CHAPTER 3

Combinatoriality

3.1 The need for an f-mental grammar

From the standpoint of communication systems in the natural world, one of the most striking facts about human language is that its users can create and understand an unlimited number of utterances on an unlimited number of topics. This productivity is possible thanks to an important design feature of language: utterances are built by combining elements of a large but finite vocabulary into larger meaningful expressions. Furthermore, the principles of combination are such as to enable users to construct arbitrarily long expressions, subject only to extrinsic limitations of memory and attention of the sort alluded to in Chapter 2.

As is observed in every introductory linguistics course, language provides many different ways of constructing elaborate utterances. Among them are successive conjunction of phrases (1a) or clauses (1b), multiple adjectival modifiers (1c), successively embedded prepositional modifiers (1d), successively embedded relative clauses (1e), successively embedded complement clauses (1f), plus free mixtures of the above (1g).

(1)

a. We ate apples and oranges and pears and pretzels and stew and ... and I can’t remember what else.

b. Ducks quack and dogs bark and cats meow and fish swim and worms wriggle and ... and I don’t know what armadillos do.

c. This is a big, new, imposing, poorly designed, cold, uncomfortable, very expensive building.

d. There’s a sty on the eye on the fly on the lamp on the wart on the frog on the bump on the log in the hole in the bottom of the sea.

e. This is the butcher that killed the ox that drank the water that quenched the fire that burned the stick that beat the dog that bit the cat that ate the goat that my father bought for two zuzim.

f. I wonder if Susan knows that Fred assured Lois that Clark would remind Pat to buy food for dinner.

g. We have a young child and our chef has twins so we know how difficult it is to find a first-rate restaurant that doesn’t shudder when you show up at the door with kids in tow.

The length and complexity of utterances involves not only the number of words and their syntactic organization. The messages that utterances convey—and their topics—are equally unlimited. In this respect human utterances contrast sharply with the long and complex songs of certain species of whales and birds, which, as far as can be told at present (Hauser 1996; Payne 2000; Slater 2000), basically convey only the message “Here I am, everyone!”

In principle, a communication system might construct arbitrarily long messages just by adding more and more new elements to the end of utterances, along the lines of a shopping list. But human language doesn’t work like that: it builds up large utterances according to structural principles or rules. What made generative grammar in the modern sense possible was the development of formal techniques for describing rules and systems of rules, deriving from work in the foundations of mathematics during the first half of the twentieth century. (The very same techniques, of course, led to the development of the digital computer.) And, although there were precursors in post-Bloomfieldian structuralism, in particular Chomsky’s teacher Zellig Harris (Harris 1951), it was Chomsky who developed and made clear the connections between this mathematical work and linguistic description.

Putting the issue of combinatoriality into a mentalist framework adds an important twist. Since the number of possible utterances of a human language is unlimited, language users cannot store them all in their heads. Rather, knowledge of language requires two components. One is a finite list of structural elements that are available to be combined. This list is traditionally called the “lexicon,” and its elements are called “lexical items”; for the moment let us suppose lexical items are words or morphemes (we will alter this substantially in Chapter 6). The other component is a finite set of combinatorial principles, or a grammar. To the extent that speakers of a language (or a dialect) are consistent with one another (see section 2.5), we can speak of the “grammar of the language” as a useful approximation to what all its speakers have in their heads.

The task of a theory of linguistic competence, then, is to model the lexicon

---

1 A terminological point: Sometimes “grammar” is taken to encompass rules plus lexicon, particularly when, as will be seen in section 3.3, the lexicon is thought to contain rules as well. From force of habit, I will no doubt fall into this inconsistency, but I will attempt to make clear what is intended when it makes a difference.
and the grammar in a language user's f-mind, using every sort of empirical
dvidence available, from speakers' grammaticality judgments to patterns of histori-
development to brain imaging. Diagrams like Fig. 1.1 are taken to reflect the
principles by which sentences are built up from (or can be analyzed into) their
constituent parts. The structure one draws comes out differently depending on
the principles one claims are involved in its construction.3

Chomsky coined the term "generative grammar" to refer to a precise formul-
ation of the combinatorial principles that characterize a speaker's competence.
He deliberately used the term ambiguously, to characterize both the principles
in the speaker's head and those formulated by the linguist, relying on context to
make clear which was intended. For example, if one speaks of "writing a gram-
mar," it is obviously the linguist's grammar; but if one speaks of "a child learning
or acquiring a grammar," the principles in the head are intended.

The next two sections will undertake a brief survey of the sorts of grammati-
cal rules that various versions of generative grammar have found it useful to
posit. With these examples before us, we will be in a better position to ask how
to construe the notion of a rule of mental grammar, a crucial issue in trying to
integrate linguistic theory into a larger theory of mind (section 3.4). The chap-
ter will conclude by considering some important implications of combinatorial-
ity for theories of brain processing.

3.2 Some types of rule

Across a broad range of formulations of linguistic theory, three major types of
rules emerge, which I will call formation rules, derivational rules, and con-
straints.3 We take up these types in the present section. In addition, many
approaches claim that the lexicon is not just a list of unstructured items. Rather,
lexical items have their own internal structure which can be characterized by
lexical rules. These come in at least two types, lexical formation rules and lex-
cal relations; a special case of the latter is inheritance hierarchies. Lexical rules
will be the topic of section 3.3.

---

3 Could there be more than one grammar that predicts the same linguistic behavior, so that it
would be a mistake to speak of the grammar of English? In principle, yes. In practice, it is usually
hard enough to find even one grammar that does the trick in sufficient detail. On the other hand,
the possibility of multiple grammars at certain points in a language's history is often posited as a
source of grammatical change over the course of a generation or two (Kiparsky 1969).

3 A terminological point: sometimes rules are taken to include only formation rules and deriv-
tional rules, contrasting with constraints. It is convenient here to lump all three types together as
rules.
This can be read as “An NP has as parts a Det, an AP, and an N, in that order.” Alternatively, it can be read “bottom-up”: “A Det followed by an AP followed by an N may be taken to constitute an NP.” Within this notation, we can think of constructing a tree as “clipping” treelets together. For instance, (4) would “clip onto” an NP node at the bottom of another treelet; in turn, a treelet with AP at the top would “clip onto” the AP node of (4). “Clipping together” can proceed from top to bottom, bottom to top, or any combination thereof.6

Some formulations of generative grammar have found it of interest to separate out the rules for constituency (the parts of an expression) from the principles for linear order. Such a procedure is useful, for instance, in characterizing languages with freer word order than English. In this case the formation rules specify only constituency (what is a part of what); principles of linear order are formulated as constraints (see below). There is no standard notation for rules of unordered constituency, and I will not bother to invent one here.

Whatever notation one adopts, an absolutely essential feature of formation rules already emerges: they must be formulated in terms of abstract categories.

A formation rule that mentioned only particular lexical items would not be especially useful. There may exist such rules in English, for instance the one that says hello and there may be combined in that order to form the utterance hello there. But this way of combining items is not going to lead to any sort of productivity. At the very least one must be able to refer to the class of words as a whole. For instance, suppose we wish to specify something as simple as “An utterance is a string of concatenated words.” This requires a formation rule which counts anything that falls into the class of words as part of an utterance—but which excludes non-word entities such as grunts, gestures, and sneezes. That is, the formation rule must contain a typed variable—a specification “anything of this particular sort.”6

Most actual formation rules in language use typed variables narrower than

5 This approach is most explicit in Tree-Adjoining Grammar (Joshi 1987; Frank and Kroskrity 1995), but it is also essentially the tack taken in all unification-based approaches such as HPSG and LFG.

