Toshiaki Nishihara, Katsuaki Nakanishi, Rachel Nordlinger, Stephan Oepen, Michael Orme, Gerald Penn, Paul Postal, Geoffrey Pullum, Emily Reyna, John Rickford, Susanne Riehemann, Sarah Risken, Jed Rose, Scott Schwenter, Jeramie Scott, Rylan Sekiguchi, Peter Sells, Yael Shrager, Luke Swartz, Ryosuke Takahashi, Stuart Tannock, Shiao Wei Tham, Ida Toivonen, Judith Tonhauser, Louise Vigeant, Rick Warren, Gert Webelhuth, Steve Wechsler, Bill Weigel, and Corinne Yates. We would also like to thank Dikran Karagueuzian, Director of CSLI Publications, for his continued support and patience, as well as Lauri Kanerva, Anubha Kothari, Michael Frank, Christine Sosa, Maureen Burke, and Tony Gee for their help in matters of production.

This book was written at Stanford's Center for the Study of Language and Information – an ideal environment for thought and writing. Thanks to Emma Pease for sustained help with all matters technological. Some of the material in this text is based on research conducted under the auspices of CSLI's Linguistic Grammars Online (LINGO) project. In that connection, we gratefully acknowledge contracts from the Bundesministerium für Bildung, Wissenschaft, Forschung, und Technologie (BMBF), who supported LINGO's participation in the VERBMOBIL project. This material is also based in part upon work supported by the National Science Foundation under Grant No. BCS-0094638. Work on the second edition was completed while Sag was a Fellow at the Center for Advanced Study in the Behavioral Sciences, in which connection we acknowledge the support of a grant (# 2000-5633) to CASBS from The William and Flora Hewlett Foundation.

Finally, we express our deep appreciation to our spouses, Penny Eckert, Judith Wasow, and Vijay Menon, for putting up with us during the long and often tedious revision process.

Sag, I, Wasow, T & E. Bender 2003

Syntactic Theory 2nd edition

1

Introduction

1.1 Two Conceptions of Grammar

The reader may wonder, why would a college offer courses on grammar – a topic that is usually thought of as part of junior high school curriculum (or even GRAMMAR school curriculum)? Well, the topic of this book is not the same thing that most people probably think of as grammar.

What is taught as grammar in primary and secondary school is what linguists call 'prescriptive grammar'. It consists of admonitions not to use certain forms or constructions that are common in everyday speech. A prescriptive grammar might contain rules like:

> Be sure to never split an infinitive.
>
> Prepositions are bad to end sentences with.

As modern linguists our concerns are very different. We view human language as a natural phenomenon amenable to scientific investigation, rather than something to be regulated by the decrees of authorities. Your seventh grade math teacher might have told you the (apocryphal) story about how the Indiana legislature almost passed a bill establishing the value of π as 3, and everybody in class no doubt laughed at such foolishness. Most linguists regard prescriptive grammar as silly in much the same way: natural phenomena simply cannot be legislated.

Of course, unlike the value of π, the structure of language is a product of human activity, and that can be legislated. And we do not deny the existence of powerful social and economic reasons for learning the grammatical norms of educated people.[1] But how these norms get established and influence the evolution of languages is a (fascinating) question for sociolinguistics and/or historical linguistics, not for syntactic theory. Hence, it is beyond the scope of this book. Similarly, we will not address issues of educational policy, except to say that in dismissing traditional (prescriptive) grammar instruction, we are not denying that attention to linguistic structure in the classroom can turn students into more effective speakers and writers. Indeed, we would welcome more enlightened grammar instruction in the schools. (See Nunberg 1983 and Cameron 1995 for insightful discussion of these issues.) Our concern instead is with language as it is used in everyday communication; and the rules of prescriptive grammar are of little help in describing actual usage.

[1] By the same token, there may well be good economic reasons for standardizing a decimal approximation to π (though 3 is almost certainly far too crude an approximation for most purposes).

So, if modern grammarians don't worry about split infinitives and the like, then what do they study? It turns out that human languages are amazingly complex systems, whose inner workings can be investigated in large part simply by consulting the intuitions of native speakers. We employ this technique throughout this book, using our own intuitions about English as our principal source of data. In keeping with standard linguistic practice, we will use an asterisk to mark an expression that is not well-formed – that is, an expression that doesn't 'sound good' to our ears. Here are some examples from English:

Example 1 The adjectives *unlikely* and *improbable* are virtually synonymous: we talk about unlikely or improbable events or heroes, and we can paraphrase *It is improbable that Lee will be elected* by saying *It is unlikely that Lee will be elected.* This last sentence is synonymous with *Lee is unlikely to be elected.* So why does it sound so strange to say **Lee is improbable to be elected*?

Example 2 The sentences *They saw Pat with Chris* and *They saw Pat and Chris* are near paraphrases. But if you didn't catch the second name, it would be far more natural to ask *Who did they see Pat with?* than it would be to ask **Who did they see Pat and?* Why do these two nearly identical sentences differ with respect to how we can question their parts? Notice, by the way, that the question that sounds well-formed (or 'grammatical' in the linguist's sense) is the one that violates a standard prescriptive rule. The other sentence is so blatantly deviant that prescriptive grammarians would never think to comment on the impossibility of such sentences. Prescriptive rules typically arise because human language use is innovative, leading languages to change. If people never use a particular construction – like the bad example above – there's no point in bothering to make up a prescriptive rule to tell people not to use it.

Example 3 The two sentences *Something disgusting has slept in this bed* and *Something disgusting has happened in this bed* appear on the surface to be grammatically completely parallel. So why is it that the first has a passive counterpart: *This bed has been slept in by something disgusting*, whereas the second doesn't: **This bed has been happened in by something disgusting*?

These are the sorts of questions contemporary grammarians try to answer. The first two will eventually be addressed in this text, but the third Will not.[2] The point of introducing them here is to illustrate a fundamental fact that underlies all modern work in theoretical syntax:

> Every normal speaker of any natural language has acquired an immensely rich and systematic body of unconscious knowledge, which can be investigated by consulting speakers' intuitive judgments.

