the advantage of the identity mapping /k^hæt/ → [k^hæt] over /kæt/ → [k^hæt] is that the former avoids a faithfulness violation (IDENT [±aspiration]). Lexicon Optimization yields (approximately) opposite results on the issue of underlying specification compared to lexical minimality: non-alternating redundant information will always end up lexically listed. Here too, one can think of empirical work that could test this hypothesis, including lines of research of the sort cited in Section 7.4.2.

7.5 Constraints on contrast

Both LP's lexical filters and the surface-oriented filters of OT are constraints on sounds and sound sequences. When interacting with rules (in LP) or faithfulness (in OT) these filters indirectly generate patterns of contrast, allophonic variation and neutralization. The alternative explored by Flemming's (1995, 2002, 2004) Dispersion Theory of Contrast (DTC; see also Padgett 2001, 2003b and references there) is that certain core constraints refer directly to properties of the *relation of contrast*, namely its distinctiveness, rather than to the quality of the sounds standing in contrast. To understand how properties of sounds differ from properties of contrasts, imagine three tonal inventories, each contrasting a relatively lower tone with one relatively higher tone: using Chao's (1930) numbers, the inventories are {2, 4}, {3, 5} and {1, 5}. Each inventory differs from the others in the absolute height of one or both tones, and thus in the properties of the sounds involved; but the first two inventories are equivalent in the relative spacing of the tones ({2, 4} and {3, 5}) and thus in the distinctiveness of the contrast defined. The third inventory ({1, 5}) requires a greater distance between contrasting tonal values: it defines a better separated, more distinctive, tonal contrast.

The following is an example (adapted from Flemming 2004:250ff. and Flemming and Johnson 2004) where reference to contrast rather than sound properties is necessary. To characterize the fact that most varieties of English lack [ɨ], other accounts (e.g. Calabrese 1995) include rules or constraints about the properties of this sound, such as *[+high, +back, −round]. The activity of such a filter is independent of that of filters on [i], [u], [y]: in other words, a standard system decides whether to let [ɨ] in, regardless of what other sounds [ɨ] will coexist with. The DTC differs on this because it predicts the absence of [ɨ] by reference to the distinctiveness of

the contrasts that would exist, if [ɨ] were allowed. Specifically, the DTC singles out the effect [ɨ] would have on decreasing the distance in perceptual space (here F2) between the pairs [ɨ]-[u], [ɨ]-[i]: these involve smaller distances in F2 compared to that between [u] and [i]. So removing [ɨ] is beneficial, not because [ɨ] possesses any inherently bad quality, but because the contrasts it would necessarily enter into would be less distinctive. (At the same time, removing [ɨ] is detrimental, because the system is left with one fewer expressive category. The formal account of dispersion, outlined below, exploits the conflict between expressiveness and dispersion in characterizing the typology.)

The DTC predicts then that the grammatical status of a sound will change depending on the system of contrasts it's embedded in: [ɨ] has detrimental effects on a system containing [i] and [u], but it fares well as the unique vowel. It is from this type of prediction that the DTC draws significant empirical support. The vowel system of American English illustrates this (15): stressed syllables contain [i] and [u], [ɪ] and [ʊ], along with other vowels, but not [ə]; stressless final syllables contain [i], [o] and [ə], again without [ɨ]; whereas stressless non-final syllables contain a single vowel quality and that is [ɨ][4] (Flemming and Johnson 2004).

(15) *American English vowel distribution*

stressed	ɪ ʊ i u	ɑ æ ɔ ɛ ʌ e o
stressless final	i o	ə
stressless non-final	ɨ	

The striking fact in (15) is that in any given context we find either an F2 contrast such as [i]-[o], or, if no such contrast exists, then [ɨ]. That's exactly what the DTC predicts: in the absence of contrast, there's nothing wrong with [ɨ].

Dispersion alone does not predict that [ɨ] is *necessary* in a one-vowel system, only that it's a possible choice there. The factor that specifically favors [ɨ] is articulatory: CɨC sequences, for most choices of Cs, avoid steep articulatory transitions better than other CVCs. This applies to stressless medial syllables, as in (15), because those are typically very short, and steep transitions relate to short durations (Flemming 2004:250ff.).

