ii. The mapping /a/ → [x] is less faithful than /a/ → [b]. That is, the highest-ranking constraint that distinguishes them is a faithfulness constraint favoring /a/ → [b] over /a/ → [x].

These requirements derive from the basic structure of the theory; ultimately, all can be understood as consequences of the Cancellation/Domination Lemma ((15) in §1.3), which itself follows from EVAL.

Clauses (1a–b) say that any unfaithful mapping requires the basic ⟦M ≫ F⟧ ranking. The markedness constraint must dominate *some* relevant faithfulness constraint, because unfaithfulness is never gratuitous; rather, it is always the price paid for concomitant improvement in markedness (§3.1.4.5). Clause (1b) also excludes the blocking configuration ⟦C_F ≫ M ≫ F⟧ (such as (14) in §1.3), where M dominates some but not all faithfulness constraints that the unfaithful mapping /a/ → [b] violates.

As was just noted, the only reason for an unfaithful mapping to be optimal is if it does better than the faithful candidate on the markedness constraints as they are ranked in the language under investigation. A fortiori, an unfaithful mapping cannot be optimal if it produces worse performance on the ranked markedness constraints. That is the import of clause (1c): if unfaithful [b] is more marked than faithful [a], then the unfaithful mapping cannot be more harmonic than the faithful one. This clause also excludes a blocking configuration, the ⟦C_M ≫ M ≫ F⟧ ranking exemplified by (13) in §1.3.

Clause (1d) recognizes the effects of homogeneity of target/heterogeneity of process (§1.3.2, §3.1.4.2). Many unfaithful mappings could in principle satisfy the markedness constraint M, but the mapping to [b] is the one that is actually observed. The various alternative mappings must be less harmonic because they involve candidates that are more marked or less faithful than [b] (again, relative to the language's particular constraint hierarchy). A concrete example, where the alternatives are less faithful, can be found in (13) and (14) of §1.3.

In summary, all of these requirements must be met to guarantee that /a/ maps most harmonically onto [b]. If (1a–c) do not hold, then /a/ will map faithfully to [a]. If (1d) is not met, then /a/ will receive unfaithful treatment, but it will end up as something other than [b].

3.1.2 Inventories and Richness of the Base

3.1.2.1 Basic Concepts

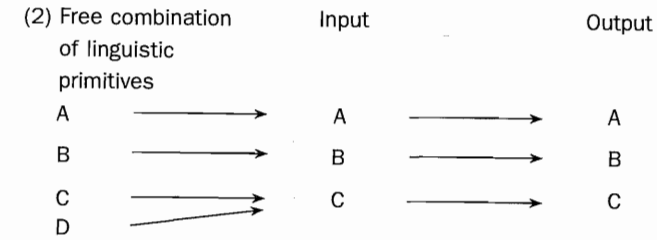
As used here, the term *inventory* refers to the set of linguistic objects that are permitted in the output representations of a language. It is often useful to speak of the inventory of objects of some specific type, such as the inventory of vowels in English or the inventory of clitic pronouns in Spanish. Some members of an inventory may have a restricted *distribution*, meaning that they are limited to (or prohibited from) appearing in certain contexts. The theory of

inventories in OT is the topic of this section, and distributional restrictions are addressed in §3.1.3.

These terms are probably used more in phonology than in syntax, but the underlying concepts are relevant throughout linguistics. In every language and at every level of analysis – phonological, morphological, or syntactic – there are limitations on what elements are permitted in surface structure and where they are permitted. Any linguistic theory needs a way of accounting for these observations and the associated typological generalizations (§3.1.5).

An observed inventory restriction can be described schematically as follows. Suppose that the free combination of primitive linguistic objects (e.g., phonological or morphosyntactic features) allows for the four-way distinction A/B/C/D. But in the language under investigation, only the three-way distinction A/B/C is actually observed in surface structures. The inventory of this language is restricted by the absence of D. In principle, this gap could be accidental, like missing *blick* in English, but let us suppose further that familiar criteria like productivity tests or typological consistency show that it is not.

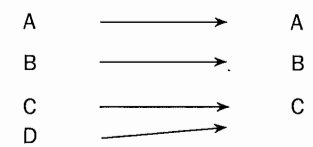
In both phonology and syntax, inventory restrictions have usually been analyzed by imposing a filter on the input side, barring D from the lexicon or other source of inputs. Lexical redundancy rules, morpheme structure constraints, or simply the lexicon itself impose structure on the freely combined linguistic primitives that we see in (2).



If the input is identified with the lexicon, we would say that the lexicon of this language systematically fails to exploit an option that UG supplies. Other languages may, of course, differ on this point by including D in the lexicon. This is a standard way to account for between-language variation in both phonology and syntax; in fact, according to one view (Chomsky 1993), this might be the *only* way of accounting for between-language variation. Some examples from English: the lexicon is subject to a phonological redundancy rule prohibiting front rounded vowels (*ü*, *ö*); the lexicon lacks a Q element, and so *wh*-phrases must be fronted (Chomsky 1995:69).[2]

Most work in OT, however, recognizes no distinction between the free combination of linguistic primitives and the input. D is absent from surface structure because input D is unfaithfully mapped to something else – either some other member of the inventory or the null output (§4.1.2). For the purpose of discussion, assume that D is mapped to C, as in (3).

(3) Free combination of linguistic primitives = Input → Output



The absence of D from surface forms is here a consequence of the unfaithful mapping of D to C, which absolutely neutralizes any possible distinction between them. Examples like this can be found throughout this book, such as (14) in §3.1.2.5 (morphosyntax), (21) in §3.1.3.5 (phonology), and (4) in §4.1.3 (syntax).

The hypothesis that the free combination of linguistic primitives and the input are identical is called *richness of the base* (ROTB).[3] 🕮§3.4¶1 Equivalently, ROTB says that there are no language-particular restrictions on the input, no linguistically significant generalizations about the lexicon, no principled lexical gaps, no lexical redundancy rules, morpheme structure constraints, or similar devices. All generalizations about the inventory of elements permitted in surface structure must be derived from markedness/faithfulness interaction, which controls the faithful and unfaithful mappings that preserve or merge the potential contrasts present in the rich base.

