# I
# Conflicts in grammars

## 1.I Introduction: goals of linguistic theory

### 1.1.1 *Universality*

The central goal of linguistic theory is to shed light on the core of grammatical principles that is common to all languages. Evidence for the assumption that there should be such a core of principles comes from two domains: language typology and language acquisition. Over the past decades our knowledge of linguistic typology has become more and more detailed, due to extensive fieldwork and fine-grained analysis of data from languages of different families. From this large body of research a broad picture emerges of 'unity in variety': core properties of grammars (with respect to the subsystems of sounds, words, phrases, and meaning) instantiate a set of universal properties. Grammars of individual languages draw their basic options from this limited set, which many researchers identify as Universal Grammar (UG). Each language thus reflects, in a specific way, the structure of 'LANGUAGE'. A second source of evidence for universal grammatical principles comes from the universally recurring patterns of first language acquisition. It is well known that children acquiring their first language proceed in remarkably similar ways, going through developmental stages that are (to a large extent) independent of the language being learnt. By hypothesis, the innateness of UG is what makes grammars so much alike in their basic designs, and what causes the observed developmental similarities.

The approach to universality sketched above implies that linguistic theory should narrow down the class of universally possible grammars by imposing restrictions on the notions of 'possible grammatical process' and 'possible inter-action of processes'. In early Generative Grammar (Chomsky 1965, Chomsky and Halle 1968), processes took the shape of *rewrite rules,* while the major mode of interaction was *linear ordering.* Rewrite rules take **as** their input a linguistic representation, part of which is modified in the output. Rules apply one after another, where one rule's output is the next rule's input. It was soon found that this rule-based theory hardly imposes any limits on the notion of 'possible rule',

nor on the notion of 'possible rule interaction'. In the late 1970s and early 1980s, considerable efforts were put into constraining both rule typology and interactions. The broad idea was to factor out universal properties of rules in the form of conditions.' While rules themselves may differ between languages, they must always respect a fixed set of universal principles. Gradually more and more properties were factored out of rules and attributed to universal conditions on rules and representations. Developments came to their logical conclusion in Principles-and-Parameters Theory (Chomsky 1981b, Hayes 1980), which has as its central claim that grammars of individual languages are built on a central core of fixed universal properties (principles), plus a specification of a limited number of universal binary choices (parameters). Examples of parameters are the side of the 'head' (left or right) in syntactic phrases, or the obligatoriness (yes/no) of an onset in a syllable. At the same time, considerable interest developed in representations, as a way of constraining rule application, mainly with respect to *locality* (examples are trace theory in syntax, and underspecification theory in phonology). Much attention was also devoted to constraining rule interactions, resulting in sophisticated theories of the architecture of UG (the 'T'-model) and its components (e.g. Lexical Phonology, Kiparsky 1982b).

### 1.1.2 *Markedness*

What all these efforts to constrain rules and rule interactions share, either implicitly or explicitly, is the assumption that universal principles can only be universal if they are actually inviolate in every language. This interpretation of 'universality' leads to a sharp increase in the abstractness of both linguistic representations and rule interactions. When some universal principle is violated in the output of the grammar, then the characteristic way of explaining this was to set up an intermediate level of representation at which it is actually satisfied. Each grammatical principle thus holds at a specific level of description, and may be switched off at other levels.

This absolute interpretation of universality is not the only one possible, however. In structuralist linguistics (Hjelmslev 1935, Trubetzkoy 1939, Jakobson 1941; cf. Anderson 1985), but also in Generative Phonology (Chomsky and Halle 1968, Kean 1975, Kiparsky 1985) and Natural Phonology (Stampe 1972, Hooper 1976), a notion of MARKEDNESS plays a key role, which embodies universality in a 'soft' sense. The idea is that all types of linguistic structure have two values, one of which is 'marked', the other 'unmarked'. Unmarked values are cross-linguistically preferred and basic in all grammars, while marked values are cross-linguistically avoided and used by grammars only to create contrast. For example,

---

[1] **For example, SUBJACENCY was proposed as a universal condition on syntactic movement rules and the OBLIGATORY CONTOUR PRINCIPLE as a universal condition on phonological rules.**

all languages have unrounded front vowels such as [i] and [e], but only a subset of languages contrast these vowels with rounded front vowels such as [y] and [ø]. Hence, the unmarked value of the distinctive feature [round] is [−round] in front vowels. At a suprasegmental level, markedness affects prosodic categories. For example, the unmarked value for syllable closure is 'open' since all languages have open syllables (CV, V), while only a subset of languages allow closed syllables (CVC, VC).[2] The notion of markedness is not only relevant to sound systems. Markedness principles have been proposed for morphological and syntactic systems as well (Chomsky 1981a).

The markedness approach of linguistic universality is built on two assumptions. First, markedness is inherently a relative concept: that is, a marked linguistic element is not ill-formed *per se,* but only in comparison to other linguistic elements. Second, what is 'marked' and 'unmarked' for some structural distinction is not an arbitrary formal choice, but rooted in the articulatory and perceptual systems. By this combination of two factors, markedness allows an interpretation of universality that is fundamentally different from Principles-and-Parameters Theory, in which markedness has no substantive status in the grammar, but functions as an external system of annotations on parameter values, evaluating a grammar's 'complexity'.[3]

## 1.2   Basic concepts **of** OT

Optimality THEORY (Prince and Smolensky 1993, McCarthy and Prince 1993a,b) turns markedness statements into the actual substance of grammars. Markedness is built into grammars in the form of universal OUTPUT CONSTRAINTS which *directly* state marked or unmarked patterns, for example: 'front vowels are unrounded' or 'syllables are open'. The universal interpretation of markedness constraints is reconciled with the observation that languages, to a certain extent at least, tolerate marked types of structures. Universal markedness constraints can be literally *untrue* for a grammar's output, or to phrase it in optimality-theoretic terms: constraints are VIOLABLE. Violation of a constraint is not a direct cause of ungrammaticality, nor is absolute satisfaction of all constraints essential to the grammar's outputs. Instead what determines the best output of a grammar is the least costly violation of the constraints. constraints are intrinsically in CONFLICT, hence every logically possible output of any grammar will necessarily violate at least some constraint. Grammars must be able to regulate conflicts between universal constraints, in order to select the 'most harmonic' or 'optimal' output form.

---

[2] Markedness may also involve scales. For example, the higher a consonant's sonority value, the more likely its occurrence in the syllable coda.

[3] For the view of markedness as a criterion external to the grammar, evaluating its complexity, see Chomsky and Halle (1968) and Kean (1975, 1981).

This conflict-regulating mechanism consists of a RANKING of universal constraints. Languages basically differ in their ranking of constraints. Each violation of a constraint is avoided; yet the violation of higher-ranked constraints is avoided 'more forcefully' than the violation of lower-ranked constraints. Accordingly, the notion of 'grammatical well-formedness' becomes a relative one, which is equivalent to the degree of satisfaction of the constraint hierarchy, or HARMONY.

OT's viewpoint of UG is fundamentally different from that of classical rule-based generative theory, where UG is defined as a set of inviolate principles and rule schemata (or 'parameters'). OT defines UG as a set of universal constraints (markedness relations and other types of constraints, as we will see below), and a basic alphabet of linguistic representational categories. In its interactions, it is limited to a single device: constraint ranking. OT still shares with its rule-based generative ancestors the central position taken by UG, as described above. OT *is* a theory of the human language capacity.

The remainder of this chapter is organized as follows. Section 1.2 will introduce basic notions of OT: conflict, constraints, and domination, which will be exemplified in section 1.3. In section 1.4, we will discuss the architecture of an OT grammar. Section 1.5 will deal with interactions of markedness and faithfulness, relating these to the lexicon in section 1.6. A factorial typology of constraint interactions will be developed in section 1.7 and applied to segment inventories in section 1.8. Finally, section 1.9 presents conclusions.

### 1.2.1   *Language as a system of conflicting universal forces*

At the heart of Optimality Theory lies the idea that language, and in fact every grammar, is a system of conflicting forces. These 'forces' are embodied by CONSTRAINTS, each of which makes a requirement about some aspect of grammatical output forms. Constraints are typically conflicting, in the sense that to satisfy one constraint implies the violation of another. Given the fact that no form can satisfy all constraints simultaneously, there must be some mechanism selecting forms that incur 'lesser' constraint violations from others that incur 'more serious' ones. This selectional mechanism involves hierarchical RANKING of constraints, such that higher-ranked constraints have priority over lower-ranked ones. While constraints are universal, the rankings are not: differences in ranking are the source of cross-linguistic variation.

But before discussing actual constraints and their rankings, let us first find out in a general way about the two major forces embodied by constraints. Two forces are engaged in a fundamental conflict in every grammar. The first is MARKEDNESS, which we use here as a general denominator for the grammatical factors that exert pressure toward *unmarked types of structure*. This force is counterbalanced by

FAITHFULNESS, understood here as the combined grammatical factors *preserving lexical contrasts*. Let us focus on both general forces to find out why they are inherently conflicting.

In sound systems, certain types of structure – segments, segment combinations, or prosodic structures – are universally favoured over others. For example, front unrounded vowels are unmarked as compared to front rounded vowels, open syllables as compared to closed syllables, short vowels as compared to long vowels, and voiceless obstruents compared to voiced obstruents. As was observed above, marked structures are avoided by all languages, while they are completely banned by some languages. Therefore the notion of markedness is inherently *asymmetrical.*

Most phonologists agree that phonological markedness is ultimately GROUNDED in factors outside of the grammatical system proper. In particular, the systems of articulation and perception naturally impose limitations on which sounds (or sound sequences) should be favoured. Yet explaining markedness relations by phonetic factors does not amount to denying the basis of phonology as a grammatical system, for two reasons. The first reason is that phonetic factors are gradient, and add up to numerical patterns, while phonological factors are categorical, producing patterns whose boundaries are clearly cut by categorical distinctions. The symmetry of phonological systems cannot be captured by the interaction of 'raw' phonetic factors. The second reason is that the relative strength of the individual markedness factors varies from language to language, which entails that there must be a language-specific system defining the balance of factors. This is the grammar, a system of ranked constraints, of which phonology is an integral part.

The major force counterbalancing markedness is *faithfulness* to lexical contrasts. A grammar that is maximally 'faithful' to a lexical contrast is one in which output forms are completely congruent with their lexical inputs with respect to some featural opposition. Or to put it differently, the total amount of lexically contrastive variation of some feature is realized in all of the grammar's output forms. For example, a lexical contrast of voicing in obstruents is preserved in output forms regardless of their phonological context (at the end of a word, between vowels, etc.). Thus one may think of faithfulness as the general requirement for linguistic forms to be realized as close as possible to their lexical 'basic forms'. From a functional angle, the importance of faithfulness is clear: to express contrasts of *meaning,* any language needs a minimal amount of formal *contrast.* Formal contrasts should be preserved in realizations of lexical items, and not be 'eroded' (or at least, not too much) by factors reducing markedness. In the realm of sound

systems (or 'phonologies'), lexical contrasts are carried by oppositions between sounds, as well as by their combinations. Phonological elements are not the only carriers of lexical contrast. (Although phonology is what we will focus on in this book.) Lexical contrasts are also expressible by word structure (morphology) or phrase structure (syntax).

Closely related to faithfulness (or preservation of lexical contrasts) is the pressure towards the shape invariability of lexically related items in various grammatical contexts. This was known in pre-generative linguistics as 'paradigm uniformity'. Shape invariance of lexical items is understandable as another priority of linguistic communication: there should be a one-to-one relation between lexical items, the 'atoms' of meaning, and the shapes which encode them.

### 1.2.2 *Conflicts between markedness and faithfulness*

Markedness and faithfulness are inherently conflicting. Whenever some lexical contrast is being preserved, there will be some cost associated in terms of markedness since in every opposition one member is marked. For example, consider the fact that English limits the possible contrasts in its vowels with respect to the dimensions of backness and rounding: no rounded front vowels stand in contrast to unrounded front vowels. This correlation of rounding and backness in vowels is not idiosyncratic to English, but it reoccurs in a great majority of the world's languages. In fact it is grounded in properties of the articulatory and perceptual systems. Yet this restriction is certainly not 'universal' in the sense that all of the world's languages respect it. Many languages do allow a contrast of rounding in front vowels, thus increasing the potential amount of lexical contrast at the expense of an increase in markedness.

Generally we find that the larger the array of means of encoding lexical contrasts, the larger the complexity of the sound system, either in terms of segmental complexity, or in terms of the combinatory possibilities between segments ('phonotactics'). A language can be maximally faithful to meaningful sound contrasts only at the expense of an enormous increase in phonological markedness. Conversely, a language can decrease phonological markedness only at the expense of giving up valuable means to express lexical contrast.

First consider what a hypothetical language would look like at one extreme of the spectrum: a language giving maximal priority to the expression of lexical contrasts, while imposing no markedness restrictions. We endow this language with the combined segment inventories of the world's languages, roughly 50 consonants and **30** vowels (Ladefoged and Maddieson 1996). We drop combinatory markedness restrictions, allowing all logically possible segment combinations to form a lexical item. Permutation of these 80 segments into lexical items of two

segments already produces some 6,400 items, including [pʰy], [ŋʌx], and [Ɵdʲ], all highly marked. But why stop at two segments per item? By sheer lack of phonotactic limitations, nothing rules out lexical items of 37 or 4,657 segments, or even longer. Now consider the fact that the number of possible lexical items increases exponentially with the number of segments (80") so that at segmental length 6 we already approximate an awesome 300 billion potential lexical items. Clearly no human language requires this number of lexical contrasts, hence there is room to impose markedness restrictions on segments and their combinations in lexical items. Since such restrictions make sense from an articulatory and perceptual point of view, we expect to find them.