6 One might wonder why we need to introduce something as technical as a typed variable in order to specify something as loose as “any word.” The answer becomes clearer when we notice that language contains contexts that permit a category broader than words. An example is And then he went ***,** where ***,** can be any sort of noise at all—or even a soundless gesture such as shrugging the shoulders.

“any word.” For example, (4) has four variables: NP, Det, AP, and N: it says “Any Det followed by any AP followed by any N can constitute an NP.” Of these variables, Det and N are categories (or types) of word; let us concentrate on them for a moment. In order for rule (4) to be able to put words together, two things are necessary. First, lexical items must be marked for what category they belong to. For instance, star and cat belong to the category Noun; the, a, and this belong to the category Determiner. Second, we need a “meta-principle” of variable instantiation, which can be seen from two perspectives. From the “top-down” perspective, it allows instances of a category to be substituted for a variable of that category, thus “satisfying” or “saturating” the variable. From a “bottom-up” perspective, it allows an individual word to be analyzed as having a role in a larger structure; this structure can be identified by substituting a variable for the individual word.

Next, consider the categories NP and AP. These are not categories of words, but categories of phrases. Such categories allow us not only to string words together, but also to build them into larger units that can then combine further. A phrasal category is more abstract than a lexical category such as Noun and Determiner, in that it is not instantiated just by pulling a single word out of the lexicon. Rather, it is instantiated by a string of one or more words and/or phrases that satisfy the variables in its formation rule (i.e. that are “clipped onto it”). Thus, for instance, NP can be instantiated by stars, the little star, this fat cat, and so forth; AP by fat, very skinny, moderately attractive, and so forth.

The reason for using phrasal variables rather than just strings of lexical types becomes clear when the same phrasal variable appears in several different formation rules. For example, NP appears as a variable in, among other things, the following formation rules of English (assuming a general introductory-course set of phrase structure rules).

In (5a), NP is the subject of a sentence; in (5b) it is the object of a verb phrase; in (5c) it is the object of a prepositional phrase. The fact that the same variable, with the same set of instantiations, shows up in multiple places is part of what convinces us we are dealing with a genuine linguistic unit and not just an analytic artifact.

In introducing formation rules as part of a grammar, we already have to face three important and interlocked empirical questions. First, what are the actual formation rules for the language under analysis? Second, and more generally, what is the available repertoire of types for variables in formation rules? For
instance, how many lexical categories are there and what are they, and how many phrasal categories are there and what are they? Third, what is the repertoire of relations induced by formation rules? Do they specify linear order of the parts, like (3), or do they specify unordered constituency? Is the relation of the head to the phrase (shown by a double line in (4) and (5)) distinct from the relation of other constituents (single lines)?

We have alluded to the possibility that some of the work of formation rules (say, linear order) is taken over by constraints. In the Minimlist Program (Chomsky 1995) this possibility has been carried to the extreme position that there is a single formation rule, called Merge: “Any word or phrase can be combined with any word or phrase, subject to all applicable constraints.” Similarly, in Optimality Theory in both phonology (Prince and Smolensky 1993) and syntax (Grimshaw 1997), there is a component called Gen (“generate”) which essentially says “Combine any units of the appropriate sort (i.e. phonological or syntactic respectively).” In Chapter 6 we will arrive at a more or less parallel position.

A second sort of formation rule specifies a category as composed from a number of more limited variables. For instance, Chapter 1 mentioned the analysis of the category NP as the composition of the three features [+N, −V, +phrasal]). In order for this composition to make sense, the grammar must contain a formation rule that stipulates the range of possibilities, for instance (6).

(6) Syntactic category = [+N, −V, +phrasal]

The variables here are the + signs, which can be instantiated as + or −. This rule has the effect of creating a set of eight syntactic categories, corresponding to all the combinations of the three variables. Thus instead of a repertoire of eight unrelated categories, we have the broader type “syntactic category” plus a repertoire of three features and their variables.7

Here is why we might want such an analysis. Suppose we find that certain categories, say NP and AP, behave in parallel fashion in various respects. We can then formulate the principles governing this parallel behavior in terms of a typed variable that ignores the difference between them—perhaps [+N, +phrasal], which leaves the value of the −V feature open. That is, feature composition rules like (6) permit us a wider range of typed variables. The empirical issue is always to find the right set of features that permits us to express the generalizations that we in fact find.8

To sum up, we find two basic kinds of formation rule: rules of constituency and rules of feature composition. Both contain typed variables. The “meta-rule” of variable instantiation connects these variables to their instantiations.

3.2.2 Derivational (transformational) rules

The most exciting innovation of early generative grammar, however, was not formation rules but derivational rules (or transformations): rules that take a fully formed structure and change some aspect of it. Sentence pairs like (7) provide a simple illustration.

(7) a. Dave really disliked that movie.
   b. That movie, Dave really disliked.

These two sentences mean essentially the same, with only perhaps a difference in emphasis. (7a) displays a more “basic” order: the thing that is disliked is in the “normal” direct object position. By contrast, in (7b), disliked is not followed by an object, as it should be, and that movie is in a curious position before the subject. So, the proposal goes, the grammar can capture the similarity between (7a) and (7b) by saying that (7b) in fact is not generated by the formation rules. Rather, it has an “underlying form” that is more or less identical to (7a) and that is generated by the formation rules. However, “after” the formation rules create the underlying form, a derivational rule moves that movie to the front of the sentence to create the “surface form.”

This approach carries with it a number of advantages. First, the semantic relation between the two sentences is explicit in underlying form: at that level they are in fact the same sentence (or at least very close). Second, the formation rules are simpler, in that they do not have to include the position at the front of the sentence. Third, the constraint that dislike requires a direct object is apparently violated in (7b)—but now this violation is only apparent, since it is explicitly observed in underlying form. These advantages are gained at a price: the introduction of a new and complex type of rule.

A derivational rule can be thought of more generally as a relation between two structures, one “more underlying” and one “more superficial.” Most parts of the structures are the same, but one or more parts are different. In the case of (6), the direct object in the more underlying form moves to initial position in the more superficial form, and everything else remains the same. Of course, in order to state this rule generally, it is necessary to use typed variables again: any NP in the underlying position can correspond to the same NP in the superficial position. Thus a very limited form of the rule responsible for (7) could be stated as
(8). (The subscripts on the NPs in (8) are notational conveniences so we can tell them apart.)

(8) More underlying:

```
   / \   
  S   / \   
 /   \ /   
NP   VP NP
    \    |
     V    
```

The application of a derivational rule can thus be notated as an ordered pair of trees. Alternatively it can be abbreviated in a single tree, say like (9).

(9) More superficial:

```
   / \   
  S   / \   
 /   \ /   
NP   NP VP
    \    
     V
```

As section 1.7 mentioned, since the middle 1970s (e.g. Chomsky 1975) it has been assumed that movement rules always leave behind a “trace,” a sort of unpronounced pronoun, whose antecedent is the moved constituent. In this notation, the application of the rule can be notated as (10), where \( T \) is the trace, and the subscripts show the relationship between the trace and its antecedent.