This material presents abundant opportunities for terminological confusion, which I here attempt to sort out. The *inventory* of a language is the set of permitted surface structures. Except for accidental gaps, the observed inventory of a language should exactly match the output of EVAL for that language's constraint hierarchy. The inventory, then, is derived or *emerges* from applying the hierarchy to a set of inputs. The *base* is the universal set of inputs. If a language's constraint hierarchy has been correctly analyzed, then applying GEN and EVAL to any input chosen from the universal base will yield some surface structure in that language's inventory. The *lexicon* should really be called the *vocabulary*: because of accidental gaps, the observed inventory is a proper subset of the inventory that emerges from EVAL. The grammar is responsible for systematic gaps (*bnick* is not a possible word of English) but not for accidental ones (*blick* is not a word of English), nor does the grammar purport to explain accidental properties of lexical meaning (*brick* is an object made of clay). These accidental properties are recorded in the lexicon, which, however, lacks the internal principles familiar from other theories of phonology and syntax. Hence the term vocabulary. There will be much more about all of this in §3.1.2.

ROTB is a natural consequence of one of the central ideas in OT – that languages differ only in constraint ranking. It is, moreover, the most parsimonious hypothesis: the input (the base or lexicon) freely combines the primitive elements of linguistic representation, and then the grammar, which is needed anyway, reduces this profusion to the observed inventory.

3.1.2.2 *The Duplication Problem, with Examples*

Apart from the conceptual arguments for assuming ROTB, there is also a powerful empirical reason that was first recognized in phonological research of the 1970s. The *Duplication Problem* is a particular kind of conspiracy (§2.1).🕮§3.4¶2 Merely eliminating D from the lexicon, as in the standard theory (2), is not enough to ensure the absence of D from the inventory. It is also necessary to take precautions against any rules of the grammar creating D's in the mapping from the lexicon to surface structure. Consider, for example, a language that, like English, has no front rounded vowels (*ü*, *ö*) in its inventory. According to the standard model, this language has the lexical redundancy rule in (4a). Now suppose that the same language also has a fronting or umlaut rule that changes *u* and *o* into *i* and *e*, respectively, when the next syllable contains *i*. That rule is exemplified in (4b) and formulated in (4c).

(4) a. Lexical redundancy rule
if [–back], then [–round]
b. Fronting rule exemplified
/put + i/ → piti
/kop + i/ → kepi
c. Fronting rule

$$V \rightarrow \begin{bmatrix} -\text{back} \\ -\text{round} \end{bmatrix} / \_\_\_\_ C_0\ i$$

There is a correlation here: *ü* and *ö* are banned from the inventory of this language, and the fronting rule produces *i*'s and *e*'s rather than *ü*'s and *ö*'s. This correlation is surely not an accident, yet it is entirely unexplained in the standard theory (2). Formally, the problem is that the [–round] specification in the output of the fronting rule duplicates the [–round] specification in the consequent of the lexical redundancy rule. Since the standard theory equates simplicity with naturalness under the Evaluation Metric (§2.1), a simpler and putatively more natural fronting rule would change *u* into *ü* and *o* into *ö*. In other words, the fronting rule must be made more complicated and consequently less natural looking, to bring it into accord with the lexical redundancy rule. This is an instance of the Duplication Problem: a lexical redundancy rule and a rule of the phonology act together in service of the same surface target.

The Duplication Problem, then, is the observation that rules of grammar often duplicate in their dynamic mappings the restrictions that are imposed statically by lexical redundancy rules. "In many respects, [lexical redundancy rules] seem to be exactly like ordinary phonological rules, in form and function" (Chomsky and Halle 1968: 382). ROTB avoids the Duplication Problem simply by denying that there is any such thing as a lexical redundancy rule or the equivalent. ROTB recognizes no distinction between the mappings that enforce static inventory restrictions and those that produce dynamic alternations; the Duplication Problem shows that this distinction is in any case illusory.

In OT, a single markedness constraint, suitably ranked, is responsible for the absence of *ü*'s and *ö*'s from the inventory of the language in (4). This constraint, call it FRT/*RND, is identical to the lexical redundancy rule (4a), but it evaluates outputs rather than inputs. If FRT/*RND is to compel unfaithful mappings, as it must if it is to affect the inventory, it must be ranked above some faithfulness constraint. Suppose that the low-ranking faithfulness constraint is IDENT(back), which requires input and output to agree in their values for the feature [back]. If FRT/*RND dominates IDENT(back), and if in addition IDENT(round) dominates IDENT(back) (cf. (1c)), then input /tük/ will map unfaithfully to *tuk*, as illustrated in (5).

(5) FRT/*RND, IDENT(round) ≫ IDENT(back)

/tük/	FRT/*RND	IDENT(round)	IDENT(back)	Remarks
a. ☞ tuk			*	Backing of /ü/
b. tük	*			Faithful
c. tik		*		Unrounding of /ü/

The rich base freely combines all of the elements of phonological representation, so it must contain the input /tük/ even if the surface inventory does not. (This is not to say that /tük/ is a literal underlying representation in the vocabulary of this language. See §3.1.2.4.) But /tük/ is mapped unfaithfully to *tuk*, so one possible source of *ü*'s in the inventory is thereby foreclosed. With the opposite ranking, putting faithfulness on top, /tük/ will survive to the surface unscathed. That is the situation in German, which does allow *ü* in its inventory and has a three-way surface contrast among *i*, *u*, and *ü*.

Input /ü/ is not the only possible source of output *ü*, since the effects of the fronting process must be contended with. Assume for the sake of discussion an ad hoc markedness constraint FRONT that penalizes back vowels before *i* (as in (6)). It is ranked above the faithfulness constraint IDENT(back).

(6) FRONT ≫ IDENT(back)

/put + i/	FRONT	IDENT(back)	Remarks
a. ☞ piti		*	Fronting of /u/ before /i/
b. puti	*		Faithful

The interesting action in OT involves constraint interaction, and this case is no exception. We already know from (5) that FRT/*RND dominates IDENT(back).

If, in addition, FRONT dominates IDENT(round), as it does in (7), then the fronting process will correctly yield *piti* rather than *püti*.