Let us now turn the tables to find out what a language at the other extreme would look like, a language giving maximal priority to markedness, and minimal priority to the expression of lexical contrasts. Let us assume that this language limits its lexical items to the general shape of CV* (sequences of consonant–vowel), with C ∈ {p,t,k} and V ∈ {i,a}.[4] The complete set of potential monosyllables contains 6 items {pi, pa; ti, ta; ki, ka}, the set of disyllables contains 36 (or $6^2$) items ({pipi, papi, kipi...}), trisyllables 216 (or $6^3$), etc. But stop! We are overlooking the fact that the unmarked length of lexical item is two syllables (this is the minimum size in many languages and by far the most frequent size in most languages). Since we are assuming that this language is maximally concerned about markedness, we should limit word size to two syllables. The bitter consequence is a mini-lexicon containing at most 36 items. Now consider the fact that the lexicon of an average natural language contains some 100,000 items.[5] It is clear that giving maximal priority to markedness implies an acute shortage of lexical contrasts, which no language can afford.

This comparison of two extremes shows that languages may, in principle at least, go astray in either of two ways: by giving blind priority to expression of lexical contrast, resulting in massive costs in terms of markedness or, at the other end of the spectrum, by giving unlimited priority to markedness reduction, resulting in a fatal lack of contrast.

[4] These limitations are actually **grounded** in speech production and perception: every consonant is maximally different from a vowel (hence, all consonants are voiceless stops). Every vowel is maximally different from other vowels (a 2-vowel set, i–a). Every consonant is maximally different from other consonants @lace of articulation restricted to labial, alveolar, and velar). Every vowel is preceded by a consonant (no word-initial vowels, no hiatus). Every consonant precedes a vowel for optimal release (hence no consonant clusters nor word-final Cs).

[5] Suppose that our hypothetical language would not respect word size restrictions, having at its disposition all possible CV*-shaped items. Here, with a maximal density of lexical contrast, all potential items up to seven syllables long would not suffice to build the required size of lexicon. This would only reach to a moderate total of (46,656 + 7,776 + 1296 + 216 + 36 + 6) = 55,986 lexical items. The average item in this language would be over six syllables long. Without doubt, speaking would become a rather time-consuming activity.
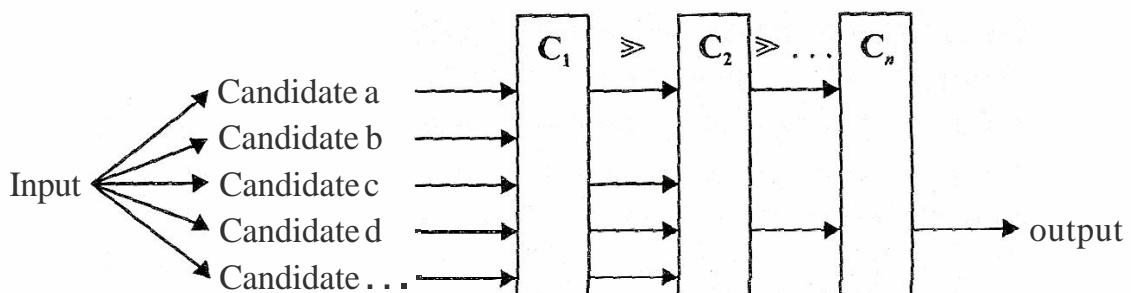
In sum, we have seen that every grammar must reconcile the inherently compet-ing forces of faithfulness to lexical contrasts (the inertness which draws output forms back to their basic lexical shapes) and markedness (minimization of marked forms). However, as we are about to find out, Optimality Theory recog-nizes no unitary or monolithic forces of faithfulness or markedness: the picture is more fragmented. In the grammars of individual languages, the overall conflict between both 'forces' assumes the form of finer-grained interactions of individual *constraints.* At this level, where individual constraints compete, languages are quite diverse in their resolutions of conflicts between 'markedness' and 'faithful-ness'. A language may give priority to faithfulness over markedness with respect to some opposition, but reverse its priorities for another opposition.

Let us now turn to the implementation of these basic ideas in Optimality Theory.

### 1.2.3  *The OT grammar as an input–output device*

The basic assumption of OT is that each linguistic output form is *optimal,* in the sense that it incurs the least serious violations of a set of conflicting constraints. For a given input, the grammar generates and then evaluates an infinite set of output candidates, from which it selects the optimal candidate, which is the actual output. Evaluation takes place by a set of hierarchically ranked constraints (C, % $C_2 \gg \ldots C_n$), each of which may eliminate some candidate outputs, until a point is reached at which only one output candidate survives. This elimination process is represented schematically:[6]

(1)        Mapping of input to output in OT grammar



The optimal output candidate is the one that is 'most harmonic' with respect to the set of ranked constraints. 'Harmony' is a kind of relative well-formedness, taking into account the severity of the violations of individual constraints, as determined by their hierarchical ranking. That is, violation of a higher-ranked

---

[6] Elimination of less-harmonic candidates is portrayed in (1) as a serial filtering process, but we will learn to view it as a parallel process, with higher-ranked constraints taking priority over lower-ranked constraints.

constraint incurs a greater cost to harmony than violation of a lower-ranked constraint. Some violations must occur in every output candidate, as constraints impose conflicting requirements. Accordingly, a lower-ranked constraint can be violated to avoid the violation of a higher-ranked one, but violation is always kept to a minimum, given the requirement of maximal harmony.

With the basic assumptions of OT in our minds, let us now turn to a finer-grained discussion of the core notions 'constraints', 'conflict', 'domination', and 'optimality'.

### 1.2.4 *Constraints: universality and violability*

Our preliminary definition of CONSTRAINT is: a *structural requirement that may be either satisfied or violated by an output form*. A form SATISFIES a constraint if it fully meets the structural requirement, while any form not meeting this requirement is said to VIOLATE it. For the moment we will assume no degrees of violation, so that output forms are simply categorized by a crude binary criterion as either satisfying or violating a constraint. Forms may satisfy constraints *vacuously,* which is the case if a constraint makes a requirement about some structural element that is not present in a particular candidate.

OT recognizes two types of constraints, *faithfulness* constraints and *markedness* constraints. Each individual constraint evaluates one specific aspect of output markedness or faithfulness. Let us now look into the general properties of both types of constraints, and into their functions in the grammar.

*Markedness* constraints require that output forms meet some criterion of structural well-formedness. As the examples below illustrate, such requirements may take the form of prohibitions of marked phonological structures, including segment types (2a), prosodic structures (2b), or occurrnces of segment types in specific positions (2c).

(2)  Examples of markedness constraints
   a. Vowels must not be nasal
   b. Syllables must not have codas
   c. Obstruents must not be voiced in coda position
   d. Sonorants must be voiced
   e. Syllables must have onsets
   f. Obstruents must be voiced after nasals

However, markedness constraints may just as well be stated positively, as in (2d–f). Note that markedness constraints refer to output forms only and are blind to the (lexical) input.

As we have seen in section 1.1, markedness is an inherently asymmetrical notion. Hence, the universal constraint inventory lacks the *antagonist* constraints of (1a-e), which make opposite requirements 'syllables must have codas', 'sonorants must be voiceless', etc.[7]

*Faithfulness* constraints require that outputs preserve the properties of their basic (lexical) forms, requiring some kind of similarity between the output and its input.

(3)        Examples of faithfulness constraints
           a. The output must preserve all segments present in the input
           b. The output must preserve the linear order of segments in the input
           c. Output segments must have counterparts in the input
           d. Output segments and input segments must share values for [voice]

Faithfulness constraints are, strictly speaking, not pure output constraints, since they take into account elements at two levels: input and output. In contrast, markedness constraints never take into account elements in the input.[8] The important thing is, however, that both kinds of constraints refer to the *output* (exclusively so in markedness, and in relation to the input in faithfulness). OT has no constraints that exclusively refer to the input. (This is a crucial difference from classical generative phonology, as we will see in chapter 2.)

From a functional viewpoint, faithfulness constraints protect the lexical items of a language against the 'eroding' powers of markedness constraints, and thereby serve two major communicative functions. First, they preserve *lexical contrasts*, making it possible for languages to have sets of formally distinct lexical items to express different meanings. Phrasing it slightly differently, with an emphasis on contrast, we may say that faithfulness is what keeps the shapes of different lexical items apart. Second, by limiting the distance between input and output, faithfulness constraints restrict the *shape variability* of lexical items. Faithfulness thus keeps the contextual realizations of a single morpheme (called its *alternants)* from drifting too far apart. This enhances the one-to-one relations of meaning and form. In sum, the overall function of faithfulness is to enforce the phonological shape of lexical forms in the output, as a sort of inertness limiting the distance between outputs and their basic shapes.

Two more assumptions are to be made about constraints in OT: they are *universal* and *violable* requirements on some aspect of linguistic output forms. Let us now focus on each of these properties of constraints. The first property is

---

[7] We will see later that some markedness constraints do have antagonists.
[8] See chapter 9 for OT models which weaken this assumption.

– Universality: constraints are universal.

In its strongest interpretation, by which all constraints are part of UG, this implies that all constraints are part of the grammars of all natural languages. This is not to say that every constraint will be equally active in all languages. Due to the language-specific ranking of constraints, a constraint that is never violated in one language may be violated but still be active in a second language, and be totally inactive in yet a third language. This strong interpretation, which leaves no room for language-specific constraints, nor for constraint variability, will be slightly relativized below.

For phonological markedness constraints, 'universality' may be established by a variety of factors, ideally in combination. The first sense of universality is typological: a constraint states a preference for certain structures over other types of structures, which reoccurs in a range of unrelated languages. Segmental markedness constraints, for example, may be validated by inspecting the relative markedness of segments in inventories on a cross-linguistic basis. (Such an overview is presented in Maddieson 1984.) However, any exclusively typology-based definition of universality runs the risk of circularity: certain properties are posited as 'unmarked' simply because they occur in sound systems with greater frequency than other 'marked' properties.

Hence, a second (non-circular) criterion of universality should ideally accompany typological criteria: phonological markedness constraints should be *phonetically* grounded in some property of articulation or perception. That is, phonetic evidence from production or perception should support a cross-linguistic preference for a segment (or feature value) to others in certain contexts. For example, there is articulatory evidence (to be reviewed in chapter 2) that voiced obstruents are preferred to voiceless obstments in a position immediately following a nasal. Indeed many languages avoid or disallow voiceless post-nasal obstruents, neutralizing voicing contrasts in this position.[9] Even though a growing number of constraints has been phonetically grounded (see the suggested readings at the end of this chapter), such grounding is still lacking for others.

It should be clear from this discussion that we should be very careful about positing any constraint lacking both typological motivation and phonetic grounding, even if there is compelling motivation for it from the language data under analysis. Nevertheless, not all constraints that have been proposed in the OT literature satisfy both criteria, indicating that the major issue of universality of constraints has not yet been resolved, since analysts do not share the same criteria. In this book, whenever we employ a constraint that strikes us as 'parochial' or

---

[9] Post-nasal voicing and its typological consequences will be discussed in detail in chapter 2.

language-specific (since it lacks both phonetic grounding and cross-linguistic motivation), this will be indicated.

However, the universality of constraints should directly be relativized somewhat. We will find that in special cases, language-specific elements may occur in constraints of otherwise universal formats. This option is typical for a class of constraints defining the interface of morphology and phonology, so-called 'alignment' constraints, matching up the edges of specific morphemes and prosodic categories. (See chapters 3 and 5.) Such interface constraints define schemata in which individual languages may substitute their specific morphemes.

We now move on to the second major property of OT constraints: their 'softness', or violability. Violability of constraints must be understood in a specific way: the general requirement is that it must be minimal:

(5)         **Violability:** constraints are violable, but violation must be minimal.

No constraint is violated without a compelling reason: avoiding the violation of another higher-ranked constraint. And even if a constraint is violated, violation must be kept to a minimum. Everything else being equal, forms with 'lesser' violations are more harmonic than forms with 'greater' violations. (Exactly how *degree of violation* is determined will be taken up in section 1.4.3.) Violability of constraints is an essential property of OT, representing a radical break away from derivational models, as well as from constraint-based theories, such as *Declarative Phonology* (Bird *1990,* Scobbie 1991), which assume that constraints are 'hard' or inviolate. (For a broad comparison with derivational theory, see chapter 2.)

This discussion of violable constraints gives rise to an important new question, to which we now turn: what is the 'optimal' candidate?

### 1.2.5   *Optimality: domination and conflict*

As mentioned before, optimality is the status of being most harmonic with respect to a set of conflicting constraints. It is now time to take a closer look at the concept of OPTIMAL in OT. The general idea is that the grammar evaluates an infinite set of candidate output forms, all analyses of a given input. From this candidate set it selects the *optimal output,* the one which 'best matches' the set of conflicting constraints. But what precisely does it mean for an output to be 'optimal'? Does it involve some sort of compromise between constraints of different strengths? Or is it perhaps the case that 'weaker' constraints are rendered 'inactive' when they come into conflict with 'stronger' constraints?

In fact optimality involves neither *compromise* nor *suppression* of constraints, but instead it is built on (strict) domination of constraints in a hierarchy.

