I
Conflicts in grammars

1.1 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 interaction 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*.[1] 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 [pʰɣ], [ɱʎx], and [Ɵɗ], 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^n) 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 ∈ {i̶,a}.[4] The complete set of potential monosyllables contains 6 items {pi̶, pa; ti̶, ta; ki̶, ka}, the set of disyllables contains 36 (or 6^2) items ({pi̶pi̶, papi̶, ki̶pi̶ . . . }), 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, i̶–a). Every consonant is maximally different from other consonants (place 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.

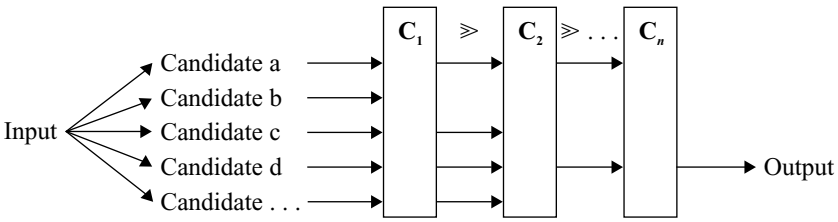
In sum, we have seen that every grammar must reconcile the inherently competing forces of faithfulness to lexical contrasts (the inertness which draws output forms back to their basic lexical shapes) and markedness (minimization of marked forms). However, as we are about to find out, Optimality Theory recognizes 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 'faithfulness'. 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 ($C_1 \gg C_2 \gg \ldots C_n$), 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 occurrences 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.