CONSTRAINT-BASED APPROACHES TO PHONOLOGY

G. N. Clements, CNRS, UA 1027, Paris

ABSTRACT

Current research in phonology has placed increasing emphasis on the importance of constraints and their interactions in phonological systems, while decreasing or eliminating the role of generative rewrite rules. The present paper offers a brief review of constraint-based approaches to phonology, considering some of their advantages over traditional models.

1. THE EMERGENCE OF CONSTRAINTS IN RECENT PHONOLOGICAL THEORY

One of the fundamental hypotheses of generative phonology since its inception in the early 1960s has been that the phonological component of a grammar consists of a set of rewrite rules that apply in sequential order to generate surface forms from underlying representations. One of its main insights has been that regular alternations in the phonological shape of morphemes could be captured by assigning each such morpheme a single underlying representation, and generating its alternants by rules which often prove to be of considerable generality. A strong constraint on rules is that they cannot access any information other than that present in the input string. Thus, in particular, they are "blind" to the effects they produce in their individual and collective output.

While this view of the organization of a phonological system is the one that continues to be presented in textbooks, it has been undermined in recent years by the increasing role played by constraints as a central feature of phonological explanation. By "constraint" I mean any statement, universal or language-particular, which has the effect of defining the set of lawful phonological representations without directly specifying a change in structure. In various guises—structure conditions, phonotactics, filters, well-formedness conditions, etc.—constraints began to appear in the literature on a sporadic basis in the 1970s, at the margins of otherwise quite orthodox analyses. Toward the beginning of the 1980s, however, some researchers began to believe that constraints play a more central explanatory role in phonology than had previously been thought. Since that time, the notion of constraint has gathered considerable momentum, and today seems in a position to replace the notion of rewrite rule altogether.

This evolution in thinking has had a variety of causes. For one thing, a similar evolution had taken place in syntactic theory, where transformational rules have come to be largely eliminated in favor of a variety of types of constraints on representations; the successful elimination of derivational, rule-based approaches in syntax has no doubt inspired linguists to explore similar approaches to phonology.

However, there are other reasons for the emergence of "constraints", having to do with the particular nature of phonological data. For one, many linguists have observed that phonological rules do not apply in a perfectly arbitrary fashion, but tend to favor certain types of outputs. For example, rules of epenthesis and deletion may apply in such a way as to produce open syllables, or clusters no longer than two consonants, depending on the language [1]. In tone languages, rules tend to assign tones to toneless syllables, and to disprefer contour tones [2]. The rules of stress systems apply in such a way as to create preferred types of stress patterns, avoiding adjacent stresses and favoring alternating stress, and placing main stresses at the extremities of words [3]. Segmental rule systems tend to avoid or eliminate adjacent identical segments [4]. The apparently goal-oriented character of such subsystems cannot be readily reconciled with the output-blind nature of rewrite rules. A further observation, which stimulated much discussion in the 1970s but no widely-agreed-upon solutions [5], was that the effect of phonological rules is often replicated by constraints holding over phoneme sequences within morphemes. For example, languages which assimilate obstruents to the voicing of a following obstruent across a morpheme boundary usually require all members of an obstructed cluster to agree in voicing within a morpheme. This duplication of the effects of morpheme structure conditions and rewrite rules is purely accidental in the standard SPE framework.

However, perhaps the single most important factor leading to the emergence of constraints has been the development of nonlinear phonology in its various forms—autosegmental, metrical, syllabic, prosodic, and so forth. What these frameworks have in common is the complexity of their representational systems compared to the simple, linear representations of standard generative phonology. Given sufficiently rich representations, many properties of surface representations that had formerly been accounted for as the effect of ordered rules can be shown to follow from purely structural features of representations. To take a simple example, the recognition of the syllable as a phonological unit allows a significant reduction in the amount of rules needed to account for alternations that are (from our current standpoint) best viewed as syllable-conditioned [6]. Perhaps most significantly, the increasing richness of representational systems imposes a new need for severely constraining the ways in which the various parts of a representation can fit together. In autosegmental phonology, for example, it has proven desirable to eliminate certain types of cross-tier association patterns (notably, those in which association lines cross) in terms of a universal Well-formedness Condition, which functions both to eliminate ill-formed underlying representations and to "police" the operation of rules so that violations are not produced in rule output [7]. In metrical phonology, it has been found that stress systems obey rather strict constraints that do not follow directly from properties of metrical representations themselves, and much work has been directed toward the goal of constraining the theory by proposing a small number of representational parameters along which only a reduced number of choices are available [8, 9]. In syllable theory, an important set of constraints on syllable types can be stated in terms of the Sonority Sequencing Generalization, originally proposed in the 19th century and rediscovered in the context of the recent renaissance of syllable theory (see [10] for a review).

