CHAPTER 2

Language as a Mental Phenomenon

2.1 What do we mean by “mental”?

The remarkable first chapter of Noam Chomsky's Aspects of the Theory of Syntax (1965) sets in place an agenda for generative linguistic theory, much of which has survived intact for over thirty-five years. The present chapter and the next two will be devoted to evaluating and rearticulating this agenda, and to replying to some of the more common and longstanding criticisms of the approach.

We follow Aspects by starting with the issue of the status of linguistic description. The standard techniques of linguistic research lead us to some posited structure, say Fig. 1.1, for the sentence *The little star's beside a big star*. How is such a structure to be understood? The fundamental claim of Aspects is that this structure is more than just a useful description for the purposes of linguists. It is meant to be "psychologically real": it is to be treated as a model of something in the mind of a speaker of English who says or hears this sentence. What does this claim mean?

Often the answer is put in these terms: Fig. 1.1 is a model of a mental representation of the sentence. Unfortunately, I have to plunge right in and attempt to wean readers away from this terminology, which I think has led to unnecessary and prolonged misunderstanding. The problem is that the term "representation" suggests that it represents *something*—and for something to represent something else, it must represent *it to someone*. But we don't want to say that Fig. 1.1 "represents the sentence to the language user": that would suggest somehow that the language user has conscious access to all the structure in the figure, or could have it with sufficient introspective effort. Nor do we want to say that the figure represents the sentence to some entity within the language user's unconscious mind: that would conjure up the notorious homunculus, the
“little person in the brain” who (to use the term of Dennett 1991) sits in the “Cartesian theater” watching the show.

“Representation” belongs to a family of related terms that pervade cognitive science and that raise parallel problems. For instance, it is customary to speak of Fig. 1.1 as part of a symbolic theory of mental representation or of brain function; written symbols such as the phoneme b or the category NP are taken to model “symbols” in the mind. Now, the written symbols do symbolize something, namely the entities in the mind. But do the entities in the mind symbolize anything? The entity b in the mind doesn’t symbolize the phoneme b, it is the mental entity that makes the phoneme what it is. Furthermore, a symbol is a symbol by virtue of having a perceiver or community of perceivers, so using this terminology implicitly draws us into the homunculus problem again.

Even the apparently innocuous term “information” is not immune: something does not constitute information unless there is something or someone it can inform. The writing on the page and the linguistic sounds transmitted through the air do indeed inform people—but the phoneme b and the category NP in the head are among the things that the writing and sounds inform people of.

As some readers will recognize, I am making all this fuss to head off the thorny philosophical problem of intentionality: the apparent “aboutness” of thoughts and other mental entities in relation to the outside world. John Searle (1980), for example, argues against the possibility of ever making sense of analyses like Fig. 1.1 in mentalistic terms, on the grounds that having such a structure in one’s mind would not ever explain how it can be about the world, how it can symbolize anything. Jerry Fodor (1987, 1998), while deeply committed to the existence of mental representations, agrees with Searle that an account of intentionality is crucial; but then (if I may summarize his serious and complex argument in a sentence) he more or less teases himself in half trying to come up with a resolution of the ensuing paradoxes. The philosophical concerns with intentionality have traditionally been addressed to meaning (semantic/conceptual structure in Fig. 1.1); we will treat them in some detail in Chapters 9 and 10. But the same difficulties pertain, if more subtly, to the “symbols” of phonological and syntactic structure.

Accordingly, I propose to avoid all such problems from the outset by replacing the intentionality-laden terms “representation,” “symbol,” and “information” with appropriately neutral terms. I’ll call Fig. 1.1 a model of a “cognitive structure,” and I’ll call components such as the phoneme b and the category NP “cognitive entities” or “structural elements.” Instead of speaking of “encoding information,” I’ll use the old structuralist term “making distinctions.” Note of course that a structural element may itself be a structure: for instance b is composed of its distinctive features.

Our revised construal of Fig. 1.1 is therefore that it models a cognitive structure in the mind of a speaker. But there is still a problem: the term “mind.” The mind is traditionally understood as the seat of consciousness and volition; the “mind–body problem” concerns the relations of consciousness and volition to the physical world. Since at least Freud, we have also become accustomed to speak of the “unconscious mind”. Common parlance, following Freud, takes the unconscious mind to be just like the conscious mind except that we aren’t aware of it. Hence it is taken to be full of thoughts, images, and so forth that are at least in principle available to conscious introspection.

This notion of the unconscious is then often taken to be as far as one can go in describing phenomena as “mental.” From there on down, it’s all “body”–brain, to be more specific. This leaves no room in the mind for elaborate structures like Fig. 1.1, which go far beyond anything ever available to introspection. It leaves room only for neurons firing and thereby activating or inhibiting other neurons through synaptic connections. This is precisely the move Searle wants to make and Fodor wants to resist. In order for us to resist it successfully, we have to open up a new domain of description, as it were in between the Freudian unconscious and the physical meat.

