9 Residual issues

9.1 Introduction
The previous eight chapters of this book have offered an overview of Optimality Theory, with an emphasis on areas in which the theory has proved successful. This chapter will add an overview of issues that have not yet been successfully resolved, and developments and modifications in the theory that are currently taking place. The issue that is perhaps most urgently in need of a solution is opacity, to be discussed in section 9.2. Remaining sections each deal with a major theme in ongoing research. First, absolute ungrammaticality will be addressed in section 9.3. Section 9.4 will evaluate strategies to deal with optionality and free variation. Next, section 9.5 will discuss positional faithfulness in relation to functional approaches to phonology. Finally, section 9.6 will re-evaluate the role of underlying forms in OT, linking this issue with phonologically driven allomorphy. Conclusions and a perspective on future developments in OT will be presented in section 9.7.

9.2 Opacity

9.2.1 Introduction
Opacity refers to the phenomenon that output forms are shaped by generalizations that are not surface-true. Opaque generalizations lurk at a level deeper than the output, which becomes apparent by ‘peeling off’ effects overlaid by other surface-true generalizations. Opacity is predicted by any theory allowing non-surface levels of description (the input, or any level mediating between input and output). However, opacity presents a potential problem for surface-oriented OT, a theory disallowing reference to preoutput levels by well-formedness constraints. (Of course, the fact that a generalization is not surface-true is, by itself, not immediately problematic for OT. Output forms are always in violation of some constraints.) Before we look into strategies for dealing with opacity in OT, we will first discuss its varieties. That is, opaque generalizations are either ‘non-surface-apparent’ or ‘non-surface-true’, following McCarthy’s (1998) terminology.

We call a generalization non-surface-apparent if it takes effect at a level concealed at the surface. A set of forms undergo a process, although they fail to match its structural description at the surface. This kind of interaction occurs in Turkish (Zimmer and Abbott 1978, Kenstowicz and Kisseberth 1979, Sezer 1981, Inkelas and Orgun 1995).

Opacity of vowel epenthesis in Turkish

a. Vowel epenthesis
/bas-mi/ bas-mi 'my head'
ljel-mi ljel-mi 'my wind'
b. Velar deletion
/a.ja.4/ a.ja.4 'his foot'
/i.ne.i/ i.ne.i 'his cow'
c. Interaction
/lajak-mi/ lajak-mi 'my foot'
/i.ne.im/ i.ne.im 'my cow'

Vowel epenthesis is motivated by forms such as [bas-mi], where a consonant cluster at the end of a word is broken up by a vowel (whose quality harmonizes with the stem vowel by a process that is irrelevant here). Next, velar deletion is apparent from [a.ja.4] where /kl/ is deleted in intervocalic position. Finally, consider the interaction of both processes in [lajak-mi]. While velar deletion applies transparently (the output contains both vowels that trigger it), epenthesis does not, since only one of the triggering consonants, namely /mi/, surfaces. The output is opaque with respect to epenthesis, because its context of application is not recoverable at the surface level. The effect is an overapplication of epenthesis. (Compare similar overapplications in reduplication and stem-based morphology—chapters 5 and 6.)

This is not an isolated example: many similar cases have been reported (for a wide variety of other languages), most of which involve well-motivated processes whose application is otherwise completely general. Our second example is from Tunica (Haas 1940, Kenstowicz and Kisseberth 1979).

Opacity of vowel harmony in Tunica

a. Vowel harmony
/pò-?aki/ pò-?aki 'she looks'
/pi-?aki/ pi-?aki 'she emerges'
b. Syncope
/hipu-?uhki/ hipu-?uhki 'he dances'
/njì-?uhki/ njì-?uhki 'he leads (someone)'

(1)

(2)
Opacity has received much attention in rule-based serial phonology since it offers a major argument for this theory. Therefore let us first take a look at opacity from a serial viewpoint. For a rule to apply, all that matters is whether its structural context is satisfied at the point of the derivation at which the rule applies. For this reason serial theory easily captures generalizations that are not surface-true or not surface-apparent, thus predicting *opacity* on principled grounds. The two types of rule interactions that produce opacity are known as 'counterbleeding' and 'counterfeeding', respectively (Kiparsky 1973).

*Counterbleeding* arises when a rule's structural context is potentially destroyed or removed by the application of a prior rule, but the ordering is such that both rules apply. The second rule, which might have destroyed the context of application for the first rule, applies 'too late' to actually do so. In the Tunica example, harmony first copies rounding and backness of the rightmost stem vowel onto the suffix vowel. This triggering vowel is then deleted by syncope:  

(4) **Counterbleeding** (Tunica and Turkish)  
\[
\begin{align*}
\text{a.} & \quad \text{hipu-} & \text{Vowel harmony} & \quad \text{ajak-m} & \text{Vowel epenthesis} \\
\text{hipu} & \text{?aki} & \text{hipu-} & \text{?aki} & \text{hipu-} & \text{?aki} \\
\text{ajak} & \text{kim} & \text{ajak} & \text{kim} & \text{ajak} & \text{kim} \\
\hline
\text{b.} & \quad \text{Intervocalic} & \text{k-deletion} & \quad \text{Intervocalic} & \text{k-deletion} \\
\text{hipu} & \text{?aki} & \text{hipu} & \text{?aki} & \text{hipu} & \text{?aki} \\
\text{ajak} & \text{im} & \text{ajak} & \text{im} & \text{ajak} & \text{im} \\
\end{align*}
\]

In both examples, a rule (Syncope, Intervocalic k-deletion) deletes the trigger of another rule (Vowel harmony, Vowel epenthesis), but both deletions apply **too late**, missing their chance to 'bleed' the earlier rule. Therefore, in both examples, both rules apply.

*Counterfeeding* represents the logically opposite situation. The structural context of a rule is potentially satisfied due to the application of a prior rule, but the **ordering** is such that only one rule applies. The second rule, which might have created the context of application for the first rule, applies 'too early' to actually feed it. Such a counterfeeding order produces opacity of approximant devoicing in Isthmus Nahuat:  

(5) **Counterfeeding** (Isthmus Nahuat)  
\[
\begin{align*}
\text{Fikakil}\text{-fikakil} & \quad \text{put it in it'} \\
\text{Fikakil} & \quad \text{Apocope} \\
\end{align*}
\]

The effect is an 'underapplication' of approximant devoicing, in terms of its output form.

It should now be clear that what distinguishes opaque cases from 'exceptions' is their systematicity. Opaque generalizations become transparent once we take into account a level 'preceding' the output. Usually this level is the input, or underlying form. Taking into account the input level, vowel epenthesis in Turkish vowel harmony in Tunica, and approximant devoicing in Isthmus Nahuat are perfectly transparent. Incidentally, however, opacity involves a level mediating between the input and the output, as we will see later.

---

1. Glottal stop /ʔ/ is transparent to harmony.
2. Syllable-final approximants may also be devoiced, under conditions that are irrelevant here.

---
Residual issues

Apocope misses its chance to establish the context for devoicing, which does not need for it? Note that the case for [a.jam] is robust enough to be independent of constraint ranking. A comparison of the violation patterns of both candidates shows that violation marks incurred by (7d) form a proper subset of those incurred by (7c). This finding destroys all hope that the problem is solvable without the help of additional constraints (or even new theoretical machinery).

The 'underapplication' type of opacity, as occurs in Isthmus Nahuat, is equally problematic. Each of the interacting processes, apocope (3a) and devoicing (3b), involves domination of a well-formedness constraint over a faithfulness constraint:

(8)

a. Apocope: FINAL-C ('Stem ends in C') \(\gg\) MAX-I0
b. Devoicing: *VOICED-Coda \(\gg\) *IDENT-IO(voice)

But no matter how we integrate the subrankings into a total ranking, the opaque candidate (9b) will never become optimal. Instead the transparent candidate (9c) is predicted:

(9)

<table>
<thead>
<tr>
<th>Input: /jikakili/</th>
<th>FINAL-C</th>
<th>MAX-I0</th>
<th>*VOICED-Coda</th>
<th>IDENT-IO (voice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. j̄ikakili</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. j̄ikakil</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>c. j̄akakil</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. (\otimes) a.jam</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Opaque (9b) and transparent (9c) only differ in their violation marks for *VOICED-Coda and IDENT-IO(voice). However, the mutual ranking of these constraints is forced, since it is independently motivated by approximant devoicing in words such as [tajo:] 'shelled corn' (3b). The incorrect prediction crucially depends on constraint ranking (rather than being independent of it, as in overapplication). This is a clear ranking paradox.

Opacity appears to be a direct empirical refutation of the surface-based evaluation of well-formedness constraints in OT. Since opacity is OT's Achilles's heel, researchers have attempted to find solutions for it which maximally preserve the theory's advantages. This section will present an overview of these attempts, reflecting on their relative merits. OT approaches to opacity can be put under various headings, all discussed below:

- Two-level well-formedness (section 9.2.2)
- Intermediate levels (section 9.2.3)
- 00-correspondence (section 9.2.4)
9.2.2 Two-level well-formedness

One approach to opacity allows reference to the input by well-formedness constraints, in a more or less direct way. By giving up the assumption that well-formedness constraints have access to output forms only, this approach preserves direct mapping of input to output, avoiding intermediate levels. Simultaneous reference to input and output by well-formedness constraints originates in pre-OT work by Koskenniemi (1983) and Karttunen (1993). We will refer to this idea as two-level well-formedness. Note that Correspondence Theory is a 'two-level model' in a more restricted way, since only faithfulness constraints (and not well-formedness constraints) can refer to the input and output simultaneously.

A conceptually related OT model is the Parse/Fill theory of Prince and Smolensky (1993), also known as Containment Theory, briefly discussed in chapter 3 as a precursor of Correspondence Theory. This model is named after its key assumption that no element may be literally removed from the input, which is thus contained in all candidate outputs. Deletion of an element is represented by leaving it prosodically unparsed, resulting in the lack of phonetic interpretation. Since deleted input elements are contained in the output representation, they can be phonologically active. In Correspondence Theory, reference to the input by well-formedness constraints has been implemented by Cole and Kisseberth (1995), McCarthy (1995b), Orgun (1995), and Archangeli and Suzuki (1997).

To illustrate the basic idea, let us consider the Tunica example. Opacity of vowel harmony is accounted for if harmony is triggered by input vowels, but takes effect in the output. This can be achieved by a well-formedness constraint that refers to the input and output simultaneously, for example:

\[ \text{HARMONY-IO} \]

If \( V_1 \ldots V_2 \) is a sequence of vowels in the input, and \( V'_2 \) is the correspondent of \( V_2 \) in the output, then \( V_1 \) and \( V'_2 \) agree in backness and rounding.

9.2 Opacity

HARMONY-IO is a two-level well-formedness constraint, stating a requirement of featural agreement between segments at different levels of representation, input and output. It can be read as follows:

\[ \text{If } V_1 \ldots V_2 \text{ is a sequence of vowels in the input, and } V'_2 \text{ is the correspondent of } V_2 \text{ in the output, then } V_1 \text{ and } V'_2 \text{ agree in backness and rounding.} \]

Featural harmony between \( V_1 \) and \( V'_2 \) can be said to be transmitted by the correspondence relation between \( V_2 \) and \( V'_2 \). Note that its dependence on correspondence does not give HARMONY-IO the status of an 10-faithfulness constraint. The output vowel is required to resemble another vowel's input, rather than its own. This is a well-formedness target, not a faithfulness target.

Two-level well-formedness constraints have been extended to express restrictions on the cooccurrence of different features. This is exemplified by a constraint proposed by Archangeli and Suzuki (1997):

\[ \text{LOWERING-IO} \]

Any output correspondent of an input long vowel must be [–high].

LOWERING has the format of a correspondence constraint, but with one crucial difference: correspondents in the input and output are not required to agree with respect to the same feature (as in IDENT-IO[F]). Like (12), LOWERING-IO is not an 10-faithfulness constraint, since feature-cooccurrence is required for different features (here, length and height), and for unrelated values. In Archangeli and Suzuki's analysis of Yokuts, LOWERING triggers a featural change from the input value of [high]. Therefore, it functions as a well-formedness constraint, although it has the 10-format of standard correspondence constraints. A theory allowing for constraints such as (10) and (12) essentially gives up the distinction between well-formedness constraints and faithfulness constraints.