(10) can be thought of roughly as that movie, Dave disliked (it), where the parenthesized pronoun is not pronounced.

```
   / \   
  S   / \   
 /   \ /   
NP   NP VP
    \    |
     V    
```

The reason I call the two structures related by (8) “more underlying” and

9 In the interest of paring things down to the essentials, I have again omitted many generally-accepted details of structure in (8).

“more superficial,” rather than simply “underlying” and “superficial,” is that derivational rules can chain up, one applying to the output of another. Let me illustrate. In approaches that make use of derivational rules (i.e. all of Chomsky’s successive frameworks), it is standard to treat the subject of a passive sentence as derived from a more underlying object, as sketched in (11). (I give only the string of words; the structure is to be understood.)

(11) This book, has been studied \( T \), by generations of linguists.

Similarly, it is standard to derive the subject of certain predicates such as \textit{seem} from a more underlying subject of a subordinate clause; that is, the underlying form of (12b) is close to the nearly synonymous (12a). The derivational rule involved is called “raising.”

(12) a. It seems [that John likes ice cream].
    b. John \textit{seems [\( T \), to like ice cream]}.

(13) can then be derived as the product of passive followed by raising.

(13) This book, seems \( [T] \), to have been studied \( T \), by generations of linguists.

In other words, we can understand \textit{this book} as the underlying object of \textit{study}, despite its considerable distance from this position. It has achieved this distance by a sequence of independently motivated movements.

Thus a full derivation for a sentence consists of the creation of an underlying form by means of the formation rules, followed by a sequence of derivational rules to create the ultimate surface form. In the \textit{Aspects} framework the underlying form was called the sentence’s “Deep Structure,” the most superficial form its “Surface Structure”; in later frameworks these terms were abbreviated to “D-structure” and “S-structure.”

Such a derivation has an inherent directionality, from underlying to surface. Students are always reminded that the notion of movement is intended “metaphorically,” and that this directionality is just a way of defining a set of well-formed structures. No claim is implied that speakers actually move phrases around in their heads—this is taken to be a matter of performance. However,

10 In some approaches, such as that of \textit{Syntactic Structures}, the Minimalist Program, and Tree-Adjoining Grammar (Joshi 1987; Kroch 1987; Frank and Kroch 1995), structures created by derivational rules can be inserted for variables in formation rules, so that this strict ordering need not be observed.
the movement metaphor is pervasive, in that it is customary to speak of one rule applying "after" another, as can be seen in the discussion above. We return to this issue in Chapter 7, in connection with processing.

Two of the important empirical questions concerning derivational rules are (a) how explicit they have to be and (b) how their order of application is determined. The tendency in syntactic theory has been to extract more and more of the special properties of particular movements and re-encode them as constraints. This trend reaches its culmination in Government-Binding Theory (Chomsky 1981), which proposes only the maximally general derivational rule Move a (i.e. move anything to anywhere), subject to heavy independent constraints.

3.2.3 Constraints

A constraint is a kind of rule that places extra conditions on structures created by formation rules and derivational rules. It may consist of conditions that structures must necessarily satisfy, or alternatively of conditions that help make a structure more "favorable" or "stable." I will briefly mention some of the different kinds that have achieved some acceptance. Note that all of them require typed variables.

- Lexical items themselves can impose constraints on the structure they inhabit. For instance, the verb dislike imposes the constraint that it must be followed by an NP (in underlying structure); this is precisely what it means to say the verb is transitive.

- Constraints can impose extrinsic requirements on a structure created by a derivational rule. Here is an illustration. As mentioned in section 1.7, a major preoccupation of syntactic theory since the middle 1960s has been how movement rules are restricted; it has turned out that all of them are subject to very similar limitations. For instance, no matter what movement rule one tries to apply, it is impossible to remove anything from the inside of the subject of a sentence. Hence putative structures like (14) are grossly ungrammatical.

\[
\begin{align*}
\text{(14)} & \quad \text{a. } *\text{Bill, a rumor about } t, \text{ is going around town.} \\
& \quad \text{(from [a rumor about Bill] is going around town)} \\
& \quad \text{b. } *\text{Which book, did a review of } t, \text{ appear in the Times?} \\
& \quad \text{(related to [a review of some book] appeared in the Times)}
\end{align*}
\]

This restriction can be purged from each individual movement rule if we extract it into a more general constraint. This constraint, generally called the Sentential Subject Constraint, applies to the relation between traces and their antecedents. A version of it appears in (15). (Note that X is a typed variable that stands for any constituent; the ellipses are typed variables that stand for any random string of constituents, including a null string.)

\[
\begin{align*}
\text{(15)} & \quad \text{The structure } X \ldots S \ldots \text{ is ungrammatical.} \\
\end{align*}
\]

- Constraints can impose conditions on the relation between different stages in the derivation. As mentioned above, the early formulations of movement lacked traces, so constraints on movement had to be stated on stages related by a derivational step. For instance, the equivalent of (15) in this earlier framework (e.g. Ross 1967) would be stated as (16).

\[
\begin{align*}
\text{(16)} & \quad \text{A derivation from } \ldots S \ldots \text{ to } X \ldots S \ldots \text{ is illegal.} \\
\end{align*}
\]

Among more contemporary theories, Optimality Theory (OT) proposes that derivations involve exactly two layers: the Input (which plays approximately the role of underlying form) and a large collection of candidate Output (or surface) structures. The basic principle of OT is to impose constraints on the relation between these two layers, choosing the candidate Output that best meets the constraints. For instance, the "Parse" constraint stipulates that every element in the Input structure should appear in the Output structure. However, constraints in OT are regarded as violable. Therefore, under certain conditions when other constraints override Parse, some element of the Input will fail to appear in the Output; this is how OT treats what would be treated as a deletion in a theory with derivational rules. A major difference between OT and a derivational theory is that OT applies all constraints simultaneously to the Input–Output pair, whereas a derivational theory sees the relation between underlying and surface forms as a potentially unlimited sequence of simple steps.

- All of the rules we have discussed so far involve syntactic structure alone. Another kind of constraint imposes conditions between two structures of disparate types. This type of constraint, which I will call a correspondence rule or an interface...
rule, plays an especially important role in what is to come here. Two simple cases are the principles stated informally as (17a, b). They can best be understood in the context of section 1.6, which discussed how corresponding constituents in the phonological, syntactic, and semantic structures in Fig. 1.1 are linked by subscripts.

(17) a. If the semantic structure of a sentence contains an Agent (the entity bringing an action about), it (normally) corresponds to the subject in syntactic structure.
   b. The linear order of elements in phonological structure (normally) corresponds to the linear order of corresponding elements in syntactic structure.

Such principles appear in a wide range of theories of grammar; we will discuss them in Chapter 5. Importantly, in many of these theories, relations such as the active–passive alternation are not captured in terms of derivational rules. Rather, the grammar contains alternative interface rules relating syntactic structure and meaning, such that active and passive syntactic structures map into essentially the same meaning. We see here one of the major divides among theories of grammar: should such semantic relations among sentences be captured by more elaborate derivational rules, or by more elaborate interface rules? (Note how this relates to the alternative treatments of Tense in section 1.6.)