(6)     **Optimality:** an output i s 'optimal' when it incurs the least serious violations of a set of constraints, taking into account their hierarchical ranking.

So we assume that each output form of the grammar is by definition the 'best possible' in terms of the *hierarchy* of constraints, rather than the form which matches all constraints at the same time. 'Perfect' output forms are principally non-existent, as every output form will violate at least some constraints. Therefore the selection of the 'optimal' output form involves setting priorities.

This is where a hierarchy comes into play. Conflicts are resolved by DOMINATION:

(7)     **Domination:** the higher-ranked of a pair of conflicting constraints takes precedence over the lower-ranked one.

This tentative definition will be refined below in section 1.4, on the basis of more complex cases.

The ranking of constraints can be demonstrated by a TABLEAU: this lists two (or any number of) output candidates vertically in random order, and constraints horizontally, in a descending ranking from left to right. The cells contain violation marks '*' incurred by each candidate for the constraint heading the column. Schematically:

(8)     A tableau for simple domination

|  | $C_1$ | $C_2$ |
|---|---|---|
| a. ☞ *candidate a* |  | * |
| b.    *candidate b* | *! |  |

The optimal candidate is marked by the index '☞'. This candidate is (8a), which has no violations of the higher-ranked constraint $C_1$, a constraint violated by its competitor (8b). Note that the optimal candidate (8a) is actually not impeccable itself: it has a violation of $C_2$, but this flaw is insignificant to the outcome. Although the pattern of violations for $C_2$ is the reverse of that for $C_1$, this does not help candidate b. Its violation of $C_1$ is already fatal, indicated by the accompanying exclamation mark '!' and the shading of cells whose violation content is no longer relevant. In sum, candidate (a) is optimal as *no candidate* is available that fares better, satisfying *both constraints* at the same time. A violation of $C_2$ is taken for granted, as long as $C_1$ can be satisfied.

We now turn to exemplification of the ideas that have been introduced thus far.

### 1.3 Examples of constraint interaction

#### 1.3.1 *Neutralization of voicing contrast in Dutch*

Among the universal and violable constraints is the following:

(9)     *VOICED-CODA
        Obstruents must not be voiced in coda position.

This is a typical ***markedness*** constraint, which bans a marked segment type (here: voiced obstruents) from the syllable coda (which is itself a marked position).[10]

Coda obstruents are voiceless in Dutch, as illustrated by the following alternation:

(10)     a. /bed/     bet     'bed'
         b. /bed-an1  bedən   'beds'

Dutch has no voicing contrast in final obstruents, neutralizing it towards voice-lessness.

Next consider the evaluation of two candidate outputs for the input /bɛd/, [bɛt], and [bɛd], with respect to *VOICED-CODA:

(11)     Evaluation of two candidates with respect to *VOICED-CODA
         a. [bɛt]  satisfies *VOICED-CODA
                (since [t] is an obstruent in a syllable coda, and [t] is
                voiceless)
         b. [bɛd]  violates *VOICED-CODA
                (since [d] is an obstruent in a syllable coda, and [d] is
                voiced)

If this constraint were the only one relevant for these forms, then things would be simple. Violators could be dismissed without second thoughts. But in actual grammars things are not that simple since constraints may make conflicting requirements about output forms.

A second constraint of the universal inventory is a typical ***faithfulness*** constraint, requiring that the input value of the feature [voice] be preserved in the output.

(12)     IDENT-IO(voice)
         The specification for the feature [voice] of an input segment must be
         preserved in its output correspondent.

---

[10] Actually *VOICED-CODA can be interpreted as the conjunction of two markedness statements, an idea to which we will return in chapter 9.

This faithfulness constraint mentions a notion 'correspondent', which is tentatively (and very informally) defined as follows:

(13)     **Correspondent:** the output segment that is the 'realization' of an input segment.

This informal definition is precise enough for our present purposes. (We will return to the important notion of 'correspondence', particularly in chapters 2 and 5.)

In a 'correspondence diagram' of the Dutch word [bet] 'bed' the input and output segments that are correspondents of one another are connected by vertical lines.

(14)     Correspondence diagram of [bɛt]
         /b ɛ d/ Input
         | | |
         [b ɛ t]  Output

This diagram indicates that IDENT-IO(voice) is violated in [bɛt]. Violation arises since [t], a voiceless segment in the output, corresponds with a voiced segment /d/ in the input, and both segments have conflicting values for voice. But at the same time, [bet] satisfies the markedness constraint *VOICED-CODA, as [t] is a voiceless obstruent in coda position.

We are, of course, looking at a simple conflict between two constraints, *VOICED-CODA and IDENT-IO(voice). Both constraints make incompatible requirements about the value of voice for any coda obstruent whose input is specified as [+voice]. An evaluation of both candidate outputs, [bed] and [bɛt], by the conflicting constraints is shown in (15):

(15)     Evaluation of two candidate outputs for the input /bɛd/
         a. [bed]   satisfies IDENT-IO(voice), but violates *VOICED-CODA
         b. [brt]   violates IDENT-IO(voice), but satisfies *VOICED-CODA

Observe the conflict: the evaluation of both output forms is different for each constraint.

This conflict requires resolution, which is the task of the constraint hierarchy. The form [bet] emerges as the *optimal* output of the grammar, given the following fragment of the phonology of the language:

(16)     Coda devoicing in Dutch
         *VOICED-CODA ≫ IDENT-IO(voice)

The symbol '≫' connecting both constraints is to be read as 'dominates'. Hence we read (16) as follows: *VOICED-CODA dominates IDENT-IO(voice). Domination ensures that the candidate outputs, [bɛd] and [bɛt], differ in their relative

well-formedness with respect to the ranking in (16). Or stated differently, [bet] is 'more harmonic' than [bed] with respect to the ranking in (16).

(17)  Harmonic ranking of two output candidates for the input /bed/ in Dutch

[bɛt] > [bed]

Since we are only considering two candidates here, the harmonic ranking directly gives us the optimal output: [bet].

The correctness of this constraint ranking can be represented in a tableau-format:

(18)  Tableau for the input /bed/, assuming the Dutch ranking

| Candidates: | *VOICED-CODA | IDENT-IO(voice) |
|---|---|---|
| a. ☞ [bɛt] | | * |
| b. [bɛd] | *! | |

The *optimal* candidate in the top row, [bet], incurs a violation of IDENT-IO(voice) while it satisfies *VOICED-CODA. *Suboptimal* [bɛd] has exactly the reverse pattern of violations: it has a violation mark for *VOICED-CODA, but none for IDENT-IO(voice).

Being presented with these two output candidates, the grammar (whose only goal is selecting an optimal output) must settle for a candidate that has a violation of a lower-ranked constraint, simply because no perfect output candidate is available, satisfying both constraints. This point can be made more general: constraints are intrinsically conflicting, hence perfect output candidates will never occur in any tableau:

(19)  **Fallacy of perfection:** no output form is possible that satisfies all constraints.

An output is 'optimal' since there is no such thing as a 'perfect' output: all that grammars may accomplish is to select the most harmonic output, the one which incurs the minimal violation of constraints, taking into account their ranking. Nothing better is available.

Observe that the result of the constraint interaction in Dutch is a neutralization of the voicing contrast in a specific context: the syllable coda. That neutralization indeed takes place can be easily shown by the following set of examples:

(20)  a.i /bɛd/      bet      'bed'
a.ii /bɛd-ən/ bɛ.dən 'beds'
b.i /bɛt/      bet      '(I) dab'
b.ii /bet-an/ bɛ.tən '(we) dab'

Neutralization of the sound shapes of two lexical items is the ultimate conse- quence of the domination of markedness over faithfulness. The lexical contrast between /bɛd/ and /bɛt/, residing in.the value of voicing of their final stem con- sonants, might (in principle at least) have been preserved in all morphological contexts in which they occur. But this is not the case, and a complete neutraliza- tion occurs, into [bɛt].

### 1.3.2  *Preservation of voicing contrast in English*

In English, as opposed to Dutch, an analogous input /bɛd/ is mapped to an output [bɛd], preserving the voicing in the final consonant. Accordingly, English con- trasts words such as *bed* and *bet.* This is due to the following fragment of the phonology of this language:

(21)     Preservation of voicing contrast in English
          IDENT-IO(voice) ≫ *VOICED-CODA

In English, IDENT-IO(voice) dominates *VOICED-CODA, which is the reverse ranking of the one we established for Dutch. Accordingly, the 'harmonic ranking[7] of the output candidates under discussion is reversed, as compared to the one of Dutch:

(22)     Harmonic ranking of two output candidates for the input /bɛd/ in
          English
          [bɛd] > [bɛt]

That is, assuming an input /bɛd/, '[bed] is *more harmonic* than [bet]' with respect to the ranking in (21).

   Again, we illustrate this ranking with the help of a tableau, evaluating the same candidates as we used in tableau (18) for Dutch. Observe that IDENT-IO(voice) and *VOICED-CODA have changed places:

(23)     Tableau for the input /bɛd/, assuming the English ranking

| Candidates: | IDENT-IO(voice) | *VOICED-CODA |
|---|---|---|
| a.      [bɛt] | *! | |
| b. ☞ [bɛd] | | * |

The net result of this ranking is that the 'index' pointing at the optimal output has shifted downwards (as compared to tableau 18) to the second candidate under consideration, that is, [bed]. Note that by this ranking, English preserves the phonological contrast between distinct lexical items, as in *bed* [bed] versus *bet*

[bɛt]. (This contrast is actually reinforced by a subsidiary vowel length difference between both words: [bɛˑd] versus [bɛt].)

### 1.3.3  *The relation between universal and language-spec@*

What we have just witnessed in the examples from Dutch and English is the universal 'pan-grammatical' conflict of markedness and faithfulness taking place on a micro-scale. In both languages, the same conflict arises with respect to preservation of a contrastive property (the feature [voice]), and its neutralization in a specific context (syllable coda). However, the outcome of this conflict is different for both languages. Dutch resolves it in the favour of markedness whereas English favours faithfulness. This shows that universal constraints are ranked in language-specific ways. OT clearly marks off the universal from the language-specific. Both constraints and the general principles of their interaction are universal, while constraint hierarchies are language-specific.

Speaking of *forces* of faithfulness and markedness is somewhat misleading, since this suggests that conflicts between these 'forces' are resolved on a super-ordinate level in the grammar of a single language. This is clearly not the case. For example, the fact that Dutch ranks markedness above faithfulness with respect to voice in coda obstruents does not imply that it selects the same ranking $(M \gg F)$ with respect to voice in other contexts, nor that it selects this ranking with respect to other features in the syllable coda. In Dutch, voice is contrastive in obstruents in onsets (even though voiced obstruents are universally marked). Also, place features are contrastive in obstruents and nasals in codas (in spite of the markedness of labials and velars). This shows that there are no monolithic 'forces' of faithfulness and markedness, but that instead finer-grained interactions occur between the context- and feature-specific versions of these classes of constraints. Still, for expository purposes, the classification of constraints into 'faithfulness' and 'markedness' constraints remains useful, as are shorthand notations such as '$M \gg F$.

These remarks bring us back to our starting point in this section: the conception of universal grammar in OT. But what exactly do we mean by 'grammar' in the first place? The OT grammar, and its architecture, will be the topic of the next section.

### 1.4  The architecture of an OT grammar

The OT grammar is an *input-output mechanism* that pairs an output form to an input form (such that each input has precisely one output). To accomplish this function, the grammar contains a division of labour between a component which maps the input onto an infinite set of candidate output forms, and another component that is burdened with *evaluating* the candidate output forms by a set

of ranked constraints, and selecting the **optimal** output among these. These two components are known under the names of GENERATOR (or Gen) and EVALUATOR (or Eval). This grammatical organization is schematically represented in a function notation as follows:

(24)    The grammar as an input–output mechanism
        *Gen* **(input)** $\Rightarrow$ {cand,, cand, ... cand,)
        *Eval* {cand,, cand, ... cand$_n$} $\Rightarrow$ **output**

That is, *Gen* is a function that, when applied to some input, produces a set of candidates, all of which are logically possible analyses of this input. Similarly, *Eval* is a function that, when applied to a set of output candidates, produces an output, the optimal analysis of the input. In addition to Gen and Eval, the grammar contains a LEXICON storing all lexical forms that are input to Gen. Recapitulating, we find the following model of the grammar:

(25)    Components of the OT grammar
        LEXICON: contains lexical representations (or underlying forms) of morphemes, which form the input to:
        GENERATOR: generates output candidates for some input, and submits these to:
        EVALUATOR: the set of ranked constraints, which evaluates output candidates as to their harmonic values, and selects the optimal candidate.

Let us now focus on some properties of the different components.

### 1.4.1    *The LEXICON, and Richness of the Base*

The LEXICON contains all contrastive properties of morphemes (roots, stems, and affixes) of a language, including phonological, morphological, syntactic, and semantic properties. The Lexicon provides the input specifications which are to be submitted to the Generator. In this connection, perhaps the most striking property of the Lexicon, as conceived of in OT, is that no specific property can be stated at the level of underlying representations:

(26)    **Richness of the Base:** no constraints hold at the level of underlying forms.

In OT grammatical generalizations are expressed as interactions of constraints at the *level* of the output, never at the input level. Markedness constraints always state requirements of output forms. Faithfulness constraints also evaluate output forms, although they refer to the input level in stating their requirements. The notion of contrast, which derivational theory locates at the level of the lexical

representation, is attributed to interactions at the output level in OT. Whether or not a feature is contrastive in some language depends on interactions of output-oriented markedness and faithfulness constraints, either preserving or overruling input specifications (see section 1.5).