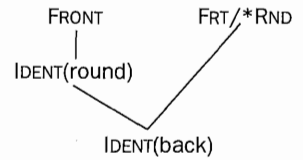
(7) FRONT ≫ IDENT(round)

/put + i/	FRONT	FRT/*RND	IDENT(round)	IDENT(back)
a. ☞ piti			*	*
b. puti	*			
c. püti		*		*

The optimal form in (7) violates both of the low-ranking faithfulness constraints, but its competitors do worse.

The rankings required in this language are summarized in (8).

(8)



Diagrams like this are probably the best way to summarize the accumulated inferences about the constraint hierarchy of a language. It shows the constraints in a partial ordering, which is often all that can be determined (§1.1.2). Higher-ranking constraints are written at the top. If there is a strictly downward path between two constraints, then the higher one dominates the lower. For example, FRONT dominates the two faithfulness constraints in (8). If there is no strictly downward path between two constraints, then no ranking between them has been established. That is the case with FRT/*RND and all other constraints except IDENT(back). Depicting the constraint hierarchy by flattening it out, though unavoidable in a tableau like (7), loses this fine structure and can be misleading.

Taken together, tableaux (5) and (7) show that the same constraints in the same hierarchy are responsible for the static restriction – *ü* and *ö* do not occur in underived contexts – and for the dynamic one – *ü* and *ö* are not created by applying processes. There is no Duplication Problem, because the observed inventory restriction is accounted for once and only once in the grammar since there are no language-particular restrictions on inputs. The overall idea is that static and dynamic restrictions on inventories have the same source as each other and as all other aspects of between-language variation in OT: the interaction of markedness and faithfulness constraints.

ROTB does not deny the possibility of *universal* restrictions on input. Like putative restrictions on GEN (§1.1.3), though, they should be approached skeptically. It undoubtedly makes sense to impose some very general restrictions on inputs, such as providing a universal alphabet of phonological or morphosyntactic features. But there are alternative interpretations of many narrower

restrictions. For example, because no known language has a contrast in syllabification between tautomorphemic *pa.ta* and *pat.a* or *pa.kla* and *pak.la*, it is often proposed that syllabification is universally absent from underlying representations.[4] But OT offers a different approach to this observation. Suppose that CON has no constraints demanding faithfulness to syllable structure. Markedness constraints will fully determine the syllabification of every input, without interference from faithfulness constraints. The non-contrastiveness of syllabification follows from this: inputs can contain all the syllabification they want, or none, or something in between, but no input syllabification will have any influence on the surface outcome if there are no syllabic faithfulness constraints to transmit that influence.

Before going on to look at a syntactic example, we need to consider an alternative solution to the Duplication Problem found in the phonological literature. Global rules or derivational constraints (Kisseberth 1970a, 1970b), linking rules (Chomsky and Halle 1968), persistent rules (Chafe 1968: 131; Halle and Vergnaud 1987: 135; Myers 1991), and underspecification (Archangeli 1984; Kiparsky 1981) share a common approach to languages like (4): details aside, they give the lexical redundancy rule (4a) a special durable status, so that it can "fix up" *ü*'s and *ö*'s, regardless of their source, by changing them into *i*'s and *e*'s. On this view, the fronting rule produces the prohibited segments *ü* and *ö* (or their underspecified counterparts), but the durable fix-up rule is immediately triggered, further changing them to *i* and *e*. Similar ideas are also common in syntactic analysis, though the fix-up theory itself (e.g., the mapping from S-Structure to PF) has until recently received less attention than its phonological counterparts.

The problem with the fix-up approach is that it accounts only for situations like (4) where the inventory restriction has a triggering effect (§1.3, §2.1, §2.3, §3.1.4). But inventory restrictions can also have blocking effects, stopping a process from applying when its output would escape from the licit inventory. Kiparsky (1982a, 1982b) calls this property structure preservation because of its resemblance to the syntactic principle with that name (Emonds 1970). In OT, whether a markedness constraint has a triggering effect, as in (7), or a blocking effect, as in (9), is a matter of interaction.

(9) Blocking from IDENT(round) ≫ FRONT

/put + i/	FRT/*RND	IDENT(round)	FRONT	IDENT(back)
a. piti		*		*
b. ☞ puti			*	
c. püti	*			*

By swapping the ranking of FRONT and IDENT(round), two distinct interactional possibilities are realized. With the ranking [[FRONT ≫ IDENT(round)]] in (7), the fronting process can go ahead even when it leads to unfaithfulness in rounding as a consequence of further interaction with FRT/*RND. With the opposite ranking in (9), the fronting process cannot proceed under those conditions; it is simply blocked. Both types of interaction are well attested, and no criteria have been discovered that can consistently predict whether a given inventory restriction will apply in triggering or blocking mode. From the OT perspective, this is exactly as expected: blocking versus triggering is a matter of constraint ranking, and constraint ranking differs across languages.

The Duplication Problem does not seem to have been recognized in the pre-OT syntactic literature, but it has figured in applications of OT to syntax. Grimshaw (1997b: 409) discusses a couple of situations where independently necessary properties of the grammar render a parallel lexical restriction superfluous. For example, English has no complementizer in embedded questions: **I wonder who **that** he saw*. This fact is standardly taken to mean that the English lexicon lacks [+wh] complementizers, but that kind of systematic, language-particular restriction on inputs is incompatible with ROTB. In OT, then, the grammar must supply the explanation for the impossibility of complementizers in embedded questions, and indeed it does. In English, heads are usually at the left edge of their phrases, unless some higher-ranking constraint compels minimal displacement (see (74) in §3.3.1). This observation shows that the edge Alignment constraint (§1.2.2.) HEAD-LEFT dominates its symmetric counterpart HEAD-RIGHT. The complementizer *that*, as head of CP, will maximally satisfy HEAD-LEFT if it is at the left edge of CP. But in embedded questions, there is a *wh*-word at the left edge of CP, so perfect satisfaction of HEAD-LEFT is not possible. In consequence, English has no complementizer at all, because HEAD-LEFT dominates OB-HD (which, short for obligatory heads, requires every projection to have a head – see (26) in §3.1.3.6).