Alongside general system constraints of these types, phonologists have recognized more parochial constraints, specific to certain languages, that further restrict the variety of representational structures available to a language.

The notion of constraint is not unique to current linguistic frameworks. In pre-generative theory, constraints often played an important role in phonological description in the guise of "laws of euphony", "phonetics", and other types of statements which specified what phoneme combinations could and could not occur in phonemic representations. Some such statements were framed in terms of a hierarchy of constituents in the modern sense; thus, Hockett [11] proposed that all languages contain sequences of syllables, and that syllables consist of ordered sequences of smaller constituents such as onset, peaks, and codas, etc. In his view, the specification of sequential constraints on phonemes in a language involves, in part, a specification of which phonemes may occur in which type of syllabic constituent.

What distinguishes current constraint-based frameworks from earlier work of this type is its retention of the generativist goal of accounting not only for static phoneme distributions, but also for phonologically-conditioned morpheme alternations. Thus, to take an example, we not only need to account for the fact that a language like LuGanda does not allow adjacent vowels in its morphemes and words (*ai, *iu, etc.), we also want to account for the related generalization that when two vowels abut as a result of morpheme combination, the first one is eliminated via glide formation if it is high (1a), and via deletion if it is low (1b).

1. (a) /li-ata/ / lyato ‘boat’
   /mu-iko/ / mwiiko ‘trowel’
   /ma+a-ato/ / maa+ato ‘boat’ (dim.)
   /ka+cezi/ / keezi ‘moon’ (dim.)

The resulting vowel is long. Note that the prefix vowels are retained before consonant-initial stems such as */-mpi/ ‘short’; cf. [li-mpi], [mu-mpi], [ma-mpi], and [ka-mpi]. (Also, all vowels are lengthened before NC clusters by a subsequence whose effect is not shown here; see [12] for fuller discussion.) To account for the surface forms in (1), it is not enough simply to
state the constraint against vowel sequences; we must also provide specific principles stating how violations of the constraint are lawfully resolved. This is not a straightforward matter, as we can see by considering the various ways that an anti-hiatus constraint can be resolved in principle: (i) by deleting the first vowel, (ii) by deleting the second vowel, (iii) by gliding the first high vowel, (iv) by gliding the second high vowel, (v) by assimilating one vowel to the other, (vi) by fusing the two vowels into a different one (coalescence), (vii) by ephenthesizing a consonant between them, etc. Early attempts to incorporate constraints into phonological descriptions often neglected this problem, and so failed to provide satisfactory solutions to the treatment of alternations. Many of the specific features of current constraint-based frameworks can be understood in terms of the need to resolve the problem of alternations in a principled way.

2. SOME CURRENT CONSTRAINT-BASED APPROACHES

Most current constraint-based theories maintain a double commitment to the goals of accounting for static regularities of distributional facts and for genuine phonologically-conditioned morpheme alternations. Other than this common core, they differ in other substantial ways. One can currently count nearly a score of well-defined and distinguishable constraint-based theories. Here we will briefly review three frameworks that have received particular attention: constraint-and-repair theory, declarative theory, and optimality theory. General overviews of these theories, containing illuminating comparisons among these (and other) approaches and further references, are given in [13, 14, 15]. Our discussion must necessarily be cursory, and we refer the reader to the fuller presentations available in these sources.

One useful basis of comparison is that between theories which are based on invariable constraints, and those which allow constraints to be violated. In the earliest discussions, constraints were usually considered invariable, the principal argument for this view being that the use of violable constraints would greatly weaken the predictive power of the model, particularly when used in conjunction with (violable) rewrite rules. However, some subsequent work has relaxed this condition, entertaining constraint violations on either a temporary or permanent basis.

We will first consider two constraint-based approaches in which constraints cannot be violated at the surface level. They differ in that the first allows constraint violations in the course of derivations, but not in surface representations, while the second, a non-derivative framework, allows no violations in any representations. We then consider a third (also non-derivative) approach which allows constraint violations in surface representations.