In other words, knowing a language involves mastering an intricate system full of surprising regularities and idiosyncrasies. Languages are objects of considerable complexity, which can be studied scientifically. That is, we can formulate general hypotheses about linguistic structure and test them against the facts of particular languages.

The study of grammar on this conception is a field in which hypothesis-testing is particularly easy: the linguist can simply ask native speakers whether the predictions

[2]For extensive discussion of the third question, see Postal 1986.

regarding well-formedness of crucial sentences are correct.[3] The term 'syntax' is often used instead of 'grammar' in technical work in linguistics. While the two terms are sometimes interchangeable, 'grammar' may also be used more broadly to cover all aspects of language structure; 'syntax', on the other hand, refers only to the ways in which words combine into phrases, and phrases into sentences – the form or structure of well-formed expressions.

Linguists divide grammar into 'syntax', 'semantics' (the study of linguistic meaning), 'morphology' (the study of word structure), and 'phonology' (the study of the sound patterns of language). Although these distinctions are conceptually clear, many phenomena in natural languages involve more than one of these components of grammar.

1.2 An Extended Example: Reflexive and Nonreflexive Pronouns

To get a feel for the sort of research syntacticians conduct, consider the following question:[4]

> In which linguistic environments do English speakers normally use reflexive pronouns (i.e. forms like *herself* or *ourselves*), and where does it sound better to use a nonreflexive pronoun (e.g. *her*, *she*, *us*, or *we*)?

To see how to approach an answer to this question, consider, first, some basic examples:

(1) a.*We like us.
b. We like ourselves.
c. She likes her. [where, she ≠ her]
d. She likes herself.
e. Nobody likes us.
f.*Leslie likes ourselves.
g.*Ourselves like us.
h.*Ourselves like ourselves.

These examples suggest a generalization along the following lines:

Hypothesis I: *A reflexive pronoun can appear in a sentence only if that sentence also contains a preceding expression that has the same referent (i.e. a preceding* COREFERENTIAL *expression); a nonreflexive pronoun cannot appear in a sentence that contains such an expression.*

[3]This methodology is not without its pitfalls. Judgments of acceptability show considerable variation across speakers. Moreover, they can be heavily influenced by context, both linguistic and nonlinguistic. Since linguists rarely make any serious effort to control for such effects, not all of the data employed in the syntax literature should be accepted without question. On the other hand, many judgments are so unequivocal that they can clearly be relied on. In more delicate cases, many linguists have begun to supplement judgments with data from actual usage, by examining grammatical patterns found in written and spoken corpora. The use of multiple sources and types of evidence is always a good idea in empirical investigations. See Schütze 1996 for a detailed discussion of methodological issues surrounding the use of judgment data in syntactic research.

[4]The presentation in this section owes much to the pedagogy of David Perlmutter; see Perlmutter and Soames (1979: chapters 2 and 3).

The following examples are different from the previous ones in various ways, so they provide a first test of our hypothesis:

(2) a. She voted for her. [she ≠ her]
b. She voted for herself.
c. We voted for her.
d.*We voted for herself.
e.*We gave us presents.
f. We gave ourselves presents.
g.*We gave presents to us.
h. We gave presents to ourselves.
i.*We gave us to the cause.
j. We gave ourselves to the cause.
k.*Leslie told us about us.
l. Leslie told us about ourselves.
m.*Leslie told ourselves about us.
n.*Leslie told ourselves about ourselves.

These examples are all predicted by Hypothesis I, lending it some initial plausibility. But here are some counterexamples:

(3) a. We think that Leslie likes us.
b.*We think that Leslie likes ourselves.

According to our hypothesis, our judgments in (3a,b) should be reversed. Intuitively, the difference between these examples and the earlier ones is that the sentences in (3) contain subordinate clauses, whereas (1) and (2) contain only simple sentences.

Exercise 1: Some Other Subordinate Clauses

Throughout the book we have provided exercises designed to allow you to test your understanding of the material being presented. Answers to these exercises can be found beginning on page 543.

It isn't actually the mere presence of the subordinate clauses in (3) that makes the difference. To see why, consider the following, which contain subordinate clauses but are covered by Hypothesis I.

(i) We think that she voted for her. [she ≠ her]
(ii) We think that she voted for herself.
(iii)*We think that herself voted for her.
(iv)*We think that herself voted for herself.

A. Explain how Hypothesis I accounts for the data in (i)-(iv).
B. What is it about the subordinate clauses in (3) that makes them different from those in (i)-(iv) with respect to Hypothesis I?

Given our investigation so far, then, we might revise Hypothesis I to the following:

Hypothesis II: *A reflexive pronoun can appear in a clause only if that clause also contains a preceding, coreferential expression; a nonreflexive pronoun cannot appear in any clause that contains such an expression.*

For sentences with only one clause (such as (1)-(2)), Hypothesis II makes the same predictions as Hypothesis I. But it correctly permits (3a) because *we* and *us* are in different clauses, and it rules out (3b) because *we* and *ourselves* are in different clauses.

However, Hypothesis II as stated won't work either:

(4) a. Our friends like us.
b.*Our friends like ourselves.
c. Those pictures of us offended us.
d.*Those pictures of us offended ourselves.
e. We found your letter to us in the trash.
f.*We found your letter to ourselves in the trash.

What's going on here? The acceptable examples of reflexive pronouns have been cases (i) where the reflexive pronoun is functioning as an object of a verb (or the object of a preposition that goes with the verb) and (ii) where the ANTECEDENT – that is, the expression it is coreferential with – is the subject or a preceding object of the same verb. If we think of a verb as denoting some sort of action or state, then the subject and objects (or prepositional objects) normally refer to the participants in that action or state. These are often called the ARGUMENTS of the verb. In the examples in (4), unlike many of the earlier examples, the reflexive pronouns and their antecedents are not arguments of the same verb (or, in other words, they are not COARGUMENTS). For example in (4b), *our* is just part of the subject of the verb *like*, and hence not itself an argument of the verb; rather, it is *our friends* that denotes participants in the liking relation. Similarly, in (4e) the arguments of *found* are *we* and *your letter to us*; *us* is only part of an argument of *found*.