- A more complex type of constraint, which might be called a “meta-constraint,” applies not between two levels of structure in the same derivation, but between two or more alternative derivations. A situation where this seems unavoidable is morphological blocking. For example, the standard procedure for producing the past tense of an English verb is to add -d;¹² this applies to anything that counts as a verb, even new made-up verbs like fax. However, this procedure does not work in the 180 or so cases where it is supplanted by an irregular form; thus we say went and shook instead of goed and shaked. In other words, the derivation of go + -d, although in principle legitimate, is blocked because there is another way to express the same combination of meanings. This sort of phenomenon is ubiquitous in principles of morphology (word structure).

Meta-constraints with a much more global spirit appear in both the Minimalist Program and Optimality Theory. In the former, there are in principle many ways to get from an underlying form to an acceptable surface form, but one of these is chosen in preference to the others in part by a meta-constraint called Economy. This constraint gives priority to the shortest derivation that satisfies all the other constraints.

In OT, a large number of different candidate Outputs can be associated with the same Input; typically each of them violates one or more of the constraints. This presents the problem of deciding which violation is least serious. In the solution proposed by OT, the grammar of the language stipulates a ranking among the constraints from most to least important. Each candidate Output, then, has a most important violation. Among the candidates, the one whose most important violation is least important is chosen as the actual Output. This principle of choice among candidates, then, is another sort of meta-constraint.¹²

3.3 Lexical rules

So far we have been thinking of the lexicon as simply a list of items that can be arranged in hierarchical structures by formation rules. This approach is encouraged by Chomsky’s (1965: 84) characterization of the lexicon as “simply an unordered list of all lexical formatives”; he cites similar characterizations by Bloomfield (1933) and Sweet (1933).

But lexical items are not just atomic undecomposable units like constants in algebra. It is necessary to say how they are built too. Hence a theory of competence must specify the repertoire of possible “sub-lexical” elements and how they combine into actual lexical items. This specification constitutes a set of lexical rules. These too fall into a number of different types, including lexical formation rules, lexical redundancy rules (or lexical relations), and inheritance hierarchies. We take these up in turn.

3.3.1 Lexical formation rules

Perhaps the most frequently cited aspect of Saussurean doctrine is that a word is an arbitrary association of a sign with a meaning. The sign has two aspects: phonological structure (how it is pronounced) and syntactic structure (how it combines with other words in sentences). Thus a basic formation rule for lexical items specifies that they are triples of phonological structure, syntactic structure, and meaning. Every framework of grammar adopts such an assumption, at least implicitly.

Within each of these structures something has to be said as well. In order for an item to be able to instantiate typed variables in syntactic rules, its lexical syntactic structure has to specify what categories it belongs to; these categories will include

¹² The idea of violable constraints that interact to produce an optimal choice of output also plays a major role in the theory of musical cognition proposed by Lerdahl and Jackendoff (1983). As pointed out by Smolensky (1999), Dell et al. (1999), and Seidenberg and MacDonald (1999), such constraint systems are attractive for implementation in terms of connectionist networks. However, see section 3.5 for independent problems with network implementations.
at least part of speech, grammatical number and gender, and so forth. In addition, section 3.2.3 observed that a lexical item can place a constraint on its syntactic environment; for instance a verb may be specified as transitive. Such a constraint is called a subcategorization feature; we will see in section 6.6 how these are specified.

Among lexical formation rules we might also include principles of morphosyntax, which make it possible for words to have internal syntactic structure. For instance, the word *perturbation* is built from the verb *perturb*, which is converted to a noun by attaching the ending -ation. It is common to analyze words like this in terms of a tree structure altogether parallel to syntactic tree structures, except that the trees are inside of a word, as in (18a).

Moreover, the trees can be hierarchical, as in (18b).

The lexical formation rules must thus include tree fragments from which such structures can be built.

Turning to phonology, an item's phonological structure determines how it is pronounced. The formation rules for lexical phonological structure have to specify the range of possible pronunciations for words in the language: the inventory of phonological segments and how they combine into syllables and larger units. To cite a well-known example, *bikl* is a possible word of English but *bnik* is not, because the syllabic formation rules for English permit the onset cluster *bl* but not *bn-*. (even though *bn-* is pronounceable and is a possible onset cluster in some other languages). As illustrated in Fig. 1.2, phonological segments themselves have a systematic decomposition into features; the possibilities for such decomposition are stated in terms of rules or formation rules parallel to those for syntactic categories such as (6).

An item's lexical semantic/conceptual structure is also conceived of in (most approaches) as combinatorial. And again, the lexical formation rules must specify the available repertoire of more basic units and how they are combined. We return to this issue in Chapters 9–11.

The reader may have already detected hints that lexical formation rules are really very similar in format to phrasal formation rules. And indeed this is the case, as we will bring out more clearly in Chapter 6.

3.3.2 Lexical redundancy rules

How should the grammar account for the systematic relations among pairs of words such as *construct/construction, impress/impression, and suggest/suggestion*? In early generative grammar, including Aspects and notably Lees (1960), such relations were treated in terms of derivational rules applying to phrases. For instance, the NP the construction of a wall was taken to be derived from an underlying clause along the lines of someone constructs a wall. It was not too long before this approach was seen to be problematic, for a variety of technical reasons detailed in Chomsky (1970).\(^{13}\)

The alternative was to admit *construct* and *construction* as separate but related forms in the lexicon.\(^{14}\) Their relation is partly idiosyncratic, but partly systematic; the systematic part is expressed by a lexical redundancy rule (or, in later parlance, simply a *lexical rule*) (Jackendoff 1975). The relevant rule here can be stated informally as (19).

\[ (19) \text{A verb pronounced /X/ that denotes an action can be related to a noun pronounced /X+som/ that denotes the performance of such an action (or, alternatively, the product of such an action).} \]

This rule involves correlations in phonology, syntax, and semantics between the paired items. Its effect is to mark the parts of the paired items that are shared, in effect noting the redundancies between them. Such a rule is of interest to the extent that it relates many different pairs of items in the same way.

---

\(^{13}\) Among the reasons:

- There are many apparently "derived" cases for which there is no apparent source. For example, if the writer of the book is derived from someone who wrote the book, what is the source for the altogether parallel the author of the book? There is no verb *auth*.
- When verbs are converted to morphologically more complex nouns, their modifying adverbs are converted to morphologically simpler adjectives: *John suddenly refused* vs. *John's sudden refusal*. Hence there is an apparent conflict in the direction of derivation.
- "Derived" nouns can appear with modifiers that lack close parallels in the sentences from which they are putatively derived, e.g. *John's three criticisms of the book vs. John (three-flying) criticized the book (three times/over years)*.
- Many "derived" nouns mean something other than a simple transformation from the verb would predict. For instance, both *recitation* and *recital* have something to do morphologically with the verb *recite*, but only the former has a meaning pertaining to reciting.

The Generative Semantics movement that flourished in the late 1960s (e.g. Lakoff 1970; McCawley 1968; Postal 1970a) was founded on the idea that one could account for all meaning relations of this sort through derivational rules. Observations like those just cited were the opening salvo in the attacks on Generative Semantics (Chomsky 1972c; Jackendoff 1974; Akmaian 1973; Bowes 1975).