The phenomena supporting the DTC are the typology of enhancement and neutralization. Both require that the grammatical system evaluate the distinctiveness of contrasts, in addition to articulatory properties of individual sounds. Thus Flemming shows that neutralization is triggered by contrasts that are insufficiently separated (see also Barnes 2002, Bradley 2001, Crosswhite 2001, Padgett 2001, 2002, Steriade 1999b, 2001b), so its proper formalization should involve explicit comparison of candidate inventories based on the distinctiveness of their contrasts. *Enhancement* (cf. Stevens et al. 1986) is the alternate remedy for insufficiently distinctive

contrasts: if x and y contrast on some dimension D_1, but are insufficiently separated on D_1, their contrast can be enhanced by making them differ also on some other dimension D_2. For instance, a voicing contrast (e.g. {t, d}) is frequently enhanced by duration and F0 differences on neighboring vowels (Kingston and Diehl 1994). A significant finding is that *only contrasts are enhanced* (Kingston and Diehl 1994:436ff; Flemming 2004:258ff): Tamil [d], a contextually voiced co-allophone of [t], does not receive the F0 properties that enhance voicing in the contrastive /d/ of English. This observation can be modeled only if the grammatical system tells apart an allophonic voicing difference from a voicing contrast. The formalization of the DTC does exactly that.

The DTC uses two classes of novel constraints: constraints that favor maximizing the number of contrasting categories on specific auditory dimensions (e.g. closure duration; F2; VOT; loudness) and those that favor maximally distinct contrasting categories. Their format is illustrated in (16)–(17), using the example of backness (F2) as a dimension of contrast. Flemming's original statements are reformulated in minor ways.

(16) *Constraints on contrast numbers: MAX-Contrast*
   i. There are at least 2 distinct categories on the F2 dimension.
   ii. There are at least 3 categories on the F2 dimension.

(17) *Constraints on minimal distance between contrasting categories: MINDIST*
   i. MinDist=F2:1 Any two categories on the F2 dimension differ by at least 1 unit.
   ii. MinDist=F2:4 Any two categories on the F2 dimension differ by at least 4 units.

The basic idea of this system is that distance between contrasting categories on a dimension is inversely related to the number of categories defined on it. The MINDIST:F2 constraints penalize less well separated contrasts and thus, indirectly, systems in which more contrasting categories are packed into the space of F2. The MaxContrast:F2 constraints push in the opposite direction. Assuming for this illustration that there are at most 6 potential categories definable on the F2 value of high vowels ({i i̱ i̱ ɨ ɯ u}), an inventory that selects just the F2 extremes {i, u} ensures a distance of 4 units between these categories and thus satisfies both (17.ii) and (17.i). The selection of {i ɨ u} reduces this distance to 1, so this system satisfies only (17.i), but it provides better satisfaction for the MAX-Contrast constraints: both (17.i) and (17.ii) are satisfied by this inventory. Contextual neutralization – e.g. the collapse of a larger vowel inventory into a small one in specific contexts – is formalized through the interaction of these constraints with constraints on articulatory effort. Thus the reduction in medial stressless syllables of the entire vowel inventory to a contextually variable vowel centered on [ɨ] is attributed to the drastic decrease in

duration that accompanies lack of stress, as sketched above. So the feature composition of sounds is an emergent property in the DTC: it emerges from the interplay of dispersion (MinDist), expressiveness (MaxContrast) and avoidance of articulatory effort.

7.6 Interactions between dimensions of contrast

Different dimensions of possible contrast interact in the case of enhancement (as in the example of voicing and vocalic F0 above) or in the related case of *a displaced contrast*, where a contrast on one dimension migrates to a related dimension (e.g. a voicing contrast, possibly enhanced by vocalic F0, becomes just a tonal contrast; Halle and Stevens 1971 Hombert et al. 1979). The notion of displaced contrast has also been used by Łubowicz (2003) to explore certain benefits of opacity (cf. McCarthy 5.4).

A more challenging sort of interaction between contrast dimensions is raised by the phenomenon of feature economy mentioned in (7.3, cf. Clements 2003). It was observed there that feature economy competes with dispersion, but that the competition is limited in certain ways, since it fails to yield certain highly uneconomical systems, such as {p t^h k' β ɾ ɡ w n̪ ŋ}. The effect of feature economy in a grammar is currently unformalized – no constraint enforces it – but it is interesting to observe that feature economy relates to an unexplored property of Flemming's MaxContrast constraints. This is mentioned here in the belief that these issues will eventually receive a unified resolution.