In modern cognitive science, essentially following Chomsky’s usage, the term “mind” (and more recently “mind/brain”) has come to denote this in-between domain of description. It might be characterized as the functional organization and functional activity of the brain, some small part of which emerges in consciousness and most of which does not. Unfortunately, this usage tempts confusion with the everyday sense of the term: “It makes no sense to say you have an NP in mind when you utter The little star is . . .” Of course it doesn’t. To stave off such misunderstanding, I will introduce the term of art “f-mind” (“functional mind”) for this sense, to make clear its distinctness from common usage.

The standard way to understand “functional” organization and activity (some people call it “subsymbolic”) is in terms of the hardware–software distinction in computers: the brain is taken to parallel the hardware, the mind the software. When we speak of a particular computer running, say, Word 97, and speak of it storing certain data structures that enable it to run that program, we

---

1 In revising terminology one is faced with a number of choices, none ideal. One can persist in using “mind,” in which case readers (especially those picking up the book in the middle) are prone to understand the term in the everyday sense. Or one can create an entirely new and opaque term such as “cognizer” that leaves everyone cold. As a middle ground, I have chosen to adopt the traditional term but with a diacritic that flags it as a special technical usage. I apologize in advance for its awkwardness.

2 I should also make clear that this sense of “functional” is unrelated to the approach to linguistic theory called “functionalism,” which seeks to derive grammatical properties from the exigencies of communication (see section 2.5).
are speaking in functional terms—in terms of the logical organization of the task the computer is performing. In physical (hardware) terms, this functional organization is embodied in a collection of electronic components on chips, disks, and so forth, interacting through electrical impulses. Similarly, if we speak of the mind/brain determining visual contours or parsing a linguistic expression, we are speaking in functional terms; this functional organization is embodied in a collection of neurons engaging in electrical and chemical interactions. There is plenty of dispute about how seriously to take the computational analogy (e.g. Searle 1980 again), but within certain bounds it has proven a robust heuristic for understanding brain processes.

There are limits on this analogy. First, no one writes the “programs” that run in our minds. They have to develop indigenously, and we call this learning and development—an issue we return to in Chapter 4.

Second, it has become clear that, unlike a standard computer, the brain (and therefore the f-mind) has no “executive central processor” that controls its activities. Rather, it comprises a large number of specialized systems that interact in parallel to build our understanding of the world and to control our goals and actions in the world. Even what seems to be a unified subsystem such as vision has been found to be subdivided into many smaller interacting systems for detecting motion, detecting depth, coordinating reaching movements, recognizing faces, and so forth.

Third, the character of the “software” and “data structures” that constitute the f-mind are far more tightly bound up with the nature of the “hardware” than in a standard computer. An early attitude toward studying the f-mind was carried over from experience with computers, where the same program could be run on physically very different machines: the functional organization of the mind was treated as a mathematical function, relatively independent of its physical instantiation in the brain (see e.g. Arbib 1964; Pylyshyn 1984). It now has become clearer that the “software” is exquisitely tuned to what the “hardware” can do (in a way that, say, Word 97 is not especially tuned to the Pentium chip).

As a consequence, discoveries about brain properties are now believed to have a more direct bearing on functional properties than was previously thought, a welcome development. As Marr (1982) eloquently stresses, though, the connection is a two-way street: if it can be demonstrated that humans must in effect compute such-and-such a function in order to perform as they do on some task, then it is necessary to figure out how the brain’s neural circuitry could compute that function.²

² I should stress that deciding exactly what function people (or animals) compute is a matter for intense experimental investigation. Such investigation often shows that the f-mind resorts to “cheap tricks” rather than the mathematically most robust solution.

Even with these understandings of the relation between functional organization and neural instantiation, there has been a concerted attack on the usefulness of a theory of functional organization, coming this time not from philosophers but from certain communities in neuroscience and computational modeling (e.g. Rumelhart and McClelland 1986a; Churchland and Sejnowski 1992; Edelman 1992). According to this school of thought, the scientific reality is lodged in the neurons and the neurons alone; hence again there is no sense in developing models like Fig. 1.1.

I can understand the impulse behind this reductionist stance. The last two decades have seen an explosion of exciting new techniques for understanding the nervous system: recordings of the activity of individual neurons and the whole brain, computational modeling of perceptual and cognitive processes, and explanation of nervous system processes in terms of biochemical activity. Such research significantly deepens our understanding of the “hardware”—a quest with which I am altogether in sympathy. Furthermore, some aspects of “mental computation” in the functional sense are quite curious from the standpoint of standard algorithmic computation, but fall out rather naturally in neural network models (see Chapter 6). So there is no reason to relinquish the “Good Old-Fashioned Artificial Intelligence” treatment of the f-mind as a variety of serial and digital Turing machine, functionally quite unlike the brain.

On the other hand, researchers working within the reductionist stance often invoke it to delegitimate all the exquisitely detailed work done from the functional stance, including the work that leads to Fig. 1.1. Yet little has been offered to replace it. All we have at the moment is relatively coarse localization and timing of brain activity through imaging and studies of brain damage, plus recordings of individual neurons and small ensembles of them. With few exceptions (primarily in low-level vision, e.g. Hube and Wiesel 1968), it is far from understood exactly what any brain area does, how it does it, and what “data structures” it processes and stores. In particular, none of the new techniques has yet come near revealing how a cognitive structure as simple as a single speech sound is explained in terms of a physical embodiment in neurons.