\(^{14}\) Chomsky (1970) actually proposed that there is a single more general lexical item which is pronounced *construct* when used as a verb and *construction* when used as a noun. So far as I know, nobody has ever really worked out this approach.
3.3.3 Inheritance hierarchies

How should the grammar account for the fact that many verbs expressing transfer (or intended transfer) appear in two possible syntactic frames?

(20) a. Beth gave/handed/sent/offered a teapot to Nancy.
    b. Beth gave/handed/sent/offered Nancy a teapot.

(21) a. Beth told a story to Nancy.
    b. Beth told Nancy a story.

In early generative grammar, this alternation was accomplished by means of a derivational rule (often called Dative Shift) that optionally turned the underlying order (20a, 21a) into the surface order (20b, 21b); a variant of this approach has been revived more recently (Larson 1988). The advantage of this treatment is that the verbs in question need to be supplied only with a single subcategorization feature instead of two; the price is this single derivational rule.

However, again for a variety of technical reasons (see references in Levin 1993; Jackendoff 1990b), most linguists have rejected the derivational approach for this alternation. But this leaves the problem of why all these verbs have two distinct subcategorization frames. One approach that has won some degree of acceptance (Pollard and Sag 1987; 1994; Michaelis and Lambrecht 1996) is that the lexicon contains not just the actual lexical items of the language but also more abstract schemata from which actual items can “inherit” properties. In the present case, the schema might be stated roughly as (22); the subscripts correlate the syntactic variables with the semantic ones, in the way they did in Fig. 1.1.

(22) Syntax:
    Part of speech: V
    Subcategorization: ___ NP₁ to NP₂
    or
    ___ NP₂, NP₁

Semantics:
    TRANSFER (X₁, TO Y₁)

This says that a verb that means “transfer” accepts both syntactic frames, with the same meaning. Because the words in (20) and (21) inherit their properties from this schema, they “cost less”: the extra subcategorization comes in a sense “for free.” New items in the language with the appropriate properties inherit these properties as well, with no further ado; for instance the new verb fax has both forms.

As in the discussion of derivational rules versus constraints, we see here an important locus of contention among alternative frameworks. Given any particular phenomenon, should it be treated as a derivational rule, a lexical redundancy rule, or an inheritance hierarchy? These are not just matters of notational preference, as they might appear at a first approximation. Rather, at the second and third approximations, different rule types lend themselves to different sorts of generalization, and so often it is only when one explores the phenomena in depth that the consequences of choosing one theory rather than another begin to emerge. Some such points were sketched in note 13 above, and others will appear throughout the course of our exposition.

3.4 What are rules of grammar?

As linguists, we may find it necessary to write rules of all these sorts in order to make sense of the patterns of linguistic use. Our focus is quite naturally on the proper formulation of rules. But with some sample rules in hand, it is worth posing a more general question that is often left to be asked by outsiders: What in the world is a rule of mental grammar supposed to be?

Like the term “knowledge” and other terms discussed in Chapter 2, “rule” has many uses in ordinary language. Is a linguistic rule like any of these? Let us for instance compare the rules discussed in the previous two sections to rules of a game. Players of a game consciously learn its rules, and can consciously invoke them. For example, there are rules of tennis that say what counts as a legal serve, what counts as winning a point, and so forth; and players can cite them. These rules have been agreed upon by some collection of people who have made up the game. By contrast, speakers of English can hardly cite the rules of English (unless perhaps they have studied linguistics): linguistic rules are essentially unconscious. Moreover, the rules of language, though shared by all speakers (to a very good approximation), are not overtly agreed upon by the speakers, and no established body makes them up.

Are linguistic rules like rules of law (e.g., traffic laws)? No, for many of the same reasons. In addition, if one breaks a rule of law, further laws spell out the consequences. By contrast, if a speaker break a rule of grammar (as I just did), the violation may provoke notice, but beyond that, the speaker just communicates less effectively. Are linguistic rules like moral rules, such that one should obey them at the risk of social opprobrium? The prescriptive rules of school grammar (“Don’t use ain’t,” “Don’t end a sentence with a preposition”) might have this quality, but the rules of the last two sections certainly do not.

Going to the other extreme from consciously invoked rules, we might try to see rules of grammar as like laws of physics: formal descriptions of the behavior

13 Though it is addressed in detail by Chomsky, for instance in Chomsky (e.g. 1975; 1980). Much of this section recapitulates his discussion, if much more briefly.
of speakers, with no implications for how this behavior is actually implemented. For instance, just as the planets do not solve internalized differential equations in order to know where to go next, we might want to say that speakers do not invoke internalized formation rules and constraints in order to construct and understand sentences.

Some linguists do affect this stance: “The rules I’m working out are just a formal description of languages; I don’t care what they have to do with the mind.” But this abandons the mentalist thesis that we set out to adopt in Chapter 2. It leads us away from the attempt to understand how people manage to be language users, and hence away from potential connections with psychology and neuroscience. Physicists who have developed insightful formal descriptions of physical behavior always go on to ask what mechanism is responsible for it: if the planets don’t compute their trajectories, then what makes the trajectories come out the way they do? The same question should be asked of rules of grammar. If they are not in the mind, then what is in the mind, such that speakers observe these regularities?

Another difference between rules of grammar and laws of physics is that rules of grammar differ from language to language, and one does not come into the world automatically following the rules of English or any other language. By contrast, laws of physics are universal and timeless. They are not acquired; they just are. One can break rules of grammar; one cannot break laws of physics.

It is sometimes suggested, especially by those of a behavioristic bent, that rules of grammar are like ingrained habits. This is closer to acceptable, as long as we accept the idea of a habit not just as a propensity to behave, but as a complex ingrained cognitive organization of perception and behavior. In particular, the “habit” of using linguistic rules plays a role in producing sentences, perceiving sentences, making grammaticality judgments, making judgments of rhyme, and solving anagrams. It is far more abstract and distant from actual behavior than, say, the habit of going to bed at 10 p.m. or the habit of taking such-and-such a route to work. And unlike these latter two habits, it cannot be acquired deliberately. As we will see in the next chapter, the acquisition of language involves a complex and subtle interaction between the child and the environment.

The proper way to understand rules of grammar, I suggest, is to situate them in the metaphysical domain between the conscious mind and the physical neurons: in the functional mind introduced in the previous chapter. Recalling that discussion, we treated linguistic structures like Fig. 1.1 as functionally characterized “data structures” in the f-mind. In those terms, the rules of grammar for a language are a general characterization of the state-space available to its users.

The lexical rules characterize possible lexical items of the language, and the phrasal rules characterize their combinatorial possibilities. What makes elements of the language rules rather than basic elements is that they contain typed variables (i.e., open places)—that is, they describe patterns of linguistic elements.

It is an open question how rules of grammar are to be incorporated into a model of performance, and hence into a theory of neural instantiation. Here are three plausible options.

- Rules are (in some sense) explicit in long-term memory within the f-mind, and the language processor explicitly refers to them in constructing and comprehending utterances. Following the computer metaphor, rules are like data structures in a computer.
- Rules are partial descriptions of the operations of the processor itself. In the computer metaphor, the rules as we write them are high-level descriptions of parts of the processor’s program.
- Rules are implicit in the f-mind; they describe emergent regularities (perhaps statistical regularities) among more basic elements, but are not themselves implemented in any direct way.