So to account for these differences, we can consider the following:

Hypothesis III: *A reflexive pronoun must be an argument of a verb that has another preceding argument with the same referent. A nonreflexive pronoun cannot appear as an argument of a verb that has a preceding coreferential argument.*

Each of the examples in (4) contains two coreferential expressions (*we, us, our*, or *ourselves*), but none of them contains two coreferential expressions that are arguments of the same verb. Hypothesis III correctly rules out just those sentences in (4) in which the second of the two coreferential expressions is the reflexive pronoun *ourselves*.

Now consider the following cases:

(5) a. Vote for us!
b.*Vote for ourselves!
c.*Vote for you!
d. Vote for yourself!

In (5d), for the first time, we find a well-formed reflexive with no antecedent. If we don't want to append an *ad hoc* codicil to Hypothesis III,[5] we will need to posit a hidden subject (namely, *you*) in imperative sentences.

Similar arguments can be made with respect to the following sentences.

(6) a. We appealed to them$_1$ to vote for them$_2$. [them$_1$ $\neq$ them$_2$]
b. We appealed to them to vote for themselves.
c. We appealed to them to vote for us.

(7) a. We appeared to them to vote for them.
b.*We appeared to them to vote for themselves.
c. We appeared to them to vote for ourselves.

In (6), the pronouns indicate that *them* is functioning as the subject of *vote*, but it looks like it is the object of the preposition *to*, not an argument of *vote*. Likewise, in (7), the pronouns suggest that *we* should be analyzed as an argument of *vote*, but its position suggests that it is an argument of *appeared*. So, on the face of it, such examples are problematical for Hypothesis III, unless we posit arguments that are not directly observable. We will return to the analysis of such cases in later chapters.

You can see that things get quite complex quite fast, requiring abstract notions like 'coreference', being 'arguments of the same verb', and 'phantom arguments' that the rules for pronoun type must make reference to. And we've only scratched the surface of this problem. For example, all the versions of the rules we have come up with so far predict that nonreflexive forms of a pronoun should appear only in positions where their reflexive counterparts are impossible. But this is not quite true, as the following examples illustrate:

(8) a. We wrapped the blankets around us.
b. We wrapped the blankets around ourselves.
c. We admired the pictures of us in the album.
d. We admired the pictures of ourselves in the album.

It should be evident by now that formulating precise rules characterizing where English speakers use reflexive pronouns and where they use nonreflexive pronouns will be a difficult task. We will return to this task in Chapter 7. Our reason for discussing it here was to emphasize the following points:

- Normal use of language involves the mastery of an intricate system, which is not directly accessible to conscious reflection.
- Speakers' tacit knowledge of language can be studied by formulating hypotheses and testing their predictions against intuitive judgments of well-formedness.
- The theoretical machinery required for a viable grammatical analysis could be quite abstract.

[5]For example, an extra clause that says: 'unless the sentence is imperative, in which case a second person reflexive is well-formed and a second person nonreflexive pronoun is not.' This would rule out the offending case but not in any illuminating way that would generalize to other cases.

1.3 Remarks on the History of the Study of Grammar

The conception of grammar we've just presented is quite a recent development. Until about 1800, almost all linguistics was primarily prescriptive. Traditional grammar (going back hundreds, even thousands of years, to ancient India and ancient Greece) was developed largely in response to the inevitable changing of language, which is always (even today) seen by most people as its deterioration. Prescriptive grammars have always been attempts to codify the 'correct' way of talking. Hence, they have concentrated on relatively peripheral aspects of language structure. On the other hand, they have also provided many useful concepts for the sort of grammar we'll be doing. For example, our notion of parts of speech, as well as the most familiar examples (such as *noun* and *verb*) come from the ancient Greeks.

A critical turning point in the history of linguistics took place at the end of the eighteenth century. It was discovered at that time that there was a historical connection among most of the languages of Europe, as well as Sanskrit and other languages of India (plus some languages in between).[6] This led to a tremendous flowering of the field of historical linguistics, centered on reconstructing the family tree of the Indo-European languages by comparing the modern languages with each other and with older texts. Most of this effort concerned the systematic correspondences between individual words and the sounds within those words. But syntactic comparison and reconstruction was also initiated during this period.

In the early twentieth century, many linguists, following the lead of the Swiss scholar Ferdinand de Saussure, turned their attention from the historical (or 'diachronic'[7]) study to the 'synchronic'[8] analysis of languages – that is, to the characterization of languages at a given point in time. The attention to synchronic studies encouraged the investigation of languages that had no writing systems, which are much harder to study diachronically since there is no record of their earlier forms.

In the United States, these developments led linguists to pay far more attention to the indigenous languages of the Americas. Beginning with the work of the anthropological linguist Franz Boas, American linguistics for the first half of the twentieth century was very much concerned with the immense diversity of languages. The Indo-European languages, which were the focus of most nineteenth-century linguistic research, constitute only a tiny fraction of the approximately five thousand known languages. In broadening this perspective, American linguists put great stress on developing ways to describe languages that would not forcibly impose the structure of a familiar language (such as Latin or English) on something very different; most, though by no means all, of this work emphasized the differences among languages. Some linguists, notably Edward Sapir and Benjamin Lee Whorf, talked about how language could provide insights into how people think. They tended to emphasize alleged differences among the thought patterns of speakers of different languages. For our purposes, their most important claim is that the structure of language can provide insight into human cognitive processes. This idea has

[6]The discovery is often attributed to Sir William Jones who announced such a relationship in a 1786 address, but others had noted affinities among these languages before him.

[7]From the Greek: *dia* 'across' plus *chronos* 'time'

[8]*syn* 'same, together' plus *chronos*.

wide currency today, and, as we shall see below, it constitutes one of the most interesting motivations for studying syntax.