Consequently, the bread-and-butter work that linguists do on, say, case-marking in Icelandic, stress in Moroccan Arabic, and reduplication in Tagalog has no home within this tradition, at least in the foreseeable future. Should linguists just put these sorts of study on ice till neuroscience catches up? I submit that it is worth considering an alternative stance that allows for insights from both approaches.

2.2 How to interpret linguistic notation mentally

Toward such an alternative: No one denies that cognitive structures subsist on a
neural substrate. And no one (I think) denies the importance of understanding how the neurons manage to accomplish language comprehension and production. So what might a result like Fig. 1.1 tell us about neural instantiation?

First, it is important to understand exactly what claims are made by the notation. It is obvious that speakers don’t have a direct counterpart of the symbol NP in their heads. Rather, what is significant about the symbol is only that it differs from the other syntactic categories, not how it is labeled. Similarly, in saying syntactic structure is modeled by a tree structure, we are not claiming that speakers literally have trees in their heads. In fact, we often replace tree structures such as (1a) with the “labeled bracket notation” illustrated in (1b). Some people even use a “box” notation like (1c), and there are still other notations.

\[
\begin{array}{c}
\text{(1) a.} \\
\text{Det} \quad \text{AP} \quad \text{N} \\
\quad \quad \quad \text{the} \quad \text{A \ star} \\
\quad \quad \quad \text{big} \\
\text{b.} \quad [\text{Det \ the}] \quad [\text{AP \ [A \ big]} \ [\text{N \ star}]] \\
\text{c.} \quad \text{Det} \quad \text{AP} \quad \text{N} \\
\quad \quad \quad \text{the} \quad \text{A \ big} \quad \text{star}
\end{array}
\]

All of these are ways of notating the theoretical claims (a) that words belong to syntactic categories; (b) that words are in linear order; and (c) that words group hierarchically into larger constituents that also belong to syntactic categories. These aspects of the notation, then, must be reflected somehow in neural instantiation. Beyond these aspects, the choice of notation is solely a matter of convenience.

With this understanding in mind, we can think of the combination of the states of all the neurons in the relevant parts of the brain as defining a “state-space” with a huge number of dimensions. When someone hears or produces the sentence *The little star’s beside a big star*, their brain can be thought of as being at some point in that state-space; it will be at another point for each different linguistic expression. The notation in which Fig. 1.1 is couched encodes hypotheses about the significant dimensions of the state-space, and each element in the notation encodes a position in one (or more) of those dimensions.

For instance, the notation $b$ encodes a position in the subspace of “phoneme choice”; in turn, the phonological distinctive feature notation (Fig. 1.2) encodes subdimensions such as consonanthood, voicing, and position of articulation within that subspace. The notation NP encodes a position in the subspace of “syntactic category.” Its feature decomposition (alluded to in section 1.4) positions it in the dimensions of that subspace: it is a “phrasal” category (as opposed to “lexical” categories such as Noun) and “nominal” (as opposed to “verbal,” “adjectival,” and so forth). The fact that $b$ is positioned in phonological structure and NP in syntactic structure encodes a hypothesis that certain dimensions of the overall state-space are more tightly linked than others, in particular that there are significant groupings of dimensions that can in functional terms be referred to as phonology and syntax. In other words, the functional approach is to be taken as making hypotheses about the significant dimensions of variation of brain states, and about the shape of the space of distinctions the brain can make.\(^3\)

At least two interesting problems arise in making this conception work. The first is that the functional state-space in language is usually taken to be discrete or categorical. A phoneme is a $b$ or a $p$ but not something in between; a syntactic category is an NP or an AP but not something in between. By contrast, neural computation appears to be somewhat graded, a matter of degree of activation and synaptic strength. However, the seeds of accommodation are already present, I think. On the neural side, it is recognized that some neurons have highly tuned responses, making very sharp distinctions between what they do and do not fire to. Still, even the sharpest tuning is not absolute.

Looking at the linguistic side: In acoustic phonetics it has been known for a long time that categorial perception of phonemes is not absolutely sharp. Rather, there is a narrow range of acoustic inputs between, say, $p$ and $b$ that produce variation and context-dependency in hearers’ judgments (e.g. Liberman and Studdert-Kennedy 1977). It has also become clear that in semantics, graded categories are altogether the norm (see section 11.6). And in syntax, indeterminacy has occasionally been recognized around the edges of categories (e.g. Ross’s (1972) “squishy categories” and Culicover’s (1999) proposal of a continuum of possible categories).

These observations in the functional domain are surely reflections of non-categorial behavior in the hardware. They force us to recognize that the dimensions of linguistic state-space are to some degree continuous rather than digital.

