

2

Some Simple Theories of Grammar

2.1 Introduction

Among the key points in the previous chapter were the following:

- Language is rule-governed.
- The rules aren't the ones we were taught in school.
- Much of our linguistic knowledge is unconscious, so we have to get at it indirectly; one way of doing this is to consult intuitions of what sounds natural.

In this text, we have a number of objectives. First, we will work toward developing a set of rules that will correctly predict the acceptability of (a large subset of) English sentences. The ultimate goal is a grammar that can tell us for any arbitrary string of English words whether or not it is a well-formed sentence. Thus we will again and again be engaged in the exercise of formulating a grammar that generates a certain set of word strings – the sentences predicted to be grammatical according to that grammar. We will then examine particular members of that set and ask ourselves: ‘Is this example acceptable?’ The goal then reduces to trying to make the set of sentences generated by our grammar match the set of sentences that we intuitively judge to be acceptable.¹

A second of our objectives is to consider how the grammar of English differs from the grammar of other languages (or how the grammar of standard American English differs from those of other varieties of English). The conception of grammar we develop will involve general principles that are just as applicable (as we will see in various exercises) to superficially different languages as they are to English. Ultimately, much of the outward differences among languages can be viewed as differences in vocabulary.

This leads directly to our final goal: to consider what our findings might tell us about human linguistic abilities in general. As we develop grammars that include principles of considerable generality, we will begin to see constructs that may have universal applicability to human language. Explicit formulation of such constructs will help us evaluate Chomsky's idea, discussed briefly in Chapter 1, that humans' innate linguistic endowment is a kind of ‘Universal Grammar’.

¹Of course there may be other interacting factors that cause grammatical sentences to sound less than fully acceptable – see Chapter 9 for further discussion. In addition, we don't all speak exactly the same variety of English, though we will assume that existing varieties are sufficiently similar for us to engage in a meaningful discussion of quite a bit of English grammar; see Chapter 15 for more discussion.

In developing the informal rules for reflexive and nonreflexive pronouns in Chapter 1, we assumed that we already knew a lot about the structure of the sentences we were looking at – that is, we talked about subjects, objects, clauses, and so forth. In fact, a fully worked out theory of reflexive and nonreflexive pronouns is going to require that many other aspects of syntactic theory get worked out first. We begin this grammar development process in the present chapter.

We will consider several candidates for theories of English grammar. We begin by quickly dismissing certain simple-minded approaches. We spend more time on a formalism known as ‘context-free grammar’, which serves as a starting point for most modern theories of syntax. Appendix B includes a brief overview of some of the most important schools of thought within the paradigm of generative grammar, situating the approach developed in this text with respect to some alternatives.

2.2 Two Simplistic Syntactic Theories

2.2.1 Lists as Grammars

The simplest imaginable syntactic theory asserts that a grammar consists of a list of all the well-formed sentences in the language. The most obvious problem with such a proposal is that the list would have to be too long. There is no fixed finite bound on the length of English sentences, as can be seen from the following sequence:

- (1) Some sentences go on and on.
 Some sentences go on and on and on.
 Some sentences go on and on and on and on.
 Some sentences go on and on and on and on and on.
 ...

Every example in this sequence is an acceptable English sentence. Since there is no bound on their size, it follows that the number of sentences in the list must be infinite. Hence there are infinitely many sentences of English. Since human brains are finite, they cannot store infinite lists. Consequently, there must be some more compact way of encoding the grammatical knowledge that speakers of English possess.

Moreover, there are generalizations about the structure of English that an adequate grammar should express. For example, consider a hypothetical language consisting of infinitely many sentences similar to those in (1), except that every other sentence reversed the order of the words *some* and *sentences*:²

- (2) An Impossible Hypothetical Language:
 Some sentences go on and on.
 Sentences some go on and on and on.
 Some sentences go on and on and on and on.
 Sentences some go on and on and on and on and on.
 *Sentences some go on and on.
 *Some sentences go on and on and on.

²The asterisks in (2) are intended to indicate the ungrammaticality of the strings in the hypothetical language under discussion, not in normal English.

*Sentences some go on and on and on and on.

*Some sentences go on and on and on and on and on.

...

Of course, none of these sentences³ where the word *sentences* precedes the word *some* is a well-formed English sentence. Moreover, no natural language exhibits patterns of that sort – in this case, having word order depend on whether the length of the sentence is divisible by 4. A syntactic theory that sheds light on human linguistic abilities ought to explain why such patterns do not occur in human languages. But a theory that said that grammars consisted only of lists of sentences could not do that. If grammars were just lists, then there would be no patterns that would be excluded – and none that would be expected, either.

This form of argument – that a certain theory of grammar fails to ‘capture a linguistically significant generalization’ – is very common in generative grammar. It takes for granted the idea that language is ‘rule governed’, that is, that language is a combinatoric system whose operations are ‘out there’ to be discovered by empirical investigation. If a particular characterization of the way a language works fails to distinguish in a principled way between naturally occurring types of patterns and those that do not occur then it’s assumed to be the wrong characterization of the grammar of that language. Likewise, if a theory of grammar cannot describe some phenomenon without excessive redundancy and complications, we assume something is wrong with it. We will see this kind of argumentation again, in connection with proposals that are more plausible than the ‘grammars-as-lists’ idea. In Chapter 9, we will argue that (perhaps surprisingly), a grammar motivated largely on the basis of considerations of parsimony seems to be a good candidate for a psychological model of the knowledge of language that is employed in speaking and understanding.

2.2.2 Regular Expressions

A natural first step toward allowing grammars to capture generalizations is to classify words into what are often called ‘parts of speech’ or ‘grammatical categories’. There are large numbers of words that behave in similar ways syntactically. For example, the words *apple*, *book*, *color*, and *dog* all can appear in roughly the same contexts, such as the following:

- (3) a. That ___ surprised me.
 b. I noticed the ___ .
 c. They were interested in his ___ .
 d. This is my favorite ___ .

Moreover, they all have plural forms that can be constructed in similar ways (orthographically, simply by adding an *-s*).

Traditionally, the vocabulary of a language is sorted into nouns, verbs, etc. based on loose semantic characterizations (e.g. ‘a noun is a word that refers to a person, place, or thing’). While there is undoubtedly a grain of insight at the heart of such definitions,

³Note that we are already slipping into a common, but imprecise, way of talking about unacceptable strings of words as ‘sentences’.

we can make use of this division into grammatical categories without committing ourselves to any semantic basis for them. For our purposes, it is sufficient that there are classes of words that may occur grammatically in the same environments. Our theory of grammar can capture their common behavior by formulating patterns or rules in terms of categories, not individual words.

Someone might, then, propose that the grammar of English is a list of patterns, stated in terms of grammatical categories, together with a lexicon – that is, a list of words and their categories. For example, the patterns could include (among many others):

- (4) a. ARTICLE NOUN VERB
b. ARTICLE NOUN VERB ARTICLE NOUN

And the lexicon could include (likewise, among many others):

- (5) a. Articles: a, the
b. Nouns: cat, dog
c. Verbs: attacked, scratched

This mini-grammar licenses forty well-formed English sentences, and captures a few generalizations. However, a grammar that consists of a list of patterns still suffers from the first drawback of the theory of grammars as lists of sentences: it can only account for a finite number of sentences, while a natural language is an infinite set of sentences. For example, such a grammar will still be incapable of dealing with all of the sentences in the infinite sequence illustrated in (1).

We can enhance our theory of grammar so as to permit infinite numbers of sentences by introducing a device that extends its descriptive power. In particular, the problem associated with (1) can be handled using what is known as the ‘Kleene star’.⁴ Notated as a superscripted asterisk, the Kleene star is interpreted to mean that the expression it is attached to can be repeated any finite number of times (including zero). Thus, the examples in (1) could be abbreviated as follows:

- (6) Some sentences go on and on [and on]*.

A closely related notation is a superscripted plus sign (called the Kleene plus), meaning that one or more occurrences of the expression it is attached to are permissible. Hence, another way of expressing the same pattern would be:

- (7) Some sentences go on [and on]⁺.