In the period around World War II, a number of things happened to set the stage for a revolutionary change in the study of syntax. One was that great advances in mathematical logic provided formal tools that seemed well suited for application to studying natural languages. A related development was the invention of the computer. Though early computers were unbelievably slow and expensive by today's standards, some people immediately saw their potential for natural language applications, such as machine translation or voice typewriters.

A third relevant development around mid-century was the decline of behaviorism in the social sciences. Like many other disciplines, linguistics in America at that time was dominated by behaviorist thinking. That is, it was considered unscientific to posit mental entities or states to account for human behaviors; everything was supposed to be described in terms of correlations between stimuli and responses. Abstract models of what might be going on inside people's minds were taboo. Around 1950, some psychologists began to question these methodological restrictions, and to argue that they made it impossible to explain certain kinds of facts. This set the stage for a serious rethinking of the goals and methods of linguistic research.

In the early 1950s, a young man named Noam Chomsky entered the field of linguistics. In the late '50s, he published three things that revolutionized the study of syntax. One was a set of mathematical results, establishing the foundations of what is now called 'formal language theory'. These results have been seminal in theoretical computer science, and they are crucial underpinnings for computational work on natural language. The second was a book called *Syntactic Structures* that presented a new formalism for grammatical description and analyzed a substantial fragment of English in terms of that formalism. The third was a review of B. F. Skinner's (1957) book *Verbal Behavior*. Skinner was one of the most influential psychologists of the time, and an extreme behaviorist. Chomsky's scathing and devastating review marks, in many people's minds, the end of behaviorism's dominance in American social science.

Since about 1960, Chomsky has been the dominant figure in linguistics. As it happens, the 1960s were a period of unprecedented growth in American academia. Most linguistics departments in the United States were established in the period between 1960 and 1980. This helped solidify Chomsky's dominant position.

One of the central tenets of the Chomskyan approach to syntax, known as 'generative grammar', has already been introduced: hypotheses about linguistic structure should be made precise enough to be testable. A second somewhat more controversial one is that the object of study should be the unconscious knowledge underlying ordinary language use. A third fundamental claim of Chomsky's concerns the biological basis of human linguistic abilities. We will return to this claim in the next section.

Within these general guidelines there is room for many different theories of grammar. Since the 1950s, generative grammarians have explored a wide variety of choices of formalism and theoretical vocabulary. We present a brief summary of these in Appendix B, to help situate the approach presented here within a broader intellectual landscape.

1.4 Why Study Syntax?

Students in syntax courses often ask about the point of such classes: why should one study syntax?

Of course, one has to distinguish this question from a closely related one: why DO people study syntax? The answer to that question is perhaps simpler: exploring the structure of language is an intellectually challenging and, for many people, intrinsically fascinating activity. It is like working on a gigantic puzzle – one so large that it could occupy many lifetimes. Thus, as in any scientific discipline, many researchers are simply captivated by the complex mysteries presented by the data themselves – in this case a seemingly endless, diverse array of languages past, present and future.

This reason is, of course, similar to the reason scholars in any scientific field pursue their research: natural curiosity and fascination with some domain of study. Basic research is not typically driven by the possibility of applications. Although looking for results that will be useful in the short term might be the best strategy for someone seeking personal fortune, it wouldn't be the best strategy for a society looking for long-term benefit from the scientific research it supports. Basic scientific investigation has proven over the centuries to have long-term payoffs, even when the applications were not evident at the time the research was carried out. For example, work in logic and the foundations of mathematics in the first decades of the twentieth century laid the theoretical foundations for the development of the digital computer, but the scholars who did this work were not concerned with its possible applications. Likewise, we don't believe there is any need for linguistic research to be justified on the basis of its foreseeable uses. Nonetheless, we will mention three interrelated reasons that one might have for studying the syntax of human languages.

1.4.1 A Window on the Structure of the Mind

One intellectually important rationale for the study of syntax has been offered by Chomsky. In essence, it is that language – and particularly, its grammatical organization – can provide an especially clear window on the structure of the human mind.[9]

Chomsky claims that the most remarkable fact about human language is the discrepancy between its apparent complexity and the ease with which children acquire it. The structure of any natural language is far more complicated than those of artificial languages or of even the most sophisticated mathematical systems. Yet learning computer languages or mathematics requires intensive instruction (and many students still never master them), whereas every normal child learns at least one natural language merely through exposure. This amazing fact cries out for explanation.[10]

Chomsky's proposed explanation is that most of the complexity of languages does not have to be learned, because much of our knowledge of it is innate: we are born knowing about it. That is, our brains are 'hardwired' to learn certain types of languages.

[9] See Katz and Postal 1991 for arguments against the dominant Chomskyan conception of linguistics as essentially concerned with psychological facts.

[10] Chomsky was certainly not the first person to remark on the extraordinary facility with which children learn language, but, by giving it a central place in his work, he has focused considerable attention on it.

More generally, Chomsky has argued that the human mind is highly modular. That is, we have special-purpose 'mental organs' that are designed to do particular sorts of tasks in particular ways. The language organ (which, in Chomsky's view, has several largely autonomous submodules) is of particular interest because language is such a pervasive and unique part of human nature. All people use language, and (he claims) no other species is capable of learning anything much like human language. Hence, in studying the structure of human languages, we are investigating a central aspect of human nature.

This idea has drawn enormous attention not only from linguists but also from people outside linguistics, especially psychologists and philosophers. Scholars in these fields have been highly divided about Chomsky's innateness claims. Many cognitive psychologists see Chomsky's work as a model for how other mental faculties should be studied, while others argue that the mind (or brain) should be regarded as a general-purpose thinking device, without specialized modules. In philosophy, Chomsky provoked much comment by claiming that his work constitutes a modern version of Descartes' doctrine of innate ideas.

