

9.3 Constraint-Based Lexicalism

We turn now to some reflections on the relationship between the sort of grammatical descriptions in this text and what is known about the mental processes underlying human language comprehension and production. Adopting the familiar terminology of Chomsky (1965), we distinguish between speakers' knowledge of their language – what Chomsky called their 'competence' – and the ways in which that knowledge is put to use in speaking and understanding – what Chomsky called 'performance'.

The way we speak and understand is clearly influenced by many things other than our linguistic knowledge. For example, we all make speech errors on occasion, reversing words or garbling our utterances in other ways; and we also sometimes misunderstand what was said. These sorts of errors are more likely to occur under certain conditions (such as a drunk speaker or a noisy environment) that have nothing to do with the interlocutors' knowledge of the language.

There are also subtler aspects of the competence/performance distinction. For example, memory limitations prevent anyone from being able to produce or understand a sentence a million words long. But we do not say that all such examples are ungrammatical, because the memory limitations that make such sentences unusable are not intrinsic to our knowledge of language. (If a speaker were to come along who could produce and understand million-word sentences of English, we would not say that that person spoke a different language from our own). Many other aspects of language use, including what people find easy and hard to understand, are generally included under the rubric of performance.

Psycholinguists are concerned with developing models of people's actual use of language. They try to figure out what sequences of (largely unconscious) steps people go through in producing and understanding utterances. They are, therefore, concerned with the types of errors people make, with what people find easy and difficult, and with how nonlinguistic factors influence language use. In short, psycholinguists study performance.

Chomsky (1965:15) wrote: 'In general, it seems that the study of performance models incorporating generative grammars may be a fruitful study; furthermore, it is difficult to imagine any other basis on which a theory of performance might develop.' We agree wholeheartedly with the idea of incorporating competence grammars into models of performance. However, at the time Chomsky wrote this, only one theory of generative grammar had been given serious consideration for modeling natural language. Since that time, a wide range of alternatives have been explored. One obvious basis for comparing these alternatives is to see how well they comport with what is known about performance. That is, theories of linguistic competence should be able to serve as a basis for testable models of linguistic performance.

We believe not only that grammatical theorists should be interested in performance modeling, but also that empirical facts about various aspects of performance can and should inform the theory of linguistic competence. That is, compatibility with performance models should bear on the design of grammars. As we will show later in this chapter, there is now a considerable body of psycholinguistic results that suggest properties that a competence theory should have, if it is to be embedded within an account of human linguistic performance. And we will argue that the theory we have been developing

does well on this criterion.⁷

Let us start with three basic observations about the grammar we have been developing:

1. It is **SURFACE ORIENTED**. Our grammar (like standard context-free grammars) provides a reasonably simple structure that is directly associated with the string of words that constitute each sentence. The ancillary structure that has to be computed to ascertain whether a given sentence is grammatical expresses information that is straightforwardly derivable from properties of the words in the string. No additional abstract structures are posited. In particular, our theory has no need for the sequences of phrase structures that constitute the derivations of sentences in transformational grammar.
2. It is **CONSTRAINT-BASED**. There are no operations that destructively modify any representations. The principles of the theory, the grammar rules, and the lexical entries are all just constraints that interact so as to define a set of phrase structures – those that simultaneously satisfy the relevant constraints of our grammar. Once generated, phrase structures are not rearranged, trimmed, or otherwise modified via transformational rules.
3. It is **STRONGLY LEXICALIST**. We have localized most grammatical and semantic information within lexical entries. These lexical entries furthermore correspond directly to the words present in the sentence, which can be viewed as the key elements that drive the construction of the syntactic and semantic structure of the sentence. As will become evident in the next few chapters, many of the relationships that transformational grammarians have analyzed using rules relating sentence types are handled in our theory via lexical rules.

Any theory that has these three design properties exemplifies a viewpoint that we will refer to as **CONSTRAINT-BASED LEXICALISM (CBL)**.

9.4 Modeling Performance

Available evidence on how people produce and comprehend utterances provides some general guidelines as to the nature of an adequate performance model. Some of that evidence is readily available to anyone who pays attention to language use. Other evidence has come out of controlled laboratory experiments, in some cases requiring sophisticated methods and equipment. The two most striking facts about language processing are the following:

- Language processing is incremental: Utterances are sequences of sounds. At any point in the production or comprehension of an utterance, language users are working on what has just been said and what is about to be said. Speakers do not wait until they have their utterances fully planned to begin speaking; and listeners do not wait until the end of an utterance to begin trying to figure out what the speaker means to say.
- Language processing is rapid: producing and understanding three words per second is no problem.

⁷Jackendoff (2002:Chapter 7) makes a similar argument. He takes a different stand on the question of modularity, discussed in Section 9.4.3, but on the whole his conclusions and ours are quite similar.