Let us now evaluate two-level well-formedness models. The first question to be asked is: 'Do such models account for opacity?', and the second question: 'If so, at what theoretical cost?' We will spot some serious empirical and theoretical problems.

Empirically, two-level well-formedness models fail to account for types of opacity that are not controlled by the input, nor by the output. An opaque generalization may hold at an intermediate level, for example with respect to prosodic...
structure. (So far we have not seen such cases.) For example, an opaque generalization referring to prosody cannot be analysed by reference to the input, assuming that prosody is not present in input forms.

A well-studied case is compensatory lengthening (CL), a situation where a vowel is lengthened 'to compensate for' the quantity of a deleted consonant (Hayes 1989). An example comes from Oromo (Lloret 1988, Sprouse 1997):

(13)

<table>
<thead>
<tr>
<th></th>
<th>/fed-/-</th>
<th>feena 'we wish'</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>/fed-na/</td>
<td>feena 'we wish'</td>
</tr>
<tr>
<td>b.</td>
<td>/fed-a/</td>
<td>feda 'I wish'</td>
</tr>
<tr>
<td></td>
<td>/fed-adfa/</td>
<td>fedadafa 'wish for self'</td>
</tr>
</tbody>
</table>

A serial derivation shows the origin of the long stem vowel in [feena] 'we wish'. A mora is assigned to the coda consonant closing the first syllable (by Weight-by-Position, chapter 4). This consonant is then deleted before a consonant-initial suffix, but its mora is preserved in the form of an extra element of length on the preceding vowel. Opacity is explained by a counterbleeding rule order (Weight-by-Position + Pre-consonantal deletion):

(14)

```
  σ σ σ σ σ
  /fed-na/ → fed na → fe na → fe na
```

As Sprouse (1997) points out, the mora that triggers the lengthening cannot be part of the input. If faithfulness to input moras were to trigger CL, and the consonant /d/ be specified as moraic in the input, then CL should occur in any context in which /d/ cannot hold on to its mora. This is incorrect; see *[feeda] (13b), where /d/ surfaces as an onset. Instead CL is restricted to the context of Weight-by-Position. Moreover, if input consonants were contrastively specified as moraic (so that weight of CVC syllables would lexically vary), then certain stems would undergo CL, while other stems would simply undergo deletion. Such contrasts are apparently not found in Oromo. Finally, as McCarthy (1998) points out input specification of moras runs into a general problem: it is at odds with Richness of the Base, a cornerstone of OT. In sum, the length-triggering mora cannot be part of the input, ruling out a two-level analysis of CL. The best a two-level analysis can achieve is to encode the context of Weight-by-Position into a two-level constraint:

(15)  

\[ \text{Project an output } \mu \text{ for every input C followed by a C.} \]

If course, this is a bad constraint. Sprouse (1997: 2) notes 'it is more of a rule, combining a well-formedness constraint (how syllables should be organized) with its repair (project a mora) [...] even when the well-formedness constraint can't be evaluated directly since the mora and segment that projects it are never present at the same level'. Finally, \( \mu \)-PROJECTION loses the generalization that a mora is projected for every coda consonant. It stipulates the opaque pattern, rather than explaining it.

The theoretical status of both earlier two-level constraints (10) and (12) is equally problematic. Both function as rules, combining a structural condition (the input structure) and a repair. A theory allowing for two-level well-formedness constraints may stipulate any type of relation between the input and output, being equivalent in this respect to rule-based theory (Lakoff 1993). This power undermines standard OT's solutions to problems inherent to rule-based serialism, in particular conspiracies and the Duplication Problem. Recall from chapter 2 that standard OT improves over rule-based theory by separating structural conditions (triggers) and structural changes (repairs), thus predicting a range of repair strategies in response to a single well-formedness requirement. However, reference to the input by well-formedness constraints implies a crucial deviation from standard OT. Triggers and repairs become once more entangled, hence the explanation of conspiracies and the Duplication Problem is fatally undermined. As Sprouse (1997) observes, blurring the borderline between well-formedness constraints and faithfulness constraints produces negative consequences for factorial typology as well.

Finally, we may ask to what extent two-level well-formedness constraints can still be considered to be grounded. Each of the constraints \( \text{HARMONY-I0} \) and \( \text{LOWERING-I0} \), interpreted as a single-level output constraint, makes a phonetically natural requirement. (Harmony reduces the number of articulatory gestures, and lowering supports the quantity contrast by matching it with the longer intrinsic duration of non-high vowels.) However, phonetic grounding does not carry over to the two-level versions of the constraints since the goals are not manifest in the output. This mismatch is, of course, inherent to opacity. But explanation of opacity will be lost if it is 'hard-wired' into the constraints, rather than reduced to interactions of grounded constraints.

### 9.2.3 Intermediate levels

Yet another approach to opacity is based on intermediate levels between input and output. The idea behind this move should be clear: the success of serial theory in accounting for opacity was based precisely on pre-surface levels. Although this gives up the standard OT assumption of direct mapping, there is also a potential pay off: preserving the standard constraint format. This offers a reverse picture as
compared to two-level well-formedness discussed in section 9.2.2, which gave up constraint format while preserving direct mapping.

Let us assume an OT grammar which is internally organized into serially ordered strata (Stratum 1–\( n \), with \( n \geq 2 \)), each containing both functions Gen and Eval. The input to each stratum is defined by the output of the previous one, except that the input to Stratum 1 is identical to the lexical representation. Within each stratum, the input–output mapping is direct. Each stratum contains its own constraint ranking:

\[
\text{(16) Multi-stratal evaluation}
\]

\[
\begin{align*}
\text{Input} & \downarrow \\
\text{Stratum 1: Gen, Eval,} & \quad \text{} \\
\text{...} & \quad \text{} \\
\text{Stratum } n & \quad \text{Gen, Eval,} \\
\text{output} & \quad \text{}
\end{align*}
\]

Multi-stratal models have their origins in derivational Lexical Phonology (Kiparsky 1982b, Mohanan 1982, Booij and Rubach 1987), where they accounted for a range of phenomena, such as word domain effects, affix ordering, structure preservation, and cyclicity. With the emergence of constraint-based theories, direct mapping became the default assumption. Nevertheless, constraint-based multi-stratal models were proposed, by Goldsmith (1993a, Harmonic Phonology) and in OT by, among others, McCarthy and Prince (1993b), Inkelas and Orgun (1995), and Sprouse (1997).

In multi-stratal OT the generation of output forms essentially follows a derivation, as in serial theory. This is illustrated below for Turkish, a case of ‘overapplying’ opacity, involving counterfeeding in serial theory. This grammar contains two strata. Stratum 1 takes the lexical representation /ajak-m/ as its input, and produces an intermediate output [a.ja.kim], where epenthesis has applied but intervocalic /k/ is preserved.

\[
\text{(17) ‘Overapplication’ in multi-stratal evaluation (Turkish)}
\]

\[
\text{Stratum 1: I, } /\text{ajak-m/} \\
\downarrow \\
\text{*Complex, Max-IO } \gg \text{ Dep-IO, *Vkv} \\
O_1 [\text{a.ja.kim}]
\]

Stratum 2: I, /a.ja.kim/  
\*Complex, *Vkv \gg \text{Dep-IO, Max-10}  
O_2 [a.ja.im]

The output of \( S_1 \), [a.ja.kim], is due to the ranking *Complex, Max-10 \gg \text{Dep-IO, *Vkv}:

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Input: } /\text{ajak-m/} & \text{*Complex} & \text{Max-IO} & \text{Dep-IO} & \text{*Vkv} \\
\hline
\text{a. a.jakm} & *! & & & \text{*} \\
\text{b. } & & & & \text{*} \\
\text{c. a.ja.im} & *! & & \text{*} & \text{*} \\
\text{d. a.jam} & & & & \text{*!} \\
\hline
\end{array}
\]

This \( S_1 \) output is then fed as an input into \( S_2 \). Ranking at \( S_2 \) is such that intervocalic /k/ is deleted (*Vkv \gg \text{Max-10}), giving an \( S_2 \) output [a.ja.im], which is the surface form.

\[
\text{(19) ‘Underapplication’ in multi-stratal evaluation (Isthmus Nahuat)}
\]

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Input: } /\text{ajak-m/} & \text{*Complex} & \text{*Vkv} & \text{Dep-IO} & \text{Max-IO} \\
\hline
\text{a. a.jakm} & *! & & \text{**!} & \text{*} \\
\text{b. a.ja.kim} & *! & & \text{*} & \text{*} \\
\text{c. } & & & & \text{*} \\
\text{d. a.jam} & & & & \text{**!} \\
\hline
\end{array}
\]

In sum, both strata have minimally different rankings, involving only reranking of a well-formedness constraint and a faithfulness constraint.'

Let us now briefly discuss a case of ‘underapplication’, which is captured in serial theory by counterfeeding. Approximant devoicing in Isthmus Nahuat is opaque, as shown earlier in this chapter. Here intermediate outputs are [jikakili] and [tajo:!,] respectively:

\[
\text{(20) ‘Underapplication’ in multi-stratal evaluation (Isthmus Nahuat)}
\]

\[
\begin{align*}
\text{Stratum 1: I, } & /\text{jikakili/} /\text{tajo:!/} \quad \text{Max-10 } \gg \text{ Final-C} \\
\downarrow & \downarrow \\
O_1 [\text{Jikakili}] & \text{[tajo:!]}
\end{align*}
\]

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Input: } /\text{jikakili/} /\text{tajo:!/} & \text{*Voiced-Coda } \gg \text{ Ident-IO(voice)} \\
\hline
\text{O_1 [Jikakili]} & \text{[tajo:!]} \\
\hline
\end{array}
\]

It is difficult to establish the ranking of Dep-IO at \( S_2 \). We may assume it to be undominated. It marks no violation for ‘epenthetic’ [i], because this vowel occurs in the input of Stratum 2.
Consequently each form undergoes only one process, either devoicing or apocope, but not both. (Note that values for [voice] as these occur in the output of Stratum 1 are preserved at Stratum 2, regardless of their values in the ‘original’ lexical input.)

In contrast to two-level well-formedness theory, an ‘intermediate-levels’ theory of opacity holds the promise of accounting for types of opacity that are not controlled by the input. Compensatory Lengthening in Oromo (13), for example, may be accounted for by an intermediate level at which Weight-by-Position has applied, but consonant deletion is still blocked. CL becomes an *identity effect* preserving quantity occurring at intermediate levels. In sum, multi-stratal evaluation has the advantage of a broad empirical coverage of opacity effects, being equivalent to serial theory in this respect.

But does it preserve OT’s advantages over serial theories, which are largely based on direct mapping and evaluation of output forms? If OT is not to be reduced to a notational variant of serial theory, one major question should be answered: ‘Can we have the best of both worlds?’ That is, can intermediate levels be restricted in a way that captures opacity, but also preserves OT’s advantages? Can the number of strata be restricted to a universal maximum? At present, it is not clear what the ultimate answers to these questions will be, due to a lack of experience with multi-stratal models. We foresee two problems, however.

Firstly, the problem of finding independent motivation for strata. Motivation must be independent in the sense that it goes beyond the desire to capture opaque interactions. If such (presumably morphological) evidence cannot be produced, then multi-stratal OT is reduced to a variant of serial theory, differing in having two mechanisms (constraint interaction and derivation) instead of one (derivation). Considerations of generality of explanation would then favour the serial model.

Secondly, multi-stratal evaluation implies that multiple rankings are active within a grammar. This predicts that stratal rankings (hypothetically, at least) can differ as widely as those of different languages. But cases under discussion show that such radically different rankings for strata do not occur. On the contrary, rankings differ only in minimal ways, typically by the reranking of a pair of constraints (a well-formedness constraint and a faithfulness constraint), or perhaps two pairs, but not by massive rerankings.

