

Sentence comprehension

Martin J. Pickering

Human Communication Research Centre, University of Glasgow, UK

Language is clearly extremely complex at all kinds of different levels. Hence, it is very striking that language comprehension is in general so efficient. People can read, listen to a speaker or hold a conversation, and, except on rare occasions, understand most of what the writer or speaker intends to convey. One aspect of language comprehension is what I call sentence comprehension: determining the meaning of a sentence as a whole on the basis of a sequence of words. If we can understand the processes and mechanisms that underlie sentence comprehension, we will have taken one step forward in understanding how people can master the use of language.

Let us first try to provide a fairly precise delimitation of the area that the term *sentence comprehension* refers to. I assume that words have been recognised, so that the processor has access to their lexical properties. Sentence comprehension is concerned with how people obtain a particular syntactic analysis for a string of words and assign an interpretation to that analysis. Thus, it is not principally concerned with word recognition, morphological processing, anaphoric resolution, figurative language, discourse coherence, and inferencing in general (see other chapters). Very roughly, it concentrates on those aspects of language comprehension that draw upon the rules and representations that are studied within generative grammar. However, it is important to stress that the goal of this process is to obtain an interpretation for a string of words, not simply to obtain a

syntactic analysis. Hence, I call this chapter *sentence comprehension* rather than simply *parsing*, which is sometimes employed in the narrow sense of syntactic analysis.

Probably the most striking and uncontroversial finding in this area is that sentence comprehension is highly incremental. In both spoken and written language, words are encountered sequentially. Experimental evidence indicates that a great deal of processing occurs immediately, before the next word is encountered (e.g. Just & Carpenter, 1980; Tyler & Marslen-Wilson, 1977). Thus, word recognition is not normally delayed (e.g. Marslen-Wilson, 1987; Rayner & Duffy, 1986), even if disambiguation occurs after the ambiguous word (e.g. Rayner & Frazier, 1989). This is a necessary precondition for incremental sentence comprehension. More importantly, there is normally no measurable delay before syntactic analysis and some aspects of semantic interpretation begin (though it is impossible to be sure that there are no circumstances under which delay occurs).

Evidence for incremental syntactic analysis comes from the vast literature showing "garden-path" effects. For example, the sentence *The horse raced past the barn fell* (Bever, 1970) is hard (in part, at least) because people assume that *raced* is an active past-tense verb, and hence that *the horse raced past the barn* is a complete sentence. When they encounter *fell*, they realise this is impossible, and reinterpret *raced* as a past participle in a "reduced relative" construction (cf. *The horse that was raced past the barn fell.*), or fail to understand the sentence entirely. In other words, they are "led up the garden path" by such a sentence. Hence, they have performed incremental syntactic analysis before reaching *fell*. Experimental evidence strongly supports this conclusion for many different sentence types, and suggests that syntactic analysis begins very rapidly (e.g. Altmann & Steedman, 1988; Frazier & Rayner, 1982; Rayner, Carlson, & Frazier, 1983; Trueswell, Tanenhaus, & Garnsey, 1994).

Evidence also supports the intuition that people start to understand sentences as they hear or read them. Most famously, Marslen-Wilson (1973, 1975) showed that participants' errors in shadowing (i.e. immediately repeating) a text at a lag of only 300 ms were constrained by semantic context. This demonstrates that the meaning of what is heard can be rapidly integrated with general knowledge, though it is conceivable that integration occurs during production rather than comprehension. Data from eye-tracking gives more direct evidence for incremental interpretation. For instance, Traxler and Pickering (1996b) found that readers were disrupted as soon as they read the word *shot* in (1):

- (1) That is the very small pistol in which the heartless killer shot the hapless man yesterday afternoon.

Hence, they must have semantically processed the sentence fragment up to *shot* when they first encounter the word. Many other experiments also provide good evidence for incremental semantic processing using various methods (e.g. Boland, Tanenhaus, Garnsey, & Carlson, 1995; Clifton, 1993; Garrod, Freudenthal, & Boyle, 1994; Holmes, Stowe, & Cupples, 1989; Swinney, 1979; Trueswell et al., 1994; Tyler & Marslen-Wilson, 1977). We can conclude that the language processing system must very rapidly construct a syntactic analysis for a sentence fragment, assign it a semantic interpretation, and make at least some attempt to relate this interpretation to general knowledge.

These experiments suggest that any delays in either syntactic analysis or associated aspects of semantic interpretation must be extremely subtle. Hence, models that assume a major delay component (e.g. Marcus, 1980) cannot be accurate. In contrast, there may sometimes be delays in other aspects of interpretation, such as anaphoric resolution (Greene, McKoon, & Ratcliff, 1992; though cf. Garrod et al., 1994; Marslen-Wilson, Tyler, & Koster, 1994) or clausal integration (e.g. Millis & Just, 1994; though cf. Traxler, Bybee, & Pickering, 1997).