OT thus abandons *Morpheme Structure Constraints* (MSCs), which in classical generative phonology (Chomsky and Halle 1968) account for prohibitions against specific types of structure at the level of the morpheme, in specific languages. MSCs were used, for example, to express prohibitions against front rounded vowels, or sequences of three or more consonants, or two labial consonants occurring within a morpheme. In the early 1970s MSCs were argued to be theoretically problematic in the sense that they duplicate information which is, independently, expressed by phonological rewrite rules, or that they globally guide the application of rules, a property called 'structure-preservingness'." By locating the burden of explanation of the lack of specific kinds of structure at the level of the output, OT, in principle at least, circumvents this *Duplication Problem.*

### 1.4.2   *The GENERATOR, and Freedom of Analysis*

The essential property of the **GENERATOR** is that it is free to generate any conceivable output candidate for some input. This property is called *Freedom of Analysis*.

(27)          **Freedom of Analysis:** Any amount of structure may be posited.

The only true restriction imposed on all output candidates generated by *Gen* is that these are made up of licit elements from the universal vocabularies of linguistic representation, such as segmental structure (features and their grouping below the level of the segment), prosodic structure (mora, syllable, foot, prosodic word, etc.), morphology (root, stem, word, affix, etc.), and syntax (X-bar structure, heads/complements/specifiers, etc.). Within these limits, 'anything goes'.

Since *Gen* generates all logically possible candidate analyses of a given input, the OT grammar needs no rewrite rules to map inputs onto outputs. All structural changes are applied in one step, in parallel. The evaluation of these candidate analyses is the function of the *Evaluator,* the component *of* ranked constraints, discussed in section **1.4.3.** There we will also discuss the issue of whether or not *Eval* is able to deal with an infinite candidate space.

### 1.4.3   *The EVALUATOR: economy, strict domination, and parallelism*

The **EVALUATOR** (henceforth *Eval)* is undoubtedly the central component of the grammar since it is burdened with the responsibility of accounting for all

---

[11] For example, rewrite rules may be blocked if their output would violate a MSC, or may be triggered to repair a violation of a MSC.

observable regularities of surface forms. Although any candidate output can be created by *Gen*, the crucial role of *Eval* is to assess the 'harmony' of outputs with respect to a given ranking of constraints.

*Eval* is structured as a (language-specific) hierarchy of universal constraints, plus devices for evaluation. The latter include the means to assess *violation marks* on candidate outputs for every constraint, and the means to *rank* an infinite set of candidate outputs for *harmony* with respect to the hierarchy of constraints, and select the most harmonic one of these as *optimal* – the actual output of the grammar. Let us now take a closer look at each of these devices: the constraint hierarchy, marking of violations, and harmony evaluation.

First, the constraint hierarchy contains all universal constraints (a set called *Con*), which are ranked in a language-specific way. We (tentatively) assume that all constraints are ranked with respect to each other, so as to exclude variable and undetermined rankings. (For cases in which two constraints cannot be ranked with respect to each other, due to a trivial lack of interaction, we nevertheless assume some ranking, arbitrarily one or the other.)

Moreover, within the hierarchy, dominance relations are transitive:

(28)      Transitivity of ranking:   *If* $C_1 \gg C_2$ *and* $C_2 \gg C_3$ *then* $C_1 \gg C_3$

This property of ranking will allow us to construct *ranking arguments,* as we will see below.

Second, with respect to *violation marks,* we assume that each output candidate is provided with as many marks as it has violations for a constraint. This number of marks potentially ranges from zero until *infinite*. However, for purposes of determining optimal outputs, an infinite number of marks is never practically relevant. The essence of minimal violation of constraints is that every violation of a constraint serves a purpose: to avoid a violation of some higher-ranked constraint. This is a property which is stated by Prince and Smolensky (1993: 27):
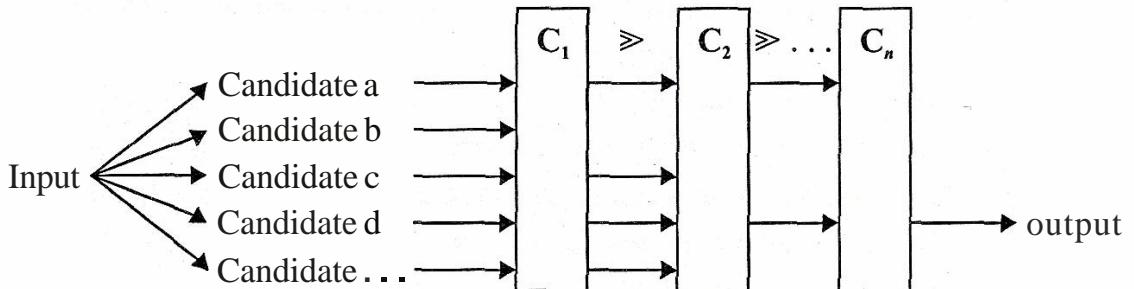
(29)      Economy: banned options are available only to avoid violations of
          higher-ranked constraints and can only be banned *minimally.*

For example, the *Generator* component is free to submit any kind of analysis of (English) /bɛd/ that is couched within the universal alphabet of representational options, including excessively unfaithful candidates such as [pɪlow] and [mætrəs]. But these candidates will be (hopefully!) ruled out regardless of constraint ranking, since they violate faithfulness constraints without compensation from reductions in markedness. This economy property of OT will be discussed in more detail in section 1.7.5.

Third, we have not yet precisely formulated in which way the evaluation of output candidates by ranked constraints proceeds. *Eval* determines the harmonic status of output candidates, and eventually the most harmonic or optimal candidate. To this end, it uses a process by which the set of candidates is reduced until the point is reached at which one output remains. This is a multi-step process, schematically repeated below from (1):

(30)

Candidate a →
Candidate b →
Input — Candidate c →
Candidate d →
Candidate ... →

$C_1$ ≫ $C_2$ ≫ ... $C_n$ → output

The major property of this evaluation process is that it applies from one state to another without looking ahead to following steps. That is, the elimination of candidate outputs by a constraint $C_n$ is never affected by a lower-ranked constraint $C_{n+m}$. Stated in a non-serial manner, this implies:

(31) **Strict domination:** violation of higher-ranked constraints cannot be compensated for by satisfaction of lower-ranked constraints.

Optimality does not involve any kind of compromise between constraints of different ranks.

To illustrate strict domination, let us return to tableau (8) – the cases of simple domination – and ask what would have happened if the first candidate had had not one but two violations of $C_2$. The following tableau shows that even here, the first candidate would still be optimal, even though its total number of violations is greater:

(32)     Strict domination: multiple violations of a lower-ranked constraint

|  | $C_1$ | $C_2$ |
|---|---|---|
| a. ☞ *candidate a* |  | ** |
| b. *candidate b* | *! |  |

No smaller amount of violations can compensate for ranking of constraints. Domination is *strict:* any candidate that incurs a violation of some higher-ranked constraint (on which another candidate incurs no violations) is mercilessly excluded, regardless of its relative well-formedness with respect to any lower-ranked constraints.

There is yet another sense in which domination is strict, which is not illustrated by (32) – constraint violations are never added for different constraints. The added violations of two lower-ranked constraints ($C_2$ and $C_3$) are not able to 'cancel' out a single violation of a higher-ranked constraint ($C_1$):

(33)        Strict domination: violations of multiple lower-ranked constraints
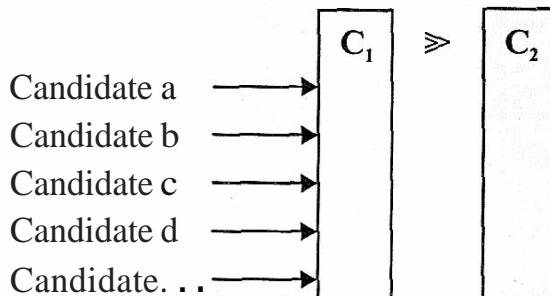
|   |  | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|---|
| a. ☞ *candidate a* |  |  | * | * |
| b. *candidate b* |  | *! |  |  |

That is, lower-ranked constraints cannot 'team up' against a higher-ranked constraint.

We see that there is no element of compromise in the notion 'optimal': evaluation of candidates by the set of constraints is based on strict domination, and accordingly, satisfaction of higher-ranked constraints has uncompromised priority over satisfaction of lower-ranked ones. Uncompromised, since no possible degree of satisfaction of lower-ranked constraints can compensate for the violation of a single high-ranked constraint.

Not all interactions of constraints are of this relatively simple kind, where an optimal candidate satisfies a high-ranked constraint that is violated by all competitors. Actually most interactions involve some degree of violation in the optimal candidate. How can this occur? Violation of a constraint is, by itself, an insufficient ground for ungrammaticality. Recall that the goal of evaluation is to single out one unique form as the most harmonic one. Elimination of all candidates in the set under consideration is therefore not allowed. This is shown in diagram below, where $C_1$ functions as a *no-pass* filter:

(34)

Candidate a
Candidate b
Candidate c
Candidate d
Candidate. . .

$C_1$ ≫ $C_2$

This must be avoided. Hence for a violation of some constraint $C_1$ to be *fatal* (eliminating from the candidate set any forms which incur it) at least one other form must occur in the candidate set that satisfies $C_1$ (without being less harmonic on higher-ranked constraints, of course). If no such form can be found, some violation must be taken for granted.

## Conflicts in grammars

In such a situation, in which all remaining candidate outputs violate a con- ,
straint (due to higher-ranked constraints), the **seriousness** of violation must be
taken into account for each individual form. That is, forms with *fewer* violation
marks of C, are preferred to forms with **more** violation marks for $C_1$. This situ-
ation may still produce a ranking argument for $C_1$ and $C_2$, as tableau (35) shows:

(35)    Amount of violation decisive

|  | $C_1$ | $C_2$ |
|---|---|---|
| a. ☞ *candidate a* | * | * |
| b. *candidate b* | **! | |
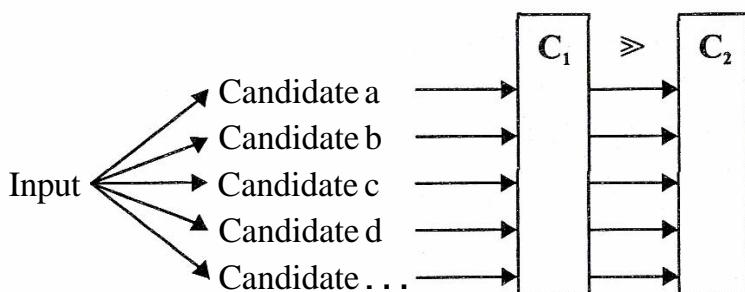
Finally, if multiple candidates have the **same** number of violations for $C_1$ (and this
equals the minimal violation in the set), then all survive and are passed on for
evaluation by the next constraint down the hierarchy, $C_2$.[12]

(36)    Tie between candidates (with lower-ranking constraint decisive)

|  | $C_1$ | $C_2$ |
|---|---|---|
| a. ☞ *candidate a* | * | *! |
| b. *candidate b* | * | |

This situation can be represented as an **all-pass** filter C,:

(37)



Of course, ties between candidates may also arise between forms that have
no violations at all, or between forms that have two, three, or any number of
violations.

Finally, we emphasize that lower-ranked constraints are not rendered 'inact-
ive', or switched off by higher-ranked constraints, but that their violation is
only avoided with less priority. Lower-ranked constraints may be violated by the

---

[12] Note that in the case of a tie, the ranking of constraints $C_1$ and $C_2$ becomes indeterminable
from the actual form – however, we assume that this ranking may be established from other
forms.

optimal output, but their violation *must be minimal.* Given the chance, any constraint (regardless of its position in the hierarchy) will be active in determining the optimal output.

(38)     Activity of a dominated constraint

|  | $C_1$ | $C_2$ | $C_3$ |
|---|---|---|---|
| a. ☞ *candidate a* |  | * | * |
| b.     *candidate b* |  | **! |  |
| c.     *candidate c* | *! |  |  |

That $C_2$ is dominated is apparent from the fact that candidate (38c) is less harmonic than (38a), even though it has no violations of $C_2$. But $C_2$ is still active, since it dominates $C_3$.

The final property of *Eval* to be briefly discussed here is:

(39)          **Parallelism:** all constraints pertaining to some type of structure interact in a single hierarchy.

In a trivial sense, it is parallelism which predicts that faithfulness constraints may interact with markedness constraints in a single hierarchy. But at a higher level of sophistication, parallelism is also the basis of explanation of phenomena involving 'interface' properties. In particular, we will see many examples in this book showing that morphological and phonological properties of an output form are mutually dependent. The most spectacular cases will come from the area of 'prosodic morphology', that is, types of morphology that depend on aspects of syllabification and metrical structure (examples being reduplication, infixation, and truncation). It is parallelism that makes information flow back and forth between 'morphological' and 'prosodic' aspects in such cases. Striking consequences of parallelism will be discussed in later chapters of this book, in particular in chapter 4 (on interactions of quantity and stress) and chapter 5 (on reduplication).

### 1.4.4   *Fear of infinity*

Freedom of Analysis may seem to pose an overwhelming *computational* problem for the basic function of a grammar, which is to provide a mapping between input and output. Perhaps the most apparent fear is that an infinite candidate space is computationally intractable. Reactions to this point focus on the nature of candidate space, on evaluation strategies which assure a more efficient processing, and on computational results booked so far in modelling OT. For an elaboration of the

arguments below, and for some others, see chapter 10 of Prince and Smolensky (1993).