9.4.1 Incremental Processing

We don't have to venture into a psycholinguistic laboratory to convince ourselves that language processing is highly incremental. We saw this already in Chapter 1, when we considered examples like (47):

- (47) 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 we hear such a sentence, we process it as it comes – more or less word by word – building structure and partial interpretation incrementally, using what nonlinguistic information we can to make the right decisions at certain points. For example, when we encounter the PP *on the atom*, we have to decide whether it modifies VP or NOM; this is a kind of ambiguity resolution, i.e. deciding which of two currently available analyses is the one intended. We make this decision 'on-line' it seems, using a plausibility assessment of the meaning that would result from each structure. Information that can resolve such a local parsing ambiguity may appear later in the sentence. If the processor makes a decision about how to resolve a local ambiguity, but information later in the sentence shows that the decision was the wrong one, we would expect processing to be disrupted.

And indeed, psycholinguists have shown us that sentence processing sometimes does go astray. GARDEN PATH examples like (48a,b) are as remarkable today as they were when they were first brought to the attention of language researchers.⁸

- (48) a. The horse raced past the barn fell.
b. The boat floated down the river sank.

On first encountering such examples, almost all English speakers judge them to be totally ungrammatical. However, after seeing them juxtaposed to fully well-formed examples like (49), speakers recognize that examples like (48) are grammatical sentences, though very hard to process.

- (49) a. The horse that was raced past the barn fell.
b. The horse taken to the hospital died.
c. The boat that was floated down the river sank.
d. The boat seen down the river sank.

Experimental researchers thought at first that these garden paths showed that certain purely linguistic processing strategies (like trying to build an S out of the NP *the horse* and a VP beginning with *raced past*) were automatic - virtually impossible to turn off. But modern psycholinguistics has a very different story to tell.

First, note that in the right context, one can eliminate the garden path effect even with the sentences in (48). The right context can even make the NOM-modifying interpretation of *raced past the barn* the most natural one:⁹

- (50) The horse that they raced around the track held up fine. The horse that was raced down the road faltered a bit. And the horse raced past the barn fell.

⁸By Bever (1970).

⁹This kind of effect is discussed by Crain and Steedman (1985).

The context here highlights the need to identify one horse among many, which in turn favors the meaning of the NOM-modifying structure of (48a).

Moreover, if we keep the same potential for ambiguity, but change the words, we can eliminate the garden path effect even without an elaborate preceding context. Consider examples like (51a,b).

- (51) a. The evidence assembled by the prosecution convinced the jury.
 b. The thief seized by the police turned out to be our cousin.

As shown in a number of studies,¹⁰ examples like these present no more processing difficulty than their unambiguous counterparts in (52):

- (52) a. The evidence that was assembled by the prosecution convinced the jury.
 b. The thief who was seized by the police turned out to be our cousin.

That is, the examples in (51), even in the absence of a prior biasing context, do not cause garden path effects.

The explanation for this difference lies in the relevant nonlinguistic information. Evidence (or, say, a particular piece of evidence) can't assemble itself (or anything else), and the sentence built out of a subject NP *the evidence* and a VP headed by *assembled* would require some such implausible interpretation. (Similarly, intransitive uses of *seize* normally take some sort of mechanical device as their subject, making a thief an unlikely subject for *seized* in (51b)). That is, it is a fact about the world that only animate things (like people, animals, and perhaps some kinds of machines or organizations) assemble, and since evidence is inanimate, that hypothesis about the interpretation of the sentence is implausible. The fact that the decision to reject that interpretation (and hence the associated sentential structure) is made so quickly as to be imperceptible (i.e. so as to produce no noticeable garden path effect) is evidence that language comprehension is working in a highly integrative and incremental fashion. Linguistic and nonlinguistic constraints on the interpretation are interleaved in real time.

9.4.2 Rapid Processing

Just how rapidly people integrate available information in processing language has become evident since the early 1990s, thanks largely to technological advances that have made possible sophisticated new methods for investigating language use.¹¹ Of particular interest in the present context are head-mounted eye trackers, whose application to psycholinguistic research was pioneered by Michael Tanenhaus of the University of Rochester. These devices show investigators exactly where a participant's gaze is directed at any given moment. By following listeners' eye movements during speech, it is possible to draw inferences about their mental processes on a syllable-by-syllable basis.

The evidence from a great many experiments using this technique can be summed up concisely as follows: listeners use whatever information is available to them, as soon as it becomes available to them, to infer the speaker's intentions. In other words, language processing rapidly draws on all available types of linguistic and non-linguistic information as such information is needed.

¹⁰See, for example, Trueswell et al. 1992, Pearlmuter and MacDonald 1992, and Tabossi et al. 1994.

¹¹However, earlier work had made similar points. See, for example, Marslen-Wilson and Tyler 1987.

In one study, for example, participants viewed a grid with several objects on it, e.g. a box, a wallet, a fork, etc. Two of the objects would normally be described with words whose initial portions sound the same, for example, a *candle* and a *candy*; such pairs are called ‘competitors’. Participants received instructions to pick up an object and to place it somewhere else on the grid. For example, they might be told, ‘Pick up the candle. Now put it above the fork’. In some cases, the object they were told to pick up had a competitor on the grid (e.g. in the example just given, a candy might be present). Comparing cases in which a competitor was present to cases without a competitor provided evidence regarding the processes of word recognition and comprehension. Participants eye movements to the objects they picked up were significantly faster in cases when no competitor was present (445 milliseconds vs. 530 milliseconds). Tanenhaus et al. (1996:466) concluded that the timing of eye movements ‘provides clear evidence that retrieval of lexical information begins before the end of a word.’