Originally the MaxContrast constraints were formulated as specific to individual dimensions of contrast, as seen in (17). A problem that arises with the original formulation was that the system is not encouraged to "fully cross" its contrasts: so the inventory (18) satisfies MaxContrast:F1=3 (i.e. "have at least three vowel height categories") as well as the less fully crossed (19) and (20) do.

(18) {a e o i u, ã ẽ õ ĩ ũ}
(19) {a e o i u, ã ĩ ũ}
(20) {a e o i u, ɑ̃}

Intuitively, MaxContrast:F1=3 should be satisfied only by (18): in (19) and (20) the height categories in the nasal system have been reduced to two and one, respectively. But if what is required is that just somewhere in the system there be three height degrees, that's equally true of all of (18)–(20). Moreover, the dispersion constraints are better satisfied by (20) as nasal vowels tend to be realized with wider formant bandwidths, so, under this interpretation, (18) and (19) are harmonically bounded by (20). That's incorrect: all three systems are instantiated, and (20), the least economical, in Clements's sense, is also by far the least well attested (Ruhlen 1978).

The problem is solved by the modified version of MAXCONTRAST which appears in Flemming (2004:240), and which is no longer dimension specific: there is now a single MAXCONTRAST constraint that is better satisfied by systems possessing more segments overall. However, it is feasible to evaluate this constraint only if we limit ourselves to a segment inventory. When we step into the larger world of sequential contrasts, accentual contrasts, contrasts in syllable numbers, and so on, it is no longer clear what kinds of additional expressions will provide an equal or better satisfaction of MAXCONTRAST compared to simply adding novel segment types. It was suggested earlier (Sec. 7.2) that the distinction between segmental and non-segmental contrasts is somewhat artificial: for this reason, among others, the real resolution to the problem posed by (18)–(20) does not seem to lie in setting aside the inventory of *segmental* contrasts and evaluating globally its expressiveness. Perhaps a more interesting solution will emerge if a revised formalization of the DTC evaluates numbers of contrastive categories on individual featural dimensions, as the original formalization did, but takes on the problem of incorporating into the grammar the violable requirement of feature economy. Feature economy – not contrast numbers – is probably the factor that allows systems to prefer (18) to (19) and both to (20).

7.7 Conclusion

This chapter has surveyed the transition from the early generative conception of an alphabet of contrasting phonemes, defined on underlying representations, to the Optimality Theoretic idea that phonemic alphabets are the result of the interaction between surface oriented constraints with faithfulness conditions. In the last sections, we have reviewed work demonstrating that neutralization and enhancement are triggered by insufficient distinctiveness, or insufficient separation in perceptual space between contrasting sounds. Grammars that evaluate the degrees of distinctiveness of candidate inventories must perform certain global comparisons – such as that of (18) to (19) to (20). The relation between such evaluations and the more familiar evaluation of mappings from UR to SR in individual utterances remains to be explored. It does appear clear from this review that there is no substitute to recognizing the role of systemic constraints – dispersion and economy – in the organization of contrast systems.

Notes

I would like to thank Paul de Lacy and Ania Łubowicz for comments on the chapter; and Adam Albright and Edward Flemming for enlightening discussion of its contents.

1 These differences on *how* to model the relation between typology and individual competence are minor in comparison with the debate on *whether* typology and grammar stand in any kind of direct relation: cf. Blevins (2004). This topic is more general than that of contrast and will not be further addressed here.
2 The notions of symmetry and pattern congruity discussed in the structuralist literature (Hockett 1955:159) reduce to feature economy.
3 In an AXB classification experiment reported by Harnsberger (1999), Malayalam subjects judged [n̪] to be *more* similar to [n] than American English subjects. This result could be an artefact of the experimental conditions, but it highlights the need for solid evidence on the distinctness judgments serving as the empirical basis of the lexical level.
4 More precisely, stressless non-final vowels are realized in a region in F1-F2 space whose center is [ɨ]. As with other reduced vowels, there is considerable contextual variation here.