Firstly, it is a well-accepted assumption among linguists that there is a distinction between the grammar (competence) and its cognitive implementation (performance). This distinction is assumed in most formal theories of grammar, and particularly in generative linguistics (Chomsky 1965). Therefore a model of grammar is adequate to the extent that it explains observed systematicities in natural languages, and the grammatical judgements of speakers. Explaining the actual processing of linguistic knowledge by the human mind is not the goal of the formal theory of grammar, but that of linguistic disciplines (such as psycholinguistics, neurolinguistics, and computional linguistics). The central point is that a grammatical model should not be equated with its computational implementation.

Secondly, turning now to computational plausibility, the fact that candidate space is infinite does not imply that the problem is logically unsolvable. You may convince yourself of this by thinking of arithmetic or any kind of numerical problem. For example, there is a unique solution to the equation $3n^2 - 3 = 45$, which you will be able to find after a moment's thought, even though the candidate set (let us say, all integers) is infinite. From a computational viewpoint, the decisive factor is that a guaranteed method (an algorithm) exists that will certainly produce a solution for any input. Therefore, no simple argument against OT as being 'computationally intractable' can be based on the observation that candidate space is infinite.[13]

Thirdly, 'smart' computational strategies may eliminate suboptimal candidates by classes, rather than on a one-by-one basis. As soon as a candidate has been excluded due to its violation of some constraint C, the evaluation process can immediately eliminate all other candidates that violate this constraint $C$ more severely. This leads us to yet another property of candidate space that might be put to use in computational evaluation models. By far the great majority of candidates proposed by Gen can never be selected as optimal, under any possible ranking of constraints. Such intrinsically suboptimal candidates can be readily identified as follows: they share with another candidate (of the same input) some set of violation marks, but have at least one additional violation of some other constraint (an example will be discussed in section 1.7.5). Sophisticated evaluation strategies may capitalize on this. Since the identification of intrinsically suboptimal candidates involves no ranked constraints, infinite candidate space may be drastically reduced by eliminating the 'worst-of-the-worst' of candidates by preprocessing prior to the evaluation by ranked constraints. Since this preprocessing

---

[13] Conversely, a finite set of candidates does not guarantee that a problem is logically solvable. This argument is due to Alan Prince (presentation at Utrecht University, January 1994).

would eliminate the great majority of candidates, the ultimately relevant remaining part of candidate space may well have quite manageable proportions, and perhaps even reduce to a finite set (Hammond 1997).

Whether or not a computational method can be established for the evaluation of an infinite candidate space in OT grammars is still largely an open issue, but encouraging results are available. For example, Karttunen (1998) demonstrates that 'the computation of the most optimal surface realization of any input string can be carried out entirely within a *finite-state* calculus, subject to the limitation (Frank and Satta (1998)) that the maximal number of violations is bounded'. Karttunen adds that '[i]t is not likely that this limitation is a serious obstacle to practical optimality computations with finite-state systems as the number of constraint violations that need to be taken into account is generally small'.

## 1.5   Interactions of markedness and faithfulness

This section will deal with important types of interactions of markedness and faithfulness constraints, building on key insights of Prince and Smolensky (1993), Kirchner (1995), and Steriade (1995b). In section 1.3.3 we studied final devoicing in Dutch as a case of positional neutralization: the feature [voice] is neutralized in a specific context (the syllable coda), as a result of a markedness constraint dominating a faithfulness constraint. Here we will first extend this simple interaction of markedness and faithfulness to a new phenomenon: allophonic variation. In section 1.6 we will look into the notion of contrast as it is defined in OT, and its consequences for lexical representation. This will allow us to define more clearly the ranking schemata of faithfulness and markedness constraints that are responsible for the various attested situations ('contrast', 'neutralization', and 'allophonic variation'). In section 1.7 we will summarize these results in the form of a 'factorial typology'. In section 1.8 we will see how segment inventories follow from interactions of faithfulness and markedness.

### 1.5.1   *Allophonic variation*

Consider a language that has no lexical contrast of oral and nasal vowels. In this language oral and nasal vowels are allophones, variants of one another which are fully predictable from the phonological contexts. For example, vowels are generally oral except when they directly precede a tautosyllabic nasal stop, in which case they are nasal. This allophonic pattern occurs in many dialects of English; see the examples below:

(40)  a.i   cat  [kæt]      b.i   can't  [kæ̃nt]
      a.ii  sad  [sad]      b.ii  sand   [sæ̃nd]
      a.iii met  [met]      b.iii meant  [mẽnt]
      a.iv  lick [lɪk]      b.iv  link   [lĩŋk]

When we say that English lacks a contrast of oral and nasal vowels, we do not imply that English completely lacks either kind of vowels, but only that no word pairs occur that are distinguished by orality/nasality of their vowels. Whatever variation there is between oral and nasal vowels is totally conditioned by the context and does not reflect lexical specification. Vowels are nasal when they precede a tautosyllabic nasal, and are oral in all other contexts. This complement-ary distribution, and the corresponding lack of word pairs that differ only in the specification of some feature, is what defines an allophonic pattern. How can the allophonic pattern in (40) be stated in terms of violable constraints?

In order to answer this question, we must first identify the set of constraints which are involved. Universally, nasal vowels are 'marked' as compared to oral vowels. Most of the world's languages completely lack nasal vowels, having oral vowels only (Maddieson 1984). Languages may have both oral and nasal vowels, but no languages have only nasal vowels. In sum, when a language has nasal vowels, it must also have oral vowels. The marked status of nasal vowels is expressed by the ***context-free*** markedness constraint in (41), which militates against nasal vowels:

(41)  $*V_{NASAL}$
      Vowels must not be nasal.

When this constraint is undominated in some language, then all of its vowels will be oral, regardless of their lexical specification, or their position in the syllable (before an oral or nasal).

Moreover, many languages tend to nasalize vowels in precisely the position where they are nasal in English: before a tautosyllabic nasal stop. The vowel thus anticipates the nasality of the following stop, a preferred state of affairs from the viewpoint of perception and articulation (Cohn 1993a).[14] Again, a markedness constraint expresses the universal markedness, ruling out oral vowels that precede a tautosyllabic nasal:

(42)  $*V_{ORAL}N$
      Before a tautosyllabic nasal, vowels must not be oral.

Observe that this constraint is ***context-sensitive,*** since it states a connection between the nasality of a vowel and a nasal stop in its context. More precisely, it is violated by an oral vowel that stands directly before a tautosyllabic nasal:

---

[14] Cohn (1993a) argues that nasalization in English vowels is gradient, and has no phonological status, as in French. For the sake of the argument, we will assume here that English nasalization is in fact categorical, although it is crucially non-contrastive.

(43)      a.  *V$_{ORAL}$N satisfied      b.  *V$_{ORAL}$N violated
          i.  æ̃n]$_\sigma$                    æn]$_\sigma$
          ii. ad],

If this constraint is undominated, underlying contrasts between oral and nasal vowels (if any) will be neutralized in positions before a tautosyllabic nasal.

### 1.5.2  *Neutralization and contrast as constraint rankings*

Now consider the consequences of the OT assumption of the *Richness of the Base,* which was stated in section 1.4.1. This says that no constraints restrict the input, or to put it differently, that lexical representations in any language are free to contain any kind of phonological contrast. Whether some surface phonetic contrast (such as that between oral and nasal vowels) is allophonic or lexically distinctive in a language depends on interactions of two basic kinds of constraints: markedness constraints, which express markedness statements, and faithfulness constraints, which penalize deviations of the surface form (output) from its lexical form (input). When markedness dominates faithfulness, the language achieves outputs that are minimally marked, at the expense of a neutralization of lexical contrasts. But when faithfulness dominates markedness, the language makes the reverse choice, realizing its input contrasts at the expense of output markedness:

(44)      a.  Markedness ≫ Faithfulness   lexical contrasts are neutralized
          b.  Faithfulness ≫ Markedness   lexical contrasts are expressed

Richness of the Base implies that English (as any other language) is allowed the option of setting up a contrast of oral and nasal vowels in its underlying representations. However, this hypothetical contrast is never realized at the surface, because with respect to nasality/orality in vowels, English happens to be a language of the type (44a), which gives priority to markedness over faithfulness. Whatever lexical contrast of nasality there might be in vowels will be *obscured* by effects of markedness. The input faithfulness constraint that is crucially dominated in English requires that surface values of nasality in vowels are identical to their underlying values:

(45)      **IDENT-IO**(nasal)
          Correspondent segments in input and output have identical values for [nasal].

In a language in which IDENT-IO(nasal) is undominated, any lexical contrast of nasality in vowels will be allowed to surface, uninhibited by the markedness constraints (41–2). Such a language is free to set up and preserve any lexical contrast between oral and nasal vowels *anywhere,* that is, without any neutralization. This

situation corresponds to the interaction (44b). But in a language in which IDENT-IO(nasal) is dominated by both of the markedness constraints (41) and (42), any (potential) contrast of orality/nasality in vowels will be fully neutralized, as is the case in allophonic variation. This is the situation (44a), found in English.

Let us now return to the allophonic pattern (40) and find out how this results from the interaction of the three constraints that were introduced earlier. In terms of constraint interaction, faithfulness to the lexical specification of a vowel is completely dominated by markedness constraints reflecting markedness of orality/nasality in vowels. In terms of ranking, IDENT-IO(nasal) is dominated by both markedness constraints:

(46)    Neutralization of lexical contrast
        Markedness        ≫ Faithfulness
        $*V_{NASAL}$, $*V_{ORAL}N$ ≫ IDENT-IO(nasal)

This is an instantiation of the schema in (44a), where markedness constraints completely dominate faithfulness.

The question which arises next is how both markedness constraints, $*V_{ORAL}N$ and $*V_{NASAL}$, are ranked with respect to each other, As we observed earlier in connection with the context-free constraint $*V_{NASAL}$, any language in which this is undominated will totally lack nasal vowels in its surface patterns. This is not the case in English, however, where nasal vowels do occur (as allophones of oral vowels) in specific positions, that is, before tautosyllabic nasal stops. We must therefore refine the ranking in (46) to that in (47):

(47)    Allophonic variation
        Contextual markedness ≫ Context-free markedness ≫ Faithfulness
        $*V_{ORAL}N$               ≫ $*V_{NASAL}$              ≫ IDENT-
                                                                  IO(nasal)

This ranking states that nasal realization of vowels before tautosyllabic nasal consonants takes priority over a total lack of nasality in vowels. In sum, both nasal and oral vowels occur at the surface, but their distribution is fixed, rather than free.

This ranking is illustrated in the tableaux (48–51). First consider the case of an oral vowel in the actual output, for example *sad* [sad]. When we assume that this has an oral vowel in its lexical representation, e.g. /sæd/, matching its surface status, we arrive at the first tableau (48). Candidate (48a) is optimal as it violates

none of the constraints in the tableau, regardless of ranking. It satisfies $*V_{ORAL}N$ since this constraint has nothing to say about vowels that stand before oral stops. It also satisfies $*V_{NASAL}$ since it has no nasal vowel. Finally it satisfies IDENT-IO(nasal) because the input and output agree in nasality.

(48)

| Input: /sæd/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. ☞ [sæd] | | | |
| b.    [sæ̃d] | | *! | * |

The losing candidate [skd] (48b) is less harmonic than (48a) in two respects. It contains a nasal vowel, fatally violating the markedness constraint $*V_{NASAL}$. It violates IDENT-IO(nasal) as well, as the nasal vowel in the output fails to match its oral correspondent in the input.

Because of Richness of the Base, we must guarantee that this correct result is not negatively affected when we make different assumptions about the nasality of vowels in the input. Indeed, the same candidate [sæd] is selected when the input would contain a nasal vowel, e.g. lskdl, here in defiance of its surface form. This is shown in tableau (49). Again markedness uniquely determines the outcome, without interference on the part of the faithfulness constraint IDENT-IO(nasal).

(49)

| Input: /sæ̃d/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. ☞ [sæd] | | | * |
| b.    [sæ̃d] | | *! | |

Note that in this case IDENT-IO(nasal) is violated in the optimal candidate. This motivates the ranking $*V_{NASAL} \gg$ IDENT-IO(nasal), a markedness constraint dominating faithfulness. That is, even if the input of *sad* were to contain a nasal vowel, its nasality would be wiped out in the surface form by markedness constraints. This is of course the central result that we need to account for allophonic variation, in a theory which assumes Richness of the Base.

We can only rightfully claim to have captured the 'complementary distribution' of oral and nasal vowels if we can prove the total 'irrelevance of the input' for words which surface with nasal vowels, for example *sand* [sæ̃nd]. Again we consider two underlying forms, one with an oral vowel and the other with a nasal vowel. Tableau (50) shows that an underlying form with an oral vowel /sænd/ results in an optimal output with a nasal vowel, [sæ̃nd]. This is due to the undominated context-sensitive markedness constraint $*V_{ORAL}N$, which requires that vowels are nasal before a tautosyllabic nasal stop:

(50)

| Input: /sænd/ | *V<sub>ORAL</sub>N | *V<sub>NASAL</sub> | IDENT-IO(nasal) |
|---|---|---|---|
| a.　　[sænd] | *! | | |
| b. ☞ [sãˑnd] | | * | * |

Observe that two markedness constraints, *V<sub>ORAL</sub>N and *V<sub>NASAL</sub>, are in conflict here. The former requires a nasal vowel in the output whereas the latter militates against it. The fact that the actual output [sãˑnd] has a nasal vowel shows that *V<sub>ORAL</sub>N dominates *V<sub>NASAL</sub>. (If the ranking had been reverse, the result would have been in favour of candidate 50a, which has an oral vowel.) Observe also that the underlying orality of the vowel in *sand* does not affect the outcome. IDENT-IO(nasal) is violated in the optimal output, since it contains a nasal vowel whereas the input contains an oral vowel. This conclusion is essential to the argument that faithfulness is dominated by *both* markedness constraints. We have already reached this conclusion for *V<sub>NASAL</sub> in tableau (49), and now we confirm it for *V<sub>ORAL</sub>N.