2.1. CONSTRAINT-AND-REPAIR APPROACHES

The family of constraint-and-repair approaches was one of the first developments of standard generative phonology in which constraints on representations have a well-defined (and in some versions, exclusive) role in monitoring derivations [16, 17, 18, 19, 20, 21, 22]. To see the relation between constraints and rules, it may be helpful to consider the logical structure of a standard rewrite rule making use of the format A —~ B / C _D. In rules of this type, the structural description is defined as the input string CAD, and the structural change as the output string CBD. Note that the structural description of an obligatory rule consists, in effect, of a description of a sequence which is ill-formed at the point in the derivation at which the rule applies, while the expression "A —~ B" specifies the way in which this violation is eliminated. In other words, a rewrite rule pairs an input constraint with an operation which has the effect of producing a locally well-formed output. Once we perform this disassociation, we find that the SP-type rule can be factored into what can be viewed as a local constraint and a local repair operation.

The particular insight of constraint-and-repair theories is not, then, to introduce the notions of constraint and repair as such, but to delink the connection between these notions which had been inseparably paired up to that time. Once delinked from a specific repair, a constraint can operate perversely, defining ill-formed sequences both in underlying representations and at subsequent levels, where such sequences may result from morpheme concatenation and from the operation of "output-blind" rules.

Moreover—and here is a crucial advantage—more than one way of repairing a given ill-formed sequence can be specified. Let us consider again the LuGanda forms presented in (1). In a traditional rewrite rule framework, the surface forms can be accounted for in terms of two rules, one turning an initial high vowel into a glide, and the other deleting an initial non-high vowel. In a constraint-and-repair framework, these two rules can be replaced by a single constraint prohibiting vowel sequences—let us call it the *VV constraint—and two repair operations. The constraint both accounts for the absence of vowel sequences in the underlying representation of morphemes, and serves to trigger appropriate repair operations when vowel sequences are created by morpheme concatenation. The repair operations required in this case are [V, -[high]] —~ G and [V, -[high]] —~ 0. (Compensatory lengthening of the second vowel must be assured by independent means.) As "repair strategies", these operations are kept in reserve, applying only when they are needed to eliminate constraint violations.

The *VV constraint, once extracted from conventional rule statements, can be recognized as expressing the familiar cross-linguistic dispreference for vowels standing in hiatus. We can consider it a member of the set of universal principles defining preferred or unmarked representations, one which is invoked in the grammar of LuGanda and in many other (but not all) languages. The repairs themselves can be assigned to a small pool of universal elementary operations, including linking, delinking, and deletion.

This treatment has several clear advantages over a standard rewrite-rule approach. First, it extracts a single anti-hiatus constraint from a set of rules which was forced, in the standard theory, to state it twice. Second, it accounts for underlying constraints, surface regularities and alternations by the same set of principles. Third, it reinterprets the structural description and structural change of two arbitrary rules in terms of a set of phonetically plausible universal constraints and repair operations.

There are a number of fairly obvious questions that a constraint-and-repair approach must address if it is to be internally coherent and descriptively adequate. First, it is commonly assumed that constraints fall into two types: those that have a blocking function, preventing rules from applying if their outputs would violate the constraint, and those that do not block rules, but rather trigger repairs after the rule has applied. Constraint-and-repair theory must provide a principled way of predicting which constraints are of which type, unless we are willing to allow each constraint to be annotated for this information on a case-by-case basis. Second, given that repairs are formally dissociated from constraints, it is no longer a straightforward matter to determine how a given constraint violation will be repaired. Two or more repairs may applicable to a given constraint violation, and if repairs are not extrinsically ordered, the current work assumes, then principles must be offered that will predict which of a set of competing repair operations will apply in any given situation. These questions can be subsumed under the general observation that constraint-and-repair theory, as a derivational approach, must provide a sufficient core of system-level principles to administer the rich sets of interactions predicted by its logical structure (see [23] for relevant recent discussion).

2.2. DECLARATIVE APPROACHES

A second family of constraint-based approaches is founded on the principle that phonological grammars consist exclusively of a pool of unordered constraints or well-formedness conditions which, taken together, associate each lexical entry with a well-formed surface representation. Such as system-level principles to administer the constraints are "declarative" in the sense that they specify conditions that must be satisfied by surface representations, rather than operations or procedures that must be applied to derive one from the other, as in standard generative phonology. Taken collectively, the constraints are generative in the sense that they completely specify
the surface form of each lexical entry, including phonologically conditioned morpheme alternations. Approaches of this type include Categorial Phonology, an extension of categorial grammar to the phonological level [24, 25], and Declarative Phonology, which similarly extends unification-based grammar to phonology [26, 27, 28].