As suggested in section 3.2, the traditional formulation of phrase structure rules like (2) and of derivational rules like (8) is conducive to viewing the rules as like a program for constructing sentences. The connotations of the term “generate” in “generative grammar” reinforce such a view. However, as already mentioned, students are always cautioned to resist this interpretation. In particular they are exhorted to view derivational movement as metaphorical: “We are, after all, describing competence, not performance.” The upshot is that the status of such rules vis-à-vis performance models is left unspecified. We return to this issue in Chapter 7.

Feature composition rules and constraints lend themselves better to a “data structure” interpretation, in the sense that they delimit a space of possibilities within which linguistic entities can be located. In the approach to be worked out here, particularly in Chapter 6, a considerable proportion of standard rules of grammar will be interpreted in this light.

On the other hand, we will also conclude in Chapter 6 that certain classes of lexical rules fall into the “implicit” category. Such rules are nowhere present in the f-mind in the form we write them. Rather, these rules are indeed just descriptions of regularities in the organization of linguistic memory. I will maintain, however, that such rules still shed light on the nature of linguistic structure, so they are worth studying with traditional linguistic techniques.
3.5 Four challenges for cognitive neuroscience

Just as linguistic structures like Fig. 1.1 are functional characterizations that require neural instantiation, so the functional regularities that we state as rules of grammar must be neurally instantiated. To repeat a point from Chapter 2: although a great deal is known about functional localization of various aspects of language in the brain, I think it is fair to say that nothing at all is known about how neurons instantiate the details of rules of grammar. In fact, we don't even have any idea of how a single speech sound such as /p/—much less a category like NP—is instantiated in neural firings or synaptic connections. The rest of this chapter will lay out four challenges that linguistic combinatoriality and rules of language present to theories of brain function—challenges that to my knowledge have not been widely recognized in the cognitive neuroscience community.

3.5.1 The massiveness of the binding problem

Consider what happens in the perception or production of our familiar sentence (23).

(23) The little star's beside a big star.

Much of the neuroscience of language has been concerned with how words stored in long-term memory are activated (“light up”) in the course of sentence perception and production (e.g. Caramazza and Miozzo 1997; Pulvermüller 1999). But activation of words alone is not sufficient to account for the understanding of sentences. If understanding (23) consisted only of activating the words, the sentence in (24a), not to mention the complete nonsense in (24b), would “light up” the same words and hence be understood the same.

(24) a. The big star's beside a little star.
    b. Beside a the big little star star's.

Clearly a sentence is more than a collection of words: the word meanings are structured into the meaning of the sentence by means of semantic relations among them. These semantic relations are to some degree signaled by the syntactic structure of the sentence, which in turn is correlated with the linear order of the phonological words.

Thus all of the structure modeled in Fig. 1.1 must be functionally present—at once—in order for this sentence to be grasped. Introspectively, the whole seems to be maintained at least for some brief period after the completion of the sentence.

During this time, all the connections within and among the structures are available: phonologically, one retains the order of words; semantically, one retains the relations among the words; and one knows which sounds (e.g. /hugstar/) correlate with which part of meaning, in particular with the spatial structure. It will not do to say that one hears or utter the words in sequence but that one does not retain the whole in memory. In order to grasp a sentence, one need not have previously memorized it: one may spontaneously utter it when asked to describe a visual configuration, or one may hear someone else utter it. Our ability to spontaneously produce and perceive it is a consequence of the productive combinatoriality of language. What one must have memorized, though, is the words the, star, big, little, beside, and a, and the clitic’s, plus the principles for putting them together.

The need for combining independent bits into a single coherent percept has been recognized in the theory of vision under the name of the binding problem (not to be confused with linguists' Binding Theory, mentioned in section 1.7). In the discussions I have encountered, the binding problem is usually stated this way: we have found that the shape and the color of an object are encoded in different regions of the brain, and they can be differentially impaired by brain damage. How is it, then, that we sense a particular shape and color as attributes of the same object? The problem becomes more pointed in a two-object situation: if the shape region detects a square and a circle, and the color region detects red and blue, how does the brain encode that one is seeing, say, a red square and a blue circle rather than the other way around? In fact, under time pressure, subjects can mismatch the features of multiple perceived objects (Treisman 1988). A proposal that has gained a certain popularity (Gray et al. 1989; Crick and Koch 1990; Singer et al. 1997) is that the different representations are phase-linked: the neurons encoding “red” and “square” fire in synchrony, and those encoding “blue” and “circle” do as well, but out of phase with the first pair.

However, the binding problem presented by linguistic structure is far more massive than this simple description. The trivially simple sentence (23) has the four independent structures shown in Fig. 1.1. Each structure has multiple parts that must be correlated; in addition, the four structures must be correlated with each other, as noted by the subscripts in Fig. 1.1. Consider just the

---

16 This section is based on unpublished work done in collaboration with Maja Matarić. It also draws a great deal on the work of Gary Marcus (1998; 2001).
prepositional phrase in the syntactic structure. To characterize it properly, the following relationships must be encoded:

(25)  
   a. It is of the type PP.
   b. It is a part of VP.
   c. It follows V.
   d. It has P and NP as parts.
   e. It corresponds to the Place-constituent in conceptual structure.
   f. It corresponds to the phonological constituent beside a big star.

For the object of the preposition:

(26)  
   a. It is of the type NP.
   b. It is a part of PP.
   c. It follows P.
   d. It has Det, AP, and N as parts.
   e. It corresponds to a particular Object-constituent in conceptual structure.
   f. It corresponds to the phonological constituent a big star.

How effectively can firing synchrony be scaled up to deal with such an interlocking web of relationships? In particular, if the PP is synchronized with its parts, and its parts are synchronized with their parts, the whole tree structure is temporally unsolvable. More generally, is there sufficient bandwidth in the temporal resolution of neural firing to distinguish all the parts of the sentence from one another? To me, these questions cast considerable doubt on the feasibility of solving the binding problem in terms of temporal synchrony alone—even if temporal synchrony is indeed an important part of the solution.18

It should be observed that this complexity is not peculiar to language perception and production. The sorts of visual experiment cited above typically involve two objects with two relevant properties apiece. But consider the complexity of an everyday visual field, say one's desk, covered with books, papers, pictures, and office paraphernalia. All these objects—their (partially occluded) shapes, parts, colors, and locations—must be bound together in visual cognition. To be sure, one can perhaps only attend to (and therefore see in detail) some small part of this at once (as observed by Dennett 1991, for example). Nevertheless, much of the unattended spatial layout must still be present in working memory, for if someone says “Now look at the picture” one’s eyes go right to it without search; similarly, one can reach for the coffee cup off in the corner without really looking at it. Thus, as in the case of sentences, a full visual field provides a complex web of transient and novel relationships that must be put together out of familiar parts.

3.5.2 The Problem of 2

In much work on language processing (e.g. Barsalou 1992; 1999; Smith and Medin 1981; Caramazza and Moscovitch 1997; Elman et al. 1996), the only mechanism available for constructing linguistic expressions is spreading activation among nodes connected in a semantic network or neural net. Thus it is assumed that “lexical retrieval” in processing—i.e. identifying what word one is hearing or speaking—amounts to activating the nodes encoding (or instantiating) that word in long-term memory. We have already seen that such an approach is silent about capturing the relations among the words of a sentence. But there is an even simpler difficulty, which I will call the Problem of 2.