The argument for the irrelevance of inputs in allophonic patterns is completed by an inspection of tableau (51), which has an underlying form with a nasal vowel, /sãˑnd/. In this tableau, the same optimal candidate is selected as in the previous one, simply because markedness uniquely determines the outcome.

(51)

| Input: /sãˑnd/ | *V<sub>ORAL</sub>N | *V<sub>NASAL</sub> | IDENT-IO(nasal) |
|---|---|---|---|
| a.　　[sænd] | *! | | * |
| b. ☞ [sãˑnd] | | * | |

A comparison of tableaux (50) and (51) reveals the complete inactivity of the faithfulness constraint IDENT-IO(nasal). We conclude that the orality/nasality of the underlying vowel is completely irrelevant to the surface distribution of oral and nasal vowels.

### 1.6　Lexicon Optimization

The main result of the preceding section is that lexical specifications for [nasal] in vowels in English are totally irrelevant to their surface realization. Should we then conclude that the English lexicon is completely unstructured for nasality in vowels, in the sense that the vowels in lexical items *sad* and *sand* are randomly specified for this feature? Or should we still insist that the lexicon be kept 'clean' from featural noise, and contain only feature values that are actually related to

output values? An answer to this question is potentially relevant to language acquisition. In order to build a lexicon, the learner must somehow be able to determine *underlying* forms, for example to infer the underlying form of *sad* on the basis of its surface form [sæd]. When we concentrate on possible values for nasality in the vowel, there is a choice of two lexical representations, /sæd/ and /sæ̃d/. We have seen that, given the ranking of $*V_{NASAL}$ over IDENT-IO(nasal), both lexical representations result in identical outputs. This ranking completely *masks* the input, obscuring empirical evidence that the learner might use to base his/her choice of an underlying form on.

It has been proposed that in the absence of empirical evidence for one input form over another, the input should be selected that is closest to the output, in this case /sæd/. That is, wherever the learner has no evidence (from surface forms) to postulate a specific diverging lexical form, (s)he will assume that the input is identical to the surface form. In terms of constraint violations, this strategy has the advantage of minimizing the violation of faithfulness, *as compared to any other hypothetical inputs producing the same output.* This strategy is called *Lexicon Optimization* in Prince and Smolensky (1993: 192):

(32) **Lexicon Optimization:** suppose that several different inputs $I_1$, $I, \ldots, I$, when parsed by a grammar G lead to corresponding outputs $O_1, O_2 \ldots, O_n$, all of which are realized as the same phonetic form $\Phi$ – these inputs are *phonetically equivalent* with respect to *G.* Now one of these outputs must be the most harmonic, by virtue of incurring the least significant violation marks: suppose this optimal one is labelled $O_k$. Then the learner should choose, as the underlying form for $\Phi$, the input $I$.

This principle is, in its turn, an elaboration of an idea of Stampe (1972), who suggested that underlying forms should always match surface forms in the absence of evidence to the contrary. (The 'masking' effect of one underlying form, /sæ̃d/, by another, /sæd/, is called 'Stampean occultation' in Prince and Smolensky 1993.)

An alternative to Lexicon Optimization is to assume that certain inputs contain no specification with respect to a feature (Kiparsky 1985, Steriade 1987, Archangeli 1988). This *underspecification* analysis of nasality in vowels is based on the idea that the burden of explanation for contrastive versus allophonic patterns is in the underlying form, rather than in the relationship between underlying form and surface form, as is the case in OT.[15]

---

[15] See Smolensky (1993), Inkelas (1995) and Itô, Mester, and Padgett (1995) for comments on underspecification in OT.

Importantly, Lexicon Optimization does not contradict the assumption of Richness of the Base, even though it may give rise to lexicons that are 'impoverished' in terms of featural 'noise'. The burden of explanation still remains on the interaction of markedness and faithfulness. More specifically, the ranking markedness ≫ faithfulness implies that it is not an accidental observation that nasality is never distinctive in vowels in English – on the contrary, this is a solid effect of constraint interactions in the English grammar.

## 1.7 A factorial typology of markedness and faithfulness

In the remainder of this chapter we will consider the consequences of reranking the three constraint types which we have assumed in the analysis of allophonic variation, that is: context-free markedness, context-sensitive markedness, and faithfulness. We will see that by reranking these three constraint types into different hierarchies, a 'factorial typology' arises which exactly matches the attested cross-linguistic variation in terms of allophonic variation, positional neutralization, and free contrast. This will provide further support for the 'markedness' approach of allophonic variation which we have used so far, as well as for the assumption of Richness of the Base.

### 1.7.1 Typological goals of OT

The important notion of factorial typology requires some explanation, before we actually construct one. The key assumption of OT is that grammars are means to resolve conflicts between universal constraints. More specifically, the grammar of an individual language is a specific way, out of many possible, to rank a set of universal and violable constraints. Differences between languages must therefore be due to different rankings of a single set of universal constraints. To state it differently, we can build one grammar out of another by a rearrangement of its basic universal material, that is, by 'reranking' the constraints.

The relative success of any theory of grammar should be measured by its ability to characterize the notion of 'possible grammar' (see again the remarks in the introduction of this chapter). Constructing grammars ('constraint hierarchies') of individual languages may tell us much about the ways in which linguistic properties are interconnected within a single linguistic system. But what we are eventually interested in are typological results of the theory, that is, the predictions it makes about clusterings of linguistic properties, on a broad cross-linguistic basis. For example, the theory should explain why no languages occur that have a contrast of oral and nasal vowels, but contextually restrict this contrast to vowels immediately preceding nasal stops (neutralizing it everywhere else). Languages

of this kind are logically possible, yet unattested. Can this situation be described by a reranking of the constraints governing nasality in vowels? Conversely, we should ask what language types would arise by reranking a number of constraints that are motivated in the analysis of an individual language. Does reranking of these constraints produce attested languages as well?

Taken quite literally, the reranking approach would predict that any new grammar that arises from a reranking of any pair of constraints will precisely correlate with one of the world's languages. This prediction is based on the deeply naive assumption that every possible ranking should be instantiated by some attested language. This is naive, just as it is deeply naive to expect that all logically possible permutations of genetic material in the human genome are actually attested in individual humans. Therefore, in order to test the typological predictions of the theory of contrast and contextual neutralization presented in this section, we will rerank *types* of constraints (rather than individual constraints) of the general types 'faithfulness', 'contextual markedness', and 'context-free markedness'. The resulting factorial typology will be matched with broad *typological diversity* between languages, along the dimensions that these constraint types represent. (Of course, this is not to deny that cases can occur in which it is more useful to compute factorial typologies of individual constraints.)

### 1.7.2 *Constructing a factorial typology*

To construct a factorial typology of a set of constraints, we sum up all logically possible rankings of this set of constraints, and compute the different outcomes. With large sets of constraints the number of possible rankings rises steeply, as with a constraint set of size *n,* we must consider all *n!* rankings. (This equals 2 rankings for 2 constraints, 6 rankings for 3, 24 for 4, 120 for 5, 720 for 6, etc.) Fortunately, many of the individual rankings in a factorial typology produce identical surface patterns. Therefore the number of predicted patterns is much smaller than the total number of logically possible rankings. Keeping these remarks in mind, we now turn to a real case.

Our goal is to construct a factorial typology of the constraint types introduced so far (faithfulness constraints, context-free and context-sensitive markedness constraints). One proviso has to be made: actual grammars rank individual constraints, not 'constraint types'. (For example, no grammar ranks all faithfulness constraints above all markedness constraints, for reasons discussed in section 1.2.2.) Nevertheless, we will *generalize* the factorial typology by a reranking of constraint types. Accordingly, the emerging factorial typology should be taken as a catalogue of general effects, which may be instantiated in an individual grammar with respect to specific features.

A factorial typology of markedness and faithfulness is presented below, result-ing from a reranking of both markedness constraint types with faithfulness. We will abbreviate these types of markedness constraints as *MC-free* (for context-free markedness constraint) and *MC-sensitive* (for context-sensitive markedness constraint).

(53)      **A** factorial typology of markedness and faithfulness
          a.  MC-free ≫ MC-sensitive, Faithfulness     Lack of variation
                                                        (unmarked)
          b.  MC-sensitive ≫ MC-free ≫ Faithfulness    Allophonic variation
          c.  MC-sensitive ≫ Faithfulness ≫ MC-free    Positional neutralization
          d.  Faithfulness ≫ MC-sensitive, MC-free     Full contrast

The attentive reader may have noted that we have only four rankings here, rather than the predicted six (or 3!). This reduction is due to the fact that in rankings (53a) and (53d), the mutual ranking of the bottom two constraints is of no import-ance to the outcome.

The following subsections discuss how these situations arise from these rank-ings, and also illustrate each ranking by tableaux for contrastive nasality in vowels.

### 1.7.3 *Neutralization: lack of variation versus allophonic variation*

Both (53a) and (53b) are situations of complete neutralization, since in both cases there is a total lack of activity of the faithfulness constraint, which is at the very bottom of the hierarchy. The difference between the rankings resides in whether or not the neutralized feature is 'contextually coloured', that is, subject to allo-phonic variation.

First consider the situation of total lack of variation, which is produced by ranking (53a), due to an undominated context-free markedness constraint for some feature [αF]. Accordingly the unmarked value of this feature (for segments of a given type), [$u$F], will always appear at the surface, regardless of its under-lying specification, and regardless of the context. This results in the complete neutralization of this feature in the direction of the unmarked value. Such total *lack of variation* for a given feature (in all segments of some type) is widely attested for different features among the world's languages.

For example, if the constraint $*V_{NASAL}$ is undominated, then all surface vowels are oral, even those vowels which are underlyingly nasal, and even those vowels which are adjacent to a nasal consonant. This is illustrated in the set of four tableaux below. Each of these tableaux takes as its input one of the four possible combinations of input nasality in vowels (nasal versus oral) and output context

the vowel (preceding a nasal [n] or an oral [l]). All four possible inputs {/pan/ ~ /pãn/ ~ /pal/ ~ /pãl/} map onto oral output vowels:

(54)  Lack of variation of nasality in vowels (total orality)
$*V_{NASAL} \gg *V_{ORAL}N$, IDENT-IO(nasal)

| (i)   Input: /pan/ | $*V_{NASAL}$ | $*V_{ORAL}N$ | IDENT-IO(nasal) |
|---|---|---|---|
| a.          pãn | *! | | |
| b.   ☞ pan | | * | |

| (ii)  Input: /pãn/ | $*V_{NASAL}$ | $*V_{ORAL}N$ | IDENT-IO(nasal) |
|---|---|---|---|
| a.          pãn | *! | | |
| b.   ☞ pan | | * | * |

| (iii)  Input: /pal/ | $*V_{NASAL}$ | $*V_{ORAL}N$ | IDENT-IO(nasal) |
|---|---|---|---|
| a.          pãl | *! | | |
| b.   ☞ pal | | | |

| (iv)  Input: /pãl/ | $*V_{NASAL}$ | $*V_{ORAL}N$ | IDENT-IO(nasal) |
|---|---|---|---|
| a.          pãl | *! | | |
| b.   ☞ pal | | | * |

Observe that the ranking of $*V_{ORAL}N$ and IDENT-IO(nasal) with respect to one another is totally irrelevant to the outcome, since the orality of the vowel is uniquely determined by $*V_{NASAL}$.

Ranking (53b) produces the typologically common case of **allophonic varia-tion,** of which we have already encountered an example in the form of vowel nasalization before tautosyllabic nasals in English. As compared to the previous ranking, (53a), this ranking maintains complete neutralization, yet it allows for some variation in output values for the relevant feature. For example, both values of nasality in vowels do occur in surface forms, although their distribution is totally determined by the context. Vowels are nasal before nasal consonants (regardless of their input specification), and they are oral in all other contexts (regardless of their input specification). Although tableaux of English examples have already been presented in section 1.5.2, we include new tableaux here for maximal clarity:

(55)  Allophonic variation of nasality in vowels

$*V_{ORAL}N \gg *V_{NASAL} \gg \text{IDENT-IO(nasal)}$

| (i) Input: /pan/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. ☞ pãn | | * | * |
| b. pan | *! | | |

| (ii) Input: /pãn/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. ☞ pãn | | * | |
| b. pan | *! | | * |

| (iii) Input: /pal/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. pãl | | *! | * |
| b. ☞ pal | | | |

| (iv) Input: /pãl/ | $*V_{ORAL}N$ | $*V_{NASAL}$ | IDENT-IO(nasal) |
|---|---|---|---|
| a. pãl | | *! | |
| b. ☞ pal | | | * |
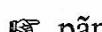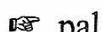
The two remaining rankings in the factorial typology, (53c) and (53d), produce varying degrees of contrastiveness, as we will see below.