Another study (also described by Tanenhaus et al. (1996)) involved sets of blocks that could differ in marking, color, and shape, so that uniquely identifying one with a verbal description would require a multi-word phrase. The stimuli were manipulated so that the target objects could be uniquely identified early, midway, or late in the production of the description. Listeners’ gaze again moved to the target object as soon as the information necessary for unique identification was uttered. What this information was depended not only on the words used, but also on what was in the visual display.

When one word in a description is contrastively accented (e.g. *the LARGE blue triangle*), the conditions for unique identification are different, since there must be another object present satisfying all but the contrasting word in the description (e.g. a small blue triangle). In some cases, this allows earlier resolution of the reference of a phrase. Eye-tracking shows that listeners use such accentual information in determining reference (Tanenhaus et al. 1996).

Similar results have been obtained under many different conditions. For example, eye movements show that resolution of prepositional phrase attachment ambiguities (*Put the apple on the towel in the box*) takes place as soon as listeners have the information needed for disambiguation, and this likewise depends on both linguistic factors and the visual display (see Tanenhaus et al. 1995).

Recent eye-tracking studies (Arnold et al. 2002) show that even disfluencies in speech are used by listeners to help them interpret speakers’ intentions. In particular, when a disfluency such as *um* or *uh* occurs early in a description, listeners tend to look at objects that have not yet been mentioned in the discourse. This makes sense, since descriptions of new referents are likely to be more complex, and hence to contain more disfluencies, than descriptions of objects previously referred to. Once again, the eye movements show the listeners using the information as soon as it becomes available in identifying (or, in this case, predicting the identification of) the objects that speakers are referring to.

It is easy to come up with many more examples showing that language comprehension proceeds rapidly and incrementally, with different types of information utilized as they are needed and available. The same is true of language production. One type of evidence for this again comes from disfluencies (see, for example, Clark and Wasow 1998 and Clark and Fox Tree 2002). The high rate of disfluencies in spontaneous speech shows that peo-

ple start their utterances before they have finished planning exactly what they are going to say and how they want to say it. And different types of disfluencies are symptoms of different kinds of production problems. For example, speakers tend to pause longer when they say *um* than when they say *uh*, suggesting that *um* marks more serious production problems. Correspondingly, *um* tends to occur more frequently at the beginnings of utterances,¹² when more planning is required, and its frequency relative to *uh* decreases later in utterances. The locations and frequencies of various types of disfluencies show that people are sensitive to a wide variety of linguistic and nonlinguistic factors in language production, just as they are in comprehension.

9.4.3 The Question of Modularity

The processing evidence cited so far also brings out the fact that people use all kinds of information – including nonlinguistic information – in processing language. Although this may strike some readers as unsurprising, it has been a highly controversial issue. Chomsky has long argued that the human language faculty is made up of numerous largely autonomous modules (see, for example, Chomsky 1981:135). Jerry Fodor’s influential 1983 book *The Modularity of Mind* elaborated on this idea, arguing that the human mind comprised a number of distinct modules that are ‘informationally encapsulated’, in the sense that they have access only to one another’s outputs, not to their internal workings.

The appeal of the modularity hypothesis stems primarily from two sources. The first is the analogy with physical organs: since various bodily functions are carried out by specialized organs (liver, kidney, pancreas, etc.), it seems plausible to posit similarly specialized mental organs to carry out distinct cognitive functions (vision, reasoning, language processing, etc.).¹³ Second, it is generally good practice to break complex problems down into simpler, more tractable parts. This is common in building computer systems, and computational metaphors have been very influential in recent theorizing about the human mind. It was natural, therefore, to postulate that the mind has parts, each of which performs some specialized function. Fodor’s version of the modularity hypothesis is not only that these mental organs exist, but that they function largely independently of each other.

According to this view, there should be severe limitations on how people combine information of different types in cognitive activities. Many psycholinguists would claim that the field has simply failed to detect such limitations, even when they use methods that can provide very precise information about timing (like the head-mounted eye tracker). These researchers would argue that linguistic processing appears to be opportunistic from start to finish, drawing on any kind of linguistic or nonlinguistic information that might be helpful in figuring out what is being communicated. Others working within the field would counter that the modularity hypothesis is not refuted by the existence of rapid information integration in sentence comprehension. Modularity can be reconciled with these results, it is argued, by assuming that informationally encapsulated language modules

¹²More precisely, at the beginnings of intonation units.

¹³The advocates of modularity are not entirely clear about whether they consider the language faculty a single mental organ or a collection of them. This is analogous to the vagueness of the notion of a physical organ: is the alimentary canal a single organ or a collection of them?

work at a finer grain than previously believed, producing partial results of a particular kind without consulting other modules. The outputs of these processors could then be integrated with other kinds of information relevant to comprehension quite rapidly. The controversy continues, hampered perhaps by a lack of general agreement about what counts as a module and what the space of hypotheses looks like in between Fodor's original strong formulation of the modularity hypothesis and the complete denial of it embodied in, for example, connectionist networks.