(10) HEAD-LEFT ≫ OB-HD

	HEAD-LEFT	OB-HD
a. ☞ I wonder [$_{CP}$ who he saw		*
b. I wonder [$_{CP}$ who that he saw	*	

By deriving this observation from the grammar, as in (10), it is related to English's general left-headedness. Compare this to theories that simply say the lexicon lacks [+wh] complementizers. This is a covert instance of the Duplication Problem: the lexicon stipulates something that could be explained in other terms. The absence of a complementizer for embedded questions from the English inventory is a fact about the grammar, not the lexicon. The rich base provides such a complementizer, but the grammar rejects it.

In summary, observed inventory restrictions are a consequence of unfaithful mappings that neutralize potential distinctions present in the rich base. The

In P&P, the lexicon is seen as the locus of all or almost all that is language particular, with the grammar tending much more strongly toward the universal. (This is an expansion of the *Aspects* thesis that the lexicon is the repository of irregularity.) In OT, however, the perspective is rather different: the base is universal, and it is the grammar – a ranking of universal constraints – that filters this rich base to yield the surface inventory. The implications of this idea for syntactic theory are now just beginning to be studied (see the references in §3.4 ¶1).

3.1.2.6 Summary and Methodological Remark

The inventory is an *emergent* property of OT grammars (cf. §3.2.2). No language-particular restrictions are imposed on the input, and so all linguistically significant generalizations about the inventory must emerge from the grammar itself. This hypothesis is necessary if OT is to be true to its boast that all typology comes from language-particular ranking and if it is to solve the Duplication Problem. Significantly, it also means that the theory of inventories is tightly integrated into OT's account of other emergent properties: distribution (§3.1.3), processes and their interaction (§3.1.4), and typological distinctions and universals (§3.1.5). This leads to a range of claims and predictions that cannot be matched by theories that attribute some or all inventory restrictions to the lexicon and not the grammar.

ROTB is a central property of this explanation; it follows from the basic architecture and typological claims of the theory. For this reason, all OT analyses need to be tested against a range of inputs that have not been restricted artificially. This methodology can seem quite alien to anyone approaching OT with a background in P&P, *SPE*, or underspecification theory – but it is nonetheless essential. The only workable research strategy is to integrate ROTB into the analysis from the outset and not attempt to graft it on at the end.

3.1.3 Distributional Restrictions

3.1.3.1 Basic Notions

Restricted distribution is another classic linguistic problem, dating back to the time of the American structuralists. The distribution of an item is the set of linguistically relevant contexts in which it appears: syllable-initially, but never syllable-finally (e.g., *h* in English); alone, but never with an auxiliary (e.g., *do*-support in English). Discussions of distribution usually focus on the relative distribution of two items, A and B, one of which may be zero. The distribution of A and B can be usefully classified by comparing the contexts C_A and C_B in which they occur, as in (15).

(15) Types of distributional restrictions

	Characterization	Description
a. Identical distribution	$C_A = C_B$	A and B have identical distributions (except for accidental gaps). The A/B distinction is maintained in all contexts where they occur.
b. Complementary distribution	$C_A \cap C_B = \emptyset$	A and B never occur in the same context. The A/B distinction is neutralized everywhere.
c. Contextual neutralization	$C_A \subsetneq C_B$	There are contexts that permit both A and B, but there are also contexts that permit only B. The A/B distinction is neutralized contextually in favor of B.

There are other distributional possibilities, but these are the most important ones.

For concreteness, we will examine the effects of oral and nasal consonants (*b* vs. *m*) on the distribution of following oral and nasal vowels (*a* vs. *ã*).[10] To simplify the exposition, several artificial limitations will be imposed: input forms and output candidates will be limited to the set {*ba*, *bã*, *a*, *ã*, *ma*, *mã*}; the only unfaithful mapping emitted by GEN will be a change in the nasality of a vowel, /a/ → *ã* or /ã/ → *a*; and CON will consist of the constraints $*V_{NAS}$, $*NV_{ORAL}$, and IDENT(nasal). $*V_{NAS}$ and $*NV_{ORAL}$ are markedness constraints. The justification for $*V_{NAS}$ is a classic Praguian markedness effect: some languages have oral vowels only, some have both oral and nasal vowels, but no language has only nasal vowels. From this we infer that UG contains a markedness constraint militating against nasal vowels – $*V_{NAS}$ – and that there is no corresponding constraint against oral vowels.[11] But in certain contexts, even oral vowels are marked, such as the position after a nasal consonant, and that is the basis for the constraint $*NV_{ORAL}$. It is violated by any sequence like *ma* (vs. *mã*).

For present purposes, I assume that the faithfulness constraint IDENT(nasal) is symmetric, meaning that it is violated by both of the unfaithful mappings /a/ → *ã* and /ã/ → *a*. Recall that these are the only unfaithful mappings permitted by our artificially limited GEN, so mappings like /b/ → *m* or /m/ → ∅ will not even be considered. (Obviously they would be addressed in a more complete analysis.)

3.1.3.2 Factorial Typology

With the modest constraint set $*V_{NAS}$, $*NV_{ORAL}$, and IDENT(nasal), it is possible and desirable to begin by computing the typology that is predicted by ranking permutation. There are just 3! = 6 possibilities, of which two pairs produce identical results, as shown in (16).

(16) Factorial typology of *V_{NAS}, *NV_{ORAL}, and IDENT(nasal)

	Ranking	Inventory
a. Identical distribution	IDENT(nasal) ≫ *NV_{ORAL} ≫ *V_{NAS} IDENT(nasal) ≫ *V_{NAS} ≫ *NV_{ORAL}	{*ba, bã, a, ã, ma, mã*}
b. Complementary distribution	*NV_{ORAL} ≫* V_{NAS} ≫ IDENT(nasal)	{*ba, a, mã*}
	*V_{NAS} ≫ *NV_{ORAL} ≫ IDENT(nasal) *V_{NAS} ≫ IDENT(nasal) ≫ *NV_{ORAL}	{*ba, a, ma*}
c. Contextual neutralization	*NV_{ORAL} ≫ IDENT(nasal) ≫ *V_{NAS}	{*ba, bã, a, ã, mã*}

We will consider each of these distributions in turn.