Consider yet again sentence (23), in which there are two occurrences of the word star. If the first occurrence simply activates the lexical entry in long-term memory, what can the second occurrence do? It cannot just activate the word a second time, since—as just argued—the word has to remain activated the first time in order for the sentence to receive a full interpretation. Nor can the second occurrence just activate the word more strongly, because that leaves the two occurrences indistinguishable. In particular, the first occurrence is bound to little and the second to big, so simultaneous binding to both would lead to the contradictory concept of a little big star.

This problem recurs at every level of linguistic structure. For example, in phonological structure, if the detection of the phoneme s consists simply of activating an s-node (or a distributed complex of nodes that together instantiate s), what happens in a word with more than one such sound, say sisyphus or sassafras?19 The same problem occurs in conceptual structure when conceptualizing a relation involving two tokens of the same type (e.g. the meaning of (23)) and in spatial structure when viewing or imagining two identical objects. It also occurs in understanding melodies: if the unit of musical memory is the individual note, then every repetition of the same note raises the problem; if the unit is

18 Singer et al. (1997) merely speculate on the possibility that their analysis in terms of synchrony will scale up to hierarchically structured relations; though see Shastry and Ajjanagadda (1993) for a proposal.

19 Rumelhart and McClelland (1986b) attempt to avoid this problem by coding words as “winkelgones,” overlapping sequences of three phonemes. But this too is subject to the Problem of 2: how is the word detector to distinguish sassafras from the nonword sassafiras, which has two occurrences of the medial sequence . . . assa . . . —or great-grandmother from great-great-grandmother, which has two occurrences of the medial sequence . . . reatgr . . . ? Pinker and Prince (1988) cite an Australian aboriginal language, Oyungand, whose words algal ‘straight’ and algolal ‘canned straight’ pose the same problem.
the interval between adjacent notes (say a major third), every repetition of the same interval raises the problem; if the unit is a motive, then every repetition of the motive (e.g., the first two occurrences of the text Happy Birthday in the birthday song) raises the problem; and so forth. Finally, the problem occurs in action patterns, for example dance patterns that involve things like “repeat such-and-such a step four times, then repeat this other step six times, then repeat the whole pattern.”

The Problem of 2 has been recognized (e.g., Pollack 1990; Marcus 2001, ch. 3 says it was well known by 1985), but it does not find a ready solution in classical network models of processing. A solution that is sometimes adopted, especially for phoneme detection, is that multiple copies of each unit exist in memory, one for each possible position it can occupy. Then the first occurrence of s in sassafras would activate [position 1: s] and the second would activate [position 2: s]; the first occurrence of star in {23} would activate [position 3: star] and the second would activate [position 6: star]. There are two difficulties with this solution, especially when trying to scale it up to multiple copies of words in sentences. First, it requires duplication of the entire repertoire of units over a large number of possible positions. This is most evident when the units in question are words: ten- to twenty-word sentences, with the words chosen out of a vocabulary of 20,000, are not uncommon. Second, and worse, there can be no generalization among these positions. There is no reason for [position 1: s] to be the same for words as [position 3: s], or for [position 3: star] to have anything to do with [position 6: star]. The two could be totally different in structural significance. (See Marcus 1998; 2001 for amplification of this argument.)

Such a “brute force” solution might be adequate for domains of brain function where there is a fixed size for the perceptual field and a limited number of distinctions to be made at each position. Primary visual cortex (V1) is a good example of such a domain: it codes a limited selection of visual features (brightness, presence of contour, orientation of contour, etc.), each detectable at each location in the visual field. But when we get to domains of more cognitive brain function—the perceived visual field full of objects, the sentence full of words—the solution does not scale up.22

20 A version of this proposal, for instance, appears in Dell et al. (1999). Their model (p. 51.3) has a bank of consonant nodes that function as syllable onsets, and another unrelated bank of consonant nodes that function as syllable codas. This is saved from obvious intractability only because they confine themselves to monosyllabic words.

22 Elman (1990), observing this difficulty of providing a large enough frame, proposes instead a solution with a recurrent network, which encodes sentential structure in terms of the recurrent dependencies among words. As pointed out by Chomsky (1957) and Miller and Chomsky (1963), though, the sequential dependencies among words in a sentence are not sufficient to determine understanding or even grammaticality. For instance, consider (i).

(i) Does the little boy in the yellow hat who Mary described as a genius like ice cream?

The fact that the italicized verb in (i) is like rather than likes is determined by the presence of does, 14 words away; and we would have no trouble making the distance even longer. However, it is not the distance in words that is significant: it is the distance in noun phrases, i.e., does is one NP away from like, whatever the length of the NP. This generalization, a typical example of what Chomsky calls the “structure-dependence” of linguistic rules, cannot be captured in Elman’s recurrent network, which only deals with word sequence. As far as I can determine, a similar objection applies to the dynamical model proposed by Tabor and Tanenhaus (1999). Steedman (1999: 619) points out, in reference to Elman’s work and several extensions of it, that “we know from work on symbolic finite-state models such as Hidden Markov Models and part-of-speech taggers [references omitted] that such approximations can achieve very high accuracy—better than 95% precision—without having any claim whatsoever to embody the grammar itself.”

A more old-fashioned computational approach to lexical access (Miller 1956; Neisser 1967; Baddeley 1986) supposes that the accessed word is copied from long-term memory (LTM) into a buffer in short-term or working memory (WM)—in this case in two different positions in WM. This approach meets the objections to the network solution, as it keeps the two occurrences of star distinct. On the other hand, how is it implemented neurally? The pure connectionist approaches with which I am familiar do not allow for the possibility of “copying” information from one “register” to another. A spreading activation equivalent of “copying” words into WM would require a set of nodes constituting WM, each of which is connected to (and can be activated by) every single word in the lexicon. This is somewhat better than the “brute force” solution, but still neurally implausible (so far as I know).

Yet another possibility would be that WM is a set of “dummy” nodes which have no content of their own but are bound to lexical items by temporal synchrony of firing, along lines discussed in section 3.5.1. Thus WM serves as a set of “pointers” to LTM, and in addition encodes relationships among the items being pointed to (for example, linear order). (Solutions of this sort are proposed by Potter (1993), citing Kahneman and Treisman (1984), and by Ballard et al. (1997).) Of course the questions raised in section 3.5.1 concerning the adequacy of temporal synchrony arise again. There is also a broader tradition of hybrid models, in which active nodes in a spreading activation network are keyed into an independent frame-and-slot component where larger structures are built (Collins and Quillian 1969; Levelt 1989; Roelofs 1997, among many examples). However, at this point the options are beyond the scope of this study; I hope that the nature of the problem has been made clear enough for those in the relevant fields. The main point is that although spreading activation may be a necessary component of memory, it is not enough for language.
3.5.3 The problem of variables

The situation is still more problematic. Consider the problem of encoding a two-place relation such as “X rhymes with Y.” The brain cannot list all the rhymes in the language. For one thing, we can acquire a new word, say ling, and know immediately that it rhymes with sting and bring but not with monk or mustard; we do not have to learn all its rhymes individually. Nor can we figure out its rhymes by analogy, reasoning for example, “Well, ling sounds sort of like the word link, and link rhymes with think, so maybe ling rhymes with think.” The only words for which such an analogical argument works are the words with which ling already rhymes—which is of course of no help.