\(^{3}\) Notice how the locution “space of distinctions the brain can make” strikes the ear differently than the more traditional “kinds of information the brain can encode.” More austere, less intentional.
They appear to have sharp categorial distinctions because there are loci of relative stability within a dimension of variation. The degree of “discreteness vs. gradedness” in a dimension then reflects the relative steepness of stability gradients in the neighborhood of such loci. The discrete categories of a functional description are therefore to be regarded as a useful approximation to the behavior of the relevant part of the neural substrate, with the recognition that there will be situations where the approximation breaks down and we must resort to more elaborate modes of functional description.  

The second problem in bringing neural and functional descriptions into potential alignment is, I think, more severe. It arises from the fact that the functional description of language is inherently combinatorial. For example, we cannot simply say that the position in state-space encoded by Fig. 1.1 is the sum of the positions of all its speech sounds. Each speech sound is a position in phonological space that excludes all other positions. To simply sum them would be to say that the phonological part of state-space invoked by the sentence is simultaneously 0, a, l, i, and so on—a total mush. Rather it is necessary to specify the overall state (in part) as 0 followed by a followed by l and so on. But it is not clear how to set up a binary relation such as “followed by” in terms of a state-space. Similar difficulties emerge in instantiating hierarchical tree structures in terms of a state-space. Moreover, because the size of utterances is unlimited, the state-space also must be unlimited. We will come back to this problem in Chapter 3, concluding that neural models for the most part have not come to grips with the depth of this problem. It thus remains a significant challenge to the possibility of melding the neural and functional approaches.

These first two sections have engaged in a very abstract discussion, not an ideal way to start an exposition of linguistic theory. But it is necessary in order to situate the inquiry properly. In 1965 life was perhaps simpler: Chomsky could just say (1965: 4): “in the technical sense, linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behavior.” He clarifies this statement in a footnote (p. 193) that states the standard functionalist line of the period.

The mentalist, in this traditional sense, need make no assumptions about the possible physiological basis for the mental reality that he studies. In particular, he need not deny that there is such a basis. One would guess, rather, that it is the mentalistic studies that

will ultimately be of greatest value for the investigation of neurophysiological mechanisms, since they alone are concerned with determining abstractly the properties that such mechanisms must exhibit and the functions they must perform.

Now, with our greater understanding of brain function at the neural level, the dependency has to be regarded as going both ways. And, although it is clear from this passage that Chomsky intends by the term “mind” a functional description, the intervening years have done little to clear up the confusion the term engenders.

In trying to purge linguistic theory of intentional terms such as “representation,” “symbol,” and “information,” and by introducing the artificial terms “f-mind” and “f-mental,” my goal is to clarify an explicit niche of description between the traditional (and Freudian) mind and the neural stance. Within this niche, linguists can do their usual research, arriving at theories of f-mental structure along the lines of Fig. 1.1. It can be recognized that these analyses ride on the back of a neural substrate whose physical structure ultimately determines the character of f-mental structure. At the same time, to the extent that cognitive structures are justified by linguistic and psycholinguistic investigation, they set boundary conditions on the character of the neural substrate that embodies them: the brain must be a structure that can compute this.

2.3 Knowledge of language

At the outset of Aspects (1965: 3–4), Chomsky says:

Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. This seems to me to have been the position of the founders of modern general linguistics... 

We thus make a fundamental distinction between competence (the speaker-hearer’s knowledge of his language) and performance (the actual use of the language in concrete situations). Only under the idealization set forth in the preceding paragraph is performance a direct reflection of competence.

This passage contains some of the most crucial points of contention between linguists and their critics, so it calls for some exegesis.

Let us start with “know” and “knowledge”; these are terms in which Chomsky sets great store. Here, for example, is a passage from the somewhat later Reflections on Language:
Thus it does not seem to me quite accurate to take “knowledge of English” to be a capacity or ability, though it enters into the capacity or ability exercised in language use. In principle, one might have the cognitive structure that we call “knowledge of English,” fully developed, with no capacity to use this structure... (Chomsky 1972: 23)

The reason for this insistence becomes clearer in the next paragraph, where he mentions “the proper way to exorcise the ghost in the machine,” alluding to a term coined by Gilbert Ryle (1949). Ryle famously stressed the distinction between “knowing that,” a cognitive relation to a proposition, and “knowing how,” a capacity, ability, habit, or disposition to perform an action. “Knowing how” was taken by Ryle to be explicable in behaviorist terms: it could be acquired on the basis of reinforcement and could be verified simply by the organism’s behavior. It hence could avoid any mention of mind, which to Ryle smacked of “the ghost in the machine” (our homunculus of Section 2.1). Chomsky, for his part, rightly insists that linguistic theory cannot be reduced to behaviorist explanation. One way for him to do so is to distance himself from any terminology such as “ability” or “habit” that exudes the slightest whiff of behaviorism, as in this passage from early in Aspects (1965: 4):

Observed use of language or hypothesized dispositions to respond, habits, and so on, may provide evidence as to the nature of this mental reality, but surely cannot constitute the actual subject matter of linguistics, if this is to be a serious discipline.

In retrospect, I would say the distinction has been somewhat overdrawn. For one thing, disavowing “knowing how” seems to draw one toward the other member of the opposition, “knowing that.” But knowing English is not really “knowing that” anything.5 To have a command of, say, the English tense system is not to be acquainted with and committed to any set of propositions (at least in the standard sense). To claim that knowledge of language is a variety of “knowing that” would seem to put it in the conscious sector, or at least in the Freudian unconscious—certainly not in the functionalist domain where we want it.