9.5 A Performance-Plausible Competence Grammar

Describing one of their eye-tracking experiments, Tanenhaus et al. write:

[T]he instruction was interpreted incrementally, taking into account the set of relevant referents present in the visual work space....That information from another modality influences the early moments of language processing is consistent with constraint-based models of language processing, but problematic for models holding that initial linguistic processing is encapsulated. (1996:466)

More generally, language understanding appears to be a process of constraint satisfaction. Competing interpretations exist in parallel, but are active to varying degrees. A particular alternative interpretation is active to the extent that evidence is available to support it as the correct interpretation of the utterance being processed. Note, by the way, that frequency can also play a significant role here. One reason the *horse raced past the barn* example is such a strong garden path is that *raced* occurs much more frequently as a finite verb form than as the passive participle of the transitive use of *race*, which is precisely what the NOM-modifying reading requires. Ambiguity resolution is a continuous process, where inherent degrees of activation (e.g. those correlating with gross frequency) fluctuate as further evidence for particular interpretations become available. Such evidence may in principle stem from any aspect of the sentence input or the local or discourse context. A garden-path sentence is one that has an interpretation strongly supported by initial evidence that later turns out to be incorrect.

The next three subsections argue that the three defining properties of Constraint-Based Lexicalism, introduced in Section 9.3, receive support from available evidence about how people process language.

9.5.1 Surface-Orientation

Our grammar associates structures directly with the string of words that the listener hears, in the form (and order) that the listener hears them. This design feature of our grammar is crucial in accounting for the word-by-word (or even syllable-by-syllable) fashion in which sentence processing proceeds. We have seen that in utterances, hearers use their knowledge of language to build partial hypotheses about the intended meaning. These hypotheses become more or less active, depending on how plausible they are, that is, depending on how well their meaning squares with the hearers' understanding of what's going on in the discourse.

Sometimes the process even takes short-cuts. We have all had the experience of completing someone else's utterance (a phenomenon that is, incidentally, far more common than one might imagine, as shown, e.g. by Wilkes-Gibbs (1986)) or of having to wait

for someone to finish an utterance whose completion had already been made obvious by context. One striking example of this is ‘echo questions’, as illustrated in the following kind of dialogue:

- (53) [Speaker A:] Señora Maria Consuelo Bustamante y Bacigalupo is coming
to dinner tomorrow night.
[Speaker B:] WHO did you say is coming to dinner tomorrow night?

In a dialogue like this, it is quite likely that Speaker A may comprehend the intent of Speaker B’s utterance well before it is complete, somewhere in the region indicated by the asterisks. Presumably, this is possible precisely because Speaker A can recognize that the remainder of B’s utterance is a repetition of A’s own utterance and can graft that bit of content onto the partial analysis A has performed through word-by-word processing of B’s utterance. What examples like this show is that a partial linguistic analysis (e.g. the partial linguistic analysis of *who did you*, *who did you say* or *who did you say is*) is constructed incrementally, assigned a (partial) interpretation, and integrated with information from the context to produce an interpretation of a complete utterance even before the utterance is complete. Amazing, if you think about it!

So if a grammar is to be realistic, that is, if it is to be directly embedded in a model of this kind of incremental and integrative language processing, then it needs to characterize linguistic knowledge in a way that allows for the efficient incremental computation of partial analyses. Moreover, the partial grammatical analyses have to be keyed in to partial linguistic meanings, because these are what interacts with other factors in processing.

The kind of grammar we are developing seems quite compatible with these performance-driven design criteria. The representation our grammar associates with each word provides information about the structure of the sentence directly, that is, about the phrases that the words are part of and about the neighboring phrases that they combine with syntactically. In addition, the words of our grammar provide partial information about the meaning of those phrases, and hence, since all phrases are built up directly from the component words and phrases in a context-free manner, there is useful partial semantic information that can be constructed incrementally, using our surface-oriented grammar.

It is not clear how to reconcile the incremental processing of utterances with transformational grammar, in which the surface ordering of elements depends on a sequence of structures and operations on them. If only the surface structures are involved in the processing model, then the transformational derivations are evidently irrelevant to performance. On the other hand, a full derivation cannot be available incrementally, because it necessarily involves all elements in the sentence.

Of course we have not actually spelled out the details of a performance model based on a grammar like ours, but the context-free-like architecture of the theory and the hybrid syntactic-semantic nature of the lexical data structures are very suggestive. Incremental computation of partial semantic structures, the key to modeling integrative sentence processing, seems to fit in well with our grammar.

9.5.2 Constraint-Based Grammar

Our grammar consists of a set of constraints that apply simultaneously to define which structures are well-formed. When this abstract model of language is applied (in a computational system, or in a model of human language processing), this simultaneity is cashed out as order independence: it doesn't matter which order the constraints are consulted in, they will always give the same collective result.