Thirdly, multi-stratal models are not learnable by Tesar and Smolensky’s algorithm, discussed in chapter 8, and it is doubtful whether they are learnable at all, due to the large increase in complexity. However, future research will have to answer this question.
9.2.4 00-correspondence
A third approach to opacity is based on 00-correspondence (McCarthy 1995a; Benua 1995, chapter 6). Opacity involves underapplication and overapplication effects, as noted in section 9.2.1. Hence, it seems logical to extend the apparatus of 00-identity constraints to opacity. If successful, this approach would have the advantage of preserving two core principles of standard OT: direct mapping and strictly output-based well-formedness constraints.

An identity-based approach to opacity requires the following pair of conditions. First, for each opaque output, there must be an output form compositionally related to it. Second, this base must be transparent with respect to the generalization. Both conditions hold for the analysis of i-Syncope in Palestinian Arabic (chapter 6, section 6.4).

i-Syncope normally deletes unstressed /i/ in open syllables (25b). It under-applies in (25c), where unstressed /i/ corresponds with a stressed vowel in the base.

(25) Regular application and underapplication of i-Syncope in Palestinian Arabic
   a. *fihiml fihim 'he understood'
   b. /fihim-na/ fhimna 'we understood'
   c. /[fihim]-na/ fhimna *fhimna 'he understood us'

Two-level well-formedness fails here since the abstract stress blocking application of i-Syncope (rendering it opaque in 25c) is not present in the input.

However, an 00-correspondence analysis works, because the 'blocking' stress is recoverable from the base. A BA-identity constraint HEADMAX-BA refers to it:

(26) HEADMAX-BA
   Every segment in the base's prosodic head has a correspondent in the affixed form.

00-correspondence reanalyses any cases of opacity accounted for by cyclic application in serial theory. Yet it fails wherever opacity is controlled by an abstract intermediate level, one that is not recoverable from a morphologically related output form.

For example, vowel epenthesis in Turkish (/ajak-m/ → [a.ja.im], section 9.2.1) cannot be analysed by 00-correspondence. Some form should exist (in the paradigm of /ajak/) that has a 'transparent' epenthetic [i] following the stem. But no such output form can exist, given the phonology of Turkish. If epenthesis applied transparently in any form related to /ajak/, then /k/ would have to resist deletion in an intervocalic context. However, this is precisely what *V_kV ≫ Max-10 excludes. We must conclude that 00-correspondence is insufficiently general to account for all types of opacity.

9.2.5 Sympathy
The most recent attack on opacity is due to McCarthy (1998), and it is called Sympathy. Its core feature is an extension of the correspondence relation to pairs of candidate forms. Faithfulness constraints require the output form to resemble another candidate of Gen, the sympathetic form, which is analogous to the abstract intermediate representation in serial theory. Sympathy preserves the key features of standard OT: direct mapping of inputs to outputs and the output-based format of well-formedness constraints. However, it implies a vast increase in the correspondence relations involved in selecting the optimal output.

To acquire an intuitive notion of 'Sympathy', we return to tableau (7) of Turkish lajak-mi (repeated below as 27). Recall that the transparent candidate (27d) is incorrectly selected as optimal over opaque (27c). This result is independent of constraint ranking, as violation marks incurred by the former are a proper subset of those incurred by the latter:

(27) Tableau (7) of Turkish lajak-mi

<table>
<thead>
<tr>
<th>Input: /ajak-m/</th>
<th>*COMPLEX</th>
<th>*VkV</th>
<th>Max-10</th>
<th>DEP-IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. a.jakm</td>
<td></td>
<td>*!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. a.ja.km</td>
<td>*!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. a.ja.im</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. a.jam</td>
<td></td>
<td></td>
<td></td>
<td>!</td>
</tr>
</tbody>
</table>

For [a.ja.im] to become more harmonic than *[a.jam], some constraint C must outrank DEP-IO, one which [a.ja.im] obeys but *[a.jam] violates. The unknown constraint is probably not a well-formedness constraint, since [a.ja.im] can boast no advantages over *[a.jam] in terms of its featural or syllabic composition. (Its extra [i] only works against it in terms of well-formedness, causing hiatus, and an onset-less syllable.) Therefore constraint C must be a faithfulness constraint.

To what form can [a.ja.im] be faithful which [a.jam] is unfaithful to? It cannot be the input /ajak-m/ for obvious reasons 00-faithfulness fails, which is precisely the problem). Nor is any transparent output form available for 00-correspondence (as we saw in section 9.2.4). McCarthy opts for a third logical possibility: the opaque candidate output is faithful to a form that is neither an input nor an output, but a candidate output itself. That is, he proposes a new type of correspondence relation defined on pairs of output candidates: this relation is called 'Sympathy'.

386
Residual issues

Let us first consider an intuitive (and informal) way of picking the 'sympathetic' candidate to which [a.ja.im] is faithful. Focussing on the candidate pair [a.ja.im] and [a.ja.kim], we find that whatever advantage of faithfulness the former may have, this must reside in its epenthetic [i]. We deduce that the yet-to-be-identified sympathetic candidate shares this vowel with [a.ja.im]. Such a candidate form is [a.ja.kim] (27b). This happens to be the form in tableau (27) that is otherwise maximally faithful to the input – this observation will be taken up directly below. This sympathetic form will be indicated by the flower ‘Ω’.

The @-candidate [a.ja.kim] matches the intermediate form in the serial analysis (4), arising after vowel epenthesis (the input to intervocalic k-deletion). In the OT analysis, [a.ja.kim] is more faithful to the input lajak-mi than opaque [a.ja.im], taking an intermediate position between both. Thus the serial analysis and OT analysis share an insight: both set up an abstract intermediate form 'connecting' the opaque output to the input.9

The next step is setting up a correspondence constraint demanding faithfulness to the @-candidate [a.ja.kim]. We need a constraint militating against 'deletion' of segments, hence a member of the MAXIMALITY family. This constraint, MAX-@O, is ranked above DEP-IO, the anti-epenthesis constraint, violation of which would have been fatal to (28c) (as indicated by ‘i’) if MAX-@O had not been there:

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Input: /ajak-m/} & \text{COMPLEX} & \text{*VkV} & \text{MAX-@O} \\
\hline
\text{@-Candidate: a.ja.kim} & \text{!*} & \text{!*} & \text{!*} \\
\hline
\end{array}
\]

The opaque candidate (28c), which would 'normally' lose to the transparent candidate (28d) because of its extra violation of DEP-IO, now becomes optimal due to its maximal faithfulness to the @-candidate. Note that one violation of MAX-@O is incurred by each segment of the @-candidate that lacks a correspondent in the output. Therefore, 'opaque' (28c) is unfaithful to the @-candidate by one segment ('deleted' /k/) whereas transparent (28d) is unfaithful by two segments ('deleted' /k/ and /i/), with the additional violation being fatal. Finally, the @-candidate itself fatally violates *VkV.10

Next, we turn to a case of 'overapplying' opacity that is not controlled by the input (which is problematic to reference-to-the-input models). Again, the @-candidate matches the intermediate representation in a serial analysis, which arises after Weight-by-Position and is the input to consonant deletion. This is [fe,d,na] (29b). The opaque output (29c) is faithful to it by satisfying MAX-μ-@O, a constraint demanding that each mora of the @-candidate have a correspondent in the output:

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Input: /fe,d-na/} & \text{V-Candidate: fe,d-na} & \text{WEIGHT-BY-} & \text{CODA} & \text{MAX-μ} & \text{DEP-μ} \\
\hline
\text{@-Candidate: fe,d-na} & \text{!*} & \text{*} & \text{*} & \text{!*} & \text{!*} \\
\hline
\end{array}
\]

Finally, consider the case of 'underapplying' opacity in Isthmus Nahuat. In a serial (counterfeeding) derivation (5) of [jikakili] (from /fikakili/), an intermediate representation occurs in which devoicing applies, but apocope has not yet applied. Here /fikakili/ fails to undergo devoicing since its final vowel is still present. Analogously, the @-candidate tableau (30) happens to be identical to the input:

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Input: /fikakili/} & \text{FINAL-C} & \text{IDENT-@O} & \text{MAX-μ} & \text{DEP-μ} \\
\hline
\text{@-Candidate: fikakili} & \text{!*} & \text{!*} & \text{!*} & \text{!*} \\
\hline
\end{array}
\]

9 McCarthy points out that such parallelism is not always exact, particularly in cases involving multiply opaque generalizations.

10 Another plausible output candidate, [aj.kim], is equally faithful to the sympathetic candidate. The difference is that an input vowel is deleted in [aj.kim] but an input consonant in [a.ja.im] (the actual output). We assume that this contrast correlates with a distinction between MAX-C and (higher-ranking)MAX-V. (See the analysis of Southeastern Tepehuan in chapter 4.) This assumption is shared by McCarthy (1998), who does not discuss the Turkish case, though.
Candidate (30a), perfectly faithful to the input and to itself (in its quality of @-candidate), is nevertheless rejected for its fatal violation of FINAŁ-@O. Opaque (30b) is optimal due to IDENT-®O(voice) ≫ "VOICED-CODA, ruling out transparent (30c).

Observe that this tableau contains two faithfulness constraints of the format IDENT(voice), a low-ranking one evaluating I-O pairs (IDENT-I0(voice)), and a top-ranking one evaluating ®-O pairs (IDENT-®O(voice)). Since the @-candidate happens to be identical to the input, both faithfulness constraints produce the same evaluation marks, at different positions in the hierarchy. We will return to this observation shortly below, in connection with the issue of how the choice of the @-candidate is made.

The @-candidate is not a priori given, and it should be determinable on the basis of positive evidence. How can this aim be achieved? In all cases of opacity discussed thus far, the @-candidate satisfies an 10-faithfulness constraint that is violated in the opaque form. In all three cases, this is MAX-TO:

\[
\begin{align*}
\text{Input: /ajak-m/} & & \text{Complex} & \text{Anchoring-I0} & \text{*VkV} & \text{MAX-I0} & \text{Dep-I0} \\
a. \text{Turkish} & a.ja.kim & a.ja.im & /k/ \rightarrow O & \& & \checkmark & & \\
b. \text{Oromo} & fə,ɛj na & fɛ,ɛj na & /d/ \rightarrow O & \& & \checkmark & & \\
c. \text{Nahuat} & fɪkəkili & fɪkəkili & /i/ \rightarrow O & \& & \checkmark & & \\
d. \text{Aja.ki.m} & a.ja.ki.m & a.ja.ki.m & & \& & \checkmark & & \\
\end{align*}
\]

We noted in section 9.2.1 that opaque generalizations are overlaid by other generalizations of the language. Now we see how the @-candidate is related to the opaque form by ‘peeling’ off a structural change (here, ‘undoing’ a segment deletion). The relation between the opaque form and the @-candidate such that the latter obeys an IO-faithfulness constraint that is violated by the former. This special IO-faithfulness constraint is called the selector.

McCarthy claims that the choice of the @-candidate is uniquely determined by the selector, given the (independently motivated) constraint ranking of the language. That is, the @-candidate is the most harmonic candidate of all candidates that satisfy the selector. For this purpose, we may think of candidate space as divided into two mutually exclusive subsets, one of candidates that satisfy the selector, and another of candidates violating it. Within the former subset, the @-candidate is the optimal candidate, given the constraint hierarchy of the language.

For Turkish this method is successful, given a number of independently motivated observations about constraint ranking. Complex and Anchoring-I0 (militating against final epenthesis) appear to be never violated in Turkish, hence they are undominated. But *VkV is actually viable in paradigms of monosyllabic roots (for example, ok 'arrow', ok-u 'his arrow', ok-um 'my arrow'). That is, no monosyllabic root ending in /k/ loses /k/ in intervocalic contexts. Therefore *VkV must be crucially dominated by some constraint enforcing root-minimality, whose exact nature is irrelevant to our discussion.

The dominated position of *VkV is confirmed by the selection of the @-candidate [a.ja.kim]. Tableau (32) shows four candidates which obey the selector constraint MAX-I0 (omitting candidates that violate this constraint, which do not qualify by definition). The independently motivated constraint ranking indeed selects (32c) as the @-candidate:

This is not the tableau selecting the actual output form but only the part that is relevant to the selection of the @-candidate. In a fully fledged tableau, the choice of the @-candidate is made 'in parallel' with the choice of the optimal output form. That is, (32) is a subtableau of (28). Note that tableau (32), unlike (28), does not include MAX-®O, the Sympathy constraint. McCarthy argues that the Sympathy constraint must indeed always be invisible to the selection of the @-candidate, to avoid circularity in selection.