Chomsky's innateness thesis and the interdisciplinary dialogue it stimulated were major factors in the birth of the new interdisciplinary field of cognitive science in the 1970s. (An even more important factor was the rapid evolution of computers, with the concomitant growth of artificial intelligence and the idea that the computer could be used as a model of the mind.) Chomsky and his followers have been major contributors to cognitive science in the subsequent decades.

One theoretical consequence of Chomsky's innateness claim is that all languages must share most of their structure. This is because all children learn the languages spoken around them, irrespective of where their ancestors came from. Hence, the innate knowledge that Chomsky claims makes language acquisition possible must be common to all human beings. If this knowledge also determines most aspects of grammatical structure, as Chomsky says it does, then all languages must be essentially alike. This is a very strong universal claim.

In fact, Chomsky often uses the term 'Universal Grammar' to mean the innate endowment that makes language acquisition possible. A great deal of the syntactic research since the late 1960s has been concerned with identifying linguistic universals, especially those that could plausibly be claimed to reflect innate mental structures operative in language acquisition. As we proceed to develop the grammar in this text, we will ask which aspects of our grammar are peculiar to English and which might plausibly be considered universal.

If Chomsky is right about the innateness of the language faculty, it has a number of practical consequences, especially in fields like language instruction and therapy for language disorders. For example, since there is evidence that people's innate ability to learn languages is far more powerful very early in life (specifically, before puberty) than later, it seems most sensible that elementary education should have a heavy emphasis on language, and that foreign language instruction should not be left until secondary school, as it is in most American schools today.

1.4.2 A Window on the Mind's Activity

If you stop and think about it, it's really quite amazing that people succeed in communicating by using language. Language seems to have a number of design properties that get in the way of efficient and accurate communication of the kind that routinely takes place.

First, it is massively ambiguous. Individual words, for example, often have not just one but a number of meanings, as illustrated by the English examples in (9).

(9) a. Leslie used a *pen*. ('a writing implement')
b. We put the pigs in a *pen*. ('a fenced enclosure')
c. We need to *pen* the pigs to keep them from getting into the corn. ('to put in a fenced enclosure')
d. They should *pen* the letter quickly. ('to write')
e. The judge sent them to the *pen* for a decade. ('a penitentiary')

(10) a. The cheetah will *run* down the hill. ('to move fast')
b. The president will *run*. ('to be a political candidate')
c. The car won't *run*. ('to function properly')
d. This trail should *run* over the hill. ('to lead')
e. This dye will *run*. ('to dissolve and spread')
f. This room will *run* $200 or more. ('to cost')
g. She can *run* an accelerator. ('to operate')
h. They will *run* the risk. ('to incur')
i. These stockings will *run*. ('to tear')
j. There is a *run* in that stocking. ('a tear')
k. We need another *run* to win. ('a score in baseball')
l. Fats won with a *run* of 20. ('a sequence of successful shots in a game of pool')

To make matters worse, many sentences are ambiguous not because they contain ambiguous words, but rather because the words they contain can be related to one another in more than one way, as illustrated in (11).

(11) a. Lee saw the student with a telescope.
b. I forgot how good beer tastes.

(11a) can be interpreted as providing information about which student Lee saw (the one with a telescope) or about what instrument Lee used (the telescope) to see the student. Similarly, (11b) can convey either that the speaker forgot how GOOD beer (as opposed to bad or mediocre beer) tastes, or else that the speaker forgot that beer (in general) tastes good. These differences are often discussed in terms of which element a word like *with* or *good* is modifying (the verb or the noun).

These two types of ambiguity interact to produce a bewildering array of (often comical) ambiguities, like these:

(12) a. Visiting relatives can be boring.
b. If only Superman would stop flying planes!
c. That's a new car dealership.
d. I know you like the back of my hand.

e. An earthquake in Romania moved buildings as far away as Moscow and Rome.
f. The German shepherd turned on its master.
g. I saw that gas can explode.
h. Max is on the phone now.
i. The only thing capable of consuming this food has four legs and flies.
j. I saw her duck.

This is not the end of the worrisome design properties of human language. Many words are used to refer to different things on different occasions of utterance. Pronouns like *them*, *(s)he*, *this*, and *that* pick out different referents almost every time they are used. Even seemingly determinate pronouns like *we* don't pin down exactly which set of people the speaker is referring to (compare *We have two kids/a city council/a lieutenant governor/50 states/oxygen-based life here*). Moreover, although certain proper names like *Sally Ride*, *Sandra Day O'Connor*, or *Condoleezza Rice* might reliably pick out the same person almost every time they are used, most conversations are full of uses of names like *Chris*, *Pat*, *Leslie*, *Sandy*, etc. that vary wildly in their reference, depending on who's talking to whom and what they're talking about.

Add to this the observation that some expressions seem to make reference to 'covert elements' that don't exactly correspond to any one word. So expressions like *in charge* and *afterwards* make reference to missing elements of some kind – bits of the meaning that have to be supplied from context. Otherwise, discourses like the following wouldn't make sense, or would at best be incomplete:

(13) a. I'm creating a committee. Kim – you're in charge. [in charge of what? – the committee]
b. Lights go out at ten. There will be no talking afterwards. [after what? – after ten]

The way something is said can also have a significant effect on the meaning expressed. A rising intonation, for example, on a one word utterance like *Coffee?* would very naturally convey 'Do you want some coffee?' Alternatively, it might be used to convey that 'coffee' is being offered as a tentative answer to some question (say, *What was Columbia's former number-one cash crop?*). Or even, in the right context, the same utterance might be used in seeking confirmation that a given liquid was in fact coffee.

Finally, note that communication using language leaves a great deal unsaid. If I say to you *Can you give me a hand here?* I'm not just requesting information about your abilities, I'm asking you to help me out. This is the unmistakable communicative intent, but it wasn't literally said. Other examples of such inference are similar, but perhaps more subtle. A famous example[11] is the letter of recommendation saying that the candidate in question has outstanding penmanship (and saying nothing more than that!).