As noted above, the order of presentation of the words in an utterance largely determines the order of the mental operations listeners perform in comprehending it. However, words are associated with many different kinds of information, and the architecture of the theory does not impose any fixed order on which kind is used first. For example, it is not the case that syntactic information (e.g. agreement information that might rule out a particular parse) is always consulted before semantic information (e.g. semantic incompatibility that would favor or disfavor some potential interpretation of an utterance). In fact, it is possible to make an even stronger claim. In examples like (54), early accessing of morphological information allows the number of sheep under discussion to be determined incrementally, and well before the nonlinguistic knowledge necessary to select the 'fenced enclosure' sense of *pen*, rather than its 'writing implement' sense.

(54) The sheep that was sleeping in the pen stood up.

In (55), on the other hand, the relevant information about the world – that sheep might fit inside a fenced enclosure, but not inside a writing implement – seems to be accessed well before the relevant morphological information constraining the number of sheep:¹⁴

(55) The sheep in the pen had been sleeping and were about to wake up.

So the information accessed in on-line language processing is typically made available in an order determined by the input stream, not by the constructs of grammatical theory. In comprehending these sentences, for example, a hearer accesses morphological information earlier in (54) and later in (55) precisely because the order of access is tied fairly directly to the order of the words being processed. A theory positing a fixed order of access – for example, one that said all strictly linguistic processing must be completed before nonlinguistic knowledge could be brought to bear on utterance interpretation – would not be able to account for the contrast between (54) and (55).

Such a theory would also be incompatible with the evidence from the head-mounted eye-tracking studies cited earlier. Those studies show that listeners use both linguistic and visual information to determine a speaker's intended meaning, and they use it as soon as the information is available and helpful to them. Hence, a theory of linguistic comprehension must allow the order of access to information to remain flexible.

Finally, we know that for the most part linguistic information functions fairly uniformly in many diverse kinds of processing activity, including comprehension, production, translation, playing language games, and the like. By 'fairly uniformly' we mean that the set of sentences reliably producible¹⁵ by a given speaker-hearer is similar – in fact bears a natural relation (presumably proper inclusion) – to the set of sentences that that speaker-hearer can comprehend. This might well have been otherwise. That there is so close and

¹⁴This pair of examples is due to Martin Kay.

¹⁵That is, sentences short enough to utter in a real language-use situation. We also intend to rule out production errors.

predictable a relation between the production activity and the comprehension activity of any given speaker of a natural language militates strongly against any theory on which the production grammar is independent from the comprehension grammar, for instance. This simple observation suggests rather that the differences between, say, comprehension and production should be explained by a theory that posits distinct processing regimes making use of a single language description. And that description should therefore be a process-neutral grammar of the language, which can serve each kind of process that plays a role in on-line linguistic activity.¹⁶ Since production involves going from a meaning to an utterance and comprehension involves going from an utterance to a meaning, a grammar that is used in both processes should not favor one order over the other.

Grammars whose constructs are truly process-neutral, then, hold the most promise for the development of processing models. Transformational grammars aren't process-neutral, because transformational derivations have a directionality – that is, an ordering of operations – built into them. To interpret a transformational grammar as a model of linguistic knowledge, then, it is necessary to abstract away from its inherent directionality, obscuring the relationship between the grammar and its role in processing. This problem can be avoided by formulating a grammar as a declarative system of constraints. Such systems of constraints fit well into models of processing precisely because they are process-neutral.

What these observations add up to is a view of grammar as a set of constraints, each expressing partial information about linguistic structures, rather than a system employing destructive operations of any kind. Moreover, we have also seen that these constraints should exhibit certain further properties, such as order-independence, if performance-compatibility is to be achieved. The grammar we've been developing has just these design properties – all the constructs of the grammar (lexical entries, grammar rules, even lexical rules and our general principles) are nothing more than constraints that produce equivalent results no matter what order they are applied in.

9.5.3 Strong Lexicalism

Our theory partitions grammatical information into a number of components whose interaction determines the well-formedness of particular examples. By far the richest locus of such information, however, is the lexicon. Our grammar rules are simple in their formulation and general in their application, as are such aspects of our formal theory as the Head Feature Principle and the Valence Principle. Most of the details we need in order to analyze individual sentences are codified in the lexical entries (though much of it need not be stipulated, thanks to lexical rules and inheritance through the type hierarchy).

However, other divisions of grammatical labor are conceivable. Indeed, a number of theories with highly articulated rule systems and relatively impoverished lexicons have been developed in considerable detail (e.g. early transformational grammar and Generalized Phrase Structure Grammar, both of which are described briefly in Appendix B).

¹⁶The fact that comprehension extends beyond systematic production can be explained in terms of differences of process – not differences of grammar. Speakers that stray far from the grammar of their language run a serious risk of not being understood; yet hearers that allow grammatical principles to relax when necessary will understand more than those that don't. There is thus a deep functional motivation for the two kinds of processing to differ as they appear to.