The choice of the selector is critical, since this determines the @-candidate, hence indirectly the choice of the opaque output. But how is the selector itself being chosen? The selector is always an IO-faithfulness constraint, violable in actual output forms of the language (crucially so, in the opaque output itself). Since most faithfulness constraints of a language are undominated, this narrows down its choice. But the bottom line is that any positive evidence regarding the choice of the selector can only reside in the opaque form itself. The learner must infer the choice of the selector from the opaque form. This may seem an unsatisfactory conclusion which has a whiff of circularity. However, an analogous conclusion holds for a serial theory of opacity: any evidence pertaining to rule ordering (bleeding or counterfeeding, feeding or counterfeeding) can only come from the opaque form itself.

Let us now briefly discuss this proposal, starting with its virtues. First, Sympathy preserves two cornerstones of standard OT – direct mapping and the output-based
format of well-formedness constraints. All interactions are parallel rather than serial, involving ranked constraints rather than ordered levels. Well-formedness constraints never refer to levels other than the output (either inputs or abstract intermediate representations). In its empirical scope, Sympathy surpasses all OT approaches to opacity seen thus far—except Intermediate Levels theory. However, it also has a number of problematic aspects.

First it weakens Correspondence Theory by extending it to candidate-to-candidate faithfulness. Standard correspondence constraints require identity with an independently established, 'concrete', form, either a lexical input (IO-faithfulness), or an actual output form (00-identity). Increasing the scope of correspondence constraints poses a potential threat to OT's restrictiveness. Faithfulness to abstract forms implies a radical increase in the number of constraint interactions, hence in computational complexity of the theory. A related issue, one which cannot be addressed without additional research, is learnability. Do the results that we discussed in chapter 7 carry over to Sympathy?

Second, it is not clear at all whether the @-candidate is uniquely determinable by the selector on the basis of only the independently motivated hierarchy of the language, as McCarthy suggests. Rankings of undominated constraints, which are not independently motivated by output forms, may become relevant to the selection of the @-candidate. Note that the set of candidates obeying the selector always contains candidates violating one or more undominated constraints. In the Turkish example, the candidate set includes a candidate [a.jukm], violating 'COMPLEX.'). What happens if, in this set of selector-obeying candidates, no candidates occur obeying all undominated constraints? Such a situation may occur if the selector imposes a requirement that implies the violation of an otherwise undominated constraint in the @-candidate. In such a case, the ranking of undominated constraints is required to break the tie. But the use of 'ranked undominated constraints' contradicts the claim that the independently motivated ranking of the language suffices to choose the @-candidate. By definition, undominated constraints are not rankable on the basis of evidence from transparent outputs, in which they are never violated.

Third, Sympathy does not offer a general theory of opacity, as it cannot deal with 'chain shifts', a kind of counterfeeding opacity that will be discussed in the next subsection.

9.2.6 Local Conjunction

A final approach to opacity in OT merits attention, partly because it is complementary to Sympathy, and partly for independent reasons. Under Local Conjunction, two constraints are conjoined as a single composite constraint which is violated if and only if both of its components are violated within some domain.

The proposal is due to unpublished work by Smolensky (1993). Consider the schematic tableau (33), which contains four candidates representing all logically possible combinations of violations of constraints \(C_1\) and \(C_2\):

\[
\begin{array}{ccc}
\text{Candidate} & C_1 & C_2 & [C_1 \& C_2] \\
1 & * & * & * \\
2 & * & * & * \\
3 & * & * & * \\
4 & * & * & * \\
\end{array}
\]

The complex constraint \([C_1 \& C_2]\) (the 'conjunction' of constraints \(C_1\) and \(C_2\)) is violated if and only if both of its components are violated. Moreover, for a violation of \([C_1 \& C_2]\) to occur, both separate violations must arise within a single domain \(\delta\) (a segment, morpheme, etc.). Evidently some domain is needed for conjunction: the severity of output ill-formedness is never increased by combinations of violations in random positions in the output. Finally, a conjoined constraint does not replace its components, but it is separately ranked. It is generally assumed that a conjoined constraint is universally ranked above the component constraints.

(34) Universal ranking schema: \([C_1 \& C_2] \Rightarrow C_1, C_2\)

We will first look at evidence for conjunction from opacity (chain shifts), and then at its wider applications.

The clearest motivation for Local Conjunction of constraints resides in chain shifts (Kirchner 1996). A chain shift is a situation in which sounds are promoted (or demoted) stepwise along some scale in some context. Crucially the chain shift does not result in neutralization, since each input takes precisely one step. This is schematized as follows:

\[
(35) \quad A \rightarrow B \quad \text{and} \quad B \rightarrow C, \quad \text{but not} \quad *A \rightarrow C
\]

That is, 'A' reoccupies the position left vacant by 'B', which itself occupies 'C', etc.

An example of a chain shift is vowel raising in Western Basque (De Rijk 1970, Kenstowicz and Kisseberth 1979, Kirchner 1996). Mid vowels and high vowels that precede another vowel are raised by one degree (mid to high and high to raised):
Kirchner assumes the following specifications for vowels for [low], [high], and [raised]:

\[
\begin{align*}
\text{Raised } & \{i^+, u^+\} \quad [-\text{low}, +\text{high}, +\text{raised}] \\
\text{High } & \{i, u\} \quad [-\text{low}, +\text{high}, -\text{raised}] \\
\text{Mid } & \{e, o\} \quad [-\text{low}, -\text{high}, -\text{raised}] \\
\text{Low } & \{a\} \quad [+\text{low}, -\text{high}, -\text{raised}] 
\end{align*}
\]

(37)

(In the Etxarr dialect, on which Kirchner focusses, low vowels are not raised in hiatus. The raising of clitic vowels /bat/ → /bet/ is conditioned by a high vowel in the stem, and we will ignore it.) First we will see that chain shifts are not problematic to serial theory, while they do pose problems to standard OT.

In serial theory the analysis is straightforward: it is a case of counterfeeding on the focus. Two counterfeeding rules raise high vowels and mid vowels, respectively:"

\[
\begin{align*}
\text{Counterfeeding analysis of chain shift} \\
\text{High vowel raising } & /\text{semie-a}/ \\
\text{Mid vowel raising } & /\text{errri}/ \\
/\text{semie} & /\text{errri}/ \\
\end{align*}
\]

Each of the counterfeeding rules applies only once per derivation.

Chain shifts, a kind of underapplying opacity, cannot be analysed under standard OT, as we will now show. Raising in hiatus is triggered by a well-formedness constraint (Kirchner 1996):

\[
\begin{align*}
\text{Hiatus-Raising} \\
\text{In } V_1, V_2, \text{ maximize height of } V_1. 
\end{align*}
\]

Kirchner argues that this constraint is evaluated gradiently, and that different values of \( V_1 \) incur the following violations marks: low ‘***’, mid ‘**’, high ‘*’.

\[
\begin{align*}
\text{\textit{Hiatus-Raising}} \\
\text{Hiatus-Raising} \\
\text{In } V_1, V_2, \text{ maximize height of } V_1. 
\end{align*}
\]

For raising to take place at all, Hiatus-Raising must be ranked above both faithfulness constraints. That is, \( e \rightarrow i \) shows that IDENT-IO(high) is dominated, while \( i \rightarrow i^+ \) shows that IDENT-IO(raised) is dominated. This is summarized in the following tableau (which, however, gives only a subset of the output candidates to be finally considered for /el/):

<table>
<thead>
<tr>
<th></th>
<th>Hiatus-Raising</th>
<th>IDENT-IO (high)</th>
<th>IDENT-IO (raised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. IDENT-IO(high)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. IDENT-IO(raised)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note, however, what happens if we add a third output candidate, /i/, to the tableau of /e/. This ranking incorrectly predicts that /e/ is raised to /i/, going two steps rather than one:

\[
\begin{align*}
\text{\textit{Hiatus-Raising}} \\
\text{Hiatus-Raising} \\
\text{In } V_1, V_2, \text{ maximize height of } V_1. 
\end{align*}
\]

The incorrect prediction is due to undominated Hiatus-Raising. No possible ranking of these three constraints allows \( i \rightarrow i^+ \), and \( e \rightarrow i \), while disallowing \( e \rightarrow i^+ \). Mid vowels simply cannot be prevented from going all the way to raised.
Residual issues

9.2 Opacity

The reasoning behind Kirchner's Local Conjunction analysis is this: the change e → i involves violation of two faithfulness constraints, while each of the individual steps i → i and e → i involves only one violation. Moreover, the faithfulness violations incurred by i → i reoccur separately in the individual steps i → i and e → i. Then what is needed is the conjunction of both faithfulness constraints into a composite constraint. This must be ranked above Hiatus-Raising in order to restrict raising to a one-step process:

(42)  
(a. High vowel raising: Hiatus-Raising $>$ Ident-IO(raised)
(b. Mid vowel raising: Hiatus-Raising $>$ Ident-IO(high)
(c. Interaction: [Ident-IO(raised) & Ident-IO(high)] $>$ Hiatus-Raising)

The composite constraint [Ident-IO(raised) & Ident-IO(high)] $>$ Hiatus-Raising is violated if and only if Ident-IO(high) and Ident-IO(raised) are violated with respect to a given segment. The domain ‘$S$’ is equal to the segment for which both faithfulness constraints are evaluated.

Tableau (43a) illustrates how the 'fell swoop' candidate e → i is eliminated by the composite constraint:

(43)  
<table>
<thead>
<tr>
<th></th>
<th>Ident-IO(high) &amp; Ident-IO(raised)] $&gt;$ Hiatus-Raising</th>
<th>Hiatus-Raising</th>
<th>Ident-IO(high)</th>
<th>Ident-IO(raised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.i</td>
<td>e → e</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a.ii $\equiv$ e → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a.iii e → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.i i → e</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.ii i → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.iii $\equiv$ i → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But in tableau (43b), the same output i → i is allowed as it involves only a single step, hence the composite constraint is not violated.

Finally, let us briefly discuss why Sympathy theory cannot deal with chain shifts. To account for opacity of raising in e → i, the output must be maximally faithful to a @-candidate that is-[raised] itself. Therefore Ident-IO(raised) must be the 'selector', and [i] the @-candidate (which also happens to be the output). Assuming Ident-$\otimes$O(raised) as the undominated @-O-faithfulness constraint, we arrive at tableau (44a):

(44)  
<table>
<thead>
<tr>
<th></th>
<th>Ident-$\otimes$O(raised)</th>
<th>Hiatus-Raising</th>
<th>Ident-IO(high)</th>
<th>Ident-IO(raised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.i</td>
<td>e → e</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
</tr>
<tr>
<td>a.ii $\equiv$ e → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
</tr>
<tr>
<td>a.iii e → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td></td>
</tr>
<tr>
<td>b.i i → e</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
</tr>
<tr>
<td>b.ii $\equiv$ i → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
</tr>
<tr>
<td>b.iii $\equiv$ i → i</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
<td>$\ast!$</td>
</tr>
</tbody>
</table>

In (44b), however, we see that the same ranking predicts that [i] is the @-candidate, hence the (incorrect) output. Reversing the ranking of Ident-$\otimes$O(raised) and Hiatus-Raising may account for i → i, but will spoil the correct outcome for input /ei/, predicting e → i. This ranking paradox cannot be resolved.

Local Conjunction has seen wider applications (outside chain shifts) in a range of phenomena, including dissimilation and Obligatory Contour Principle (OCP) effects (Alderete 1997, Itô and Mester 1998) and word stress (Kager 1994, Crowhurst and Hewitt 1997). The shared property of all these phenomena is that multiple violations of 'basic' constraints within some domain are banned, while a single violation of each of the constraints individually is tolerated. A specific case is that where multiple violations of one constraint within a domain are banned: self-conjunction.

Here we will illustrate this approach for OCP effects. Itô and Mester (1998) discuss various restrictions on the cooccurrence of segments within Japanese morphemes, which ban cooccurring segments sharing the same marked value of some feature. One of these restrictions is known as Lyman's Law:

(45)  
Lyman's Law: 'Stems must not contain more than one voiced obstruent.'