We shall employ these, as well as two common abbreviatory devices. The first is simply to put parentheses around material that is optional. For example, the two sentence patterns in (4) could be collapsed into: ARTICLE NOUN VERB (ARTICLE NOUN). The second abbreviatory device is a vertical bar, which is used to separate alternatives.⁵ For example, if we wished to expand the mini-grammar in (4) to include sentences like *The dog looked big*, we could add the pattern ARTICLE NOUN VERB ADJECTIVE and collapse it with the previous patterns as: ARTICLE NOUN VERB (ARTICLE NOUN)|ADJECTIVE. Of

⁴Named after the mathematician Stephen Kleene.

⁵This is the notation standardly used in computer science and in the study of mathematical properties of grammatical systems. Descriptive linguists tend to use curly brackets to annotate alternatives.

course, we would also have to add the verb *looked* and the adjective *big* to the lexicon.⁶

Patterns making use of the devices just described – Kleene star, Kleene plus, parentheses for optionality, and the vertical bar for alternatives – are known as ‘regular expressions’.⁷ A great deal is known about what sorts of patterns can and cannot be represented with regular expressions (see, for example, Hopcroft et al. 2001, chaps. 2 and 3), and a number of scholars have argued that natural languages in fact exhibit patterns that are beyond the descriptive capacity of regular expressions (see Bar-Hillel and Shamir 1960, secs. 5 and 6). The most convincing arguments for employing a grammatical formalism richer than regular expressions, however, have to do with the need to capture generalizations.

In (4), the string ARTICLE NOUN occurs twice, once before the verb and once after it. Notice that there are other options possible in both of these positions:

- (8) a. *Dogs chase cats.*
 b. *A large dog chased a small cat.*
 c. *A dog with brown spots chased a cat with no tail.*

Moreover, these are not the only positions in which the same strings can occur:

- (9) a. Some people yell at *(the) (noisy) dogs (in my neighborhood)*.
 b. Some people consider *(the) (noisy) dogs (in my neighborhood)* dangerous.

Even with the abbreviatory devices available in regular expressions, the same lengthy string of symbols – something like (ARTICLE) (ADJECTIVE) NOUN (PREPOSITION ARTICLE NOUN) – will have to appear over and over again in the patterns that constitute the grammar. Moreover, the recurring patterns are in fact considerably more complicated than those illustrated so far. Strings of other forms, such as *the noisy annoying dogs*, *the dogs that live in my neighborhood*, or *Rover, Fido, and Lassie* can all occur in just the same positions. It would clearly simplify the grammar if we could give this apparently infinite set of strings a name and say that any string from the set can appear in certain positions in a sentence.

Furthermore, as we have already seen, an adequate theory of syntax must somehow account for the fact that a given string of words can sometimes be put together in more than one way. If there is no more to grammar than lists of recurring patterns, where these are defined in terms of parts of speech, then there is no apparent way to talk about the ambiguity of sentences like those in (10).

- (10) a. We enjoyed the movie with Cher.
 b. The room was filled with noisy children and animals.
 c. People with children who use drugs should be locked up.
 d. I saw the astronomer with a telescope.

⁶This extension of the grammar would license some unacceptable strings, e.g. **The cat scratched big*. Overgeneration is always a danger when extending a grammar, as we will see in subsequent chapters.

⁷This is not intended as a rigorous definition of regular expressions. A precise definition would include the requirement that the empty string is a regular expression, and would probably omit some of the devices mentioned in the text (because they can be defined in terms of others). Incidentally, readers who use computers with the UNIX operating system may be familiar with the command ‘grep’. This stands for ‘Global Regular Expression Printer’.

In the first sentence, it can be us or the movie that is ‘with Cher’; in the second, it can be either just the children or both the children and the animals that are noisy; in the third, it can be the children or their parents who use drugs, and so forth. None of these ambiguities can be plausibly attributed to a lexical ambiguity. Rather, they seem to result from different ways of grouping the words.

In short, the fundamental defect of regular expressions as a theory of grammar is that they provide no means for representing the fact that a string of several words may constitute a unit. The same holds true of several other formalisms that are provably equivalent to regular expressions (including what is known as ‘finite-state grammar’).

The recurrent strings we have been seeing are usually called ‘phrases’ or ‘(syntactic) constituents’.⁸ Phrases, like words, come in different types. All of the italicized phrases in (8)–(9) above obligatorily include a noun, so they are called ‘Noun Phrases’. The next natural enrichment of our theory of grammar is to permit our regular expressions to include not only words and parts of speech, but also phrase types. Then we also need to provide (similarly enriched) regular expressions to provide the patterns for each type of phrase. The technical name for this theory of grammar is ‘context-free phrase structure grammar’ or simply ‘context-free grammar’, sometimes abbreviated as CFG. CFGs, which will also let us begin to talk about structural ambiguity like that illustrated in (10), form the starting point for most serious attempts to develop formal grammars for natural languages.

2.3 Context-Free Phrase Structure Grammar

The term ‘grammatical category’ now covers not only the parts of speech, but also types of phrase, such as noun phrase and prepositional phrase. To distinguish the two types, we will sometimes use the terms ‘lexical category’ (for parts of speech) and ‘nonlexical category’ or ‘phrasal category’ to mean types of phrase. For convenience, we will abbreviate them, so that ‘NOUN’ becomes ‘N’, ‘NOUN PHRASE’ becomes ‘NP’, etc.

A context-free phrase structure grammar has two parts:

- A LEXICON, consisting of a list of words, with their associated grammatical categories.⁹
- A set of RULES of the form $A \rightarrow \varphi$ where A is a nonlexical category, and ‘ φ ’ stands for a regular expression formed from lexical and/or nonlexical categories; the arrow is to be interpreted as meaning, roughly, ‘can consist of’. These rules are called ‘phrase structure rules’.

The left-hand side of each rule specifies a phrase type (including the sentence as a type of phrase), and the right-hand side gives a possible pattern for that type of phrase. Because

⁸There is a minor difference in the way these terms are used: linguists often use ‘phrase’ in contrast to ‘word’ to mean something longer, whereas words are always treated as a species of constituent.

⁹This conception of a lexicon leaves out some crucial information. In particular, it leaves out information about the meanings and uses of words, except what might be generally associated with the grammatical categories. While this impoverished conception is standard in the formal theory of CFG, attempts to use CFG to describe natural languages have made use of lexicons that also included semantic information. The lexicon we develop in subsequent chapters will be quite rich in structure.

phrasal categories can appear on the right-hand sides of rules, it is possible to have phrases embedded within other phrases. This permits CFGs to express regularities that seem like accidents when only regular expressions are permitted.

A CFG has a designated ‘initial symbol’, usually notated ‘S’ (for ‘sentence’). Any string of words that can be derived from the initial symbol by means of a sequence of applications of the rules of the grammar is licensed (or, as linguists like to say, ‘generated’) by the grammar. The language a grammar generates is simply the collection of all of the sentences it generates.¹⁰

To illustrate how a CFG works, consider the following grammar: (We use ‘D’ for ‘Determiner’, which includes what we have up to now been calling ‘articles’, but will eventually also be used to cover some other things, such as *two* and *my*; ‘A’ stands for ‘Adjective’; ‘P’ stands for ‘Preposition’.)