We have argued for strong lexicalism on the basis of linguistic adequacy (along with general considerations of elegance and parsimony). It turns out that the psycholinguistic evidence on language processing points in the same direction. Investigations of syntactic ambiguity resolution in general and garden path effects in particular have shown that the choice of words can make a big difference. That is, the difficulty listeners exhibit in resolving such ambiguities (including overcoming garden paths) is influenced by factors other than the structure of the tree. Processing is critically affected by semantic compatibility and pragmatic plausibility, type and valence of the words involved, and the frequencies with which individual words occur in particular constructions. Our earlier discussion of eye-tracking studies describes some of the evidence to this effect, and there is considerably more (see Tanenhaus and Trueswell 1995 for a survey of relevant results).

To give another kind of example, a sentence beginning with the sequence NP₁-V-NP₂ can be continued in a number of ways. NP₂ could be the object of the verb, or it could be the subject of a complement sentence. This is illustrated in (56a), which can be continued as in (56b) or (56c):

- (56) a. Lou forgot the umbrella . . .
 b. Lou forgot the umbrella was broken.
 c. Lou forgot the umbrella in the closet.

Hence a listener or reader encountering (56a) must either postpone the decision about whether to attach the NP *the umbrella* to the VP, or decide prematurely and then potentially have to reanalyze it later. Either way, this places a burden on the parser in at least some cases. Various experimental paradigms have been used to verify the existence of this parsing difficulty, including measuring reading times and tracking the eye movements of readers.

However, not all verbs that could appear in place of *forgot* in (56a) can appear in both of the contexts in (56b) and (56c). This is illustrated in (57):

- (57) a. Lou hoped the umbrella was broken.
 b.*Lou hoped the umbrella in the closet.
 c.*Lou put the umbrella was broken.
 d. Lou put the umbrella in the closet.

The increased parsing load in (56a) is reduced greatly when the valence of the verb allows for no ambiguity, as in (57). This has been demonstrated via the methods used to establish the complexity of the ambiguity in the first place (see Trueswell et al. 1993). This provides strong evidence that people use valence information associated with words incrementally as they process sentences.

Similarly, listeners use semantic and pragmatic information about the verb and the following NP to choose between possible attachment sites for the NP. For example, though *learn* may take either an NP object or a sentential complement, as illustrated in (58),

- (58) a. Dana learned the umbrella was broken.
 b. Dana learned a new theorem in class.

when the immediately following NP is not the sort of thing one can learn, people do not exhibit the level of complexity effects in parsing that show up in (56).

The same sort of effect of lexical meaning on parsing shows up with PP attachment ambiguities, like those in (59):

- (59) a. The artist drew the child with a pencil.
 b. Lynn likes the hat on the shelf.

In (59a), the pencil could be either the artist's instrument or something in the child's possession; in (59b), *on the shelf* could identify either Lynn's preferred location for the hat, or which hat it is that Lynn likes. The structural ambiguity of such sentences causes parsing complexity, but this is substantially mitigated when the semantics or pragmatics of the verb and/or noun strongly favors one interpretation, as in (60):

- (60) a. The artist drew the child with a bicycle.
 b. Lynn bought the hat on the shelf.

In short, lexical choices have a substantial influence on processing. Moreover, the information that we have been led to posit in our lexical entries has independently been found to play a role in language processing. After reviewing a number of studies on the factors that influence syntactic ambiguity resolution, MacDonald et al. (1994) discuss what information they believe needs to be lexically specified to account for the psycholinguistic results. Their list includes:

- valence;
- 'coarse-grained semantic information' (i.e. the sort of information about who did what to whom that is given in our SEM feature); and
- 'grammatically relevant features' such as 'tense...', 'finiteness...', 'voice (active or passive), number...', 'person...', and 'gender...'.

They also mention grammatical category, which we represent in our lexical entries by means of types (specifically, the subtypes of *pos*). In short, the elements in the MacDonald et al. list correspond remarkably well to the information that we list in our lexical entries.

9.5.4 Summary

In this section we have seen how the design features of our grammar are supported by evidence from language processing. A grammar must be SURFACE-ORIENTED to account for the incremental and integrative nature of human language processing. The fact that different kinds of linguistic information and even non-linguistic information are accessed in any order, as convenient for the processor, suggests a CONSTRAINT-BASED design of grammar. This is further motivated by the process-neutrality of knowledge of language. Finally, STRONG LEXICALISM and the particular kinds of information associated with words in our lexical entries are supported by psycholinguistic evidence from garden paths, eye-tracking experiments, and tests of parsing complexity.

9.6 Universal Grammar: A Mental Organ?

In the preceding section we have argued that the design features of our grammatical theory comport well with existing evidence about how people process language. There is yet another psycholinguistic consideration that has played a central role in much work in generative grammar, namely, learnability. In this section, we briefly address the question of evaluating our theory by this criterion.

As noted in Chapter 1, Chomsky has argued that the most remarkable fact about human language – and the one he thinks linguists should be primarily concerned with explaining – is that virtually all children become fluent speakers of a language, with little apparent effort or instruction. The puzzle, as Chomsky sees it, is how people can come to know so much about language so quickly and easily. His solution in a nutshell is that people’s knowledge of language is for the most part innate, not learned. This entails that much linguistic structure – namely, those aspects that are innate – must be common to all languages. Consequently, a central goal of much work in modern syntactic theory has been to develop a conception of universal grammar rich enough to permit the descriptions of particular languages to be as simple as possible.