A very important consequence of incrementality is that the processor often makes decisions about syntactic and semantic analysis when a sentence fragment is syntactically ambiguous. Most experimental research in sentence comprehension in the 1980s and 1990s has focused on such "local" ambiguities. Through this research, psycholinguists have attempted to understand the organisation of the language processor.

The rest of this chapter discusses both experimental research on syntactic ambiguity resolution and theoretical models of sentence comprehension. I first ask whether the processor can consider only one analysis at a time, or whether it can consider different analyses at the same time. I then outline different sources of information that are relevant to parsing, and argue that the core question in parsing research is how the processor manages to integrate them. Current accounts can broadly be split into two types: restricted accounts, in which the processor can draw upon some sources of information during initial processing but not others; and unrestricted accounts, in which the processor can draw upon all relevant sources of information without delay. I interpret different restricted and unrestricted accounts in light of a range of empirical data. I then discuss a special topic, the processing of unbounded dependencies, within this general framework, and draw some conclusions.

Given the large amount of recent work on sentence comprehension, this review is necessarily selective and incomplete. A fuller review would not focus so overwhelmingly on initial stages of analysis. There is now considerable interest in the question of how the parser performs reanalysis if the initial analysis turns out to be wrong (Ferreira & Henderson, 1991;

Gorrell, 1995; Inoue & Fodor, 1995; Pickering & Traxler, 1998; Pritchett, 1992; Rayner et al., 1983; Sturt & Crocker, 1996; Sturt, Pickering, & Crocker, 1999). I have also entirely ignored formal aspects of semantic processing, such as the resolution of quantifier-scope ambiguities, as such topics have largely been neglected (though see Kurtzman & MacDonald, 1993). The review is probably biased towards research in reading rather than listening, in part because evidence concerning the role of prosody in ambiguity resolution is reviewed by Warren (Chapter 6). Finally, I have ignored individual differences in sentence comprehension (e.g. Just & Carpenter, 1992; MacDonald, Just, & Carpenter, 1992; Pearlmuter & MacDonald, 1996; Waters & Caplan, 1996).

PARALLEL AND SERIAL MODELS OF PROCESSING

When a fragment of a sentence is compatible with only one syntactic analysis, the evidence for incremental processing suggests that the analysis is computed and interpreted. But what happens when a fragment is compatible with more than one analysis? Does the processor compute all analyses in parallel? If so, does it retain all of these analyses or does it drop some? Does it foreground some analyses and background others? Or, alternatively, does it only compute one analysis initially, but have the capacity to reanalyse? These questions are fundamental to determining the strategy that the processor uses in resolving ambiguity. They also help us address the even more fundamental question of the basic architecture of the language processor. Unfortunately, this question has not been resolved, and it is very hard to imagine that a particular series of experiments will provide conclusive evidence on this question.

In a serial model, one analysis is selected. In Bever's sentence, the processor normally adopts the main clause analysis for *The horse raced past the barn*. If this analysis becomes impossible (e.g. following *fell*), then the processor must abandon this analysis and start again. Serial accounts are therefore broadly compatible with data demonstrating the existence of garden-path effects.

A parallel account can consider multiple analyses at the same time. Consider first what we can call *pure unrestricted parallelism*, whereby the processor initially constructs all possible syntactic analyses in parallel, and regards all analyses as being of equal importance (e.g. Forster, 1979). For instance, after *The horse raced* in Bever's sentence, the processor would represent both the main clause and the reduced relative analyses. After *The horse raced past the barn fell*, the processor would drop the main clause analysis, and would continue with the reduced relative analysis, without experiencing any difficulty. However, we know that this account cannot be correct, because the reduced relative analysis causes a garden-

path effect. It therefore could not have been as available as the main clause analysis.

In a *ranked-parallel* model, one analysis is foregrounded, and any others are backgrounded. In Bever's sentence, the main clause analysis is foregrounded, and the reduced relative analysis backgrounded. If the main clause analysis becomes impossible, then the parser must change its ranking of analyses. In this case, it will foreground the reduced relative analysis; the main clause analysis may be dropped entirely. Like serial accounts, ranked parallel accounts are broadly compatible with current evidence.

Parallel models differ in many other respects, depending on how many analyses are maintained, what kind of ranking is employed, how long the different analyses are considered for, or whether parallelism is only employed under certain conditions or with certain constructions. Currently, however, the most influential kind of parallel model is the constraint-based account (e.g. MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell et al., 1994; Trueswell, Tanenhaus, & Kello, 1993), discussed in detail later. According to this account, different analyses are weighted on the basis of how compatible they are with a range of constraints. For example, an analysis will be foregrounded if it is highly frequent, highly plausible, highly compatible with the prosody employed, and so on. As the sentence progresses, new information will cause analyses to change their weightings, and so a different analysis may be foregrounded.

Alternatives to this kind of continuous competition of alternative analyses have been proposed. For instance, Gibson (1991) proposed a "beam search" mechanism in which analyses which are close enough in complexity to the simplest analysis are retained. Analyses are then dropped if their complexity, measured in a way proposed by Gibson, exceeds the complexity of the simplest analysis by some threshold value (cf. Jurafsky, 1996).