(11) a. Rules:

S → NP VP
 NP → (D) A* N PP*
 VP → V (NP) (PP)
 PP → P NP

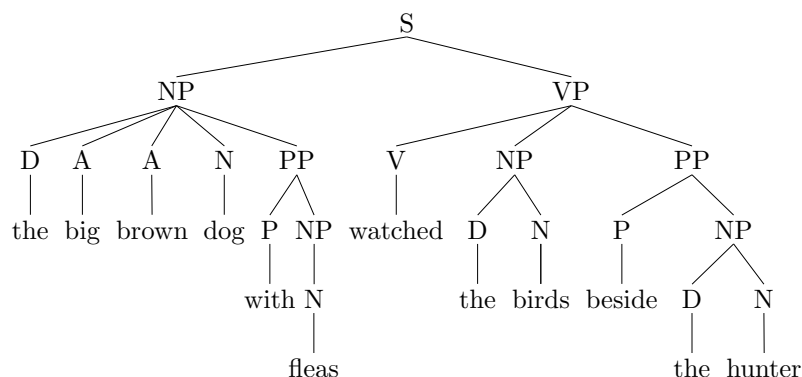
b. Lexicon:

D: the, some
 A: big, brown, old
 N: birds, fleas, dog, hunter
 V: attack, ate, watched
 P: for, beside, with

This grammar generates infinitely many English sentences. Let us look in detail at how it generates one sentence: *The big brown dog with fleas watched the birds beside the hunter*. We start with the symbol S, for ‘Sentence’. This must consist of the sequence NP VP, since the first rule is the only one with S on the left-hand side. The second rule allows a wide range of possibilities for the NP, one of which is D A A N PP. This PP must consist of a P followed by an NP, by the fourth rule, and the NP so introduced may consist of just an N. The third rule allows VP to consist of V NP PP, and this NP can consist of a D followed by an N. Lastly, the final PP again consists of a P followed by an NP, and this NP also consists of a D followed by an N. Putting these steps together the S may consist of the string D A A N P N V D N P D N, which can be converted into the desired sentence by inserting appropriate words in place of their lexical categories. All of this can be summarized in the following figure (called a ‘tree diagram’):

¹⁰Our definition of CFG differs slightly from the standard ones found in textbooks on formal language theory. Those definitions restrict the right-hand side of rules to finite strings of categories, whereas we allow any regular expression, including those containing the Kleene operators. This difference does not affect the languages that can be generated, although the trees associated with those sentences (see the next section) will be different in some cases.

(12)

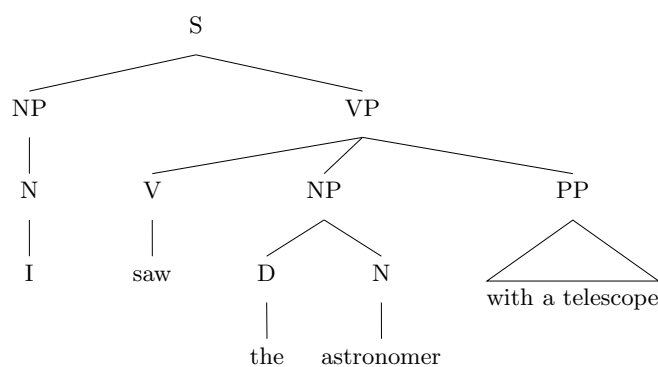


Note that certain sentences generated by this grammar can be associated with more than one tree. (Indeed, the example just given is one such sentence, but finding the other tree will be left as an exercise.) This illustrates how CFGs can overcome the second defect of regular expressions pointed out at the end of the previous section. Recall the ambiguity of (13):

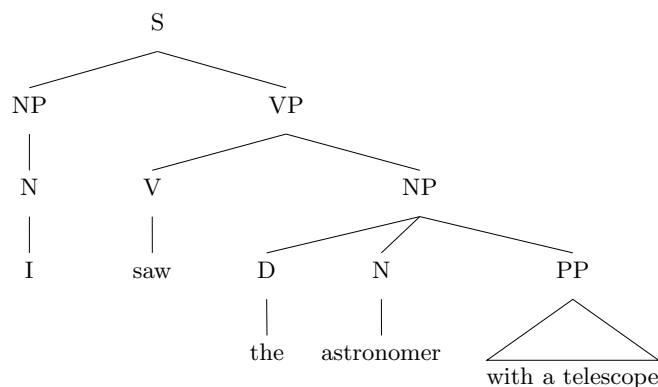
(13) I saw the astronomer with a telescope.

The distinct interpretations of this sentence ('I used the telescope to see the astronomer'; 'I saw the astronomer who had a telescope') correspond to distinct tree structures that our grammar will assign to this string of words. The first interpretation corresponds to (14a) and the latter to (14b):

(14) a.



b.



CFG thus provides us with a straightforward mechanism for expressing such ambiguities, whereas grammars that use only regular expressions don't.

The normal way of talking about words and phrases is to say that certain sequences of words (or categories) 'form a constituent'. What this means is that these strings function as units for some purpose (for example, the interpretation of modifiers) within the sentences in which they appear. So in (12), the sequence *with fleas* forms a PP constituent (as does the sequence P NP), *the big brown dog with fleas* forms an NP, and the sequence *dog with fleas* forms no constituent. Structural ambiguity arises whenever a string of words can form constituents in more than one way.

Exercise 1: Practice with CFG

Assume the CFG grammar given in (11). Draw the tree structure for the other interpretation (i.e. not the one shown in (12)) of *The big brown dog with fleas watched the birds beside the hunter*.

2.4 Applying Context-Free Grammar

In the previous sections, we introduced the formalism of context-free grammar and showed how it allows us to generate infinite collections of English sentences with simple rules. We also showed how it can provide a rather natural representation of certain ambiguities we find in natural languages. But the grammar we presented was just a teaching tool, designed to illustrate certain properties of the formalism; it was not intended to be taken seriously as an attempt to analyze the structure of English. In this section, we begin by motivating some phrase structure rules for English. In the course of doing this, we develop a new test for determining which strings of words are constituents. We also introduce a new abbreviatory convention that permits us to collapse many of our phrase structure rules into rule schemas.

2.4.1 Some Phrase Structure Rules for English

For the most part, we will use the traditional parts of speech, such as noun, verb, adjective, and preposition. In some cases, we will find it useful to introduce grammatical categories that might be new to readers, and we may apply the traditional labels somewhat differently than in traditional grammar books. But the traditional classification of words into types has proved to be an extremely useful categorization over the past two millennia, and we see no reason to abandon it wholesale.

We turn now to phrases, beginning with noun phrases.

Noun Phrases

Nouns can appear in a number of positions, e.g. those occupied by the three nouns in *Dogs give people fleas*. These same positions also allow sequences of an article followed by a noun, as in *The child gave the dog a bath*. Since the place of the article can also be filled by demonstratives (e.g. *this, these*), possessives (e.g. *my, their*), or quantifiers (e.g. *each, some, many*), we use the more general term 'determiner' (abbreviated D) for

this category. We can capture these facts by positing a type of phrase we'll call NP (for 'noun phrase'), and the rule $NP \rightarrow (D) N$. As we saw earlier in this chapter, this rule will need to be elaborated later to include adjectives and other modifiers. First, however, we should consider a type of construction we have not yet discussed.

Coordination

To account for examples like *A dog, a cat, and a wombat fought*, we want a rule that allows sequences of NPs, with *and* before the last one, to appear where simple NPs can occur. A rule that does this is $NP \rightarrow NP^+ \text{ CONJ } NP$. (Recall that NP^+ means a string of one or more NPs).

Whole sentences can also be conjoined, as in *The dog barked, the donkey brayed, and the pig squealed*.¹¹ Again, we could posit a rule like $S \rightarrow S^+ \text{ CONJ } S$. But now we have two rules that look an awful lot alike. We can collapse them into one rule schema as follows, where the variable 'X' can be replaced by any grammatical category name (and 'CONJ' is the category of conjunctions like *and* and *or*, which will have to be listed in the lexicon):

$$(15) X \rightarrow X^+ \text{ CONJ } X.$$

Now we have made a claim that goes well beyond the data that motivated the rule, namely, that elements of any category can be conjoined in the same way. If this is correct, then we can use it as a test to see whether a particular string of words should be treated as a phrase. In fact, coordinate conjunction is widely used as a test for constituency – that is, as a test for which strings of words form phrases. Though it is not an infallible diagnostic, we will use it as one of our sources of evidence for constituent structure.