Chomsky’s strong claims about the role of innate knowledge in language acquisition are by no means uncontroversial among developmental psycholinguists. In particular, many scholars disagree with his position that the human language faculty is highly task-specific – that is, that people are born with a ‘mental organ’ for language which is distinct in its organization and functioning from other cognitive abilities (see, for example, Bates and MacWhinney 1989, Tomasello 1992 and Elman et al. 1996 for arguments against Chomsky’s position; but see also Hauser et al. 2002).

There can be little doubt that biology is crucial to the human capacity for language; if it were not, family pets would acquire the same linguistic competence as the children they are raised with. There is no doubt that humans are quite special, biologically, though the details of just what is special remain to be worked out. It is far less clear, for example, that the human capacity for language is as independent of other systems of knowledge as has sometimes suggested. A range of views on this issue are possible. At one end of the spectrum is the idea that the language faculty is a fully autonomous module, unrelated to general cognitive capacity. At the other end is the idea that there are no specifically linguistic abilities – that our capacity to learn language arises essentially as a side-effect of our general intelligence or of other abilities. Chomsky’s view is close to the former;¹⁷ Tomasello (1992) argues for something close to the latter. Other scholars have defended views somewhere in between.

The participants in this debate often seem to be talking past one another. Opponents of task-specificity tend to take a simplistic view of linguistic structure, emphasizing basic communicative functions while ignoring the intricacies of syntax that are the bread and butter of generative grammar. On the other hand, proponents of task-specificity have a tendency to leap from the complexity of their analyses to the conclusion that the knowledge involved must be innate and unique to language.

We find much of the argumentation on both sides of this controversy unconvincing, and hence we take no position in this book. Nevertheless, the theory presented here can contribute to its resolution. Explicit syntactic and semantic analyses can facilitate more precise formulations of what is at issue in the debate over task-specificity. Moreover, formal representations of data structures and their interactions makes it possible to see more clearly where there could be analogues in other cognitive domains. Our position is that the grammatical constructs we have been developing in this text are well suited to a theory of universal grammar, whether or not that theory turns out to be highly task-specific, and that the explicitness of our proposals can be helpful in resolving the

¹⁷But see Hauser et al. 2002 for what seems to be a striking switch in Chomsky’s position.

task-specificity question.

To justify this claim, we will consider various components of our theory, namely: the phrase structure rules, the features and their values, the type hierarchy with its feature declarations and constraints, the definition of phrasal licensing (incorporating the Head Feature Principle, the Valence Principle, and the two semantic principles), the Binding Theory, and the lexical rules. We will find that most of these have elements that are very likely universal, and that our formulations do not prejudice the issue of task-specificity.

Phrase Structure Rules Our grammar rules (with the exception of the Imperative Rule) are sufficiently general that, aside from their linear ordering of the constituents, they are natural candidates for universality. It would not be hard to factor out the ordering, so that versions of these rules could be posited as part of universal grammar.

The sort of hierarchical structure induced by the rules, which we represent with trees, is arguably not unique to language: it also seems appropriate, for example, to aspects of mathematical reasoning. On the other hand, the concepts of ‘head’, ‘complement’, ‘specifier’, and ‘modifier’, which are crucial to our formulation of the rules, appear to be specialized to language. If it should turn out, however, that they can be shown to be instances of some more generally applicable cognitive relations, this would in no way undermine our analysis.

Features and Values Most of the features we have posited have obvious cross-linguistic application. It seems at least plausible that a more fully worked out version of the theory presented here could include an inventory of features from which the feature structures of all languages must be constructed. In later chapters, we will identify the values of some features with particular English words, a practice inconsistent with saying that the set of possible feature values is part of universal grammar. It might be possible, however, to restrict feature values to come from either the set of morphological forms of the language or a universally specifiable set.

Some features (e.g. PER, GEND, COUNT) clearly reflect properties of the world or of human thought, whereas others (e.g. CASE, FORM) seem specifically linguistic. Our treatment is neutral on the question of whether grammatical features will ultimately be reducible to more general aspects of cognition, though the general data type of features with values certainly has applications beyond linguistics.

Types and the Type Hierarchy The types we have proposed could arguably be drawn as well from a fixed universal inventory. The feature declarations associated with the types are likewise probably quite similar across languages. The constraints introduced by some types (such as SHAC), on the other hand, appear to be more specific to the particular language. Some of the (subtype and supertype) relations in the type hierarchy (e.g. that *siv-lxm* is a subtype of *verb-lxm*) are surely universal, whereas others (e.g. the hierarchy of subtypes of *agr-cat*) may vary across languages.