Moreover, we now believe that any sort of practiced and overlearned ability or expertise, from swimming to playing the clarinet to playing chess, requires complex cognitive structures to support the overt behavior. Motor paralysis may mean that one cannot exercise that ability, but, whatever Ryle may have thought, the underlying cognitive capacity need not dissolve. So the distinction between language and other abilities is hardly as sharp as Chomsky wants to make it.

Again, “knowledge” is one of those intentional terms, and this accounts for a lot of the trouble it has raised and that Chomsky has had to fend off. Sometimes (e.g. 1986: 265–69) he suggests substituting a term of art such as “cognize”; then he shows that this term would be used essentially the same way as “know” and, concluding we are free to use any term we feel comfortable with, chooses to stay with “know.” Much of his discussion in this vein is in connection with critics who fail to recognize the distinctness of the f-mental domain in which linguistic structure resides. To keep matters straight, I will append the obnoxious “f-” to the term when necessary, speaking of “f-knowledge.”

### 2.4 Competence versus performance

We now turn our attention to the heavily loaded distinction between competence and performance.6 In the quote at the beginning of the previous section, Chomsky alludes to the concern with competence—the speaker-hearer’s f-knowledge of his language—as “the position of the founders of modern general linguistics.” What is behind this assertion, I suspect, is that he is trying to justify doing what linguists have always done, namely analyze things like case systems, relative clauses, and stress placement—but in an explicitly mentalistic framework. Speaking strictly pragmatically, this sort of inquiry has yielded massive results, some of which are encapsulated in Fig. 1.1; there seems no reason to abandon it.

---

5 Except under a heavily referentially opaque interpretation of know. One can for instance say that Beth knows that every is a Determiner in English, even if she has never heard the term “Determiner”; one bases this assertion on Beth’s linguistic behavior, including perhaps her grammaticality judgments. But this is parallel to saying that Beth knows that the fifth note of Happy Birthday is a perfect fourth above the first note, based on the fact that she sings the song in tune—even though she knows no music theory and wouldn’t have the slightest idea what a “perfect fourth” is. This is hardly “knowing that” in the sense that Ryle intended.

6 It is important not to conflate the competence-performance distinction with two other distinctions in the literature. One is Chomsky’s own distinction between I-language (“internalized language”) and E-language (“externalized language”), emphasized in his 1986 book Knowledge of Language. I-language is the structure of language regarded mentalistically; it coincides more or less with competence. E-language, however, is (as I understand it) not the mechanisms that speakers use to exhibit linguistic behavior (i.e. performance), but either (a) external linguistic behavior of individuals or (b) language regarded as an object external to human minds, as an abstract object that subsists “in the community.” While Chomsky thinks studies of performance are potentially of interest (at least in Aspects), he maintains that the study of E-language will yield nothing of theoretical significance.

Another distinction sometimes conflated with competence/performance is Saussure’s (1915) often-cited distinction between langue and parole. Parole is supposed to be individuals’ speech acts; langue is supposed to be the language abstracted away from the individuals that speak it. Thus nothing here corresponds to either competence or performance in Chomsky’s sense; rather both terms correspond to aspects of E-language.
Chomsky makes the competence–performance distinction in part to ward off alternative proposals for how linguistics must be studied. In particular, he is justifiably resisting the behaviorists, who insisted that proper science requires counting every cough in the middle of a sentence as part of linguistic behavior. His resistance was particularly urgent at the time, because of the overwhelming influence of the behaviorist stance on American structuralist linguistics of the period (see Chomsky's remarks on Twaddell (1935) on p. 193 of Aspects, for example). Judging from remarks in Syntactic Structures (especially pp. 19–20), Chomsky is also trying to defend traditional linguistic practice against finite-state Markov theories that generate or analyze sentences in serial order, moving from state to state; he cites Shannon and Weaver (1949) and Hockett (1955) as instances of this approach. (Such approaches have not gone away: Elman (1990) is a recent reincarnation.)

The factors that Chomsky consigns to performance—the things he thinks a theory of linguistic competence should not be concerned with—are quite heterogeneous, and it is worth reviewing some of them to see how they fare in a more contemporary light.

A first factor is memory limitations. Presumably memory limitations account for the impossibility of producing a 6000-word sentence, every part of which is locally grammatical. One could not possibly keep all that in mind at once. For a more extreme case, one would die before completing a 60-million-word sentence. The point is that the theory of linguistic competence does not have to bother to rule such things out—there are other and more obvious extraneous factors.

A second factor in Chomsky's notion of performance is "distractions, shifts of attention and interest, and errors (random or characteristic)." There is now a strong tradition of studying speech errors and repairs (e.g. Fromkin 1971; Garrett 1975; Levelt 1989; Dell 1986), which has been useful to help justify the division of linguistic structure into domains along the lines shown in Fig. 1.1. If the domains picked out by such inquiry do not coincide with those proposed by linguistic theory, there should be a certain amount of concern. The theories of competence and performance should line up. We take up these issues in Chapter 7.