Lyman's Law is exemplified by the minimal pairs (46a–c) and systematic gaps (46d):

(46)  
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a.i</td>
<td>kaki 'persimmon'</td>
<td>a.ii</td>
<td>futa 'lid'</td>
</tr>
<tr>
<td>b.i</td>
<td>kagi 'sign'</td>
<td>b.ii</td>
<td>fuda 'sign'</td>
</tr>
<tr>
<td>c.i</td>
<td>gaki 'writing'</td>
<td>c.ii</td>
<td>buta 'pig'</td>
</tr>
<tr>
<td>d.i</td>
<td>*gagi (non-existent)</td>
<td>d.ii</td>
<td>*buda (non-existent)</td>
</tr>
</tbody>
</table>
This is arguably a case of 'positional neutralization' – voicing is distinctive in obstruents, but up to a certain limit: maximally one voiced obstruent may occur in a stem. Moreover, it can be seen as a static kind of 'dissimilation' – opposite values for voicing are preferred to cooccurrences of [+voice].

Earlier autosegmental analyses of Lyman's Law assumed the tier-adjacency of the feature [voice], excluding adjacent occurrences of [+voice]. Such analyses involved the OCP ('At the melodic level, adjacent identical elements are prohibited'). The fact that [+voice] is targeted by the OCP, rather than [-voice], is not an accidental property of Japanese. Cross-linguistically, dissimilations are triggered by adjacent segments specified for marked values for a given feature. For this reason, OCP analyses assume the underspecification of unmarked feature values. (See Steriade 1995a for a criticism of such approaches.)

It6 and Mester (1998) take a different view, and argue that the markedness content of Lyman's Law should be captured directly, without underspecification. This involves a constraint banning cooccurrences of marked [+voice] within a morpheme. This constraint is an extension of VOP, the basic markedness constraint banning [+voice] in obstruents (known from chapter 1, section 1.7.5):

\[
\text{Voiced Obstruent Prohibition (VOP)}
\]

\* [+voice, -son]

Since voicing is distinctive in Japanese obstruents (within the limits imposed by Lyman's Law), VOP must be dominated by \( \text{IDENT-IO(voice)} \). However, the absolute ban on double occurrences of [+voice] in a stem requires the conjunction of VOP with itself, dominating \( \text{IDENT-IO(voice)} \). It6 and Mester propose the following schema for self-conjunction:

\[
\text{Self-conjunction of constraints: } [C, & C_1]_6 \text{ with } C_1 = C_1
\]

Evaluation of \( [C, & C_1]_6 \): [C, & C_1]_6 is violated in domain 6 if there is more than one violation of C, in domain 6.

Applying self-conjunction to VOP, this schema produces \( [\text{VOP} & \text{VOP}]_6 \) or \( \text{VOP}^2 \):

\[
\text{VOP}^2
\]

No cooccurrence of voiced obstruency with itself.

In order to fulfill its function of 'positional neutralizer', \( \text{VOP}^2 \) must be ranked above the basic sequence \( \text{IDENT-IO(voice)} \gg \text{VOP} \). Considering a hypothetical input that has a dual occurrence of [+voice, -son], one of these is erased, while the other is preserved. Which of them is erased cannot be determined here, given the fact that there are no alternations:

\[
\begin{array}{|c|c|c|}
\hline
\text{Input: /gagi/} & \text{VOP}^2 & \text{IDENT-IO (voice)} \\hline
\text{a. gagi} & \ast ! & \ast \ast ! \\hline
\text{b. gregation} & \ast & \ast \\hline
\text{c. igration} & \ast & \ast \\hline
\text{d. kaki} & \ast \ast ! & \ast \\hline
\end{array}
\]

Minimal violation of \( \text{IDENT-IO(voice)} \) guarantees that only one occurrence of [+voice] is erased (rather than both, as in 50d).

Another advantage of Local Conjunction is that it rationalizes existing contextual markedness constraints as conjunctions of two basic constraints (Smolensky 1995, It6 and Mester 1998). For example, *VOICED-CODA ('coda obstruents are voiceless') is construed as the Local Conjunction of the context-free markedness constraint VOP ('obstruents are voiceless') and the prosodic markedness constraint NoCODA ('no codas'). Violation of *VOICED-CODA implies a combined violation of VOP and NoCODA in a single segment. Hence, \( [\text{NoCODA } & \text{VOP}] \) is equivalent to *VOICED-CODA. The example is from Dutch (see chapter 1):

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Input: /bed/} & \text{[NoCODA ] & MAX-IO } & \text{IDENT-IO (voice)} & \text{NoCODA } \\hline
\text{a. bed} & \ast ! & \ast \ast \\hline
\text{b. uration} & \ast \ast ! & \ast \\hline
\text{c. pet} & \ast \ast ! & \ast \\hline
\text{d. be} & \ast \ast \\hline
\end{array}
\]

The original constraint is somehow unsatisfactory because it restates two independently needed markedness constraints. In cases such as these, Local Conjunction helps to reduce the set of basic constraints, leading to greater generalization, and decreasing arbitrariness in the statement of individual constraints.

Should Local Conjunction be granted the status of basic theoretical device of OT? It certainly has its positive sides. First, it fills the empirical gap left by Sympathy in the analysis of opacity (chain shifts). Second, in its self-conjunction...
The essence of OT is that grammatical outputs are the best possible compromise between conflicting needs. This implies that absolute well-formedness of output forms cannot be a criterion for grammaticality. (Some constraints are necessarily violated in every output.) This approach to well-formedness predicts that for every minimal violation of basic constraints, and (b) Local Conjunction.

Secondly, the fact that Local Conjunction seems to undermine strict domination, a core principle of OT. Under strict domination, violation of higher-ranked constraints cannot be compensated for by satisfaction of lower-ranked constraints. (See chapter 1.) But under Local Conjunction, two constraints A and B, each of which is ranked too low to force the violation of C, can nevertheless dominate C by joining forces in a conjoined constraint [A & B]. Such a situation, in which two constraints 'team up' against a third constraint, conflicts with the intuition behind strict domination (even though Local Conjunction and strict domination are not formally inconsistent).

Finally, questions arise with respect to the huge increase of possible constraints in Gen which is implied by Local Conjunction. For example, can any pair of constraints be conjoined? This predicts constraints of a typologically doubtful status, such as [ONSET & NOCODA], or [+round] & PARSE-SYL]. Can faithfulness constraints be conjoined with markedness constraints? If so, this predicts even more bizarre constraints, such as [**N & IDENT-IO(Place)]. Can any number of constraints be conjoined, or is there a (binary?) upper limit? Without any upper limit, the number of constraints in Gen becomes infinite, with fatal effects on learnability. But even under maximally restrictive assumptions about Local Conjunction, a vast increase occurs in the amount of constraint interaction.

9.3 Absolute ungrammaticality
The essence of OT is that grammatical outputs are the best possible compromise between conflicting needs. This implies that absolute well-formedness of output forms cannot be a criterion for grammaticality. (Some constraints are necessarily violated in every output.) This approach to well-formedness predicts that for every minimal violation of basic constraints, and (b) Local Conjunction.

The best-studied examples involve the blocking of word formation processes to avoid the violation of some phonological well-formedness constraint. How does OT deal with such cases? Does absolute ungrammaticality motivate a fundamental revision of the model?

We will discuss an example from stress-based conditions on affixation in English, analyzed by Raffelsiefen (1996). Raffelsiefen argues that the verbalizing suffix -ize may not attach to adjectives ending in a stressed syllable, while it productively attaches to adjectives ending in an unstressed syllable:

(52) Input → Grammar → Ø

The best-studied examples involve the blocking of word formation processes to avoid the violation of some phonological well-formedness constraint. How does OT deal with such cases? Does absolute ungrammaticality motivate a fundamental revision of the model?

We will discuss an example from stress-based conditions on affixation in English, analysed by Raffelsiefen (1996). Raffelsiefen argues that the verbalizing suffix -ize may not attach to adjectives ending in a stressed syllable, while it productively attaches to adjectives ending in an unstressed syllable:

(53) Blocking of -ize after bases with final stress in English
a.i rámond rámond-ize  a.ii foreign foreign-ize
b.i corrúpt *corrúpt-ize  b.ii obscene *obscène-ize

Note that -ize is a stressed suffix (it has secondary stress wherever it appears), and that its attachment to a stem with final stress would create a 'stress clash', a situation of adjacent stresses. Intuitively, 'clash avoidance takes priority over affixation'. But how can this be captured in OT; a theory of viable constraints? If word formation were simply blocked if its optimal output violated any phonological constraint, then no word formation would be possible at all. To avoid wholesale nullification of morphology, the lexical input must be realized by the grammar, that is, be mapped onto some output. Instead of blocking the input, the grammar selects its optimal output analysis, possibly by a repair. In the case in hand, repair may involve violation of accentual faithfulness (Max-Ft-Io 'input stresses occur in the output'), or well-formedness (*Clash). In sum, absolute ungrammaticality poses a challenge to OT. In contrast, a theory of invariable constraints has no problems: an output filter (with nullifying power) will reject any outputs violating *Clash.

Raffelsiefen employs an analytic resource that was originally proposed by Prince and Smolensky (1993). The grammar always produces an output for every input. However, by Freedom of Analysis, Gen supplies one special output, which is identical to the input: a set of disjoint morphemes lacking any morphological or phonological cohesion. This Null Parse is left unpronounced, due to its lack of morpho-phonological word status. It will be optimal, though, if no other candidate is available that is more harmonic:
The question then is, how can the Null Parse be optimal? Due to its lack of phonological and morphological word status, it has one substantial advantage over all other candidates: it cannot violate any well-formedness constraints pertaining to the word level. However, the Null Parse violates the following constraint (McCarthy and Prince 1993b):

The ranking of MPARSE in the grammar determines the ‘robustness’ of word formation. If it is undominated, some (non-null) output will be produced, since anything is better than the null parse. But any constraint that dominates MPARSE has the power of ‘annihilating’ affixation. This is seen in tableau (56), in which MPARSE is dominated by an 0-identity constraint IDENT (‘The stem of the derived word must be identical to the base’), as well as by *CLASH (‘Two adjacent stresses are prohibited’).

The Null Parse (56c) is optimal since all its competitors violate undominated constraints. Hence it is better to have no (tangible) output at all than an ill-formed output.

Orgun and Sprouse (1997: 4) point out empirical problems for MPARSE theory. They argue that ‘...there are cases of ungrammaticality in which the ungrammatical candidate could be repaired by violating a constraint independently known to be violable in other (grammatical) output forms in the language’. That is, the constraint ranking required to make the Null Parse win is demonstrably inconsistent with the ranking required for non-null forms of the language. If these objections (and the analyses on which they are based) are valid, then they pose serious challenges to MPARSE theory.

Yet another unsatisfactory aspect of the MPARSE model is the assumption that the Null Parse, in spite of its lack of phonetic realization, violates no faithfulness constraints. This assumption is required to prevent the undesired result that the Null Parse is rejected for violating the same constraints that exclude its non-null (ungrammatical) competitors. For example, adding a violation mark for IDENT to the Null Parse (56c), penalizing its lack of phonetically realized stress, would render it less harmonic than all its competitors. Alternatively, it might be stipulated that identity constraints are only activated for outputs that are phonetically interpreted, but that seems equally ad hoc.

Alternatives to MPARSE have been proposed, changing the basic design of the OT grammar, by introducing inviolate constraint components. Orgun and Sprouse (1997) argue that a component of inviolate constraints (Control) is serially ordered after Gen and Eval as an output filtering device. As in standard OT, Eval selects an optimal candidate, which is submitted to Control. Unlike Eval, Control has the power of marking as ungrammatical any outputs violating its constraints.

Secondly, Control is a rather blunt tool, which is principally unable to account for different types of blocking within a single language. For example, a language may have two ‘defective’ affixations, each of which is blocked in a separate phonological context. Here Control would be of no value because it predicts that the affixations share gaps in the same contexts. In contrast, MPARSE theory may deal with such cases by splitting up MPARSE into different affix-specific versions. In sum, ‘absolute ungrammaticality’ is another challenge which OT has not yet completely mastered.  