3.1.3.3 *Identical Distribution*

If A and B have identical distributions, then nothing about the A/B distinction is predictable. Unpredictability is a sure sign of activity by faithfulness constraints, which act to protect the free combination of elements that make up the rich input.

Concretely, nasal and oral vowels have identical distribution when faithfulness stands at the zenith of the hierarchy (16a). With IDENT(nasal) top ranked, the markedness constraints cannot be active over the candidate sets being considered here. In that situation, all of the potential contrasts present in the rich base arrive at the surface unscathed (see (17)).

(17) Mappings for identical distribution

/ba/ → *ba*
/bã/ → *bã*
/a/ → *a*
/ã/ → *ã*
/ma/ → *ma*
/mã/ → *mã*

Because all the mappings are faithful, the relative markedness of *ma* versus *mã* is never an issue, and so the ranking of *NV_{ORAL} with respect to *V_{NAS} cannot be determined. For that reason, (16a) includes two different permutations that produce identical results.

3.1.3.4 *Complementary Distribution*

Complementary distribution of A and B means full predictability, with lack of respect for faithfulness when it conflicts with markedness. Two different situations of complementary distribution are covered by the rankings in (16b). The first, with faithfulness at the bottom and *NV_{ORAL} dominating *V_{NAS}, is the classic situation of complementary distribution: nasalized vowels occur only when needed, where "when needed" is defined by the context-sensitive markedness constraint *NV_{ORAL}. The second ranking in (16b) – actually a pair of rankings

that produce identical results – is a trivial situation of complementary distribution, with oral vowels occurring in all contexts and nasal vowels in no context.[12]

Here is an example of nontrivial complementary distribution from Madurese (Austronesian, Java), slightly simplified.[13] Oral and nasal vowels are in complementary distribution: oral vowels occur only after oral consonants (*ba*, **bã*) or after no consonant at all (*a*, **ã*); nasal vowels occur only after nasal consonants (*nã*, **na*). The rich base includes, as inputs, both the forbidden and the permitted output forms: /ba/, /bã/, /a/, /ã/, /ma/, and /mã/. Example (18) presents the required mappings.

(18) Mappings in Madurese

/ba/ ↘
/bã/ ↗ *ba*

/a/ ↘
/ã/ ↗ *a*

/ma/ ↘
/mã/ ↗ *mã*

As in §3.1.2, (18) shows mappings to the output from a rich base and not overt alternations. This diagram establishes that nasality and orality in Madurese vowels are fully under grammatical control, which means that markedness constraints are dispositive in all contexts, so faithfulness inevitably suffers. Two unfaithful mappings are observed from the rich base: /a/ → *ã* after a nasal consonant and /ã/ → *a* elsewhere. The analysis must contend with both.

At this point, readers might find it helpful to review the conditions for an unfaithful mapping given in (1) of §3.1.1, keeping in mind that complementary distribution requires two unfaithful mappings. The first condition (1a) says that unfaithful mappings are based on ⟦M ≫ F⟧ rankings. Two such rankings, shown in (19a–b), are relevant to Madurese.

(19) a. *V_{NAS} ≫ IDENT(nasal)

/bã/	*V_{NAS}	IDENT(nasal)
i. ☞ ba		*
ii. bã	*	

b. *NV_{ORAL} ≫ IDENT(nasal)

/ma/	*NV_{ORAL}	IDENT(nasal)
i. ☞ mã		*
ii. ma	*	

The rich base includes inputs like /bã/ or /ma/, with the "wrong" vowel. These inputs must be unfaithfully mapped to outputs that conform to the observed distribution. The tableaux in (19) exemplify that, under ROTB, complementary distribution requires at least two unfaithful mappings.[14]

Another of the prerequisites for an unfaithful mapping, (1b), says that any unfaithful mapping must result in an improvement in markedness relative to the markedness constraints of UG as ranked in the language under discussion. Mapping /bã/ to *ba* is an obvious markedness improvement, as (19a) shows, but mapping /ma/ to *mã* must also improve markedness. The argument in (20) shows how the two markedness constraints are ranked in Madurese.

(20) $*NV_{ORAL} \gg *V_{NAS}$

/ma/	$*NV_{ORAL}$	$*V_{NAS}$
a. ☞ mã		*
b. ma	*	

Because $*NV_{ORAL}$ dominates $*V_{NAS}$, *mã* is less marked than *ma* in this language. Of course, with the opposite ranking of these constraints, there would be no nasalized vowels in any context.

The remaining ranking prerequisite, (1c), says that a specific unfaithful mapping is guaranteed only if all other ways of satisfying the relevant markedness constraint(s) are foreclosed by other constraints. For expository purposes, I have assumed that GEN offers no other options, but in real life we would need to call on appropriate markedness or faithfulness constraints to rule out mappings like /bã/ → *b* or /ma/ → *ba*.

That completes the picture of Madurese. With the ranking $[\![*NV_{ORAL} \gg *V_{NAS} \gg$ IDENT(nasal)$]\!]$, nasalized vowels appear only when needed, as demanded by top-ranked $*NV_{ORAL}$. Otherwise, vowels are oral, in obedience to $*V_{NAS}$. Orality, then, is the default (§3.2.2.3).

To sum up the essence of complementary distribution: the $[\![M \gg F]\!]$ rankings must be sufficient to dispose of all A's occurring in B's context and all B's occurring in A's context. Suppose CON supplies two markedness constraints, M(A ≻ B) and M(B ≻ A), where M(X ≻ Y) means "M favors X over Y; M assigns fewer violation-marks to X than Y." Often, these two constraints will conflict. M(A ≻ B) favors A over B generally, with M(B ≻ A) favoring B over A in a specific context (like $*V_{NAS}$ and $*NV_{ORAL}$, respectively). Nontrivial complementary distribution will only be achieved if faithfulness is at the bottom and contextually restricted M(B ≻ A) dominates context-free M(A ≻ B). In traditional terms, A is the default relative to B: B occurs in limited circumstances, as defined by M(B ≻ A), and complementarily A occurs everywhere else.📖§3.4¶6

Keep in mind that, since both A and B are present in the rich base, where they can be found in all possible contexts, complementary distribution requires *both* /A/ and /B/ to be unfaithfully mapped in some contexts. This can be a source of confusion, since traditional approaches to complementarity take various precautions, such as lexical redundancy rules or underspecification, to make sure that the contrast between /A/ and /B/ is not available in the lexicon (§3.1.2).