Verb Phrases

Consider (16):

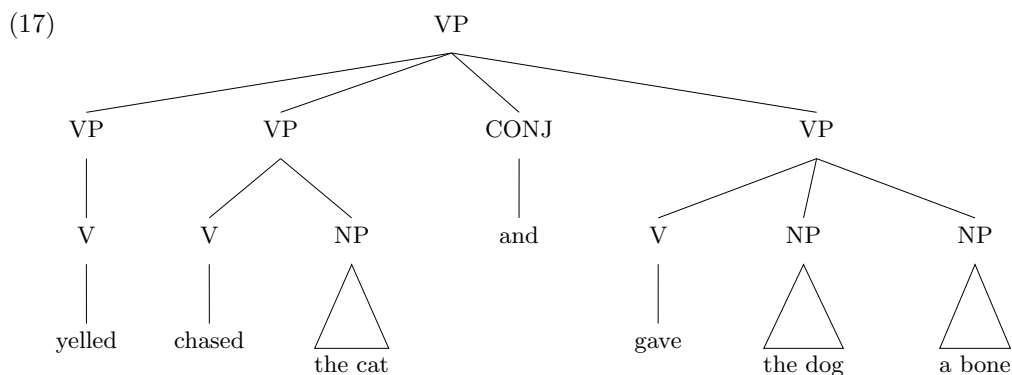
(16) A neighbor yelled, chased the cat, and gave the dog a bone.

(16) contains the coordination of strings consisting of V, V NP, and V NP NP. According to (15), this means that all three strings are constituents of the same type. Hence, we posit a constituent which we'll call VP, described by the rule $VP \rightarrow V (NP) (NP)$. VP is introduced by the rule $S \rightarrow NP VP$. A tree structure for the coordinate VP in (16) would be the following:

¹¹There are other kinds of coordinate sentences that we are leaving aside here – in particular, elliptical sentences that involve coordination of nonconstituent sequences:

- (i) Chris likes blue and Pat green.
- (ii) Leslie wants to go home tomorrow, and Terry, too.

Notice that this kind of sentence, which will not be treated by the coordination rule discussed in the text, has a characteristic intonation pattern – the elements after the conjunction form separate intonational units separated by pauses.



Prepositional Phrases

Expressions like *in Rome* or *at noon* that denote places or times ('locative' and 'temporal' expressions, as linguists would say) can be added to almost any sentence, and to NPs, too. For example:

- (18) a. The fool yelled at noon.
 b. This disease gave Leslie a fever in Rome.
 c. A tourist in Rome laughed.

These are constituents, as indicated by examples like *A tourist yelled at noon and at midnight, in Rome and in Paris*. We can get lots of them in one sentence, for example, *A tourist laughed on the street in Rome at noon on Tuesday*. These facts can be incorporated into the grammar in terms of the phrasal category PP (for 'prepositional phrase'), and the rules:

- (19) a. $PP \rightarrow P\ NP$
 b. $VP \rightarrow VP\ PP$

Since the second rule has VP on both the right and left sides of the arrow, it can apply to its own output. (Such a rule is known as a **RECURSIVE** rule).¹² Each time it applies, it adds a PP to the tree structure. Thus, this recursive rule permits arbitrary numbers of PPs within a VP.

As mentioned earlier, locative and temporal PPs can also occur in NPs, for example, *A protest on the street in Rome on Tuesday at noon disrupted traffic*. The most obvious analysis to consider for this would be a rule that said: $NP \rightarrow NP\ PP$. However, we're going to adopt a slightly more complex analysis. We posit a new nonlexical category, which we'll call NOM (for 'nominal'), and we replace our old rule: $NP \rightarrow (D)\ N$ with the following:

- (20) a. $NP \rightarrow (D)\ NOM$
 b. $NOM \rightarrow N$
 c. $NOM \rightarrow NOM\ PP$

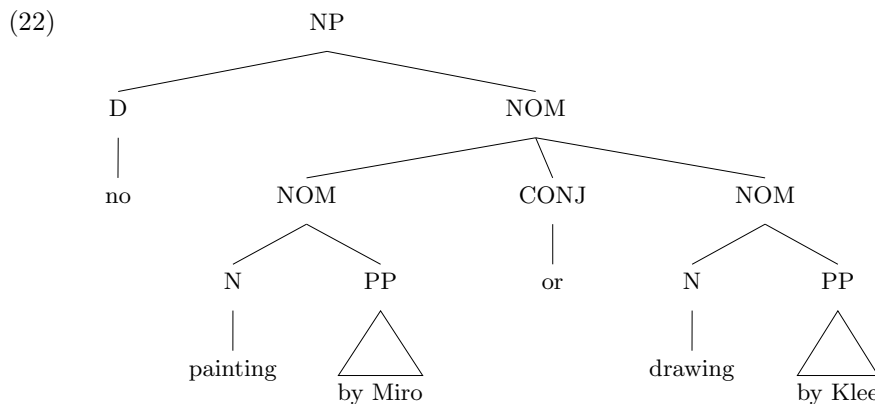
¹²More generally, we use the term **RECURSION** whenever rules permit a constituent to occur within a larger constituent of the same type.

The category NOM will be very useful later in the text. For now, we will justify it with the following sentences:

(21) a. The love of my life and mother of my children would never do such a thing.

b. The museum displayed no painting by Miro or drawing by Klee.

(21b) means that the museum displayed neither paintings by Miro nor drawings by Klee. That is, the determiner *no* must be understood as ‘having scope’ over both *painting by Miro* and *drawing by Klee* – it applies to both phrases. The most natural noun phrase structure to associate with this interpretation is:



This, in turn, is possible with our current rules if *painting by Miro or drawing by Klee* is a conjoined NOM. It would not be possible without NOM.

Similarly, for (21a), *the* has scope over both *love of my life* and *mother of my children* and hence provides motivation for an analysis involving coordination of NOM constituents.

2.4.2 Summary of Grammar Rules

Our grammar now has the following rules:

- (23)
- $S \rightarrow NP VP$
 - $NP \rightarrow (D) NOM$
 - $VP \rightarrow V (NP) (NP)$
 - $NOM \rightarrow N$
 - $NOM \rightarrow NOM PP$
 - $VP \rightarrow VP PP$
 - $PP \rightarrow P NP$
 - $X \rightarrow X^+ CONJ X$

In motivating this grammar, we used three types of evidence for deciding how to divide sentences up into constituents:

- In ambiguous sentences, a particular division into constituents sometimes can provide an account of the ambiguity in terms of where some constituent is attached (as in (14)).

- Coordinate conjunction usually combines constituents, so strings that can serve as coordinate conjuncts are probably constituents (as we argued for VPs, PPs, and NOMs in the last few pages).
- Strings that can appear in multiple environments are typically constituents.

We actually used this last type of argument for constituent structure only once. That was when we motivated the constituent NP by observing that pretty much the same strings could appear as subject, object, or object of a preposition. In fact, variants of this type of evidence are commonly used in linguistics to motivate particular choices about phrase structure. In particular, there are certain environments that linguists use as diagnostics for constituency – that is, as a way of testing whether a given string is a constituent.

Probably the most common such diagnostic is occurrence before the subject of a sentence. In the appropriate contexts, various types of phrases are acceptable at the beginning of a sentence. This is illustrated in the following sentences, with the constituent in question italicized, and its label indicated in parentheses after the example:

- (24) a. Most elections are quickly forgotten, but *the election of 2000*, everyone will remember for a long time. (NP)
 b. You asked me to fix the drain, and *fix the drain*, I shall. (VP)
 c. *In the morning*, they drink tea. (PP)

Another environment that is frequently used as a diagnostic for constituency is what is sometimes called the ‘cleft’ construction. It has the following form: *It is* (or *was*) ___ *that ...* For example:

- (25) a. It was *a book about syntax* that she was reading. (NP)
 b. It is *study for the exam* that I urgently need to do. (VP)
 c. It is *after lunch* that they always fall asleep. (PP)

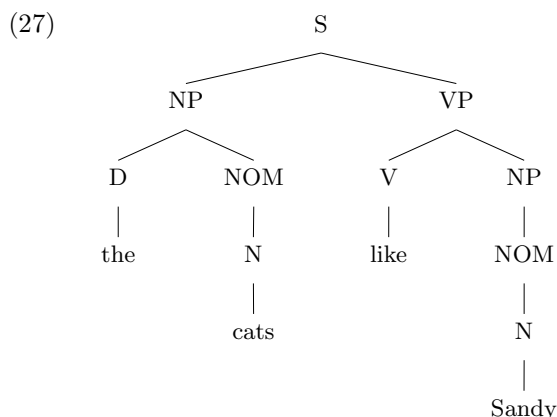
Such diagnostics can be very useful in deciding how to divide up sentences into phrases. However, some caution in their use is advisable. Some diagnostics work only for some kinds of constituents. For example, while coordination provided some motivation for positing NOM as a constituent (see (21)), NOM cannot appear at the beginning of a sentence or in a cleft:

- (26) a.*Many artists were represented, but painting by Klee or drawing by Miro the museum displayed no.
 b.*It is painting by Klee or drawing by Miro that the museum displays no.