Broekhuis (forthcoming) argues that absolute ungrammaticality can be dealt with by a component of inviolate constraints preceding Eval. This proposal is comparable to the assumption that Gen contains universally inviolable constraints preceding Eval. Broekhuis deviates from this idea, however, in his definition of ‘input’. In syntax, as we have seen in chapter 8, the input is less clearly defined than in phonology. Broekhuis, for example, defines the input to Gen as a complete syntactic representation developed in a component preceding Gen, whose requirements (‘checking’) are inviolable. Whether such a model can deal with absolute ungrammaticality in phonology remains to be seen, however.
9.4 Free variation

Another unresolved issue is variation. We will focus on variation of a specific type: the case of a single input being mapped onto two outputs, each of which is grammatical. This is ‘free variation’, also known as ‘optionality’:

\[(57) \quad \text{Input} \rightarrow \text{Grammar} \rightarrow \text{Output}_1, \text{Output}_2\]

By definition, the distribution of both outputs cannot be under grammatical control, since that would restore (ranking-based) determinism in the choice of both variants.

Examples of free variation are abundant in natural languages. Examples below are from English phonology (vowel reduction, 58a) and syntax (complementizer, 58b)

\[(58) \quad \begin{align*}
\text{a. sentimen} & \text{tality} \quad \rightarrow \quad \text{sentim} & \text{nality} \\
\text{b. I know that John will leave} & \quad \rightarrow \quad \text{I know John will leave}
\end{align*}\]

The fact that variation is ‘free’ does not imply that it is totally unpredictable, but only that no grammatical principles govern the distribution of variants. Nevertheless, a wide range of extragrammatical factors may affect the choice of output variant over the other, including sociolinguistic variables (such as gender, age, and class), and performance variables (such as speech style and tempo). Perhaps the most important diagnostic of extragrammatical variables is that they affect the choice of occurrence of one output over another in a stochastic way, rather than deterministically. Here we will focus on the consequences of free variation for the grammar.

Why does free variation pose a challenge to OT? An OT grammar is essentially an input–output mapping device. The grammar is deterministic, in the sense that each input is mapped onto a single output – the most harmonic candidate for a constraint hierarchy. Given a single deterministic competition, how can two candidates ever both be optimal? If two output candidates O and O’ are different in terms of constraint, then this difference must be relevant to some constraint(s) of the hierarchy. This implies that O and O’ do not share the same violations marks hence one is more harmonic than the other with respect to the hierarchy. The question is: how can free variation be reconciled with the deterministic nature of the grammar?\[^{15}\]

To be fair, the problem of non-unique outputs is not specific to OT. It reoccurs in derivational theory, or in fact any theory of grammar that is faced with the situation (57). A solution at a terminological level is to redefine the ‘output of the grammar’ for a given input as a set of forms \(\{O_0, O_1, \ldots, O_n\}\), rather than a unique form. But of course this does not solve the problem of how to generate sets of output forms. In derivational rule-based theory, specific rules may be marked as ‘optional’. In a derivation, an optional rule may be applied (in which case some output results), or be left unapplied (in which case another output results). But what is the counterpart of ‘optional rules’ in OT? Contrary to rules, constraints are not language-specific devices, but elements of universal grammar that are potentially active in every grammar. Therefore a solution to the ‘free variation problem’ can only reside in constraint ranking.

A radical method is splitting up the grammar into multiple constraint hierarchies, or co-phonologies, each of which selects its own optimal candidate by its own ranking. Such an approach is reminiscent of ‘multi-stratal evaluation’, discussed in section 9.2.3 in relation to opacity. It is not too difficult to imagine an extension of this stratal idea to free variation. If we assume that strata can be organized in parallel (in addition to serially, as in section 9.2.3), an input can be fed into two parallel co-phonologies, giving two outputs:

\[(59) \quad \text{Input} \rightarrow \text{Co-phonology 1} \rightarrow \text{Output}_1, \text{Co-phonology 2} \rightarrow \text{Output}_2\]

In section 9.2.3 we mentioned weaknesses in models of multi-stratal evaluation, one of which has a direct counterpart in parallel co-phonologies. Splitting a grammar up into subgrammars (‘strata’ or ‘co-phonologies’) makes the prediction that each subgrammar is independent, hence that subgrammars can be radically different. This prediction is clearly incorrect for free variation, where both outputs are mostly similar, and differ only in a minor respect. (Co-phonologies are better suited for lexical variation resulting from strata; see Inkelas and Orgun 1995 and 1997 and Orgun 1995; and exceptionality; see Inkelas, Orgun, and Zoll 1997.)

A less radical alternative is to maintain a single constraint hierarchy, while giving up the idea of a fixed ranking of constraints. Throughout this book we have assumed that constraints are strictly ranked. Two conflicting constraints \(C_i\) and \(C_j\) are ranked in either of two ways: \(C_i\) strictly dominates \(C_j\) or \(C_j\) strictly dominates \(C_i\). This assumption was upheld even for cases in which \(C_i\) and \(C_j\) cannot be ranked with respect to each other due to a lack of interaction (so that no empirical evidence can determine the ranking). In such cases we ranked \(C_i\) and \(C_j\) in the same position (called ‘stratum’ in chapter 7). But we never considered the

\[^{15}\] It has been argued that some cases of free variation indeed derive from inability on the part of the constraint inventory to distinguish different outputs (Hammond 1994, Smolensky 1996). That is, the constraint inventory simply lacks constraints discriminating between two outputs \(O_0\) and \(O_1\). However, even if this approach is feasible for the particular cases for which it has been proposed, it is clearly not generalizable to all cases of free variation. Most involve presence or absence of grammatical structure which is ‘visible’ to some constraint (ultimately to *STRUC ‘no phonological structure’).
9.5 Positional faithfulness

Linguistic structures exhibit asymmetries with respect to contexts in which phonetic contrasts can be realized. For example, in chapters 1–3, we have seen that, generally speaking, codas display a more restricted set of contrasts than onsets do. We assumed that such positional neutralization is due to contextual markedness constraints (mitigating against a feature value [αF] in specific positions), interacting with 'context-free' faithfulness constraints. But a priori, what we refer to as 'coda neutralization' may equally well be described as 'faithfulness in onsets'. Here we will sketch an alternative view, positional faithfulness, which radically revises the roles of markedness and faithfulness constraints. Resistance to neutralization is attributed to constraints that license features in specific positions, an interaction with context-free markedness constraints.

(62) Two possible views of positional neutralization:

a. Context-free faithfulness ⇔ Positional markedness
b. Positional faithfulness ⇔ Context-free markedness

---

Residual issues

In most cases it is simply impossible to find evidence for one view or the other, due to the fact that the contexts of neutralization and faithfulness are both positively characterized by mutually exclusive labels ('onset', 'coda', etc.). In current literature both views have been adopted. Regardless of the issue of whether one or both views are adequate, two general arguments for positional faithfulness can be put forward.

Firstly, positional faithfulness is supported by functional considerations. It is well known that contrasts are best realized in perceptually salient positions (Nootboom 1981, Hawkins and Cutler 1988, Ohala 1990, Ohala and Kawasaki 1984). Salient positions include word-initial consonants, prevocalic (or released) consonants, stressed vowels, and vowels in initial syllables. For example, it may be assumed that all vowels are subject to the same general forces of reduction ('minimize articulatory effort'), regardless of their position or stressing. Yet stressed vowels, by their inherent perceptual salience (tonal and durational) are best equipped to realize featural distinctions, hence to resist general reduction forces.

Recent OT studies have implemented perception-based asymmetries in the notion of positional faithfulness. Phonetic grounding renders faithfulness constraints sensitive to perceptually prominent positions. More specifically, a faithfulness constraint for a feature [F] referring to a prominent position dominates the general faithfulness constraint for [F]. This allows a 'sandwiching' of markedness constraints in between positional and general faithfulness constraints, producing positional neutralization:

(63) Ranking schema for positional neutralization

\begin{align*}
10\text{-Faithfulness (prominent positions)} & \gg \text{Markedness} \\
\text{IO-Faithfulness (general)} & \gg \\
\end{align*}

Some researchers (Steriade 1995b, Flemming 1995, Jun 1995, Kirchner 1995, 1997) have argued that perceptual principles should be directly stated in grammars, thus allowing reference to gradient and non-contrastive phonetic features. Others (Selkirk 1995, Lombardi 1995b, Beckman 1997a, b) maintain a strict separation of phonology and phonetics, in the sense of avoiding reference to gradient features.

The second (empirical) argument for positional faithfulness is that it captures and unifies a number of phonological patterns in different languages. More specifically, there is an overwhelming typological tendency for neutralizing assimilation to preserve feature values of segments in 'salient' positions (onsets, initial syllables, root segments, etc.), at the expense of segments in other positions (codas, medial syllables, affix segments, etc.). In this section we will look into one of these cases – the licensing of the feature [+round] by initial syllables – drawing on non-OT work by Steriade (1995a) and inspired by the OT analysis of Shona vowel harmony by Beckman (1997a).

A typologically common restriction regarding [+round] is that it occurs on vowels in specific positions. For example, several Altaic languages allow rounded vowels only in the initial syllable of the word. This is a case of contextual neutralization: the rounding contrast is suppressed in segments in non-initial syllables. However, an analysis of this pattern by a contextual markedness constraint banning [+round] from non-initial syllables is awkward, as the context of neutralization ('non-initial syllables') is not a natural class. This defect is avoided by a positional faithfulness analysis. A constraint IDENT-IO(round, [u]) militates against the loss of [round] in initial syllables:

(64) \text{IDENT-IO}([\text{round}, [\sigma])

An output segment standing in the initial syllable has the same value for [round] as its input correspondent.

This constraint is (nearly) identical to the general faithfulness constraint IDENT-IO(round) (65), the only difference being that the latter fails to refer to initial syllables:

(65) \text{IDENT-IO}([\text{round})

An output segment has the same value for [round] as its input correspondent.

It is assumed that pairs of faithfulness constraints such as (65–6) are universally ranked such that the position-specific constraint is ranked above the general, position-insensitive constraint. (Compare 'harmony scales' discussed in chapter 1, section 1.8.) Hence:

(66) Harmony scale

\begin{align*}
\text{IDENT-IO}([\text{round}, [\sigma]) & \gg \text{IDENT-IO}([\text{round})} \\
\end{align*}

This installs the faithfulness part of the interaction. Let us now look into the markedness part, which is very simple. The faithfulness constraints compete with a general, context-free markedness constraint *[+round], which penalizes round vowels. In Altaic, rounding is lost in all vowels except those that are licensed by an initial syllable. This is expressed in the following ranking:

(67) Licensing of rounding in initial syllables

\begin{align*}
\text{IDENT-IO}([\text{round}, [\sigma]) & \gg *[+\text{round}] \gg \text{IDENT-IO}([\text{round})} \\
\end{align*}
Let us assume a hypothetical input with rounding in both of its vowels. Candidate (68c) fatally violates positional faithfulness in its initial syllable. Of the two remaining candidates, (68b) is optimal since it minimally violates *[+round]:

This analysis offers an empirical argument for positional faithfulness: it is able to capture contextual neutralization of [round], without reference to a non-natural class 'non-initial syllables' as the context of neutralization.

However, an even better argument can be made for positional faithfulness on the basis of rounding harmony. In Yokuts (Newman 1944), [+round] is contrastive in the first syllable only, as in Altaic, but it also spreads beyond its directly licensed position, into affix vowels.

Rounding in non-initial syllables is indirectly licensed by the initial syllable. This pattern is a variation on the direct licensing pattern seen in Altaic (67). A markedness constraint HARMONY ('vowels agree in their values of [round]') dominates *[+round]:

Rounding harmony controlled by initial syllable
IDENT-IO(round, [σ]), HARMONY \(\gg\) *[+round] \(\gg\) IDENT-IO(round)

For a hypothetical input /dub-hin/, this ranking is illustrated by the following tableau:

<table>
<thead>
<tr>
<th>Input: /u-u/</th>
<th>IDENT-IO (round, [σ])</th>
<th>*[+round]</th>
<th>IDENT-IO (round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. u-o</td>
<td>*</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td>b. u-i</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>c. i-i</td>
<td>!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This analysis offers an empirical argument for positional faithfulness: it is able to capture contextual neutralization of [round], without reference to a non-natural class 'non-initial syllables' as the context of neutralization.