3.1.3.5 *Contextual Neutralization*

A contrast may be preserved in some contexts and neutralized in others. The responsible ranking, as in (16c), is one where the specific, context-sensitive constraint M(B≻A) is ranked above faithfulness, while the general, context-free constraint M(A≻B) is ranked below faithfulness. Yoruba (Niger-Congo, Nigeria) is the contextually neutralized counterpart of Madurese, and its mappings are shown in (21).

(21) Mappings in Yoruba

/ba/	→	*ba*
/bã/	→	*bã*
/a/	→	*a*
/ã/	→	*ã*
/ma/	↘	*mã*
/mã/	↗	

The distinction between oral and nasal vowels is neutralized in the context after a nasal consonant, but it is faithfully preserved elsewhere.

Because /bã/ and /ã/ are mapped faithfully in Yoruba, it is immediately apparent that IDENT(nasal) dominates $*V_{NAS}$. The real action, then, involves the ranking of $*NV_{ORAL}$ relative to faithfulness. Since /ma/ maps unfaithfully to *mã*, the ranking must be as in (22).

(22) $*NV_{ORAL} \gg$ IDENT(nasal)

/ma/	$*NV_{ORAL}$	IDENT(nasal)
a. ☞ mã		*
b. ma	*	

With $*V_{NAS}$ ranked below IDENT(nasal), the latent *a/ã* contrast is maintained, except in a postnasal context, where the demands of $*NV_{ORAL}$ take precedence.

The analysis of Yoruba just sketched typifies one kind of approach to contextual neutralization. It is based on positing context-sensitive markedness constraints like $*NV_{ORAL}$. Another, less obvious line of attack is to derive context sensitivity from interaction of context-free markedness constraints with

positional faithfulness constraints.📖§3.4 ¶7 The central idea behind positional faithfulness is that faithfulness constraints may be relativized to certain prominent positions, such as stressed syllables or root morphemes.

Here is an example. In Nancowry (Austro-Asiatic, Nicobar Islands) stress falls predictably on the final syllable of the root. In stressed syllables, there is a contrast between nasal and oral vowels, but in unstressed syllables, all vowels are oral. The required mappings, faithful and unfaithful, are in (23).[15]

(23) Mappings in Nancowry

/bata/	→	*batá*
/batã/	→	*batã́*
/bata/	↘	*batá*
/bãta/	↗	

In other words, an output oral vowel in a stressed syllable can be reliably projected back to an input oral vowel, but an oral vowel in an unstressed syllable cannot. Stressed syllables and other prominent positions are a locus of particularly robust faithfulness.

The idea is that UG distinguishes between a stressed-syllable-specific version of IDENT(nasal) and its nonspecific counterpart. (This is a stringency relation – see §1.2.3.) In Nancowry, these constraints are ranked with $*V_{NAS}$ in between: [[$\text{IDENT}_{\acute{\sigma}}$(nasal) ≫ $*V_{NAS}$ ≫ IDENT(nasal)]]. In this way, $*V_{NAS}$ can compel unfaithfulness to nasality in unstressed syllables, but not in stressed ones, as we see in (24a–b).

(24) a. /batã/ → *batã́*

/batã/	$\text{IDENT}_{\acute{\sigma}}$(nasal)	$*V_{NAS}$	IDENT(nasal)
i. ☞ batã́		*	
ii. batá	*		*

b. /bãta/ → *batá*

/batã/	$\text{IDENT}_{\acute{\sigma}}$(nasal)	$*V_{NAS}$	IDENT(nasal)
i. ☞ batá			*
ii. bãtá		*	

Tableau (24a) shows why the nasal/oral contrast is preserved in stressed syllables (because of [[$\text{IDENT}_{\acute{\sigma}}$(nasal) ≫ $*V_{NAS}$]]), and tableau (24b) shows why this contrast is neutralized in unstressed syllables (because of [[$*V_{NAS}$ ≫ IDENT(nasal)]]).

ordering: the last rule to get its hands on the representation has precedence, in the sense that it reliably states a surface-true generalization. In OT, however, precedence relations among constraints are accounted for by ranking: the highest-ranking constraint has precedence, in the same sense that it reliably states a surface-true generalization.[55] There is, then, some overlap, though certainly not equivalence, in the functions of constraint ranking and rule ordering.

Since OT has constraint ranking anyway, it makes sense to start from the assumption that this is the only way to encode precedence relations in the grammar. In other words, the parallel, global architecture in (68) is the null hypothesis for implementing OT. Alternative implementations and the evidence for them will be discussed in §3.3.2.6 and §3.3.3, but for now we will stick to exploring the results obtained from the basic model (68).

3.3.2 Exemplification

3.3.2.1 Consequences of Globality

The basic OT architecture in (68) is global in the sense that EVAL applies a single language-particular constraint hierarchy H to all constructions from all inputs. A strictly global theory would be fully integral, with no modularity whatsoever. By common consent, all research in OT assumes a more limited globality, distinguishing at least between phonological and syntactic modules. I will ignore this largely irrelevant complication in subsequent discussion, but modularity questions will arise again in §3.3.3.

The main consequences of globality can be presented fairly quickly, since they are also discussed in §3.1.4.5–3.1.4.9. In those sections, several architectural imperatives of OT are noted, all of which presuppose globality (or integrality, as the same property is referred to in that context). Harmonic ascent says that the output must be either identical to a fully faithful analysis of the input or less marked than it. Restricted process-specificity says that it is not generally possible to isolate blocking effects on different processes. And construction-independence of evaluation says that constraint interactions must in principle generalize to all applicable linguistic structures.

All three of these universals depend upon generalizing over the results of evaluation with a single constraint hierarchy. For example, harmonic ascent could easily be subverted if distinct hierarchies, with different rankings of markedness constraints, were operative under different conditions or at different stages of a derivation. The same goes for the other two universals, showing that nontrivial empirical claims follow from the globality property of the basic OT architecture. (There will be more to say about globality in §3.3.2.8.)