Summing all this up, what we have just seen is that the messages conveyed by utterances of sentences are multiply ambiguous, vague, and uncertain. Yet somehow, in spite of this, those of us who know the language are able to use it to transmit messages to one another with considerable precision – far more precision than the language itself would seem to allow. Those readers who have any experience with computer programming or with mathematical logic will appreciate this dilemma instantly. The very idea of designing a programming language or a logical language whose predicates are ambiguous or whose variables are left without assigned values is unthinkable. No computer can process linguistic expressions unless it 'knows' precisely what the expressions mean and what to do with them.

[11] This example is one of many due to the late H. Paul Grice, the philosopher whose work forms the starting point for much work in linguistics on problems of PRAGMATICS, how people 'read between the lines' in natural conversation; see Grice 1989.

The fact of the matter is that human language-users are able to do something that modern science doesn't understand well enough to replicate via computer. Somehow, people are able to use nonlinguistic information in such a way that they are never even aware of most of the unwanted interpretations of words, phrases, and sentences. Consider again the various senses of the word *pen*. The 'writing implement' sense is more common – that is, more frequent in the language you've been exposed to (unless you're a farmer or a prisoner) – and so there is an inherent bias toward that sense. You can think of this in terms of 'weighting' or 'degrees of activation' of word senses. In a context where farm animals are being discussed, though, the weights shift – the senses more closely associated with the subject matter of the discourse become stronger in this case. As people direct their attention to and through a given dialogue, these sense preferences can fluctuate considerably. The human sense selection capability is incredibly robust, yet we have only minimal understanding of the cognitive mechanisms that are at work. How exactly does context facilitate our ability to locate the correct sense?

In other cases, it's hard to explain disambiguation so easily in terms of affinity to the domain of discourse. Consider the following contrast:

(14) a. They found the book on the table.
b. They found the book on the atom.

The preposition *on* modifies the verb in (14a) and the noun in (14b), yet it seems that nothing short of rather complex reasoning about the relative size of objects would enable someone to choose which meaning (i.e. which modification) made sense. And we do this kind of thing very quickly, as you can see from (15):

(15) After finding the book on the atom, Sandy went into class, confident that there would be no further obstacles to getting that term paper done.

When you finish reading this sentence, you do not need to go back and think about whether to interpret *on* as in (14a) or (14b). The decision about how to construe *on* is made by the time the word *atom* is understood.

When we process language, we integrate encyclopedic knowledge, plausibility information, frequency biases, discourse information, and perhaps more. Although we don't yet know exactly how we do it, it's clear that we do it very quickly and reasonably accurately. Trying to model this integration is probably the most important research task now facing the study of language.

Syntax plays a crucial role in all this. It imposes constraints on how sentences can or cannot be construed. The discourse context may provide a bias for the 'fenced enclosure' sense of *pen*, but it is the syntactic context that determines whether *pen* occurs as a noun or a verb. Syntax is also of particular importance to the development of language-

processing models, because it is a domain of knowledge that can be characterized more precisely than some of the other kinds of knowledge that are involved.

When we understand how language processing works, we probably will also understand quite a bit more about how cognitive processes work in general. This in turn will no doubt enable us to develop better ways of teaching language. We should also be better able to help people who have communicative impairments (and more general cognitive disorders). The study of human language-processing is an important sub-area of the study of human cognition, and it is one that can benefit immensely from precise characterization of linguistic knowledge of the sort that syntacticians seek to provide.

1.4.3 Natural Language Technologies

Grammar has more utilitarian applications, as well. One of the most promising areas for applying syntactic research is in the development of useful and robust natural language technologies. What do we mean by 'natural language technologies'? Roughly, what we have in mind is any sort of computer application that involves natural languages[12] in essential ways. These include devices that translate from one language into another (or perhaps more realistically, that provide translation assistance to someone with less than perfect command of a language), that understand spoken language (to varying degrees), that automatically retrieve information from large bodies of text stored on-line, or that help people with certain disabilities to communicate.

There is one application that obviously must incorporate a great deal of grammatical information, namely, grammar checkers for word processing. Most modern word processing systems include a grammar checking facility, along with a spell-checker. These tend to focus on the concerns of prescriptive grammar, which may be appropriate for the sorts of documents they are generally used on, but which often leads to spurious 'corrections'. Moreover, these programs typically depend on superficial pattern-matching for finding likely grammatical errors, rather than employing in-depth grammatical analysis. In short, grammar checkers can benefit from incorporating the results of research in syntax.

Other computer applications in which grammatical knowledge is clearly essential include those in which well-formed natural language output must be generated. For example, reliable software for translating one language into another must incorporate some representation of the grammar of the target language. If it did not, it would either produce ill-formed output, or it would be limited to some fixed repertoire of sentence templates.

Even where usable natural language technologies can be developed that are not informed by grammatical research, it is often the case that they can be made more robust by including a principled syntactic component. For example, there are many potential uses for software to reduce the number of keystrokes needed to input text, including facilitating the use of computers by individuals with motor disabilities or temporary impairments such as carpal tunnel syndrome. It is clear that knowledge of the grammar of English can help in predicting what words are likely to come next at an arbitrary point in a sentence. Software that makes such predictions and offers the user a set of choices for the next word or the remainder of an entire sentence – each of which can be

[12]That is, English, Japanese, Swahili, etc. in contrast to programming languages or the languages of mathematical logic.

inserted with a single keystroke – can be of great value in a wide variety of situations. Word prediction can likewise facilitate the disambiguation of noisy signals in continuous speech recognition and handwriting recognition.

But it's not obvious that all types of natural language technologies need to be sensitive to grammatical information. Say, for example, we were trying to design a system to extract information from an on-line database by typing in English questions (rather than requiring use of a special database query language, as is the case with most existing database systems). Some computer scientists have argued that full grammatical analysis of the queries is not necessary. Instead, they claim, all that is needed is a program that can extract the essential semantic information out of the queries. Many grammatical details don't seem necessary in order to understand the queries, so it has been argued that they can be ignored for the purpose of this application. Even here, however, a strong case can be made for the value of including a syntactic component in the software.