More generally, these tests should be regarded only as heuristics, for there may be cases where they give conflicting or questionable results. Nevertheless, they can be very useful in deciding how to analyze particular sentences, and we will make use of them in the chapters to come.

2.5 Trees Revisited

In grouping words into phrases and smaller phrases into larger ones, we are assigning internal structure to sentences. As noted earlier, this structure can be represented as a tree diagram. For example, our grammar so far generates the following tree:

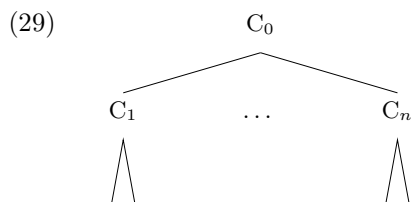


A tree is said to consist of **NODES**, connected by **BRANCHES**. A node above another on a branch is said to **DOMINATE** it. The nodes at the bottom of the tree – that is, those that do not dominate anything else – are referred to as **TERMINAL** (or **LEAF**) nodes. A node right above another node on a tree is said to be its **MOTHER** and to **IMMEDIATELY DOMINATE** it. A node right below another on a branch is said to be its **DAUGHTER**. Two daughters of the same mother node are, naturally, referred to as **SISTERS**.

One way to think of the way in which a grammar of this kind defines (or generates) trees is as follows. First, we appeal to the lexicon (still conceived of as just a list of words paired with their grammatical categories) to tell us which lexical trees are well-formed. (By ‘lexical tree’, we simply mean a tree consisting of a word immediately dominated by its grammatical category.) So if *cats* is listed in the lexicon as belonging to the category N, and *like* is listed as a V, and so forth, then lexical structures like the following are well-formed:



And the grammar rules are equally straightforward. They simply tell us how well-formed trees (some of which may be lexical) can be combined into bigger ones:



is a well-formed nonlexical tree if (and only if)

C_1, \dots, C_n are well-formed trees, and



$C_0 \rightarrow C_1 \dots C_n$ is a grammar rule.

So we can think of our grammar as generating sentences in a ‘bottom-up’ fashion – starting with lexical trees, and then using these to build bigger and bigger phrasal trees, until we build one whose top node is *S*. The set of all sentences that can be built that have *S* as their top node is the set of sentences the grammar generates. But note that our grammar could just as well have been used to generate sentences in a ‘top-down’ manner, starting with *S*. The set of sentences generated in this way is exactly the same. A CFG is completely neutral with respect to top-down and bottom-up perspectives on analyzing sentence structure. There is also no particular bias toward thinking of the grammar in terms of generating sentences or in terms of parsing. Instead, the grammar can be thought of as constraining the set of all possible phrase structure trees, defining a particular subset as well-formed.

Direction neutrality and process neutrality are consequences of the fact that the rules and lexical entries simply provide constraints on well-formed structure. As we will suggest in Chapter 9, these are in fact important design features of this theory (and of those we will develop that are based on it), as they facilitate the direct embedding of the abstract grammar within a model of language processing.

The lexicon and grammar rules together thus constitute a system for defining not only well-formed word strings (i.e. sentences), but also well-formed tree structures. Our statement of the relationship between the grammar rules and the well-formedness of trees is at present rather trivial, and our lexical entries still consist simply of pairings of words with parts of speech. As we modify our theory of grammar and enrich our lexicon, however, our attention will increasingly turn to a more refined characterization of which trees are well-formed.

2.6 CFG as a Theory of Natural Language Grammar

As was the case with regular expressions, the formal properties of CFG are extremely well studied (see Hopcroft et al. 2001, chaps. 4–6 for a summary). In the early 1960s, several scholars published arguments purporting to show that natural languages exhibit properties beyond the descriptive capacity of CFGs. The pioneering work in the first two decades of generative grammar was based on the assumption that these arguments were sound. Most of that work can be viewed as the development of extensions to CFG designed to deal with the richness and complexity of natural languages. Similarly, the theory we develop in this book is in essence an extended version of CFG, although our extensions are rather different in kind from some of the earlier ones.

In 1982, Geoffrey Pullum and Gerald Gazdar published a paper showing that the earlier arguments against the adequacy of CFG as a theory of natural language structure all contained empirical or mathematical flaws (or both). This led to a flurry of new work on the issue, culminating in new arguments that natural languages were not describable by CFGs. The mathematical and empirical work that resulted from this controversy substantially influenced the theory of grammar presented in this text. Many of the central papers in this debate were collected together by Savitch et al. (1987); of particular interest are Pullum and Gazdar’s paper and Shieber’s paper in that volume.

While the question of whether natural languages are in principle beyond the generative capacity of CFGs is of some intellectual interest, working linguists tend to be more

concerned with determining what sort of formalisms can provide elegant and enlightening accounts of linguistic phenomena in practice. Hence the arguments that tend to carry the most weight are ones about what formal devices are needed to capture linguistically significant generalizations. In the next section and later chapters, we will consider some phenomena in English that suggest that the simple version of CFG introduced above needs to be extended.

Accompanying the 1980s revival of interest in the mathematical properties of natural languages, considerable attention was given to the idea that, with an appropriately designed theory of syntactic features and general principles, context-free phrase structure grammar could serve as an empirically adequate theory of natural language syntax. This proposition was explored in great detail by Gazdar et al. (1985), who developed the theory known as ‘Generalized Phrase Structure Grammar’ (or GPSG). Work in phrase structure grammar advanced rapidly, and GPSG quickly evolved into a new framework, now known as ‘Head-driven Phrase Structure Grammar’ (HPSG), whose name reflects the increased importance of information encoded in the lexical heads¹³ of syntactic phrases. The theory of grammar developed in this text is most closely related to current HPSG. See Appendix B for discussion of these and other modern theories of grammar.

2.7 Problems with CFG

Two of our arguments against overly simple theories of grammar at the beginning of this chapter were that we wanted to be able to account for the infinity of language, and that we wanted to be able to account for structural ambiguity. CFG addresses these problems, but, as indicated in the previous section, simple CFGs like the ones we have seen so far are not adequate to account for the full richness of natural language syntax. This section introduces some of the problems that arise in trying to construct a CFG of English.

2.7.1 Heads

As we have seen, CFGs can provide successful analyses of quite a bit of natural language. But if our theory of natural language syntax were nothing more than CFG, our theory would fail to predict the fact that certain kinds of CF rules are much more natural than others. For example, as far as we are aware, no linguist has ever wanted to write rules like those in (30) in describing any human language:

(30) Unnatural Hypothetical Phrase Structure Rules

$$\text{VP} \rightarrow \text{P NP}$$

$$\text{NP} \rightarrow \text{PP S}$$

What is it that is unnatural about the rules in (30)? An intuitive answer is that the categories on the left of the rules don’t seem appropriate for the sequences on the right. For example, a VP should have a verb in it. This then leads us to consider why we named NP, VP, and PP after the lexical categories N, V, and P. In each case, the phrasal category was named after a lexical category that is an obligatory part of that kind of phrase. At least in the case of NP and VP, all other parts of the phrase may sometimes be absent (e.g. *Dogs bark*).

¹³The notion of ‘head’ will be discussed in Section 2.7.1 below.

The lexical category that a phrasal category derives its name from is called the HEAD of the phrase. This notion of ‘headedness’ plays a crucial role in all human languages and this fact points out a way in which natural language grammars differ from some kinds of CFG. The formalism of CFG, in and of itself, treats category names as arbitrary: our choice of pairs like ‘N’ and ‘NP’, etc., serves only a mnemonic function in simple CFGs. But we want our theory to do more. Many phrase structures of natural languages are headed structures, a fact we will build into the architecture of our grammatical theory. To do this, we will enrich the way we represent grammatical categories, so that we can express directly what a phrase and its head have in common. This will lead eventually to a dramatic reduction in the number of grammar rules required.