3.3.2.2 Consequences of Parallelism: Overview

Parallelism is a more complicated business than globality and requires a correspondingly greater amount of attention. In the basic OT architecture (68), there is only one pass through GEN and EVAL. GEN has the property called

freedom of analysis or inclusivity (§1.1.3), meaning that it can construct candidates that differ in many ways from the input. The candidates emitted by GEN will therefore include some that are changed in several different ways at once. These candidates would have required several derivational steps to reach in a rule-based theory. The whole of this diverse candidate set is then submitted to EVAL, which selects its most harmonic member as the final output. This is a parallel theory because, given these assumptions about GEN and EVAL, the effects of notionally distinct linguistic operations are evaluated together, in parallel. In comparison, rule-based serial theories perform one operation at a time, stepping through a succession of intermediate stages. Parallelism, then, is the submission of complete output candidates to the grammar, without intermediate stages.

The known consequences of parallelism in OT can be loosely grouped into the four overlapping categories in (75), which will serve as the basis for subsequent discussion.

(75) Consequences of parallelism
- a. *Chicken-egg effects.* The application of process A depends on knowing the output of process B, and the application of process B depends on knowing the output of process A. Under parallelism, the effects of both processes can and must be considered simultaneously.
- b. *Top-down effects* (noncompositionality). Constituent X dominates constituent Y, and the well-formedness of X depends on Y (bottom up), but the well-formedness of Y is also influenced by X (top down). Under parallelism, there is no distinction between top-down and bottom-up effects, because various candidate parsings into X and Y constituents are evaluated.
- c. *Remote interaction.* Because fully formed output candidates are evaluated by the whole grammar, remote interactions are expected. Remoteness refers here not only to structural or stringwise distance but also to derivational remoteness, when two competing candidates differ in substantial ways from one another.
- d. *Globality effects.* Some further consequences of globality also depend on parallelism. This is shown by examining the predictions of a global but serial implementation of OT.

Chicken-egg and top-down effects (75a–b) are pretty much the same thing, but in different empirical or analytic domains. They include many of the ordering paradoxes in the literature on rule-based serialism, where there is inconsistent ordering of two rules. Remote interaction (75c) is a consequence of the kinds of candidates that GEN supplies and how they are evaluated. As we will see, although there are some compelling examples of remote interaction, there are also some problematic predictions. Finally, the effects of globality that depend on parallelism (75d) can be identified by decoupling the two, positing an architecture identical to (68) except that the output of EVAL is looped back into GEN. Each of these topics is addressed in the following sections.

3.3.2.3 *Consequences of Parallelism I: Chicken-Egg Effects*

Chicken-egg effects involve two or more notionally distinct processes that mutually depend on each other's output. If these processes are expressed by separate rules in a serial derivation, there is a problem: no ordering of the rules will work.📖§3.4¶22

There is a chicken-egg effect in the morphophonology of Southern Paiute (Uto-Aztecan, Utah). This language imposes strong restrictions on coda consonants. The only permitted codas are the first half of a doubled consonant or a nasal that shares place of articulation with a following stop or affricate: *tuŋ.qōn.nuq.q^{w}I* 'Paiute name'. (The syllable boundaries are shown by ".".) Typologically, this restricted syllable structure is quite common (e.g., Japanese). The restriction of nasals to positions before stops and affricates is also typical, since assimilation of nasals to continuants is somewhat unusual (cf. English *impose* vs. *infer*).[56]

This limited syllable structure carries over to the reduplicative morphology of Southern Paiute. The reduplicative prefix usually copies the first consonant and vowel of the root: ***ma**-maqa* 'to give', ***qa**-qaiva* 'mountain', ***wi**-winni* 'to stand'. But the second consonant of the root is copied only if two conditions are both met: the first consonant of the root is a stop or affricate and the second consonant is a nasal. Examples of this CVN-reduplication include ***pim**-pinti* 'to hang onto', ***ton**-tonna* 'to hit', and ***tun**-tuŋqutto* 'to become numb'. The generalization is that CVN-reduplication is possible only when it produces an independently permitted consonant cluster consisting of a nasal and stop that share place of articulation.

This generalization cannot be captured in a serial derivation. Two basic processes are at work: reduplicative copying and nasal assimilation. The problem, in chicken-egg terms, is that it is impossible to know how much to copy until nasal assimilation has applied, but it is impossible to apply nasal assimilation unless the nasal has been copied, so neither ordering works, as we see in (76a–b).

(76) Southern Paiute serially

a. Underlying representation	/Redup + pinti/	/Redup + winni/
Reduplication	**pi**-pinti	**wi**-winni
Nasal assimilation	*does not apply*	*does not apply*
Output	***pi**-pinti	**wi**-winni
b. Underlying representation	/Redup + pinti/	/Redup + winni/
Nasal assimilation	*does not apply*	*does not apply*
Reduplication	**pi**-pinti	**wi**-winni
Output	***pi**-pinti	**wi**-winni

The *n* of *pinti* is not copyable because it is not homorganic with the initial *p*. Nasal assimilation would make it homorganic, but nasal assimilation never sees

the requisite *n-p* sequence that it needs in order to apply, no matter how it is ordered relative to reduplication.[57]

No such problem arises in parallel OT. The effects of the copying and assimilation operations are evaluated simultaneously. The winning candidate is one that copies maximally (satisfying the base-reduplicant identity constraint MAX$_{BR}$) while still obeying the undominated syllable-structure constraints (CODA-COND) (see (77a–b)).[58]

(77) a. /Redup + pinti/ → *pim-pinti*

	CODA-COND	MAX$_{BR}$
i. ☞ pim-pinti		**
ii. pin-pinti	*	**
iii. pi-pinti		***

b. /Redup + winni/ → *wi-winni*

	CODA-COND	MAX$_{BR}$
i. ☞ wi-winni		***
ii. win-winni	*	**
iii. wim-winni	*	**

The constraint MAX$_{BR}$ assigns one violation-mark for each uncopied segment. It therefore favors maximality of copying but only within the limits set by undominated CODA-COND. The latter constraint only permits nasal codas when they are followed by a homorganic stop. So it chooses the assimilated candidate *pim-pinti* over unassimilated **pin-pinti*, while rejecting both unassimilated **win-winni* and unassimilated **wim-winni* because the following consonant is not a stop. Crucially, the winning candidate *pim-pinti* shows the simultaneous effects of two processes, reduplication and assimilation, and those effects are evaluated in parallel by the constraint hierarchy under EVAL.