Harmony is satisfied in both (71a) and (71c) but only the latter violates positional faithfulness in its initial syllable. This analysis is highly elegant, and it also captures the relation with the direct licensing of rounding in the Altaic languages by a simple reranking. (In Altaic, *[+round] dominates HARMONY.) Finally, a positional faithfulness analysis is superior to one based on contextual markedness. With only a 'general' faithfulness constraint IDENT-IO(round) to penalize changes of [round], the inertness of the initial syllable must be due to contextual markedness. Even if we assume a constraint militating against [+round] in non-initial syllables (ranking it above *[+round]), its interaction with HARMONY is highly problematic. For example, HARMONY \(\gg\) *[+round] \(\gg\) IDENT-IO(round) predicts a harmony pattern *dib-hin, regardless of where we rank the contextual markedness constraint. This empirical argument for positional faithfulness complements the typological argument that was discussed earlier.

Steriade (1995a: 162) points out 'extensions of the idea of indirect licensing to the cases of local assimilation in which onsets spread place or laryngeal features onto adjacent coda consonants'. In many languages, codas cannot carry independent values for [voice] (as in Dutch, chapter 1) or place of articulation (recall the discussion of the Coda Condition in chapter 3). In a subset of these languages, coda neutralization is accompanied by coda-to-onset assimilation, resulting in an 'indirect licensing' of features by the onset.

<table>
<thead>
<tr>
<th>Coda</th>
<th>Onset</th>
<th>IDENT-IO (round, [σ])</th>
<th>*[+round]</th>
<th>IDENT-IO (round)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>dub-hun</td>
<td></td>
<td></td>
<td>**</td>
</tr>
<tr>
<td>b.</td>
<td>dub-hin</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td>dib-hin</td>
<td>!</td>
<td></td>
<td>* !</td>
</tr>
</tbody>
</table>

Steriade's observations on indirect licensing in local assimilation carry over to positional faithfulness in OT, as demonstrated in recent work by Jun (1995), Lombardi (1995b), and Beckman (1997a, b). Similar onset–coda asymmetries have been observed with respect to consonant deletions (see chapter 3, section 3.6). Again, positional faithfulness offers an explanation for such asymmetries (explaining why coda deletion is favoured over onset deletion).

\footnote{We abstract away from the interaction with vowel height, as well as from opacity (section 9.2).}
Let us now evaluate the notion *positional faithfulness* by criteria that have played a major role earlier in this chapter. These are constraint format and factorial typology.

First, does positional faithfulness maintain the strict separation between faithfulness constraints and markedness constraints? Recall that in section 9.2.2 we criticized two-level well-formedness for blurring this distinction, and giving constraints rule-like power. Positional faithfulness constraints resemble two-level well-formedness constraints in their reference to prosodic positions of output segments. Unlike two-level well-formedness constraints, they cannot be violated if the output segment fails to occupy the relevant position, and so can't trigger changes. In this respect, compare (73) and (74) stated in their correspondence formats:

(73)  
Positional faithfulness  
If an output segment S is [around] and S stands in an initial syllable, then its input correspondent S' must be [around].  
(not violated if S is not in an initial syllable)

(74)  
Two-level well-formedness  
If an input segment S is [around], then its output correspondent S' must be [around] and stand in an initial syllable.  
(violated if S is not in an initial syllable)

To avoid the problems inherent to two-level well-formedness constraints (including their 'rule'-like nature, mentioned in section 9.2.2), constraints of the type (74) should be disallowed.

A second criterion is factorial typology. We have seen that positional faithfulness is promising in several respects, but nonetheless it appears that contextual markedness, as discussed in chapter 1, cannot be simply eliminated. For that would produce the fatally impoverished typology below:

(75)  
A factorial typology of markedness and faithfulness  
a. P-Faithfulness $\gg$ Markedness $\gg$ \textbf{Faithfulness} Positional neutralization  
b. Faithfulness, P-Faithfulness $\gg$ Markedness Full contrast  
c. Markedness $\gg$ Faithfulness, P-Faithfulness Lack of variation

What is lacking from this factorial typology, as compared to the one of chapter 1, is the case of 'allophonic variation'. This arises under a complete domination of faithfulness for a feature [F] by markedness constraints for [F], with contextual markedness dominating context-free markedness. Positional faithfulness fails to reproduce this pattern since total domination of faithfulness obscures any contextual effects of positional faithfulness. The conclusion must be that positional faithfulness cannot replace contextual markedness and that both are necessary. This finding implies a more complex factorial typology, which we will not present here, though.

9.6 Underlying Representations versus allomorphy

Classical generative phonology (Chomsky and Halle 1968) is based on three assumptions: (i) all contextual variants of a morpheme derive from a single Underlying Representation, (ii) by rewrite rules (A + B /X__Y), (iii) applying in a serial derivation, based on linear ordering. As we have seen, both the second and third assumptions have been abandoned by OT, but so far we have still assumed that the input of an OT grammar is an Underlying Representation (UR) in the classical sense: a unique lexical form underlying all alternants of a morpheme. The notion of UR may have changed as a result of Lexicon Optimization, but it still functions as the input of Gen: all candidate outputs evaluated are analyses of a single UR. The standard correspondence model, which includes URs, is given below:

(76)  
Standard model (with Underlying Representations)  
+-----------------+-----------------+-----------------+  
| Output           | Base            | Affixed form    |  
| UR               | (affixed) UR    |                 |  
| Base             | UR              |                 |  

Base-identity is a priority of language: it enhances the one-to-one relation between lexical items, atoms of meaning, and the sound shapes encoding them. We will call this uniform exponence. An overlap in functions exists between base-identity and UR: both maximize uniform exponence. In this sense standard Correspondence Theory, which has two means for a single goal, is conceptually redundant.

Recently attempts were made to eliminate UR from phonological theory (Burzio 1996, forthcoming, Hayes forthcoming). This development can be seen as yet another extension of the role of output forms in explaining alternations, as well as a logical continuation of ideas on base-identity (Benua 1995, McCarthy 1995a, Kenstowicz 1996).

On the standard view, the input equals an abstract UR, which is mapped onto the output, an actual surface representation. UR is abstract as it need not surface in its input form, but serves to supply phonotactically unpredictable shape aspects
of alternants. If we eliminate UR, ignoring alternations for the moment, then abstractness of the input is no longer necessary, since every input will be identical to its actual output. On this new UR-less view, the input is simply an intended surface form, whose relative harmony is to be determined by Eval. If the grammar's optimal output matches the input, then the input is 'grammatical'. The role of the grammar as an input-output mapping mechanism changes into a checking mechanism.

This checking role of the grammar in fact originates in the standard model, where it forms the cornerstone of the theory of neutralization, allophonic variation, and contrast. Under Richness of the Base, any imaginable input may be proposed to Eval, but only a subset of inputs actually surface in their original shape. However, if no alternations occur in a morpheme's shape, the learner will never postulate an input deviating from the actual observable output form. Due to Lexicon Optimization, the input simply equals the output unless there is reason to deviate (chapter 1, section 1.6).

Let us now extend this checking view to cases in which alternations do exist, that is, cases for which the standard model uses an abstract UR. We start by making the major assumption that each surface alternant of a morpheme has its own input, which is simply identical to its output under Lexicon Optimization. (That is, we temporarily ignore the fact that alternants are systematically related in their shapes.) The assumption of multiple inputs for each morpheme (allomorphs) thus allows the grammar to continue to function as a checking device.

In the Dutch example, the input of the singular [brt] 'bed' is simply {bet}, while that of the plural [brdɛn] is simply {brdɛn}. (We will use the notation {...} for inputs to avoid confusion with the standard UR notation /.../.) Still, under Richness of the Base, an alternative singular input (brd) might have been proposed, but this would not have been able to make it to the surface due to a constraint interaction (**\textit{VOICEDCODA} \Rightarrow IDENT-IO(voice)) neutralizing it into the output [bet]. Therefore the optimal lexical input for [brt] is {brt}, due to Lexicon Optimization. The grammar does not impose such a neutralization on the input {brdɛn}, which makes it to the surface without change:

\begin{align*}
\text{(77)} & \quad \text{Input} \{\text{bet}\} \rightarrow \text{Output} \{\text{bet}\} \quad \text{Input} \{\text{bedan}\} \rightarrow \text{Output} \{\text{brdɛn}\} \\
& \quad \text{Input} \{\text{brd}\} \rightarrow \text{Output} \{\text{brdɛn}\}
\end{align*}

In this UR-less model, the input (lexical shape) simply equals the output (surface shape). 10-faithfulness maintains its original function of reinforcing parts of the input, protecting it against the neutralizing forces of markedness.

We now have a conceptually simple model, which accounts for neutralization as well as allophonic variation, but which evidently fails to account for alternations. On this simple model, shapes of morphologically related outputs ([bet] / [brdɛn]) are not related in any phonological sense, since they do not share a common input (UR). To deal with alternations, morphologically related output forms must be subjected to constraints which enforce 'uniform exponence', limiting the phonological dissimilarity between alternants. This is where 00-correspondence comes into play: it eliminates the function of UR in capturing phonological shape similarities between morphologically related output forms.

As in the standard model, identity between 'base' and affixed forms is enforced by 00-correspondence. However, the key difference from the standard model resides in a new role of the lexicon. The lexicon no longer supplies a unique UR for each morpheme, but instead it supplies a set of shape variants of the morpheme, allomorphs, chunks ready for insertion in various morphological contexts (base or affixed form). The role of 10-faithfulness is revised accordingly: this checks faithfulness of output allomorphs to their lexical shapes. Finally, well-formedness constraints preserve their function of evaluating output forms. That is, we replace model (76) with model (78).

\begin{align*}
\text{(78)} & \quad \text{Allomorphic model (without Underlying Representations)} \\
& \quad \text{00-Correspondence} \\
& \quad \text{10-Faithfulness} \\
& \quad \text{Output Base} \Leftrightarrow \text{Affixed form} \\
& \quad \text{Input} \quad \text{Base allomorph, Contextual allomorph}
\end{align*}

Such a model has been advocated in Burzio (1996, forthcoming), elaborating ideas on word-based morphology as proposed by Aronoff (1976), Bybee (1988, 1995) and others. Two sets of faithfulness constraints are assumed in this model. 00-correspondence is active in its standard role of checking identity in the network of morphologically related output forms, or the paradigm. 10-faithfulness checks identity between allomorphs in the lexical input and their output counterparts.

Let us come to understand this allomorphic model by an example of an alternation that we have studied before, the case of Dutch final devoicing. In a theory without URs, this alternation depends on allomorphs which have different specifications for the feature [voice]. Here we will abstract away from any other possible shape variation. Furthermore, to avoid complications, we focus on alternations of
Residual issues

voice in very simple paradigms which consist of the singular stem and the plural. Therefore we assume that the input is a mini-paradigm, a set of allomorph pairs, each containing one base allomorph (for insertion into the singular stem) and one contextual allomorph (for insertion into the affixed plural). As usual, Gen supplies a set of candidate output forms for each input pair, that is, a paradigm consisting of an actual singular stem and an actual affixed plural.

First we will look at the evaluation of an input paradigm \{bet \sim bed-an\}, which the grammar should approve of without changes, producing \[bet \sim [bed-an]\]. Gen proposes an infinite set of candidate output paradigms, of which four are shown in (79).

(79) Function of Gen in allomorphic model
Input \{bet \sim bed-an\} \rightarrow Gen \\
\[bet \sim [bed-an]\] \\
\[bet \sim [bet-an]\] \\
\[bed \sim [bed-an]\] \\
\[bed \sim [bet-an]\]

All four are submitted to Eval, which selects output paradigm (80a), indeed identical to the input:

(80) Inflation of paradigm
Lexical \{bed\} \rightarrow \[bet\] (Output allomorph 1, affixed form, lexically supplied) \\
\rightarrow \[bed\] (Output allomorph 2, base form, added by grammar)

The grammar (actually Gen, by Freedom of Analysis) 'produces' an allomorph (\[bet\]) for insertion in the base output if the lexicon fails to supply one. This is the first example of a lexical paradigm being overruled by the grammar: an undersized paradigm is 'inflated' to match the requirements of output well-formedness:

(81)

<table>
<thead>
<tr>
<th>Input: {bed \sim bed-an}</th>
<th>*VoicedCoda</th>
<th>IDENT-IO</th>
<th>IDENT-OO</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. [bet \sim bed-an]</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b. [bet \sim bet-an]</td>
<td></td>
<td>*<em>!</em></td>
<td></td>
</tr>
<tr>
<td>c. [bed \sim bed-an]</td>
<td>!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. [bed \sim bet-an]</td>
<td>!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simultaneously, we also observe a neutralization of the input paradigms (80) and (81) into a single output paradigm \[bet \sim [bed-an]\].