Another reason rhymes cannot be listed is that one can judge rhymes involving sounds that are not even part of one’s language. English speakers can, without knowing German, judge that Fach rhymes with Bach, despite the fact that the ch sound is not a sound of English, and despite the fact that they learned about rhymes entirely through experience with English.

A third reason is that there is a phonological process in Yiddish-influenced dialects of English that creates expressions of sarcasm by rhyming with a nonsense word whose onset is shm: “Oedipus-Shmedipus! Just so you love your mother!” That is, new rhymes can be created on the spot.

Finally, people can judge rhymes that are created through combinatorial, for example those in (27).

(27) a. try and hide/cyanide (Tom Lehrer)
   b. tonsillectomy/come direct to me/send a check to me
       (Groucho Marx’s Dr. Hackenbush)
   c. a lot o’ news/hypotenuse
      din afore/Pinafore
      I’m more wary at/commissariat
      (Gilbert and Sullivan’s Major-General Stanley)

One certainly has no significant experience with rhyming “things that sound like tonsillectomy”, so these can’t be done by any sort of analogy either.

Rather, the rhyming relation has to be encoded in the form of a pattern with two typed variables: any phonological string rhymes with any other phonological string if everything from the stressed vowel to the end is identical in the two strings, and the onset preceding the stressed vowel is different (since normally ring does not rhyme with ring).

Marcus (1998, 2001) takes up about the simplest possible two-place relation: total identity (A rose is a rose; a daisy is a daisy; a dahlia is a ...: Fill in the blank.) He demonstrates that even this case cannot be formulated without the use of typed variables. His agenda is to defuse the extravagant claims that have been made on behalf of a particular class of spreading activation models, the multi-layer perceptrons of the sort common in connectionist modeling (e.g. Rumelhart and McClelland 1986a, b; Elman et al. 1996). He shows that for principled reasons these models cannot encode variables of the sort necessary for two-place relations such as “X is identical with Y,” “X rhymes with Y,” and “X is the regular past tense of Y.” Space precludes my repeating his arguments here, not to mention his replies to the many reactions his work has produced. Suffice it to say that Marcus has tested all the relevant networks in the literature on the data sets for which he predicts they will fail, and indeed they fail.22

This principled failure is fatal to unadorned spreading activation models of language, for, as we saw in section 3.2, all combinatorial rules of language—formation rules, derivational rules, and constraints—require typed variables. Again, this does not mean that spreading activation plays no role in the brain’s storage and processing of linguistic knowledge; in fact it likely does (see Chapter 7). But some further technical innovation is called for in neural network models, which will permit them to encode typed variables and the operation of instantiating them. I think that upon the development of such an innovation, the dialogue between linguistic theory and neural network modeling will begin to be more productive.

3.5.4 Binding in working memory vs. long-term memory

As alluded to earlier, contemporary neuroscience tends to see transient (short-term) connections among items in memory as instantiated either by spreading activation through synapses or by the “binding” relation, often thought of in terms of firing synchrony. By contrast, lasting (long-term) connections are usually thought of as encoded in terms of strength of synaptic connections. However, the combinatoriality of language presents the problem that the very same relation may be encoded either in a transient structure or in one stored in memory.

Consider, for instance, the idiom kick the bucket. This has to be stored in long-term memory, since one cannot predict its meaning from the meaning of its parts. At the same time, it has the syntactic structure of an ordinary verb phrase such as lift the shovel, which is built up combinatorially. Hence, when kick the

22 Marcus’s arguments concerning variables do not appear to be adequately appreciated in the connectionist literature; e.g. I find no reference to them in a recent issue of Cognitive Science dedicated to “Connectionist Models of Human Language Processing: Progress and Prospects” (Christiansen et al. 1999).
bucket is retrieved in the course of sentence perception or production, it ought to have the very same sort of instantiation as lift the shovel in the parts of the brain responsible for syntactic structure. It does not seem correct to posit that the connections of kick the bucket are encoded as synaptic weights and those of lift the shovel as firing synchrony.

A different sort of example: By now you have undoubtedly committed our original sample sentence, The little star . . . , to memory, so it is stored in your brain as a unit. In the old view of working memory as a buffer, one could claim that its contents are simply shipped off to long-term memory. But this option is not available in a view where the words of a sentence being perceived are related only by temporal synchrony. One cannot store a memorized sentence in long-term memory by constantly reactivating the transient bindings among the words—as it were, constantly rehearsing it. The transient bindings are used by the brain to store the sentence one is currently processing, not for the ones one has laid up against future need. But how are transient connections converted to synaptic weights in the course of memorizing the sentence?

It is usually argued that transient connections have the effect of gradually adjusting synaptic weights (so-called Hebbian learning). But what about cases in which one trial is sufficient for learning? For example, you say to me, I’ll meet you for lunch at noon. I reply, OK, and indeed I do show up as agreed. My long-term memory has been laid in on the basis of one trial; there hasn’t been any opportunity to adjust synaptic weights gradually.

Again, this is not just a problem for the theory of language. It occurs any time there is one-time learning of a novel configuration of known elements. For instance, to choose a task that a non-linguistic primate might find useful: One is walking along a trail in the woods and suddenly spots some ripe figs up in a tree. The next morning, as soon as one gets up, one may head right back to that spot to look for more. There is no previous association of figs with this location, so the knowledge is acquired on a single trial.

For another case: An experimenter says to a subject: “You will see some words appear on the screen. If a word has an f in it, stick out your tongue.” Subjects can perform faultlessly nearly immediately; and when they leave the experiment, they do not continue sticking out their tongue every time they see a word with an f in it. This is the typical paradigm of a psychological experiment: a subject is confronted with a novel task made of familiar parts—and the task has typed variables! Given all the experiments psychologists do, I find it intriguing that no one ever seems to ask how people assimilate the instructions, turning themselves temporarily into specialized stimulus–response machines. For present purposes, though, the important point is that on the basis of a transient input, a lasting combinatorial connection has been forged (understanding the task) which then is used to formulate transient connections (i.e. the actual responses to stimuli in the experiment).

More generally, any sort of episodic memory (in the sense of Tulving 1972) raises this problem: an episodic memory is by hypothesis something that is remembered on the basis of one occurrence, and it usually involves objects, places, and people with which one is familiar. One cannot just encode an episodic memory by gradually strengthening the associations among the involved characters: the precise relations among the characters are crucial, and they are established at once.

I will offer no speculations on how this transfer from working memory to structured long-term memory is accomplished, given that one can no longer speak simply of “shipping information off to some other area of memory.” As with the other three cases of this section, I leave it as a challenge for neuroscience. And as with the other cases, it is not something that can simply be disregarded by ignoring language.

To sum up, a theory of how language is instantiated in the brain must grapple with four problems that arise from the combinatoriality of language: the massiveness of binding in linguistic structure, the problem of multiple instances of a known unit in a novel structure, the necessity for encoding and instantiating typed variables, and the relation between long-term and short-term memory encodings of structure. These problems are not exclusive to language, but they certainly come to the fore in dealing with the linguistic phenomena that linguists deal with every day.

A further issue arises from combinatoriality: that of learning the principles that govern it. This requires a whole chapter of its own.