A third factor in performance is the distinction between knowledge and processing. Chomsky says (1965: 9), "When we say that a sentence has a certain derivation with respect to a particular generative grammar, we say nothing about how the speaker or hearer might proceed, in some practical or efficient way, to construct such a derivation. These questions belong to the theory of language use—the theory of performance." This basic distinction between the grammatical structure of sentences and the logic of sentence processing has remained essential to grammatical theory up to the present. For example, Fig.

1.1 models a cognitive structure in the f-mind of someone who either hears or speaks the sentence The little star's beside a big star. A hearer presumably constructs this structure first by deriving a phonological structure from the auditory signal and then using that to arrive at the syntactic and conceptual structures. But a speaker presumably starts out with a meaning to express and develops a syntactic and phonological structure from it. So although the structure is the same, the hearer and speaker produce its parts in different orders.

This is, I think, the essential difference between competence theories and processing theories. Competence theories are concerned with what the total structure is for either speaker or hearer. Processing theories are concerned with how the structure is built in real time, so they naturally bifurcate into different theories for the speaker and the hearer.

Chomsky's intuition that "the study of performance models incorporating generative grammars may be a fruitful study" has of course been borne out: there is now a thriving inquiry into the course of sentence processing and sentence production (sources too numerous to mention, but see Chapter 7), based on the structures uncovered by competence theory.

A final factor in Chomsky's conception of performance has to a degree fallen by the wayside: "intonational and stylistic factors" (Chomsky 1965: 10). For instance, Chomsky cites the natural intonation of (2) as a performance error: "the intonation breaks are ordinarily inserted in the wrong places (that is, after 'cat' and 'rat,' instead of where the main brackets [i.e. the syntactic boundaries] appear" (p. 13).

(2) This is the cat that caught the rat that stole the cheese.

(2) is now normally recognized as an example of mismatch between prosodic structure (one of the tiers of phonological structure) and syntactic structure, i.e. as not an error at all. We will return to this case in Chapter 5.

Chomsky also cites the following examples of grammatical sentences that are "unacceptable" for performance reasons (1965: 11):

(3) a. "I called the man who wrote the book that you told me about up.

b. "The man who the boy who the students recognized pointed out is a friend of mine.

(3a) would now most likely be treated as a competence error. The violation comes from the fact that the particle up at the end grammatically requires its own intonation contour; however, a short intonation contour at the end of a sentence, especially following a long contour, is highly dispreferred. This is part of the competence grammar of English, but in the prosodic rather than the syntactic domain.
(3b) is the well-known case of a doubly center-embedded sentence. A language that allows nouns to be freely modified by relative clauses cannot help but come up with cases like this, among the many other possible configurations. It would be artificial for the description of relative clauses to single out this particular configuration as ungrammatical. However, Miller and Chomsky (1963) showed that under certain reasonable assumptions about language processing, this configuration, unlike other relative clause constructions, would fortuitously create a major strain on working memory. They therefore concluded that it is a performance violation, not a competence violation—that is, it is one of those extrinsic factors that the competence theory need not mention.

It appears to me that the problem in (3b), like that in (3a), is at least in part prosodic. To see this, notice that the examples in (4) are parallel to (3b) in having doubly center-embedded relative clauses, but sound much better; (4c) in particular is nearly identical to (3b). Hence the problem with (3b) cannot be solely a matter of its syntactic parse. (I suggest reading these out loud; the bracketed commas indicate intonation breaks.)

(4) a. The movie that everyone I know raved about turned out to be lousy.
   b. That professor that the girl you brought home fell in love with won
   the Nobel Prize last week.
   c. The man who the boy we recognized pointed out is a friend of mine.

These sentences evidently sound better than (3a) because their innermost relative clauses are relatively shorter and therefore can form intonational units with the nouns they modify. For instance, everyone I know in (4a) can be an intonational unit in a way that the boy who the students recognized in (3b) cannot. Such phenomena involving length are reminiscent of prosodic violations like (3a). This is not to say that the problem with (3b) is all prosodic, but more aspects of competence appear to be involved than Chomsky realized in 1965.7

To sum up, what Chomsky lumps into performance actually constitutes a wide variety of phenomena. Some fall into basic facts about memory limitations; some into different aspects of the theory of sentence processing; and some now are subsumed under competence theory.

7 A detailed contemporary analysis of center-embedding as a processing violation, with massive reference to the psycholinguistic literature, appears in Gibson (1998). Christiansen and Chater (1999) offer a treatment of center-embedding difficulties in a connectionist framework. However, the task set to their networks is simply to get subject–verb agreement correct. Hence as far as their networks are concerned, a right-embedded structure such as Bill sees the boys [who like the girl [who eats ice cream]] is identical to a non-embedded structure such as Bill sees the boys like [something]; the girl eats ice cream; that is, for them, right-embedding is not embedding at all. In the light of this basic error in the linguistic analysis, it is hard to evaluate Christiansen and Chater’s claims about center-embedding.