Up to this point, no activity of uniform exponence has been visible. But of course the grammar must allow for such activity, so as to restrict the difference in shape between allomorphs in a paradigm. Without constraints enforcing uniform exponence, any pair of input allomorphs would be able to surface (with the only restriction imposed by high-ranked well-formedness constraints).

For example, Dutch allows allomorph pairs such as \[bet \sim [bed-an]\], whereas pairs such as *\[bet \sim [pet-an]\], with voice alternations in word onset position, are systematically excluded. Such candidate output paradigms are rejected for violation of an undominated positional faithfulness constraint:

(83)

\textbf{IDENT-OO}(voice, [C])
Let \(a\) be a segment in the base, and \(\beta\) be a correspondent of \(a\) in the affixed form.
If \(a\) is in word onset position and \([yvoice]\), then \(\beta\) is \([yvoice]\). 'Corresponding segments in the word onset of the base and the affixed form must agree in voicing.'

This constraint penalizes any alternations between allomorphs with respect to voicing in the word onset. Any input paradigm is neutralized towards a surface
paradigm respecting uniform exponence with respect to word onset voicing. Since this neutralization could occur at the expense of the voicing value of either the base or the affixed form, a 'tie-breaking' constraint is required. In tableau (84), this is the low-ranked constraint VOP:

This example demonstrates the interaction of the three major types of constraints in an allomorphy model: 00-identity, 10-faithfulness, and well-formedness.

This is the second example of a lexical paradigm being overruled by the grammar, the case of an 'oversized' paradigm that is 'shrunk' to match the requirements of uniform exponence.

(85) Shrinking of paradigm

\[
\text{Input: } \{\text{bet} \sim \text{an}\} \quad \text{Output allomorph 1, lexically supplied}
\]

\[
\{\text{pet}\} \quad \text{Output allomorph 2, rejected by grammar}
\]

In sum, this allomorphically model captures systematic variation in the shapes of alternants, and it also restricts the variation in shape between alternants. That is, input paradigms are not simply 'copied' to the output, but undergo the filtering function of the grammar. This means that certain input paradigms are 'inflated', adding an allomorph to match the needs of high-ranked well-formedness constraints, while other input paradigms are 'shrunk', removing input variation between allomorphs that exceeds the limits imposed by uniform exponence.

An extension of this model may be considered under which the distribution of the allomorphs is also brought under control of the grammar. The Dutch lexicon supplies the set of two allomorphs (bet ~ bed), without marking one for insertion into 'stem' and the other for insertion into 'affixed form'. The distribution of allomorphs is fully predictable from well-formedness constraints, primarily *VoicedCodas but possibly also Inter-V-Voice, which, in spite of its low ranking in Dutch, may still exert its influence as an 'emergence of the unmarked':

Conceptually, this extension of the allomorph model is quite satisfactory, since it makes the distribution of allomorphs predictable by the same constraints which also check their shape. It remains to be seen, however, whether all types of allomorphy can be subsumed under it. Similar proposals for allomorph distribution governed by well-formedness constraints have been made by Mester (1994) and Kager (1996).

It is simply too early to evaluate the results of a theory that seeks to eliminate URs. However, some major points can be made even on the basis of the simple model outlined above. As usual, the model has pros and cons.

On the positive side, this model reduces the abstractness of lexical representation by eliminating the UR. This is a noteworthy result, since reduction of abstractness entails an increase in cognitive plausibility, reducing the role of the learner in inferring patterns that are remote from actual surface patterns. While it has been claimed in the 'classical' generative literature that such abstractness is 'just what it takes to shape an explanatory theory of alternations', any reduction of abstractness that leaves the generalizing power of the theory unaffected is very welcome. (Of course it has to be determined whether such a model indeed matches or even exceeds the explanatory power of UR theory, a point that we will return to below.)

The second point in favor of the allomorphic model is that it presents a uniform analysis of all types of alternations, productive and non-productive alike. The lexicon is the place where idiosyncratic properties of a language are listed, so nothing guarantees that every morpheme will indeed exhibit the full range of allomorphs as compared to other morphemes. For example, English is known to have exceptions to a process of trisyllabic laxing (op[et]lue ~ op[æ]lity, but ob[ɪt]se ~ ob[ɪt]sity, Kiparsky 1982b). On an allomorphic model the difference between both morphemes is simply this: opaque has two allomorphs alternating in vowel length, while obes has only one, hence no length alternation. There are no high-ranked well-formedness constraints, nor 00-identity constraints pertaining to vowel length to enforce this allomorphy. Thus complete symmetry is not expected. Or to state it differently, the grammar allows space for lexical variation, but the lexicon allots this space in the form of allomorphy. The grammar partly controls allomorphy, but does not fully predict it. (Richness of the Base was never intended to fill all the accidental gaps in the lexicon of a language.)
These advantages are counterbalanced by a number of potential problems. First, it may not be wise to lump all kinds of alternations (productive and non-productive) into a single category, given the psycholinguistic evidence. Research in speech production gives evidence for a distinction between two kinds of alternations, which can be summarized as on line production versus lexical look up. The allomorphy model claims that alternations all depend on look up, even though the grammar defines the notion of possible allomorph (as we saw above). It is unclear how distinctions of productivity between alternations can be captured in this model.

Second, related to the previous point, it may be overly optimistic to assume that a general network of uniform exponence constraints may restrict all types of allomorphy in a language. For example, if the drastic shape variation in allomorphs such as go → went is indicative of the general position of 00-correspondence constraints for consonant and vowel identity in English, then it is no longer evident that 00-correspondence does any work at all. This poses a difficult dilemma: allomorphic variation must either be radically restricted by high-ranked uniform exponence (losing the go → went case), or left more or less free (predicting random allomorphic variation all over). A solution may be found in a stratified model, in which different parts of the lexicon set their own specific balance of faithfulness and uniform exponence. However, it is far from clear whether this is feasible.

9.7 Conclusion: future perspectives

At the end of the final chapter of this book, it is time to sketch some future perspectives of OT, partly based on developments sketched in the previous sections.

One development that can be foreseen is that an even greater role will be assigned to surface forms in explaining alternations, moving away from abstract underlying forms. This emphasis on surface forms has been a cornerstone of OT in dealing with phenomena as diverse as contrast, neutralization, allophone distribution, conspiracies, overapplication and underapplication (in reduplication and stem-based morphology), and transderivational ‘cyclic’ dependencies. Extrapolation of this development may lead to the elimination of underlying representations in favour of an ‘allomorphic’ model, as we have seen in section 9.6. Grammatically adequate and cognitively plausible allomorphic models may be developed which overcome the problems mentioned there. More on a speculative note, allomorphy may also offer an alternative account of opacity, a phenomenon for which current OT has no genuinely surface-based analysis.

A second major development that is likely to continue is a reduction of the role of representations in favour of constraint interaction. Underspecification, once a cornerstone of theories of segment inventories and assimilation, has come under severe pressure from OT analysts (starting with Smolensky 1993, and elaborated by Inkelas 1995, Itô, Mester, and Padgett 1995, and Steriade 1995b) who argue that constraint interactions offer superior accounts. Similar reductionist arguments were made against representational assumptions of feature geometry (Padgett 1995, Ní Chiosáin and Padgett 1997) and syllable structure (Steriade 1995b).

An issue closely related to that of representations concerns the boundary between phonology and phonetics. In work by Steriade, Flemming, Kirchner, Hayes, and others, it is argued that constraints should be able to refer to much more phonetic detail (including non-contrastive features and numerical values of acoustic parameters) than is allowed on classical generative assumptions, which maintain a strict separation between phonology and phonetics. (An illustration of this development is the title of Flemming’s 1995 UCLA dissertation: ‘Auditory representations in phonology’.) This blurring of the phonology–phonetics boundary goes hand in hand with an increased role for functional explanations. For example, Jun (1995) and Steriade (1995b), in the footsteps of phoneticians such as Lindblom, Kohler, and Ohala, argue that speakers make more effort to preserve features in contexts where they are most salient. The articulatory and perceptual basis of phonological constraints is emphasized by Archangeli and Pulleyblank (1994), Gafo (1996), Hayes (1996a), Myers (1997b), and others. No doubt, real progress can be made by this approach. An increased role for functional explanations in grammatical theory matches well with a major goal of OT, which is to encode directly markedness in the grammar, an enterprise that has been crucial to OT’s typological achievements. The merging of the phonetic and phonological components into a single hierarchy of constraints comes tantalizingly close to a single-step mapping, another authentic goal of OT.

However, freely mixing phonology and phonetics comes at a certain cost. There is overwhelming support for a level coinciding with the output of phonology and the input to phonetic interpretation. First, while phonological specifications are categorical (that is ‘on’ or ‘off’), phonetic specification of non-distinctive features is gradient and based on interpolation (Keating 1988, Cohn 1993a; see Kirchner 1997 for a theory of contrast in a one-step mapping model). Second, there is surprisingly little evidence that grammatical generalizations refer to non-distinctive feature specifications. Third, a relatively abstract (phonetically underspecified) output level of the phonology is confirmed as a cognitive reality by various kinds of independent evidence (language games, speech errors, second language acquisition, etc.: Fromkin 1971, Stemberger 1991). Apart from the evidence for levels, there is yet another potential stumbling block for a direct reductionist approach. A theory reducing phonology to interactions of raw functional (perceptual and articulatory) factors has difficulties in accounting for symmetry as a property of phonological systems, since such factors, by definition, interact in a gradient
fashion. (But see Hayes 1996a for a theory of learning of symmetrical patterns on the basis of gradient information.) In sum, a theory which gives up the separation of phonology and phonetics holds much explanatory potential, but also faces two fundamental problems in dealing with cognitively plausible levels of representation.

When numerical values (of phonetic parameters) are incorporated into phonology, it seems only logical to take another step into numerical (gradient) types of interaction, so that constraints exert influence according to their relative strength, expressed in an index. In several constraint-based models (for example, Harmonic Grammar, Legendre, Miyata, and Smolensky 1990), numerical ranking indeed replaces the principle of strict domination, partly bridging the gap with connectionism (Rumelhart and McLelland 1986, Goldsmith and Larson 1990). It is not clear at all whether an OT model based on numerical ranking can preserve the advantages of standard OT. For example, the phenomenon of ‘the emergence of the unmarked’ depends on strict domination. For discussion see Prince and Smolensky (1993, chapter 10).

In sum, in the foreseeable future phonology may look quite different from what it is today. It has been suggested that there are some striking resemblances between OT and Structuralism (with respect to views of allophonic patterns and contrast, surface patterns, functional considerations, and allomorphy), the dominant framework in the pre-generative era. Is history circular? Perhaps, but if so, there are important differences from the earlier days. Due to the generative legacy, there remains a strong emphasis on formal precision in grammatical analysis, combined with the necessity to restrict the descriptive power of linguistic theory. Both theoretical priorities are solidly integrated into OT. Moreover, in-depth phonological analyses of many languages have given us a much better insight into cross-linguistic tendencies and typology than we had half a century ago. The emphasis on formal accuracy and explanatory adequacy, coupled to a still expanding typological basis, provides the background that is necessary to move successfully forward into areas that were previously considered to be ‘too surface-oriented’ for phonologists to deal with. The concept of grammar is strong enough to survive this continuing enterprise.

### SUGGESTIONS FOR FURTHER READING

#### Opacity and OT


#### Local Conjunction


#### Absolute ungrammaticality


#### Free variation


#### Positional faithfulness


#### Allomorphy
Residual issues


Functional approaches to phonology


REFERENCES