Our types are arranged in a default inheritance hierarchy, a kind of structure that very likely plays an important role in how people organize many kinds of information. Indeed, the use of such hierarchies in linguistics was inspired by earlier work in artificial intelligence, which suggested this sort of structure for taxonomies of concepts. The particular types we have posited appear task-specifically linguistic, though we leave open the possibility that some of them may be more general.

Phrasal Licensing Our definition of phrasal licensing involves both universal and English-specific elements. As noted earlier, the Argument Realization Principle may well differ across languages. And clearly, the Case Constraint as we have formulated it applies only to English. On the other hand, the Head Feature Principle and the two semantic principles are intended to apply to all languages.

Some parts of the phrasal licensing definition make reference to specifically linguistic constructs (such as grammar rules, heads, and particular features), but the idea of unifying information from diverse sources into a single structure has nonlinguistic applications as well.

Binding Theory All languages evidently have some binding principles, and they are quite similar. Characteristically, there is one type of element that must be bound within a local domain and another type that cannot be locally bound. But there is cross-language variation in just what counts as ‘local’ and in what can serve as the antecedents for particular elements. Our particular Binding Theory is thus not part of universal grammar. Ideally, a grammatical theory would delineate the range of possible binding principles, of which the ones presented in Chapter 7 would be instances.

While these principles appear to be quite language-specific, it is conceivable that they might be explained in terms of more general cognitive principles governing identity of reference.

Lexical Rules The lexical rules presented in the previous chapter are clearly parochial to English. However, our characterizations of derivational, inflectional, and post-inflectional lexical rules seem like plausible candidates for universality. More generally, our formulation of lexical rules as feature structures lays the groundwork for developing a more articulated inheritance hierarchy of types of lexical rules. Although formulating a general theory of what kinds of lexical rules are possible is beyond the scope of this book, our grammatical framework has a way of expressing generalizations about lexical rules that are not language-particular.

The contents of these rules are quite specific to language, but their general form is one that one might expect to find in many domains: if a database contains an object of form X, then it also contains one of form Y.

To sum up this superficial survey of the components of our theory: it contains many elements (the grammar rules, the definition of Well-Formed Tree Structure, the features and types) that are plausible candidates for playing a role in a theory of universal grammar. Moreover, some elements (the binding principles, some lexical rules) probably have close analogues in many other languages. Although our central purpose in this book is to present a precise framework for the development of descriptively adequate grammars for human languages, rather than to account for the puzzle of language learnability through the development of a theory of universal grammar, the framework we have presented here is nevertheless quite compatible with the latter goal.

Further, our grammatical theory suggests a number of parallels between the kinds of information structures needed to account for linguistic competence and those employed in other cognitive domains. However, we need not commit ourselves on the question of task-specificity; rather, we offer the hope that increasingly precise linguistic descriptions

like those that are possible within the framework developed here will help to clarify the nature of this controversy and its resolution.

9.7 Summary

Chomsky's famous distinction between knowledge of language ('competence') and use of language ('performance') has allowed syntacticians to concentrate on relatively tractable problems, by abstracting away from many features of the way people actually speak. But most generative grammarians agree that an optimal theory of competence will play a role in explaining many features of linguistic performance. To the extent that a theory of grammar attains this ideal, we call it 'realistic'.

We have argued in this chapter that the theory we are developing in this book does well by this criterion. Our theory, by virtue of being surface-oriented, constraint-based, and strongly lexicalist, has properties that fit well with what we know about how people process utterances and extract meaning from them. Our understanding of the mechanisms that underlie linguistic performance is incomplete at present, and many of the points discussed in this chapter remain controversial. Nevertheless, a preliminary examination of what is known about processing provides grounds for optimism about our approach to syntactic theory. Considerations of learnability also support such a favorable assessment.

9.8 Further Reading

Many of the issues raised in this chapter are discussed at a relatively elementary level in the essays in Gleitman and Liberman 1995. Important discussions of issues raised in this chapter can be found in the following works: Chomsky 1965, Bever 1970, Bates and MacWhinney 1989, Tomasello 1992, MacDonald et al. 1994, Pinker 1994, Tanenhaus and Trueswell 1995, Elman et al. 1996, Marcus 2001, Jackendoff 2002, Hauser et al. 2002, and Marcus 2004.

9.9 Problems

Problem 1: Inflectional Lexical Rules With No Morphological Effect

The Singular Noun Lexical Rule, the Non-3rd-Singular Verb Lexical Rule, and the Base Form Lexical Rule are all inflectional lexical rules (that is, rules of type *i-rule*) which have no effect on the shape (i.e. the phonology) of the word.

- A. Explain why we need these rules anyway.
 - B. Each of these rules have lexical exceptions, in the sense that there are lexemes that idiosyncratically don't undergo them. Thus, there are some nouns without singular forms, verbs without non-third-person singular present tense forms, and verbs without base forms. List any you can think of. [*Hint: The nouns without singular forms are ones that must always be plural; these aren't too hard to think of. The exceptional verbs are much harder to come up with; we only know of two (fairly obscure) exceptions to the Non-3rd-Singular Verb Lexical Rule and a small (though frequently used) class of exceptions to the Base Form Lexical Rule. In short, parts of this problem are hard.*]
-