It is interesting how the competence–performance distinction has been misunderstood in the literature. For instance, a 1984 edited volume (Bever et al. 1984) contains three papers with nearly identical pronouncements on competence vs. performance, by Walter Kintsch, Charles Osgood, and Roger Schank and Lawrence Birnbaum. Each of them rejects competence models on the grounds that they are about syntax, whereas a really adequate theory should deal with semantics and pragmatics as well (or instead!). Each then proposes his own theory of language processing that in fact incorporates implicitly a theory of language structure, a theory that, while not a Chomskyan generative grammar, is recognizably a competence theory.

These authors are correct in observing that generative grammar’s emphasis on syntax is not sufficient to account for language understanding; but so far as I know, Chomsky has never proposed leaving semantic interpretation out of the eventual competence theory. (Pragmatics may be a different story, though.) In fact, their objections to generative grammar are more properly attributed to its “syntacticocentrism,” the assumption that the combinatorial power of language stems from syntax alone (see Chapter 5).

In the quote at the beginning of the previous section, Chomsky views competence as an idealization abstracted away from the full range of linguistic behavior. As such, it deserves as much consideration as any idealization in science: if it yields interesting generalizations it is worthwhile. Still, one can make a distinction between “soft” and “hard” idealizations. A “soft” idealization is acknowledged to be a matter of convenience, and one hopes eventually to find a natural way to re-integrate the excluded factors. A standard example is the fiction of a frictionless plane in physics, which yields important generalizations about forces and energy. But one aspires eventually to go beyond the idealization and integrate friction into the picture. By contrast, a “hard” idealization denies the need to go beyond itself; in the end it cuts itself off from the possibility of integration into a larger context.

It is my unfortunate impression that, over the years, Chomsky’s articulation of the competence–performance distinction has moved from relatively soft, as in the quote above that introduces it, to considerably harder, as suggested by the flavor of passages like this one:

It has sometimes been argued that linguistic theory must meet the empirical condition that it account for the ease and rapidity of parsing. But parsing does not, in fact, have these properties. . . . In general, it is not the case that language is readily usable or “designed for use.” (Chomsky and Lasnik 1993: 18)

There turn out to be good reasons that Chomsky has made this shift, emanating from the character of his version of competence theory. The approach to be
developed in Part II maintains the competence–performance distinction but lends itself better to integration with performance theory. I take such a relatively soft idealization to be a desideratum. (Bresnan and Kaplan (1982) and Pollard and Sag (1994) argue for a similar position; the theories of syntax they advocate lend themselves to direct interpretation in parsing terms.)

To sum up, I think the competence–performance distinction acknowledges the value of the sort of work linguists do in their day-to-day research, while recognizing that this work eventually must be placed in a broader psychological context. But I regard it as a pragmatic division of labor, a methodological convenience, not as a firewall to protect a certain form of inquiry.

Combining this discussion with that in section 2.1, we might best view the enterprise of understanding the human language capacity as naturally dividing into three lines of inquiry:

- Theory of competence: the functional characterization of the “data structures” stored and assembled in the f-mind in the course of language use.
- Theory of performance: the functional characterization of the use of these data structures in the course of language perception and production.
- Theory of neural instantiation: how the data structures and the processes that store and assemble them are realized in the brain.

Again, the distinctions between these theories are “soft” or methodological divisions, not ideological ones. One should welcome cross-talk among them, and in practice today we indeed find a steadily increasing amount of it.

2.5 Language in a social context (all too briefly)

Of course, as has often been observed, language does not subsist in the f-minds of individuals alone; it also exists in a social context. Some would say it exists only in a social context. In a sense this is true—if there were no other individuals with whom one wished to communicate, there would be little point to language as we know it. But on the other hand, the use of a language in a community presumes that the individuals have the cognitive capacity to produce and comprehend the signals they are sending each other. This cognitive capacity is what is being studied in the mentalistic framework.

Herbert Clark (1996) emphasizes that linguistic communication is not a one-way street, a speaker making utterances and a hearer passively taking them in. Rather, virtually all communication (but especially face to face) involves a delicate negotiation between speaker and hearer in a joint effort for both to be assured that the intended message gets across. Drawing on much previous research, Clark shows that many aspects of live speech that are often taken to be matters of “performance”, for example hesitations, repairs, and interjected um and like are often metalinguistic signals designed to help guide the hearer through the process of interpretation, and to elicit feedback from the hearer as to whether the message is getting across. He also shows how gestures, facial expressions, and direction of gaze are often used to amplify the message beyond what the spoken signal conveys.

My interpretation of Clark’s work is that it adds to the domains in which it is necessary and possible to describe linguistic competence. In addition to sentence structure itself, language users need cognitive structures that permit them to understand the goals of communication and to attach significance to the associated metalinguistic signals. As Clark emphasizes, many aspects of this communicative competence are subsumed under a larger theory of how people manage to carry out any sort of cooperative activity. Thus this kind of research is a bridge between strictly linguistic competence and more general social competence. And within this domain, we need as well a theory of performance that explains how people create and receive such metalinguistic signals in real time.