### 1.7.4  *Contrast: positional neutralization versus full contrast*

Ranking (53c) produces a *positional neutralization* of underlying feature values. This is a situation in which an underlying contrast is freely realized in most contexts, but where it is neutralized in a specific context. For example, nasality is contrastive in vowels, except in the context before a nasal consonant, where all vowels are nasal:

(56)  Positional neutralization of nasality in vowels before nasal consonants

$*V_{ORAL}N \gg \text{IDENT-IO(nasal)} \gg *V_{NASAL}$
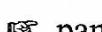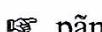
| (i) Input: /pan/ | $*V_{ORAL}N$ | IDENT-IO(nasal) | $*V_{NASAL}$ |
|---|---|---|---|
| a. ☞ pãn | | * | * |
| b. pan | *! | | |

| (ii)  Input: /pãn/ | *V$_{ORAL}$N | IDENT-IO(nasal) | *V$_{NASAL}$ |
|---|---|---|---|
| a.  ☞ pãn | | | * |
| b.  pan | *! | * | |

| (iii)  Input: /pal/ | *V$_{ORAL}$N | IDENT-IO(nasal) | *V$_{NASAL}$ |
|---|---|---|---|
| a.  pãl | | *! | * |
| b.  ☞ pal | | | |

| (iv)  Input: /pãl/ | *V$_{ORAL}$N | IDENT-IO(nasal) | *V$_{NASAL}$ |
|---|---|---|---|
| a.  ☞ pãl | | | * |
| b.  pal | | *! | |

**Finally,** the logically opposite situation of (53a) is that produced by ranking (53d), where a faithfulness constraint governing a feature dominates all markedness constraints (governing this feature). This produces a pattern in which input feature specifications are freely realized, that is, a situation of *full contrast* for the relevant feature.

For nasality, this ranking produces a situation in which underlying specifications in vowels are realized at the surface, regardless of their adjacency to nasal consonants:

(57)    Full contrast of nasality in vowels
       IDENT-IO(nasal) ≫ *V$_{NASAL}$, *V$_{ORAL}$N

| (i)  Input: /pan/ | IDENT-IO(nasal) | *V$_{NASAL}$ | *V$_{ORAL}$N |
|---|---|---|---|
| a.  pãn | *! | * | |
| b.  ☞ pan | | | * |

| (ii)  Input: /pãn/ | IDENT-IO(nasal) | *V$_{NASAL}$ | *V$_{ORAL}$N |
|---|---|---|---|
| a.  ☞ pãn | | * | |
| b.  pan | *! | | * |

| (iii)  Input: /pal/ | IDENT-IO(nasal) | *V$_{NASAL}$ | *V$_{ORAL}$N |
|---|---|---|---|
| a.  pãl | *! | * | |
| b.  ☞ pal | | | |

| (iv)  Input: /pãl/ | IDENT-IO(nasal) | *V$_{\text{NASAL}}$ | *V$_{\text{ORAL}}$N |
|---|---|---|---|
| a.        ☞ pãl |  | * |  |
| b.             pal | *! |  |  |

Again, this free contrast is a cross-linguistically common type of situation.

### 1.7.5  *Positional neutralization of voice in Dutch*

It will now be clear that Dutch final devoicing is an example of positional neu-
tralization: it produces a neutralization of the feature [voice] in obstruents in the
specific context of a syllable coda, while leaving unaffected the lexical distri-
bution of [voice] in obstruents in other contexts. To fit the Dutch case into the
ranking schemata of the previous section, we must first determine which context-
free markedness constraint is involved in the ranking.

The unmarked value for the feature [voice] in obstruents is [–voice], as stated
in VOICED OBSTRUENT PROHIBITION (58c, henceforth VOP, after It6 and Mester
1998), which is accompanied by the other two constraints relevant to the Dutch
devoicing pattern:

(58)        a. *VOICED-CODA          (context-sensitive markedness constraint)
               Coda obstruents are voiceless.
            b. **IDENT-IO**(voice)        (*faithfulness* constraint)
               The value of the feature [voice] of an input segment must be
               preserved in its output correspondent.
            c. VOP                         (*context-free markedness* constraint)
               *[+voi, –son]
               No obstruent must be voiced.

These three constraints are ranked in the following way in Dutch, instantiating the
pattern of positional neutralization (53c) with respect to the feature [voice]:

(59)        Ranking producing positional neutralization of voice in Dutch
            MC-sensitive   ≫ Faithfulness      ≫ MC-free
            *VOICED-CODA ≫ IDENT-IO(voice) ≫ VOP

This ranking states that a voiceless realization of obstruents in coda position takes
priority over preservation of [voice] in coda obstruents. However, preservation of
input values of [voice] takes priority over the complete devoicing of obstruents.
In sum, the contrast of voiced and voiceless obstruents is positionally neutralized
in the syllable coda. Elsewhere, a contrast is possible – input values of [voice] are
preserved in the output.

In terms of concrete examples, this ranking correctly predicts that the output of /bɛd/ is [bet], which is unfaithful to input values for [voice] *only in its coda consonant*. But the voiced onset consonant /b/ is protected from the complete devoicing of obstments required by the context-free markedness constraint VOP (hence, *[pet]). This interaction is shown by tableau (60), containing all four logically possible combinations of [voice] in the onset and coda consonants:

(60)

| Input: /bɛd/ | *VOICED-CODA | IDENT-IO(voice) | VOP |
|---|---|---|---|
| a. ☞ [bɛt] | | * | * |
| b.    [pɛt] | | **! | |
| c.    [bɛd] | *! | | ** |
| d.    [pɛd] | *! | * | * |

Two candidates (60c–d) are eliminated by undominated *VOICED-CODA, as each contains a voiced obstruent in coda position. Both remaining candidates (60a–b) satisfy *VOICED-CODA, hence both are passed on for evaluation by the next-lower-ranked constraint in the hierarchy, IDENT-IO(voice). Although both (60a) and (60b) violate IDENT-IO(voice), the former is selected since it violates IDENT-IO(voice) *minimally.* It has only one violation, while its contestant (60b) incurs two violations, one more than is strictly necessary. This result reflects an important property of the architecture of OT: a constraint can be 'active' even when it is dominated by one or more other constraints. *Constraints may be violated, but violation must be minimal.* This property of constraint interactions will reoccur many times in this book.

Another major property of constraint interaction is also illustrated by tableau (60). This is that *some candidates can never emerge as optimal, regardless of the ranking of constraints.* To see this, consider output candidate (60d), [prd], which preserves the input value for [voice] in its coda consonant, but is unfaithful to [voice] in its onset consonant. This incurs violations for each of the three constraints in the tableau: it violates *VOICED-CODA as it has a voiced coda obstruent [d], it violates IDENT-IO(voice) as it is unfaithful to the input value of [voice] in one of its consonants (the onset [p]), and finally it violates VOP because it contains a voiced obstruent [d]. Under what constraint ranking might this candidate be selected as optimal? The surprising answer is: 'under no ranking', since all logically possible rankings of the three constraints evaluate (60d) as *suboptimal to some other candidate.* To prove this point, we need not go through all tableaux of all possible rankings, although this method will certainly lead to the same conclusion. A more general proof is available. To mark a candidate $cand_1$ as

'intrinsically suboptimal', it suffices to identify a rivalling candidate *cand,* which shares with *cand,* the violation marks for every constraint, except for at least one constraint *C,* on which *cand,* is more harmonic. If such a constraint *C* exists, then *cand,* must be a better candidate than *cand,* regardless of the ranking of *C,* since the minimal difference in violation marks always works in its favour, even if C were to dwell at the very bottom of the hierarchy.

There happens to be such a candidate *cand,* in tableau (60): candidate (60a), [bet], shares with candidate [ped] (60d) one violation mark for IDENT-IO(voice), and one for VOP, yet it minimally improves over (60d) with respect to *VOICED-CODA. Hence:

(61)          [bet] ≻ [ped]  For input /bɛd/, *irrespective* of ranking.

This does *not* imply that [bɛt] is the optimal candidate under *any ranking:* it clearly is not (only consider rankings in which either *VOICED-CODA or VOP is undominated). It does imply, however, that [ped] is 'intrinsically suboptimal' – which means that it will never be selected as optimal under any logically possible ranking of the three constraints under consideration.

This result, although apparently limited to the interaction of the three constraints in tableau (60), in fact has broader typological implications. A prediction follows from it, which is stated in general terms as follows. Assume a context-free markedness constraint banning one value of a feature [αF], and another context-sensitive markedness constraint banning the same value [αF] in a specific context. The prediction is that no language can have a contrast of [±F] *exclusively* in the context where a context-sensitive markedness constraint bans [αF]. This seems to be correct, although further testing may be required.

One particular language type excluded is one that has a lexical contrast of voicing exclusively in syllable codas. See the following hypothetical pattern of contrast:

(62)          A hypothetical language that is predicted not to occur
              a.  a contrast of voice in syllable codas
                    lap ~ lab, pot ⁻ pod, muuk ~ muug
              b.  but no contrast of voice elsewhere
                    paa (*baa), ma.tol (*ma.dol), tol.ku (*tol.gu)

Such a language would preserve a contrast of voice in the coda, but neutralize it elsewhere. That is, it would map an input /bɛd/ onto an optimal output [prd]. But we have just seen that such a mapping is ruled out on principled grounds, since it involves the selection of an intrinsically suboptimal candidate. The asymmetry between onsets and codas is due to a context-sensitive markedness constraint

*Voiced*-Coda which rules out [+voice] in the syllable coda, while there is no analogous context-free markedness constraint which rules out any feature of voice *specifically* in the onset.

Alternative theories which do not assume markedness to be the actual substance of the grammar fail to derive this general prediction. For example, a rule-based theory in which the notion of 'markedness' is an external criterion fails to predict that hypothetical languages such as (62) should not exist. This is because phonological rules that neutralize a contrast of voice are 'natural' in *any context,* regardless of whether they apply in onset or in coda position. A rule neutralizing voicing in onsets is 'natural' in this general sense, and no language that has this rule is committed to having a second rule neutralizing voicing in codas as well. Therefore a grammar which neutralizes a voicing contrast in all contexts except in codas should be possible, even though it would be 'complex' (in the sense that different rules would be employed, instead of a single general one).

### 1.7.6 Typology: some preliminary conclusions

To wind up this section let us now summarize the results. At the heart of OT is the notion that grammars of individual languages instantiate general ranking schemata of constraints of different types. The basic method of checking the typological predictions made by the theory is that of constructing a factorial typology by the reranking of constraints of different types. In this section we have constructed a basic factorial typology of faithfulness and markedness, and found that all predicted types of input–output relationships are attested. By varying the ranking of faithfulness with respect to (context-free and context-sensitive) markedness constraints, we found a factorial typology which ranged from a situation of total neutralization on the one hand, to that of total freedom of contrast on the other hand. In between these extremes, we identified two intermediate situations: allophonic variation (a specific kind of neutralization which allows two values of some feature in the output), and positional neutralization (a situation in which a feature is contrastive, except in a specific context, in which it is neutralized). We elaborated on positional neutralization of voice in Dutch to demonstrate two typical properties of constraint interaction in OT. In the first place, we found that dominated constraints may still be active, in the sense that a constraint, even when it is violated, must be minimally violated. Secondly, we found that some output candidates are intrinsically suboptimal to others, regardless of ranking. This captures certain typological observations with respect to positional neutralization.

## 1.8  On defining segment inventories

This section will show how segment inventories result from interactions of faithfulness constraints and markedness constraints. The discussion is related to the

notion of Lexicon Optimization (section 1.6). As before, we will draw heavily or Prince and Smolensky (**1993).**

Recall that the main type of conflict in OT is that between markedness constraints and faithfulness constraints. Faithfulness constraints militate against any loss of contrast, enforcing identity between the input and the output. Markedness constraints are natural antagonists of faithfulness constraints, militating against marked structures in the output. They may produce the effect of a loss of a feature value present in the input, if that value is a 'marked' value.

### 1.8.1   *Markedness constraints on scales*

Markedness is intrinsically a relative notion. That is, which types of elements are 'marked' and which are 'unmarked' can be established only in comparison to other elements. For example, we cannot say that nasalized vowels are intrinsically marked; they are marked in relation to oral vowels only. Moreover, markedness often involves a hierarchy of segment types, each member of which is more marked than successive members of the hierarchy. For example, it has been argued in the literature on segmental markedness that *coronals* are universally less marked than *labials* (Paradis and Prunet **1991).**

To capture this relative markedness of segment types, Prince and Smolensky **(1993)** introduce the idea that markedness relations can be organized in a scalar fashion, as in the 'harmony scale' Cor > Lab. Moreover, the ranking of constraints that govern markedness relations along a single dimension (such as place of articulation) is universally fixed. In the case in hand the constraints governing the relative markedness of labials and coronals are intrinsically ranked as below:

(63)      Universal ranking for markedness constraints governing place of articulation
          *[lab] ≫ *[cor]

Presumably this ranking is universal, hence respected by every constraint hierarchy. Note that, by itself, this partial ranking makes no predictions about which segments are actually attested in a segment inventory of a particular language, and which are not. Inventories emerge from interactions of markedness constraint hierarchies (as in **63)** with faithfulness constraints, which serve to preserve input segmental contrasts in the output. Generally speaking, the higher faithfulness constraints are ranked, the larger the segment inventory will be. And vice versa, the lower faithfulness is ranked, the smaller the inventory.