Declarative approaches do not employ rules or other types of procedural statements, and do not impose extrinsic ordering on their constraints. An important consequence of these properties is that such approaches are necessarily non-derivational, in the sense that they associate full representations to lexical entries without passing through a series of derivational steps. Another consequence is that structure-changing operations are prohibited, including deletion; lexical entries must be properly or entirely contained in their surface representations (monotonicity). In contrast to constraint-and-repair approaches, the constraints of this family of theories are absolutely inviolable; this means that they must be formulated with enough precision to assure that only one of two or more potentially conflicting constraints can be satisfied by any given surface form.

To continue discussion of the LuGanda example, since constraints are inviolable in surface representations, a declarative analysis cannot allow any surface violations of the *VV constraint. But in the absence of deletion rules, how can we relate a surface form like [keeni] to its underlying representation /ka+ezi/, containing two vowels? In the case of alternating segments like the prefix vowel, declarative approaches typically underspecify, or declare as optional, any information that does not appear in all alternants. For example, since [a] does not appear in all the alternants of /ka/-, we may parenthesize it in lexical entries, as follows: /k(a)/. The parentheses indicate that the vowel is present only if it is not included by the constraint system. The *VV constraint requires the presence of a vowel in the surface form [k(a)+ezi], but correctly does not exclude this vowel in [k(a)+mpl/short/, where it is retained in the surface form [kampi]. The analysis of forms like [leyato], in which the prefix vowel is realized as a glide, proceeds in principle along similar lines, which we will not attempt to work out here. (Again, compensatory lengthening of the second vowel must be assured by independent constraints.)

It will be noted that unlike the constraint-and-repair approach, the declarative account of LuGanda is non-derivational, in the sense that the surface form is not built up, step by step, by applying a series of rules or repair operations. Rather, the constraint system defines the full set of surface alternants that corresponds to each lexical representation.

This brief discussion, though incomplete, is sufficient to show that some of the problems that potentially face constraint-and-repair approaches do not appear in declarative approaches. Since declarative approaches do not make use of rules and repairs, and do not admit constraint violations, the problem of predicting which constraints have a blocking and which a repair-triggering function, or of determining which of several applicable repairs takes precedence in a given constraint violation, simply does not arise. On the other hand, several new questions must be addressed.

For example, declarative approaches resemble the traditional morpheme-alternant models of pre-generative linguistic theory in certain potentially problematic respects. Such theories do not derive the alternants of a morpheme from a single base form, but instead state distributional rules which predict which member of the set will be selected in any given context. There are well-known analytical problems confronting such theories, which have been discussed, for instance, by Kent wig and Kissell [30, pp. 180-96], and these must be resolved if declarative approaches are to capture the same range of linguistic generalizations that traditional rule-based (and constraint-and-repair) theories have succeeded in accounting for.

A further potential problem concerns the formulation of constraints. In case a lexical entry may potentially satisfy several conflicting constraints, principles must be provided to determine which takes precedence. One solution [31] is to require that constraints be stated in sufficient detail that no two constraints will ever compete for the same form, except for the special case in which their interaction can be predicted by the Koar'sky's Elsewhere Condition [32]. However, if constraints are stated in enough detail to eliminate conflicts in a given grammar, they quickly become complex, highly language-particular, and phonetically arbitrary. A result is that constraints in declarative systems cannot in general be related to universal constraints in any straightforward way, lending themselves subject to much the same sort of objections that were earlier raised against the arbitrariness of SPE-type rewrite rules.

2.3. OPTIMALITY THEORY

The leading idea of Optimality Theory as proposed by Prince, Smolensky, and McCarthy [33, 34] is that Universal Grammar consists in part of a set of constraints on representational well-formedness which are contained in all grammars. These constraints are highly conflicting and make sharply contradictory claims about the relative well-formedness of most representations. Unlike the approaches discussed up to now, the constraints posited by Optimality Theory are typically violated in the surface forms of any language. To resolve conflicting claims, each grammar ranks the constraints in a strict dominance hierarchy, such that each constraint has absolute priority over all those it dominates in the hierarchy. It is the relative ranking of the constraints on the hierarchy that determines which candidates, among possible alternatives, are selected as actual surface representations. The preferred candidate is the one that satisfies the conflicting constraint set not absolutely, but relatively better than all others. In other words, although all candidates will typically violate some constraints, the optimal (and hence selected) candidate is the one which violates the lowest-ranked constraints.