These accounts assume that different analyses are retained for an extended period. Other accounts assume momentary parallelism. The "referential" or "incremental-interactive" account of Altmann and Steedman (1988; cf. Crain & Steedman, 1985) is of this latter kind. Here, alternative analyses are proposed in parallel, and contextual information chooses between them immediately, on the basis of how felicitous the analyses are with respect to discourse context (see later for discussion of the actual model). After an initial parallel stage, processing becomes serial. Momentary parallel accounts are similar in spirit to many models of lexical ambiguity resolution (e.g. Swinney, 1979), where all alternative meanings of a word are proposed, but all but the most contextually appropriate (or frequent) meaning is rapidly abandoned. Empirically, it

has proved extremely difficult to distinguish between serial and different kinds of ranked-parallel accounts. Though this issue is absolutely central to the development of processing models, it has in a sense remained behind the front line; most of the conflicts in this area have focused on the role of different information sources in parsing.

MODULARITY AND INFORMATION SOURCES

Traditionally, researchers have asked whether language comprehension is *modular* or not. This interest stems largely from J.A. Fodor's (1983) book, *The modularity of mind*. In it, he argued that certain mental faculties, basically consisting of the senses and aspects of language, were modules. Modules are specialised components of the mind, separate from general cognition or "central processes". Fodor defines a number of properties that he claims that all modular systems share; for instance, they are innate and employ a fixed neural architecture. Perhaps their most important property is that they are *encapsulated*: The internal workings of a module cannot be affected by anything external to the module. The output of a module is dependent on the input to the module and the internal structure of the module, and is unaffected by central processes or other modules. Fodor claimed that aspects of language comprehension constituted a module (see also Forster, 1979). Opponents of this position have argued that there are no modules, or that language comprehension in particular is not a module (e.g. Tyler & Marslen-Wilson, 1977; see Garfield, 1987).

Psycholinguists often still ask whether language comprehension is a modular process, but the current emphasis is rather different from Fodor's. In particular, the focus is entirely on encapsulation; few claims are made about, for instance, the developmental or neuroscientific status of the language processor. The main question is whether there is an encapsulated language processor, and, if so, precisely what aspects of language are contained within it.

It is worth distinguishing *representational modularity* from *processing modularity* (Trueswell et al., 1994). Representational modularity claims that sources of information like syntactic and semantic knowledge are represented separately. This assumption is standard to most generative grammars (e.g. Chomsky, 1965, 1981; Pollard & Sag, 1993), though it is not universally accepted (e.g. Lakoff, 1986). Similarly, it is assumed in the great majority of psycholinguistic research (though cf. McClelland, St. John, & Taraban, 1989). Some experimental evidence may bear on this question. For example, it is possible to prime syntactic structure in a manner probably independent of semantic factors, but the best evidence for this comes from language production (Bock, 1986). Also, evidence

from event-related brain potentials (where electrical activity from the brain is measured as participants perform a task) provides some evidence that syntactic and semantic processing are distinct. Semantically anomalous words elicit a negative-going brain wave about 400 ms after the stimulus (Kutas & Hillyard, 1980; see also Garnsey, Tanenhaus, & Chapman, 1989). After *I drink my coffee with cream and*, the wave for the anomalous *dog* is negative compared with the wave for *sugar*. In contrast, syntactically anomalous words produce a positive-going wave around 600 ms after the stimulus, whether the anomaly is due to ungrammaticality or a garden-path construction (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994; Osterhout & Mobley, 1995; see Osterhout, 1994). (However, it is important to note that wave-forms for different kinds of syntactically anomalous sentences may also differ.) Thus, there is experimental as well as linguistic evidence for representational modularity.

Most researchers in sentence processing follow standard linguistic theory in assuming representational modularity. The main debate concerns the status of processing modularity. The central question is whether all potentially relevant sources of information can be employed during initial processing or not. Let us now outline some of these sources of information. The first two of these are particularly problematic, as theories differ on how they are organised and the relationship between them.

Syntactic category information. We assume that category information forms part of the lexical entry for each word. For example, the entry for *loves* states that it is a verb and that it is transitive (i.e. it takes both a subject and an object). An important question is whether this constitutes two different sources of information: (major) category (e.g. verb, noun, adjective) and subcategory (e.g. transitive verb, intransitive verb). If so, then the processor might base initial processing decisions on major category information alone (e.g. Ferreira & Henderson, 1990; Mitchell, 1987). But if there is no distinction between category and subcategory, then this option would not be available to the processor.

Many words are ambiguous as to their category (e.g. *rose* can be a noun or a verb) or their subcategory (e.g. *eat* can be transitive or intransitive). The frequency with which each category or subcategory is used affects processing, and therefore forms part of this source of information. For example, people have less difficulty with a sentence that employs a verb used with a more frequent subcategory than a verb used with a less frequent one (e.g. Mitchell & Holmes, 1985). One important current debate is whether this information affects initial parsing decisions (e.g. Trueswell et al., 1993); see below.