To see why, imagine that we are using a database in a law office, containing information about the firm's past and present cases, including records of witnesses' testimony. Without designing the query system to pay careful attention to certain details of English grammar, there are questions we might want to ask of this database that could be misanalyzed and hence answered incorrectly. For example, consider our old friend, the rule for reflexive and nonreflexive pronouns. Since formal database query languages don't make any such distinction, one might think it wouldn't be necessary for an English interface to do so either. But suppose we asked one of the following questions:

(16) a. Which witnesses testified against defendants who incriminated them?
b. Which witnesses testified against defendants who incriminated themselves?

Obviously, these two questions will have different answers, so an English language 'front end' that didn't incorporate some rules for distinguishing reflexive and nonreflexive pronouns would sometimes give wrong answers.

In fact, it isn't enough to tell reflexive from nonreflexive pronouns: a database system would need to be able to tell different reflexive pronouns apart. The next two sentences, for example, are identical except for the plurality of the reflexive pronouns:

(17) a. List all witnesses for the defendant who represented himself.
b. List all witnesses for the defendant who represented themselves.

Again, the appropriate answers would be different. So a system that didn't pay attention to whether pronouns are singular or plural couldn't be trusted to answer correctly.

Even features of English grammar that seem useless – things that appear to be entirely redundant – are needed for the analysis of some sentences that might well be used in a human-computer interaction. Consider, for example, English subject-verb agreement (a topic we will return to in some detail in Chapters 2–4). Since subjects are marked as singular or plural – *the dog* vs. *the dogs* – marking verbs for the same thing – *barks* vs. *bark* – seems to add nothing. We would have little trouble understanding someone who always left subject agreement off of verbs. In fact, English doesn't even mark past-tense verbs (other than forms of *be*) for subject agreement. But we don't miss agreement in the past tense, because it is semantically redundant. One might conjecture, therefore, that an English database querying system might be able simply to ignore agreement.

However, once again, examples can be constructed in which the agreement marking on the verb is the only indicator of a crucial semantic distinction. This is the case with the following pair:

(18) a. List associates of each witness who speaks Spanish.
b. List associates of each witness who speak Spanish.

In the first sentence, it is the witnesses in question who are the Spanish-speakers; in the second, it is their associates. These will, in general, not lead to the same answer.

Such examples could be multiplied, but these should be enough to make the point: Building truly robust natural language technologies – that is, software that will allow you to interact with your computer in YOUR language, rather than in ITS language – requires careful and detailed analysis of grammatical structure and how it influences meaning. Shortcuts that rely on semantic heuristics, guesses, or simple pattern-matching will inevitably make mistakes.

Of course, this is not to deny the value of practical engineering and statistical approximation. Indeed, the rapid emergence of natural language technology that is taking place in the world today owes at least as much to these as it does to the insights of linguistic research. Our point is rather that in the long run, especially when the tasks to be performed take on more linguistic subtlety and the accuracy of the performance becomes more critical, the need for more subtle linguistic analysis will likewise become more acute.

In short, although most linguists may be motivated primarily by simple intellectual curiosity, the study of grammar has some fairly obvious uses, even in the relatively short term.

1.5 Phenomena Addressed

Over the next fifteen chapters, we develop theoretical apparatus to provide precise syntactic descriptions. We motivate our formal machinery by examining various phenomena in English. We also address the applicability of our theory to other languages, particularly in some of the problems.

The following is a brief overview of the most important phenomena of English that we deal with. We omit many subtleties in this preliminary survey, but this should give readers a rough sense of what is to come.

- Languages are infinite. That is, there is no limit to the length of sentences, and most utterances have never been uttered before.
- There are different types of words – such as nouns, verbs, etc. – which occur in different linguistic environments.
- There are many constraints on word order in English. For example, we would say *Pat writes books*, not **Writes Pat books*, **Books writes Pat*, or **Pat books writes*.
- Some verbs require objects, some disallow them, and some take them optionally. So we get: *Pat devoured the steak*, but not **Pat devoured*; *Pat dined*, but not **Pat dined the steak*; and both *Pat ate the steak*, and *Pat ate*.
- Verbs agree with their subjects, so (in standard English) we wouldn't say **Pat write books* or **Books is interesting*.

- There is also a kind of agreement within noun phrases; for example, *this bird* but not **this birds*; *these birds* but not **these bird*; and *much water* but not **much bird* or **much birds*.
- Some pronouns have a different form depending on whether they are the subject of the verb or the object: *I saw them* vs. **Me saw them* or **I saw they*.
- As was discussed in Section 1.2, reflexive and nonreflexive pronouns have different distributions, based on the location of their antecedent.
- Commands are usually expressed by sentences without subjects, whose verbs show no agreement or tense marking, such as *Be careful!*
- Verbs come in a variety of forms, depending on their tense and on properties of their subject. Nouns usually have two forms: singular and plural. There are also cases of nouns and verbs that are morphologically and semantically related, such as *drive* and *driver*.
- Sentences with transitive verbs typically have counterparts in the passive voice, e.g. *The dog chased the cat* and *The cat was chased by the dog*.
- The word *there* often occurs as the subject of sentences expressing existential statements, as in *There is a unicorn in the garden*.
- The word *it* in sentences like *It is clear that syntax is difficult* does not refer to anything. This sentence is synonymous with *That syntax is difficult is clear*, where the word *it* doesn't even appear.
- Certain combinations of words, known as idioms, have conventional meanings, not straightforwardly inferable from the meanings of the words within them. Idioms vary in their syntactic versatility. Examples of idioms are *keep tabs on* and *take advantage of*.
- Pairs of sentences like *Pat seems to be helpful* and *Pat tries to be helpful*, though superficially similar, are very different in the semantic relationship between the subject and the main verb. This difference is reflected in the syntax in several ways; for example, *seems* but not *tries* can have the existential *there* as a subject: *There seems to be a unicorn in the garden* vs. **There tries to be a unicorn in the garden*.
- There is a similar contrast between the superficially similar verbs *expect* and *persuade*: *We expected several students to be at the talk* and *We persuaded several students to be at the talk* vs. *We expected there to be several students at the talk* but **We persuaded there to be several students at the talk*.
- Auxiliary ('helping') verbs in English (like *can*, *is*, *have*, and *do*) have a number of special properties, notably:
  - fixed ordering (*They have been sleeping* vs **They are having slept*)
  - occurring at the beginning of yes-no questions (*Are they sleeping?*)
  - occuring immediately before *not* (*They are not sleeping*)
  - taking the contracted form of *not*, written *n't* (*They aren't sleeping*)
  - occurring before elliptical (missing) verb phrases (*We aren't sleeping, but they are*)