The notion of headedness is a problem for CFG because it cuts across many different phrase types, suggesting that the rules are too fine-grained. The next two subsections discuss problems of the opposite type – that is, ways in which the syntax of English is sensitive to finer-grained distinctions among grammatical categories than a simple CFG can encode.

2.7.2 Subcategorization

The few grammar rules we have so far cover only a small fragment of English. What might not be so obvious, however, is that they also overgenerate – that is, they generate strings that are not well-formed English sentences. Both *denied* and *disappeared* would be listed in the lexicon as members of the category V. This classification is necessary to account for sentences like (31):

- (31) a. The defendant denied the accusation.
 b. The problem disappeared.

But this classification would also permit the generation of the ungrammatical examples in (32):

- (32) a.*The defendant denied.
 b.*The teacher disappeared the problem.

Similarly, the verb *handed* must be followed by two NPs, but our rules allow a VP to be expanded in such a way that any V can be followed by only one NP, or no NPs at all. That is, our current grammar fails to distinguish among the following:

- (33) a. The teacher handed the student a book.
 b.*The teacher handed the student.
 c.*The teacher handed a book.
 d.*The teacher handed.

To rule out the ungrammatical examples in (33), we need to distinguish among verbs that cannot be followed by an NP, those that must be followed by one NP, and those that must be followed by two NPs. These classes are often referred to as INTRANSITIVE, TRANSITIVE, and DITRANSITIVE verbs, respectively. In short, we need to distinguish subcategories of the category V.

One possible approach to this problem is simply to conclude that the traditional category of ‘verb’ is too coarse-grained for generative grammar, and that it must be

replaced by at least three distinct categories, which we can call IV, TV, and DTV. We can then replace our earlier phrase structure rule

$$\text{VP} \rightarrow \text{V} (\text{NP}) (\text{NP})$$

with the following three rules:

- (34) a. $\text{VP} \rightarrow \text{IV}$
 b. $\text{VP} \rightarrow \text{TV NP}$
 c. $\text{VP} \rightarrow \text{DTV NP NP}$

2.7.3 Transitivity and Agreement

Most nouns and verbs in English have both singular and plural forms. In the case of nouns, the distinction between, say, *bird* and *birds* indicates whether the word is being used to refer to just one fowl or a multiplicity of them. In the case of verbs, distinctions like the one between *sing* and *sings* indicate whether the verb's subject refers to one or many individuals. In present tense English sentences, the plurality marking on the head noun of the subject NP and that on the verb must be consistent with each other. This is referred to as SUBJECT-VERB AGREEMENT (or sometimes just 'agreement' for short). It is illustrated in (35):

- (35) a. The bird sings.
 b. Birds sing.
 c. *The bird sing.¹⁴
 d. *Birds sings.

Perhaps the most obvious strategy for dealing with agreement is the one considered in the previous section. That is, we could divide our grammatical categories into smaller categories, distinguishing singular and plural forms. We could then replace the relevant phrase structure rules with more specific ones. In examples like (35), we could distinguish lexical categories of N-SG and N-PL, as well as IV-SG and IV-PL. Then we could replace the rule

$$\text{S} \rightarrow \text{NP VP}$$

with two rules:

$$\text{S} \rightarrow \text{NP-SG VP-SG}$$

and

$$\text{S} \rightarrow \text{NP-PL VP-PL}$$

But since the marking for number appears on the head noun and head verb, other rules would also have to be changed. Specifically, the rules expanding NP and VP all would have to be divided into pairs of rules expanding NP-SG, NP-PL, VP-SG, and VP-PL. Hence, we would need all of the following:

- (36) a. $\text{NP-SG} \rightarrow (\text{D}) \text{NOM-SG}$
 b. $\text{NP-PL} \rightarrow (\text{D}) \text{NOM-PL}$
 c. $\text{NOM-SG} \rightarrow \text{NOM-SG PP}$

¹⁴There are dialects of English in which this is grammatical, but we will be analyzing the more standard dialect in which agreement marking is obligatory.

- d. NOM-PL \rightarrow NOM-PL PP
- e. NOM-SG \rightarrow N-SG
- f. NOM-PL \rightarrow N-PL
- g. VP-SG \rightarrow IV-SG
- h. VP-PL \rightarrow IV-PL
- i. VP-SG \rightarrow VP-SG PP
- j. VP-PL \rightarrow VP-PL PP

This set of rules is cumbersome, and clearly misses linguistically significant generalizations. The rules in this set come in pairs, differing only in whether the category names end in ‘-SG’ or ‘-PL’. Nothing in the formalism or in the theory predicts this pairing. The rules would look no less natural if, for example, the rules expanding -PL categories had their right-hand sides in the reverse order from those expanding -SG categories. But languages exhibiting this sort of variation in word order do not seem to exist.

Things get even messier when we consider transitive and ditransitive verbs. Agreement is required regardless of whether the verb is intransitive, transitive, or ditransitive. Thus, along with (35), we have (37) and (38):

- (37) a. The bird devours the worm.
- b. The birds devour the worm.
- c.*The bird devour the worm.
- d.*The birds devours the worm.
- (38) a. The bird gives the worm a tug.
- b. The birds give the worm a tug.
- c.*The bird give the worm a tug.
- d.*The birds gives the worm a tug.

If agreement is to be handled by the rules in (39):

- (39) a. S \rightarrow NP-SG VP-SG
- b. S \rightarrow NP-PL VP-PL

then we will now need to introduce lexical categories TV-SG, TV-PL, DTV-SG, and DTV-PL, along with the necessary VP-SG and VP-PL expansion rules (as well as the two rules in (39)). What are the rules for VP-SG and VP-PL when the verb is transitive or ditransitive? For simplicity, we will look only at the case of VP-SG with a transitive verb. Since the object of the verb can be either singular or plural, we need two rules:

- (40) a. VP-SG \rightarrow TV-SG NP-SG
- b. VP-SG \rightarrow TV-SG NP-PL

Similarly, we need two rules for expanding VP-PL when the verb is transitive, and four rules each for expanding VP-SG and VP-PL when the verb is ditransitive (since each object can be either singular or plural). Alternatively, we could make all objects of category NP and introduce the following two rules:

- (41) a. NP \rightarrow NP-SG
- b. NP \rightarrow NP-PL

This would keep the number of VP-SG and VP-PL rules down to three each (rather than seven each), but it introduces extra noun phrase categories. Either way, the rules are full of undesirable redundancy.

Matters would get even worse when we examine a wider range of verb types. So far, we have only considered how many NPs must follow each verb. But there are verbs that only appear in other environments; for example, some verbs require following PPs or Ss, as in (42).

- (42) a. Terry wallowed in self-pity.
 b.*Terry wallowed.
 c.*Terry wallowed the self-pity.
 d. Kerry remarked (that) it was late.
 e.*Kerry remarked.
 f.*Kerry remarked the time.

Exercise 2: Wallowing in Categories

- A. Provide examples showing that the verbs *wallow* and *remark* exhibit the same agreement patterns as the other types of verbs we have been discussing.
 B. What additional categories and rules would be required to handle these verbs?
-

When a broader range of data is considered, it is evident that the transitivity distinctions we have been assuming are simply special cases of a more general phenomenon. Some verbs (and, as we will see later, some other types of words as well) occur only in the environment of particular kinds of constituents. In English, these constituents characteristically occur after the verb, and syntacticians call them **COMPLEMENTS**. Complements will be discussed in greater detail in Chapter 4.

It should be clear by now that as additional coverage is incorporated – such as adjectives modifying nouns – the redundancies will proliferate. The problem is that we want to be able to talk about nouns and verbs as general classes, but we have now divided nouns into (at least) two categories (N-SG and N-PL) and verbs into six categories (IV-SG, IV-PL, TV-SG, TV-PL, DTV-SG, and DTV-PL). To make agreement work, this multiplication of categories has to be propagated up through at least some of the phrasal categories. The result is a very long and repetitive list of phrase structure rules.