But on the other hand, a theory of communicative competence and/or performance doesn’t eliminate the need for a theory of grammatical structure. No matter how well speakers can coordinate their activity, they still have to put the verbs in the right place in their sentences. (I don’t think Clark would deny this, but one does sometimes see claims to the effect that one doesn’t need a theory of grammar because language is all about communicative effectiveness.)

Let us next turn back to Chomsky’s idealization of a homogeneous speech community. This has always been a soft idealization: generative linguists have never hesitated to discuss situations where speakers do not all have the same cognitive structures associated with the utterances they present as linguistic output. Such situations occur all the time in language and dialect contact, and above all when adults talk with young children. In these situations, Clark’s point becomes even clearer. Communicating requires not just putting a signal out there for the listener and hoping for the best, but mutually verifying that the message has gotten across to both participants’ satisfaction.

However, once we acknowledge that people do not necessarily have identical internalized cognitive structures for language, the question arises of what constitutes, say, “English”—or even “Standard American English.” I suggest that the use of language names is a harmless reification of the commonality in the linguistic f-knowledge of a perceived community of speakers. When we get down to dialect or individual differences, we can drop the idealization insofar as necessary without any problem. Again, this is common practice in linguistics.

Imagine, then, that to speak of a language in linguistics is a bit like speaking of a species in biology: one acknowledges that members of a species are not
genetically identical; and cases sometimes arise where what is apparently one species shades off imperceptibly over some geographical range into another. Does that mean there are no species? Some biologists think so. But as long as we regard the term as a convenient first approximation, there seems no harm in it.

It is worth mentioning, though, that this first-approximation reification of language very easily passes over unnoticed into a harder idealization, especially in everyday parlance. It is this idealization that, for instance, leads people to say that "the language" is degenerating because teenagers don't know how to talk anymore (they were saying that in the eighteenth century too!). It is also behind seeing the dictionary as an authority on the "correct meanings" of words rather than as an attempt to record how words are understood in the speech community. Even linguists adopt this stance all the time in everyday life (especially as teachers of students who can't write a decent paragraph). But once we go inside the heads of speakers to study their own individual cognitive structure, the stance must be dropped.

Now suppose, counterfactually, that speakers did have identical cognitive structures associated with their linguistic communication. There would still be reasons why communication cannot be perfect. For one thing, people are in different states of knowledge: they have different histories and different goals. In addition, one can never be absolutely certain, from the other's behavior, that one's message has been understood the way one intends. Rather, one settles for more or less certainty, depending on the situation. Telling someone how to perform surgery is more exacting than relating a personal anecdote in a bar, and one adjusts one's communicative expectations accordingly.

Moreover, the act of communication presents two conflicting desiderata. One goal is to get the meaning across with a minimum of physical effort on the part of both speaker and hearer. This creates a pressure towards brevity and abbreviation. But another goal is to convey the meaning as clearly as possible, which creates a pressure towards length and redundancy. Individual speech acts as well as the grammatical structure of languages reflect the tension between these two goals.

Of course, the more that both participants know they can rely on context to fix the intended content, the briefer the actual utterance has to be. Under such conditions, local conventions can arise, some extremely parochial and context-bound. For instance, in the Cafe Espresso Royale in Urbana, Illinois in the summer of 1999, regular customers could be heard to utter, "I'll have a day"—meaning, to those behind the counter, "a coffee of the day." This is incomprehensible out of context, but in its narrow context it served its purpose.

These conflicting pressures of brevity and clarity are present all the time, and are generally considered to be one of the important forces motivating historical change in language. As I understand the functionalist school of grammar (e.g. Bates and MacWhinney 1982, Givón 1995), much of linguistic structure is claimed to be a consequence of the constant problem of balancing these two pressures against one another. Still, this does not eliminate the need for analyses like Fig. 1.1, as is sometimes asserted. In a particular utterance, speaker and hearer both have plenty of structure that has to be accounted for, whatever its historical source.

The inhomogeneity of linguistic populations and the possibility of innovation lead to a fascinating question: how does a particular linguistic usage (word, construction, or grammatical feature) come to be relatively stabilized in a large dispersed population? We have some evidence for how innovations spread through a population over time (e.g. Labov 1994) for phonetic changes, Maling and Sigurjónsdóttir (1997) for a syntactic change, but this doesn't really answer why innovations spread. There are doubtless more basic principles of social dynamics that apply not only to language but to cultural innovations as well. Kirby (1998) and Steels (1998) can be understood as exploring the dynamics whereby communicative devices achieve a stable form within a community.

Finally, I should not leave this extremely brief survey of social factors without acknowledging all the issues of social identity (e.g. gender, class, and ethnicity) tied up with language use. These concerns about language certainly arouse greater passions than does the proper formulation of subject–auxiliary inversion. However, they are relatively remote from my concerns here—though not because of lack of intrinsic importance. And to be sure, research on the cognitive structure associated with language has been brought to bear on these issues, for example in the social legitimation of African-American varieties of English (e.g. Labov 1972; Rickford 1999) and the signed languages (Stokoe 1960; Klima and Bellugi 1979). The connections are there for those who wish to pursue them.