In the phonological literature, there have been various attempts to deal with cases like this by grafting some form of parallelism onto a basically serial theory (cf. Calabrese 1995; Myers 1991; Paradis 1988a). In very general terms, the idea is to segregate all operations into two basic types, which are sometimes called rules and repairs (cf. §2.1). Rules apply sequentially, but repairs are applied in parallel with rules, automatically bringing rule outputs into conformity with general structural constraints. In Southern Paiute, for instance, reduplicative copying would be a rule, but nasal assimilation would be a repair, able to fly in under the radar, so to speak, to help effectuate reduplicative copying. In principle, this line of analysis might be promising, but in practice

it encounters significant difficulties. The architecture of such a theory has never been described in detail and may turn out to be unattainable (cf. the discussion of the triggering problem in phonology in §2.3). And a principled basis for the rule/repair split has proven elusive. In OT, all unfaithful mappings are in some sense repairs, an essential thesis if homogeneity of target/heterogeneity of process is to be accounted for (§3.1.4.2).

To sum up, the argument for parallelism from chicken-egg effects is based on the observation that sometimes there is no possible serial ordering of two notionally distinct operations. Parallel derivation looks like the only viable alternative in these cases. The balance has now shifted to Occam's other foot: since parallel derivation is *sometimes* required, is it *all* that is required? More on this question in §3.3.3.

3.3.2.4 Consequences of Parallelism II: Top-Down Effects

The argument from top-down effects is a variant of the chicken-egg argument but with special relevance to the well-formedness of hierarchical constituent structure.📖§3.4¶23 In a serial theory, the naive expectation is that hierarchical structures should be constructed from the bottom up, with each layer of structure derived by a distinct step of a serial derivation. Conditions on well-formedness are enforced by rules that apply as each level is constructed, with no backtracking.

Consider, for example, how bottom-up serialism would apply in sentence phonology. On this view, the structures of sentence phonology would be built in successive stages corresponding to the levels of the prosodic hierarchy: phonological words (PWd), then phonological phrases (PPh), then intonation phrases (IPh), and so on. Rules creating structures at level n would depend on the presence, position, number, or size of structures at level $n - 1$, but by the nature of the derivation the properties of structures at level $n - 1$ could not depend in any way on the properties of level n structures. Bottom-up effects should be observed, but never top-down ones – or at the very least, top-down effects should be highly unusual.

Syntactic theory has mostly developed along strict bottom-up lines. The generalized transformations of Chomsky (1975) and the Strict Cycle Condition of Chomsky (1973) are ways of excluding or strictly limiting top-down effects. But contemporary phonological theory countenances many top-down effects, contrary to naive expectation about the consequences of serial derivation. In fact, some top-down effects in phonology are the modal situation, with strict bottom-up derivation being rare or even unknown.

Top-down effects in phonology typically involve nonuniformity of metrical or prosodic structure (§3.2.1.2), such as the unstressable word syndrome or the prosody of function words. To follow up on the latter example, the prosodic structure of a function word depends on the larger context in which it finds itself.[59] Take, for instance, the difference between reduced *tŏ* [tə] in (78a) and unreduced, stressed *tó* [tú] in (78b).

(78) a. Reduced *tŏ*

I gave the book **tŏ** Bill.
I went **tŏ** Boston.
Tŏ add **tŏ** his troubles . . .

b. Stressed *tó*

Who did you give the book **tó**?
I talked **tó**, and eventually persuaded, my most skeptical colleagues.
I went **tó** – and here I must dispense with modesty – tŏ very great lengths indeed assisting him in his search for employment. Alas, tŏ no avail.

The general rule is that monosyllabic function words (other than object pronouns) are stressed and consequently unreduced before an intonation break, and they are otherwise reduced in normal speech.

Here is how these facts are usually and no doubt correctly interpreted (also see §3.2.1.2). A stressed function word is a freestanding phonological word: $[tó]_{\text{PWd}}$. It is stressed for reasons having to do with the prosodic hierarchy: every PWd must contain a foot, to serve as its head; and every foot must contain a stressed syllable, to serve as *its* head. So $[tó]_{\text{PWd}}$ is stressed because it is a head all the way down. An unstressed function word is a clitic rather than an independent PWd. In English, function words are normally proclitic to a following lexical word: $[tŏ\ Bill]_{\text{PWd}}$. There is no imperative to supply proclitic *tŏ* with a foot, and so it is unstressed and its vowel becomes [ə] in accordance with general properties of English phonology.

The analysis of (78), then, reduces to the following question: under what conditions are function words in English analyzed as independent PWd's versus clitics? The answer: they are analyzed as PWd's only when they have to be. In language typology generally and in English specifically, monosyllabic function words are preferentially cliticized. In English, cliticization has a directional bias, favoring pro- over enclisis. Stressed $[tó]_{\text{PWd}}$ appears only when there is nothing to procliticize onto, because no PWd follows in the same intonation phrase. An IPh-final function word presents a conundrum: it cannot be procliticized, so should it be encliticized or promoted to PWd status? Standard English takes the latter option, though my own most casual register favors the former.

The analysis just sketched has an obvious translation into OT, since it is already couched in the language of constraint interaction. In fact, a version of this analysis can be seen in (61) of §3.2.1.2. The constraint PWdCon is violated by any PWd that, like $[tó]_{\text{PWd}}$, contains no lexical words. As part of UG, this constraint accounts for the typologically justified unmarkedness of cliticized function words. The ranking [[Align-R(Lex, PWd) ≫ Align-L(Lex, PWd)]] favors proclisis – $[book]_{\text{PWd}}$ $[tŏ\ Bill]_{\text{PWd}}$ – over enclisis – $*[book\ tŏ]_{\text{PWd}}$ $[Bill]_{\text{PWd}}$. Violation of PWdCon is compelled by Align-R (shown in (79)).