### 1.8.2   *Interaction of markedness scales and faithfulness*

Now consider the faithfulness constraint militating against differences of place features in the input and output:

(64)       IDENT-IO(Place)
           The specification for place of articulation of an input segment must
           be preserved in its output correspondent.

This correspondence constraint is satisfied by (65a); but it is violated by (65b):

(65)       a. /p/ Input      b. /p/ Input
              |                  |
           [p] Output        [t]  Output

Let us now consider two grammars differing only in the ranking of IDENT-IO(Place) with respect to the markedness constraints of (63). One grammar that we will consider ranks IDENT-IO(Place) above both markedness constraints:

(66)       A grammar that is maximally faithful to place of articulation
           IDENT-IO(Place) ≫ *[lab] ≫ *[cor]

This grammar is maximally faithful to its input place of articulation, due to high-ranked IDENT-IO(Place). It is more important to be faithful to the input place of articulation of a segment than to its output markedness. The tableaux (67.i–ii), one for each input segment /p/ and /t/, illustrate this:

(67.i)

| /...p.../ | IDENT-IO(Place) | *[lab] | *[cor] |
|---|---|---|---|
| a. ☞ [...p...] |  | * |  |
| b.    [...t...] | *! |  | * |

(67.ii)

| /...t.../ | IDENT-IO(Place) | *[lab] | *[cor] |
|---|---|---|---|
| a.    [...p...] | *! | * |  |
| b. ☞ [...t...] |  |  | * |

These tableaux can be summarized as follows: whatever place of articulation is specified at the lexical level will reach the surface level. Accordingly, the segment inventory of this language will contain two places of articulation: (labial, coronal).

   Next consider a second grammar, which ranks IDENT-IO(Place) more modestly:

(68)       A grammar that is less faithful to place of articulation
           *[lab] ≫ IDENT-IO(Place) ≫ *[cor]

This grammar is less faithful (than the one in 66) to its input place features. It blocks the surfacing of *any labial in the input* since the markedness constraint

militating against this segment type outranks the faithfulness constraint IDENT IO(Place). Accordingly, any input labial surfaces as a coronal, if it surfaces at all:[16]

(69.i)

| /...p.../ | *[lab] | IDENT-IO(Place) | *[cor] |
|---|---|---|---|
| a.   [...p...] | *! | | |
| b. ☞ [...t...] | | * | * |

(69.ii)

| /...t.../ | *[lab] | IDENT-IO(Place) | *[cor] |
|---|---|---|---|
| a.   [...p...] | *! | * | |
| b. ☞ [...t...] | | | * |

This amounts to a **neutralization** (loss of contrast) of place-of-articulation fea tures, in the favour of coronals. In fact the surface level of such a language woul present no evidence whatsoever for input labials. For the learner, it would b senseless to set up a lexical contrast among labials and coronals, since such contrast would be entirely overruled by markedness effects. Prince and Smolen sky (1993) refer to the obscuring of input contrasts as 'Stampean occultation' stating it as a principle of **Lexicon Optimization** (see section **1.6**).

### 1.8.3   *Conclusions and predictions*

In sum, we have found that the resulting consonant inventories depend on th ranking of faithfulness constraints with respect to the markedness constraints With respect to place of articulation, we have found the following relation be tween ranking and inventory:

(70)        Ranking                                          Inventory
        a. IDENT-IO(Place) ≫ *[lab] ≫ *[cor]   {p, t}
        b. *[lab] ≫ IDENT-IO(Place) ≫ *[cor]   {t}

The interesting consequence of this approach is that segment inventories need n longer be stipulated at the level of lexical forms, as in derivational theory. Instead the ranking of faithfulness constraints amongst markedness constraints is par of the grammar, hence it is automatically respected at the level of the output

---

[16] Of course one might also consider the logical possibility that input labials are simply deletec – there is no way of telling the difference between both possibilities since it is impossible tc establish the presence of input labials.

Derivational theory arrives at the same result by stipulating that the output of (lexical) rules must contain no segments that are not part of the input inventory: this is referred to as the 'structure-preserving' property of phonological rules (Kiparsky 1985).

Another interesting consequence of this theory of segmental markedness is that it predicts that 'unmarked' segments will emerge wherever faithfulness constraints are put 'out of control'. Some segment types in every inventory (for example, coronals) are less marked than other segments, even though the grammar allows both more and less marked segments at the surface level, due to high-ranked faithfulness constraints. However, given the chance, the grammar will still favour unmarked segments over marked segments. This occurs in special situations in which, for some reason, input faithfulness requirements no longer hold. This prediction has been confirmed robustly for a wide range of situations in a wide range of languages.

Consider, for example, *epenthesis:* a segment appears in the output exclusively for phonotactic reasons, as in the case of a vowel inserted to break up a consonant cluster. By its very nature, the inserted vowel lacks a counterpart in the input. It cannot be subject to input faithfulness, so that its featural content is fully determined by markedness factors. The prediction is that epenthetic segments are segmentally unmarked or easily influenced by segments in their contexts. This, and other related observations, is presented in (71).

(71)     a. ***Epenthetic segments*** are less marked than ***'lexically sponsored'
            segments.***
            (Explanation: epenthetic segments have no input counterparts,
            hence they are 'free' of faithfulness constraints. See chapter **3.**)
        b. ***Segments in reduplicants*** are less marked than ***segments in their
            bases.***
            (Explanation: 'copied' segments have no input counterparts,
            hence they are 'free' of faithfulness constraints. See chapter 5.)
        c. ***Segments in affixes*** are less marked than ***segments in roots.***
            (Explanation: 'affix faithfulness' is intrinsically lower-ranked than
            'root faithfulness'. See chapters 5 and 9.)

The observations in (71a–c) will all be confirmed in later chapters, in the light of insights into faithfulness–markedness interactions.

## 1.9   Conclusion

In this chapter we have laid out the foundations on which OT is built, and pointed out the most important linguistic phenomena that fall in its scope. In the discussions throughout this chapter, one aspect of OT stood out: the interaction of

faithfulness and markedness. We have seen that all general phonological phe-
nomena discussed here are variations on this theme: the notion of contrast, an4
related to this, allophonic variation, neutralization, and lexical representations,
and finally, the notion of segment inventory. In every case, a phonological pattern
resulted from (more or less complex) interactions of constraints that preserve
lexical input properties (faithfulness), and others that reduce output markedness.
In later chapters of this book, we will maintain this perspective, and generalize it
to a range of other linguistic phenomena.

The following eight chapters of this book will each be devoted to a research topic
in which OT has left its marks. Chapter 2 addresses functional relations among
processes, comparing OT with rule-based theory from this perspective. Chapter 3
discusses syllable structure and related phenomena, such as syllabically governed
epenthesis and deletion. Chapter 4 deals with metrical phenomena, particularly
word stress and quantity effects. Chapter 5 addresses the morphology–phonology
interface, focussing on reduplication, and also extends the notion of 'correspond-
ence' beyond relations of input and output. Chapter 6 further extends correspond-
ence to relations between morphologically related output forms, covering
paradigm regularity. Chapter 7 is devoted to the issues of learnability and acqui-
sition. Chapter 8 contains applications of OT outside phonology in syntax.
Finally, chapter 9 discusses residual issues, focussing on issues that deserve further
research (in particular, opacity) as well as on current theoretical developments.

## SUGGESTIONS FOR FURTHER READING

### General introductions to phonology
Goldsmith, John (1990) *Autosegmental and metrical phonology*. Oxford: Basil
    Blackwell.
Kaye, Jonathan (1990) *Phonology: a cognitive view*. Millsdale: *LEA*.
Kenstowicz, Michael (1994a) *Phonology in generative grammar*. Oxford: Basil
    Blackwell.
Kenstowicz, Michael and Charles Kisseberth (1979) *Generative phonology:
    description and theory*. San Diego: Academic Press.
Roca, Iggy (1994) *Generative phonology*. London and New York: Routledge.

### Constraints in phonology
Hayes, Bruce (1986) Inalterability in CV Phonology. *Language* **62.** 32*1*–5*1.
McCarthy, John (1986) OCP effects: gemination and antigemination. *Linguistic
    Inquiry* **17.** 207–63.

Odden, David (1986) On the role of the Obligatory Contour Principle in phonological theory. *Language 62.* 353–83.

Paradis, Carole (1988) On constraints and repair strategies. *The Linguistic Review* **6.** 71–97.

Yip, Moira (1988) The Obligatory Contour Principle and phonological rules: a loss of identity. *Linguistic Inquiry 19.* 65–100.

**General introductions to OT**

Archangeli, Diana (1997) Optimality Theory: an introduction to linguistics in the 1990s. In D. Archangeli and D. T. Langendoen (eds.), *Optimality Theory: an introduction.* Oxford: Blackwell. 1–32.

Burzio, Luigi (1995) The rise of Optimality Theory. *Glot International* **1:6.** 3–7.

Prince, Alan and Paul Smolensky (1997) Optimality: from neural networks to universal grammar. *Science* 275. 1604–10.

Sherrard, Nicholas (1997) Questions of priorities: an introductory overview of Optimality Theory in phonology. In Roca. 43–89.

**Founding papers of OT**

McCarthy, John (1993) A case of surface constraint violation. In C. Paradis and D. LaCharite (eds.), *Constraint-based theories in multilinearphonology,* special issue of *Canadian Journal of Linguistics* **38.** 169–95.

Prince, Alan and Paul Smolensky (1993) Optimality Theory: constraint interaction in generative grammar. Ms., Rutgers University, New Brunswick, and University of Colorado, Boulder. RuCCS-TR-2. [To appear, Cambridge, Mass.: MIT Press.]

**Phonetic grounding of constraints**

Archangeli, Diana and Douglas Pulleyblank (1994) *Grounded phonology.* Cambridge, Mass.: MIT Press.

Hayes, Bruce (1996a) Phonetically driven phonology: the role of Optimality Theory and inductive grounding. Ms., UCLA. [ROA-158, http://ruccs.rutgers.edu/roa.html]

Myers, Scott (1997b) Expressing phonetic naturalness in phonology. In Roca. 125–52.

Steriade, Donca. (1995b) *Positional neutralization.* Ms., University of California, Los Angeles.

**Markedness and underspecification**

Archangeli, Diana (1988) Aspects of underspecification theory. *Phonology* **5.**183–207.

Inkelas, Sharon (1995) The consequences of optimization for underspecifica-
tion. In J. Beckman (ed.), Proceedings of NELS 25. 287–302. [ROA-40,
http://ruccs.rutgers.edu/roa.html]

Itô, Junko, R. Armin Mester, and Jaye Padgett (1995) Licensing and under-
specification in Optimality Theory. *Linguistic Inquiry* **26.** 571–613.

Steriade, Donca (1995a) Underspecification and markedness. In Goldsmith. 114–
74.

### OT in computational phonology

Eisner, Jason (1997) Efficient generation in primitive Optimality Theory. *Pro-
ceedings of the 35th Annual Meeting of the Association for Computational
Linguistics.* [ROA-206, *http://ruccs.rutgers.edu/roa.html*]

Ellison, Mark T. (1994) Phonological derivation in Optimality Theory. *Proceed-
ings of the fifteenth International Conference on Computational Linguistics,
II.* Kyoto, Japan. 1007–13. [ROA-75, http://ruccs.rutgers.edu/roa.html]

Hammond, Michael (1997) Parsing syllables: modeling OT computationally. Ms.,
University of Arizona. [ROA-222, http://ruccs.rutgers.edu/roa.html, parser
code available from author, http://www.u.arizona.edu/~hammond]

Walther, Markus (1996) OT SIMPLE – A construction-kit approach to Optimal-
ity Theory implementation. Ms., Heinrich Heine Universität, Düsseldorf.
[Software and paper can be downloaded from http://www.phil-fak.uni-
duesseldorf.de/~walther/otsimple.html. Also ROA-152, http://ruccs.rutgers.
edu/roa.html]

## EXERCISES

### 1 Japanese

Consider the following distribution of [g] and [ŋ] in Japanese (Itô and Mester
1997):

| (i) | geta | *ŋeta | 'clogs' |
|-----|------|-------|---------|
| | giri | *ŋiri | 'duty' |
| | guchi | *ŋuchi | 'complaint' |
| | go | *ŋo | '(game of) Go' |
| (ii) | kaŋi | *kagi | 'key' |
| | kaŋo | *kago | 'basket' |
| | kaŋŋae | *kaŋgae | 'thought' |
| | tokaŋe | *tokage | 'lizard' |

a. State the generalization for the distribution of [g] and [ŋ]. Categorize this distribution in terms of the typology discussed in section 1.7.
b. Account for this generalization by a set of ranked constraints.
c. Support your analysis by tableaux of **geta** and **kaŋi.**

## 2 English

Consider the following English word pairs, some of which display alternations of voice:

| | | | |
|---|---|---|---|
| cat | [kæt] | cats | [kæts] |
| dog | [dɔg] | dogs | [dɔgz] |
| hen | [hen] | hens | [henz] |
| twelve | [twelv] | twelfth | [twelfθ] |
| eight | [eɪt] | eighth | [eɪtθ] |
| ten | [ten] | tenth | [tenθ] |

a. What are the underlying forms of the suffixes in (i) and (ii)?
b. To account for these alternations, you need a new constraint. State this constraint (as generally as possible). To what extent is this constraint phonetically grounded?
c. Rank the constraints, motivating each individual ranking by at least one form. Support your analysis by tableaux of **'cats, dogs, hens, twelfth,** and **eighth.**