What we need is a way to talk about subclasses of categories, without giving up the commonality of the original categories. That is, we need a formalism that permits us to refer straightforwardly to, for example, all verbs, all singular verbs, all ditransitive verbs, or all singular ditransitive verbs. In the next chapter, we introduce a device that will permit us to do this.

2.8 Transformational Grammar

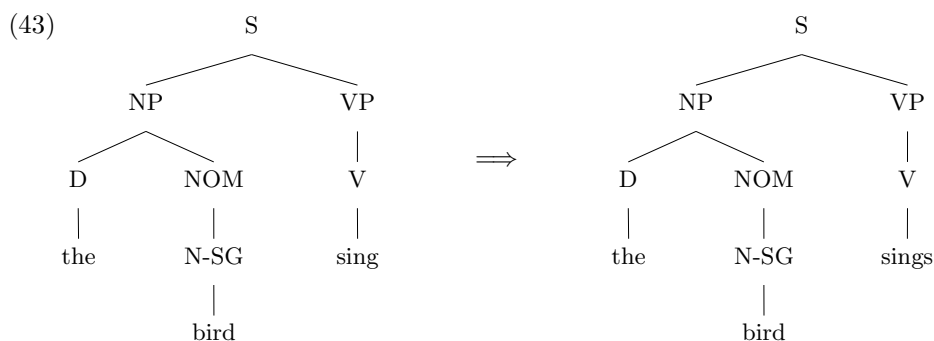
As noted in Section 2.6, much of the work in generative grammar (including this book) has involved developing extensions of Context Free Grammar to make it better adapted to the task of describing natural languages. The most celebrated proposed extension

was a kind of rule called a ‘transformation’, as introduced into the field of generative grammar by Noam Chomsky.¹⁵ Transformations are mappings from phrase structure representations to phrase structure representations (from trees to trees, in our terms) that can copy, delete, and permute parts of trees, as well as insert specified new material into them. The initial trees were to be generated by a CFG. For example, in early work on transformations, it was claimed that declarative and interrogative sentence pairs (such as *The sun is shining* and *Is the sun shining?*) were to be derived from the same underlying phrase structure by a transformation that moved certain verbs to the front of the sentence. Likewise, passive sentences (such as *The cat was chased by the dog*) were derived from the same underlying structures as their active counterparts (*The dog chased the cat*) by means of a passivization transformation. The name ‘transformational grammar’ is sometimes used for theories positing rules of this sort.¹⁶

In a transformational grammar, then, each sentence is associated not with a single tree structure, but with a sequence of such structures. This greatly enriches the formal options for describing particular linguistic phenomena.

For example, subject-verb agreement can be handled in transformational terms by assuming that number (that is, being singular or plural) is an intrinsic property of nouns, but not of verbs. Hence, in the initial tree structures for sentences, the verbs have no number associated with them. Subsequently, a transformation changes the form of the verb to the one that agrees with the subject NP. Such an analysis avoids the proliferation of phrase structure rules described in the preceding section, but at the cost of adding an agreement transformation.

As an illustration of how this would work, consider again the contrast in (35)¹⁷. Instead of creating separate singular and plural versions of NP, VP, NOM, N, and V (with the corresponding phrase structure rules in (36)), a transformational analysis could limit this bifurcation of categories to N-SG and N-PL (with the rules $\text{NOM} \rightarrow \text{N-SG}$ and $\text{NOM} \rightarrow \text{N-PL}$). In addition, an agreement transformation (which we will not try to formalize here) would give the verb the correct form, roughly as follows:



¹⁵The original conception of a transformation, as developed in the early 1950s by Zellig Harris, was intended somewhat differently – as a way of regularizing the information content of texts, rather than as a system for generating sentences.

¹⁶See Appendix B for more discussion of varieties of transformational grammar.

¹⁷The analysis sketched in this paragraph is a simplified version of the one developed by Chomsky (1957). It has long since been superseded by other analyses. In presenting it here (for pedagogical purposes) we do not mean to suggest that contemporary transformationalists would advocate it.

Notice that in a theory that posits a passivization transformation (which, among other things, would move the object NP into subject position), something like the agreement transformation described in the previous paragraph would be required. To make this more concrete, consider examples like (44):

- (44) a. Everyone loves puppies.
 b. Puppies are loved by everyone.

Substituting the the singular form of the verb in (44b) results in ill-formedness:

- (45) *Puppies is loved by everyone.

In a transformational analysis, *puppies* only becomes the subject of the sentence following application of the passivization transformation. Since agreement (in English) is consistently with the subject NP, if transformations are permitted to change which NP is the subject, agreement cannot be determined until after such transformations have applied.

In general, transformational analyses involve such rule interactions. Many transformational derivations involve highly abstract underlying structures with complex sequences of transformations deriving the observable forms.

Because versions of transformational grammar have been so influential throughout the history of generative grammar, many of the phenomena to be discussed have come to be labeled with names that suggest transformational analyses (e.g. “raising”, discussed in Chapter 12).

This influence is also evident in work on the psychology of language. In contemplating the mental processes underlying language use, linguists naturally make reference to their theories of language structure, and there have been repeated efforts over the years to find evidence that transformational derivations play a role in at least some aspects of language processing.

In later chapters, we will on occasion be comparing our (nontransformational) analyses with transformational alternatives. We make no pretense of doing justice to all varieties of transformational grammar in this text. Our concern is to develop a theory that can provide rigorous and insightful analyses of a wide range of the structures found in natural languages. From time to time, it will be convenient to be able to consider alternative approaches, and these will often be transformational.

2.9 What Are Grammars Theories Of?

In the opening paragraphs of Chapter 1, we said that linguists try to study language scientifically. We then went on to describe some of the grammatical phenomena that we would be investigating in this book. In this chapter, we have taken the first steps towards formulating a precise theory of grammar, and we have presented evidence for particular formulations over others.

We have not, however, said much about what a grammar is taken to be a theory of. Chapter 1 discussed the view, articulated most forcefully by Chomsky, that one reason for studying language is to gain insight into the workings of the human mind. On this view – which is shared by many but by no means all linguists – choosing one form of grammar over another constitutes a psychological hypothesis. That is, a grammar is a theory about the mental representation of linguistic knowledge.

As we noted, there are other views. Some linguists point out that communicating through language requires that different people share a common set of conventions. Any approach to language that seeks to represent only what is in the mind of an individual speaker necessarily gives short shrift to this social aspect of language.

To begin to get a handle on these issues, consider a concrete example: Pat says, “What time is it?” and Chris answers, “It’s noon”. The two utterances are physical events that are directly observable. But each of them is an instance of a sentence, and both of these sentences have been uttered many times. As syntacticians, we are interested in only some properties of these utterances; other properties, such as where they were uttered and by whom, are not relevant to our concerns. Moreover, there are many other English sentences that have never been spoken (or written), but they still have properties that our grammar should characterize. In short, the subject matter of our theory is sentences, which are abstractions, rather than observable physical events. We are interested in particular utterances only as evidence of something more abstract and general, just as a biologist is only interested in particular organisms as instances of something more abstract and general, such as a species.

A grammar of English should characterize the structure and meaning of both Pat’s utterance and Chris’s. So we need to abstract across different speakers, too. This raises some difficult issues, because no two speakers have exactly the same linguistic knowledge. In fact, linguistic differences among individuals and groups of individuals make it notoriously difficult to draw boundaries between languages. The conventional labels applied to languages (such as English, Chinese, or Arabic) are determined as much by political facts as by linguistic ones.¹⁸ It is largely for this reason that Chomsky and many other linguists say that their object of study is the mental representations of individual speakers.

Of course, similar difficulties arise in drawing boundaries between species, but few biologists would say on those grounds that biology should only be concerned with the DNA of individual organisms. Just as biologists seek to generalize across populations of heterogeneous individuals, we want our grammar to characterize something more general than what is in one person’s mind. Occasionally, we will deal with phenomena which are not uniform across all varieties of English (see especially Chapter 15).