Optimality Theory provides two general mechanisms to implement this approach. One is a principle GEN which associates each unprosodified lexical entry with a typically infinite set of fully prosodified candidate output forms. This principle is subject to a principle of containment requiring that each lexical entry is properly contained in each output candidate; additional structure may also be postulated. A second mechanism is the principle EVAL, which selects the optimal candidate from among the set created by GEN. It proceeds by assessing the constraint violations presented by each candidate, eliminating candidates on a worst-first basis until only one is left.

Like the declarative approaches with which it shares a number of assumptions, Optimality Theory is a nonderivational theory of phonological form; it posits no rules or repairs that map one form into another in a step-by-step, deterministic fashion. In distinction to declarative theories, however, the optimal candidate is selected from the candidate pool created by GEN with no further reference to the structure of the original lexical entry; that is, even though some constraints are conditional in form, the precondition of any such constraint is not defined on the (lexical) input but upon the (surface) output. It is therefore only the principle of containment which links output forms to specific inputs.

Let us see how these principles might be applied to our LuGanda example /ka+ezi/. On the basis of the unprosodified lexical representation, GEN creates a set of candidate forms, of which we consider three for purposes of illustration: one which contains the violation of the hiatus constraint, and two which eliminate it. Notice that the form that contains the violation is not necessarily eliminated from consideration; it will in fact be selected as the optimal candidate if the other forms from the candidate pool violate higher-ranked constraints. Therefore, in order to insure that EVAL selects the correct output [keeni], we must determine how LuGanda ranks the members of the universal constraint set. Let us assume for purposes of illustration that this set contains, in addition to a *VV constraint, the following additional constraints:

FILLX: every skeletal position must dominate segmental material
PARSE: every segment must be incorporated into syllable structure

FILLX has the effect of ruling out openess, viewed as introducing empty consonant positions into the CV- or mora-skeleton (such positions, if present in surface representations, are viewed as...
filled at a different level of representation, perhaps in the phonetics). PARSE requires all segments to be syllabified; it is assumed that un syllabified segments are not phonetically realized (but not deleted); it will be recalled that by the containment principle, no material can be deleted. In order to select the correct output, in which the prefix vowel is un pa rased (in violation of PARSE), the constraints *VV and FILLX must outrank PARSE. These rankings are sufficient to select the correct output (c) over its two competitors in the tableau shown in (2), showing a selection of candidate surface representations for kadzei.

(2) *VV FILLX PARSE
a. kaCZI
b. kACEZI
    c. kaCZI

Asterisks in any row indicate constraint violations, brackets indicate an unparsed segment, and C represents an un filled consonant position. Since PARSE is the lowest-ranked constraint, candidate (c) (the least bad) is elected, as shown by the arrow.

This simplifies example is intended, as before, only to give an idea of the strategic approach of optimality theory, and any actual analysis will necessarily be more complex. In the present case, we have not discussed the treatment of glide formation, or of compensatory lengthening, which may require some further enrichment of the representational system of this framework [35]. Grammars of different languages are viewed as differing not in their selection of constraints (since all members of the constraint set are present in the grammars of all languages), but in terms of the rankings they impose on them. For example, a language that ranked PARSE above FILLx and *VV would reject output (c) and select (a) or (b), depending on which of the remaining constraints is the higher-ranked.

Optimality theory has attracted much attention and is still undergoing development. We can see that it addresses most of the potential problems raised in regard to preceding frameworks, of which it provides, to a certain extent, a synthesis. Prince and Smolensky have pointed out: "Among the principles of Universal Grammar (i.e., the universal constraint

- GNC) are cognates of those formerly thought to be no more than loose typological and markedness generalizations. Formally sharpened, these principles now provide the very material from which grammars are built" ([33], 219]. While the incorporation of substantive constraints into the morph of the grammar is a desirable goal, it is apparent that the set of universal constraints required to account for the full range of phonological diversity will prove to be quite large, and will necessarily contain a sizable number of arbitrary constraints having limited cross-linguistic generality. And it is difficult to see in what sense a proposed constraint such as *P/a (["a" does not form a syllable peak]), essential to the Prince and Smolensky system, can be regarded as universal, in view of the fact that [a] is an optimal syllable peak in all known languages. It can be expected that such questions, and others, will be addressed as research proceeds.

3. CONCLUSION

It is now apparent that not only can constraint-based systems of phonology account for many (or most of) the phenomena that theories based on rewrite rules could account for, they can do so in many respects in a much more principled way. This fact by itself justifies the considerable attention being devoted to constraint-based phonologies at the present time. On the other hand, the diversity of current ideas suggests that theoretical models are still in evolution. For this very reason, however, added to its preliminary achievements, this direction of research must be continued, and prosperous, in the near future.