- There is considerable dialectal variation in the English auxiliary system, notably British/American differences in the use of auxiliary *have* (*Have you the time?*) and the existence of a silent version of *is* in African American Vernacular English (*She the teacher*).
- A number of constructions (such as '*wh*-questions') involve pairing a phrase at the beginning of a sentence with a 'gap' – that is, a missing element later in the sentence. For example, in *What are you talking about? what* functions as the object of the preposition *about*, even though it doesn't appear where the object of a preposition normally does.

These are some of the kinds of facts that a complete grammar of English should account for. We want our grammar to be precise and detailed enough to make claims about the structure and meanings of as many types of sentence as possible. We also want these descriptions to be psychologically realistic and computationally tractable. Finally, despite our focus on English, our descriptive vocabulary and formalization should be applicable to all natural languages.

1.6 Summary

In this chapter, we have drawn an important distinction between prescriptive and descriptive grammar. In addition, we provided an illustration of the kind of syntactic puzzles we will focus on later in the text. Finally, we provided an overview of some of the reasons people have found the study of syntax inherently interesting or useful. In the next chapter, we look at some simple formal models that might be proposed for the grammars of natural languages and discuss some of their shortcomings.

1.7 Further Reading

An entertaining (but by no means unbiased) exposition of modern linguistics and its implications is provided by Pinker (1994). A somewhat more scholarly survey with a slightly different focus is presented by Jackendoff (1994). For discussion of prescriptive grammar, see Nunberg 1983, Cameron 1995, and Chapter 12 of Pinker's book (an edited version of which was published in *The New Republic*, January 31, 1994). For an overview of linguistic science in the nineteenth century, see Pedersen 1959. A succinct survey of the history of linguistics is provided by Robins (1967).

Among Chomsky's many writings on the implications of language acquisition for the study of the mind, we would especially recommend Chomsky 1959 and Chomsky 1972; a more recent, but much more difficult work is Chomsky 1986b. There have been few recent attempts at surveying work in (human or machine) sentence processing. Fodor et al. 1974 is a comprehensive review of early psycholinguistic work within the Chomskyan paradigm, but it is now quite dated. Garrett 1990 and Fodor 1995 are more recent, but much more limited in scope. For a readable, linguistically oriented, general introduction to computational linguistics, see Jurafsky and Martin 2000.

1.8 Problems

⚠ *This symbol before a problem indicates that it should not be skipped. The problem either deals with material that is of central importance in the chapter, or it introduces something that will be discussed or used in subsequent chapters.*

⚠ Problem 1: Judging Examples

For each of the following examples, indicate whether it is acceptable or unacceptable. (Don't worry about what prescriptivists might say: we want native speaker intuitions of what sounds right). If it is unacceptable, give an intuitive explanation of what is wrong with it, i.e. whether it:

a. fails to conform to the rules of English grammar (for any variety of English, to the best of your knowledge),
b. is grammatically well-formed, but bizarre in meaning (if so, explain why), or
c. contains a feature of grammar that occurs only in a particular variety of English, for example, slang, or a regional dialect (your own or another); if so, identify the feature. Is it stigmatized in comparison with 'standard' English?

If you are uncertain about any judgments, feel free to consult with others. Nonnative speakers of English, in particular, are encouraged to compare their judgments with others.

(i) Kim and Sandy is looking for a new bicycle.
(ii) Have you the time?
(iii) I've never put the book.
(iv) The boat floated down the river sank.
(v) It ain't nobody goin to miss nobody.
(vi) Terry really likes they.
(vii) Chris must liking syntax.
(viii) Aren't I invited to the party?
(ix) They wondered what each other would do.
(x) There is eager to be fifty students in this class.
(xi) They persuaded me to defend themselves.
(xii) Strings have been pulled many times to get people into Harvard.
(xiii) Terry left tomorrow.
(xiv) A long list of everyone's indiscretions were published in the newspaper.
(xv) Which chemical did you mix the hydrogen peroxide and?
(xvi) There seem to be a good feeling developing among the students.

⚠ **Problem 2: Reciprocals**

English has a 'reciprocal' expression *each other* (think of it as a single word for present purposes), which behaves in some ways like a reflexive pronoun. For example, a direct object *each other* must refer to the subject, and a subject *each other* cannot refer to the direct object:

(i) They like each other.
(ii)*Each other like(s) them.

A. Is there some general property that all antecedents of reciprocals have that not all antecedents of reflexives have? Give both grammatical and ungrammatical examples to make your point.

B. Aside from the difference noted in part (A), do reciprocals behave like reflexives with respect to Hypothesis III? Provide evidence for your answer, including both acceptable and unacceptable examples, illustrating the full range of types of configurations we considered in motivating Hypothesis III.

C. Is the behavior of reciprocals similar to that of reflexives in imperative sentences and in sentences containing *appeal* and *appear*? Again, support your answer with both positive and negative evidence.

D. Consider the following contrast:

They lost each other's books.
*They lost themselves' books.

Discuss how such examples bear on the applicability of Hypothesis III to reciprocals.

[*Hint: before you answer the question, think about what the verbal arguments are in the above sentences.*]

Problem 3: Ambiguity

Give a brief description of each ambiguity illustrated in (12) on page 11, saying what the source of ambiguity is – that is, whether it is lexical, structural (modificational), or both.

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'.

[1] Of 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.