In short, we want our grammar to characterize the syntax of English. This involves multiple levels of abstraction from what is directly observable, as well as some attention to variation among speakers. Our object of study is not purely a matter of individual psychology, nor is it exclusively a social phenomenon. There are some aspects of language that are primarily manifestations of individual speakers’ mental representations and others that critically involve the interactions of multiple language users. Just as molecular biology and population biology both contribute to our understanding of species, linguists need not make an exclusive choice between an internal and an external perspective.

2.10 Summary

In this chapter, we began our search for an adequate model of the grammar of one natural language: English. We considered and rejected two simple approaches to grammar,

¹⁸Linguists sometimes joke that a ‘language’ is simply a ‘dialect’ with an army and a navy.

including a theory based on regular expressions (‘finite-state grammar’). The theory of context-free grammars, by contrast, solves the obvious defects of these simple approaches and provides an appropriate starting point for the grammatical description of natural language. However, we isolated two ways in which context-free grammars are inadequate as a theory of natural language:

- CFGs are arbitrary. They fail to capture the ‘headedness’ that is characteristic of many types of phrase in natural language.
- CFGs are redundant. Without some way to refer to kinds of categories rather than just individual categories, there is no way to eliminate the massive redundancy that will be required in order to analyze the agreement and subcategorization patterns of natural languages.

For these reasons, we cannot accept CFG alone as a theory of grammar. As we will show in the next few chapters, however, it is possible to retain much of the character of CFG as we seek to remedy its defects.

2.11 Further Reading

The standard reference work for the basic mathematical results on formal languages (including regular expressions and context-free languages) is Hopcroft et al. 2001. Partee et al. 1990 covers much of the same material from a more linguistic perspective. Classic works arguing against the use of context-free grammars for natural languages include Chomsky 1963 and Postal 1964. Papers questioning these arguments, and other papers presenting new arguments for the same conclusion are collected in Savitch et al. 1987. For (somewhat dated) surveys of theories of grammar, see Sells 1985 and Wasow 1989. A more detailed presentation of GPSG is Gazdar et al. 1985. The history of generative grammar is presented from different perspectives by Matthews (1993), Newmeyer (1986), Harris (1993), and Huck and Goldsmith (1995).

Perhaps the best discussions of the basic phrase structures of English are to be found in good descriptive grammars, such as Quirk et al. 1972, 1985, Huddleston and Pullum 2002, or Greenbaum 1996. Important discussions of the notion of ‘head’ and its role in phrase structure can be found in Chomsky 1970 and Gazdar and Pullum 1981. A detailed taxonomy of the subcategories of English verbs is provided by Levin (1993).

2.12 Problems



Problem 1: More Practice with CFG

Assume the grammar rules given in (23), but with the following lexicon:

- | | |
|-------|-----------------------------------|
| D: | a, the |
| V: | admired, disappeared, put, relied |
| N: | cat, dog, hat, man, woman, roof |
| P: | in, on, with |
| CONJ: | and, or |

- A. Give a well-formed English sentence that this grammar sanctions and assigns only one structure to. Draw the tree structure that the grammar assigns to it.
 - B. Give a well-formed English sentence that is structurally ambiguous according to this grammar. Draw two distinct tree structures for it. Discuss whether the English sentence has two distinct interpretations corresponding to the two trees.
 - C. Give a sentence (using only the words from this grammar) that is not covered by this grammar but which is nonetheless well-formed in English.
 - D. Explain what prevents the example in (C) from being covered.
 - E. Give a sentence sanctioned by this grammar that is not a well-formed English sentence.
 - F. Discuss how the grammar might be revised to correctly exclude your example in (E), without simultaneously excluding good sentences. Be explicit about how you would change the rules and/or the lexicon.
 - G. How many sentences does this grammar admit?
 - H. How many would it admit if it didn't have the last rule (the coordination schema)?
-

Problem 2: Structural Ambiguity

Show that the grammar in (23) can account for the ambiguity of each of the following sentences by providing at least two trees licensed for each one, and explain briefly which interpretation goes with which tree:

- (i) Bo saw the group with the telescope.
- (ii) Most dogs and cats with fleas live in this neighborhood.
- (iii) The pictures show Superman and Lois Lane and Wonder Woman.

[*Note: We haven't provided a lexicon, so technically, (23) doesn't generate any of these. You can assume, however, that all the words in them are in the lexicon, with the obvious category assignments.*]

Problem 3: Infinity

The grammar in (23) has two mechanisms, each of which permits us to have infinitely many sentences: the Kleene operators (plus and star), and recursion (categories that can 'dominate themselves'). Construct arguments for why we need both of them. That is, why not use recursion to account for the unboundedness of coordination or use Kleene star to account for the possibility of arbitrary numbers of PPs?

[*Hint: Consider the different groupings into phrases – that is, the different tree structures – provided by the two mechanisms. Then look for English data supporting one choice of structure over another.*]

Problem 4: CFG for Japanese

Examples (i)–(x) give examples of grammatical Japanese sentences and strings made up of the same words which are not grammatical Japanese sentences.

- (i) Suzuki-san-ga sono eiga-wo mita.
Suzuki-NOM that movie-ACC saw
'Suzuki saw that movie.'
- (ii)*Mita Suzuki-san-ga sono eiga-wo.
Saw Suzuki-NOM that movie-ACC
- (iii)*Suzuki-san-ga mita sono eiga-wo.
Suzuki-NOM saw that movie-ACC
- (iv)*Suzuki-san-ga eiga-wo sono mita.
Suzuki-NOM movie-ACC that saw.
- (v) Suzuki-san-ga sono omoshiroi eiga-wo mita.
Suzuki-NOM that interesting movie-ACC saw
'Suzuki saw that interesting movie.'
- (vi)*Suzuki-san-ga sono eiga-wo omoshiroi mita.
Suzuki-NOM that movie-ACC interesting saw
- (vii)*Suzuki-san-ga omoshiroi sono eiga-wo mita.
Suzuki-NOM interesting that movie-ACC saw
- (viii) Suzuki-san-ga Toukyou e itta.
Suzuki-NOM Tokyo to went.
'Suzuki went to Tokyo.'
- (ix)*Suzuki-san-ga e Toukyou itta.
Suzuki-NOM to Tokyo went.
- (x)*Suzuki-san-ga itta Toukyou e.
Suzuki-NOM went Tokyo to.

- A. Using the lexicon in (xi), write phrase structure rules that will generate the grammatical examples and correctly rule out the ungrammatical examples.

[Notes: The data presented represent only a very small fragment of Japanese, and are consistent with many different CFGs. While some of those CFGs would fare better than others when further data are considered, any answer that accounts for the data presented here is acceptable. The abbreviations 'NOM' and 'ACC' in these examples stand for nominative and accusative case, which you may ignore for the purposes of this problem.]

- (xi) N: Suzuki-san-ga, eiga-wo, Toukyou
D: sono
P: e
A: omoshiroi
V: mita, itta

- B. Draw the trees that your grammar assigns to (i), (v), and (viii).

Problem 5: Properties Common to Verbs

The rules in (34) embody the claim that IVs, TVs, and DTVs are entirely different categories. Hence, the rules provide no reason to expect that these categories would have more in common than any other collection of three lexical categories, say, N, P, and D. But these three types of verbs do behave alike in a number of ways. For example, they all exhibit agreement with the subject of the sentence, as discussed in Section 2.7.3. List at least three other properties that are shared by intransitive, transitive, and ditransitive verbs.

**Problem 6: Pronoun Case**

There are some differences between the noun phrases that can appear in different positions. In particular, pronouns in subject position have one form (referred to as NOMINATIVE, and including the pronouns *I*, *he*, *she*, *we*, and *they*), whereas pronouns in other positions take another form (called ACCUSATIVE, and including *me*, *him*, *her*, *us*, and *them*). So, for example, we say *He saw her*, not **Him saw she*.

- A. How would the category of NP have to be further subdivided (that is, beyond NP-SG and NP-PL) in order to account for the difference between nominative and accusative pronouns?
 - B. How would the rules for S and the various kinds of VPs have to be modified in order to account for the differences between where nominative and accusative pronouns occur?
-