1 Introduction

1.1 A processing approach to syntax

Since the publication of Chomsky (1957) we have seen the development of a rich tradition of generative research into syntax. This tradition began by incorporating structuralist insights (cf. Saussure 1916; Bloomfield 1933; Harris 1951) into the kinds of grammatical rules that were widely employed in traditional grammars such as Jespersen (1924), thereby making these latter more explicit and more predictive in relation to grammaticality data (cf. Chomsky 1956). It led to a hypothesis about the role of grammatical knowledge within a larger theory of performance (cf. Chomsky 1965). It has contributed fundamental new insights into basic properties of natural language syntax and also into semantics, e.g. “constituent-command” (c-command) (cf. Reinhart 1983). It has elevated the goals of grammatical analysis beyond the facts of individual languages towards the construction of a theory of Universal Grammar (or UG), with built-in parameters of variation (cf. Chomsky 1981, 1982). And it has led to the formulation of numerous alternative models based on competing foundational concepts, whose empirical consequences are currently being worked out (cf. e.g. Aoun 1985, 1986; Gazdar, et al. 1985; Chomsky 1986; Blake 1990; and the papers in Baltin and Kroch 1989).

The philosophy that has been proposed as an explanatory background to, and motivation for, these generative descriptions is the innateness hypothesis (cf. e.g. Chomsky 1968). The universals of grammar are claimed to be innately represented in the human species and to be part of our genetic endowment, and it is this that is argued to explain the remarkable rapidity of language acquisition by children, despite the impoverishment and insufficiency of the input data to which the child is exposed (cf. Hoekstra and Kooij 1988).

Contrasting with these generative theories are approaches to syntax that are generally referred to as “functional.” The basic tenet of works such as
2 Introduction

Givon (1979), Hopper and Thompson (1980), and DuBois (1987) is that fundamental properties of syntax, for example the grammatical properties of transitivity or of ergativity, are shaped by their functions in discourse, i.e. syntax is the way it is because of the functions that it is used to perform in actual communication. Functional approaches may appeal also to a more cognitive explanation for grammatical structure, whereby properties of the real world, as interpreted by human cognitive mechanisms, are reflected in natural language syntax, e.g. in grammatical areas for which Haiman (1985) has proposed an iconicity explanation. Some have questioned whether it makes sense for discourse as such to be causally prior, and have recast Hopper and Thompson's (1980) discourse basis for transitivity in more cognitive and semantic terms, cf. DeLancey (1987) and Croft (1990).

These functional approaches are typically associated with the typological paradigm of Universal Grammar (cf. Hawkins 1988a for a summary). But in fact the equation of functionalism with typology fails in both directions: there can be, and are, functional approaches to generative syntax; and typology, as a theory and enumeration of different language types, is as compatible with a generative paradigm as it is with a discourse-based or cognitive one.

A functional approach to generative syntax is simply one that seeks to explain the properties of formal syntactic rules, principles, or constructions in terms of language use. Not all such properties, and not even all universals of language, will necessarily be explainable in innateness terms, as Newmeyer (1990a) points out, and hence the generative enterprise does not exclude functional explanations in principle. Moreover, there is a class of functional explanations that are particularly appropriate for generative syntactic descriptions, namely those that derive from the processing of syntax in real time, as exemplified in the work of Janet Fodor and Lyn Frazier (cf. section 1.2). And Newmeyer (1990b, 1991) even suggests that such processing considerations might explain why some grammatical properties have become bioligized and innate in the evolution of our species, whereas others have not. As for the nature of typology, the parameters of generative theory and their correlated properties define possible and impossible language types, just as implicational universals do within the tradition of Greenberg (1966) (cf. further Hawkins 1988a).

As we survey the current syntactic scene, therefore, there is no shortage of syntactic models for describing the grammatical details of various languages in terms of proposed universal primitives. There is also a large body of functional proposals that seek to provide motivations for syntactic phenomena in discourse and cognitive-semantic terms. But what has not yet received
A processing approach to syntax

the systematic attention it deserves is the question: to what extent is syntax the way it is because of processing? Sentences need to be produced and understood in real time. As psycholinguists develop models of production and comprehension, and as more and more data become available especially about how language is recognized on-line, it becomes natural to ask what role such processing constraints impose on grammars. That is the purpose of the present book. I will seek to show that the explanatory role of this particular functional consideration is considerable, that it can potentially replace many principles of UG (i.e. there is a performance explanation for much of what is regarded as the innate and universal core of syntax), and moreover that it is often processing, and not discourse or pragmatics, that shapes syntax when there is a functional explanation.

This book explores the processing basis of syntax, therefore, with particular reference to the order of meaningful elements, and with reference to certain general properties of constituent structure. I believe that this approach can illuminate a number of facts and regularities that are currently mysterious, or apparently arbitrary, by placing them in the context of the mechanisms by which language is produced and understood. We will proceed from questions such as: how can listeners work out what the syntactic structure of the intended sentence is, given the phonetic speech stream provided by the speaker? and what structures do speakers and hearers find more or less complex to manipulate in real time? When we view the data from this perspective, explanations can often be given for what are currently pure stipulations in the grammar, and one descriptive approach to the facts can be motivated over another.

1.2 Some proposed processing explanations for grammar

There is a handful of proposed processing explanations for grammatical phenomena in the current research literature, and these need to be reviewed at the outset. The assumption that underlies them is that there are general psychological mechanisms by which language is produced (cf. e.g. Levelt 1989) and comprehended (cf. e.g. Fodor, et al 1974) in real time, and that explain how it is possible for language users to communicate information rapidly and efficiently using the language code. Since its inception, the field of psycholinguistics has provided a growing body of experimental findings that shed light on these mechanisms. A further widely held assumption is that there are innate neural structures for the processing of language of the same basic kind that regulate other cognitive and perceptual capacities, such as
Introduction

vision. Achieving efficiency and rapidity in comprehension is argued by Jerry Fodor (1983) to require a reflex-like and neurally specialized response to the incoming speech signal and to the mental representations derived from it, in the manner of his modularity thesis. For example, speech sounds can be recognized at an extremely rapid rate, up to 25 phonemes per second, whereas non-speech sounds cannot be identified as rates exceeding 7–9 items per second (cf. Lieberman 1992). Hence it is reasonable to assume that we are biologically pre-programmed to perceive human speech. There are also evidently neural mechanisms for recognizing and accessing individual words in the incoming speech signal, and these mechanisms are what underlie the extremely rapid and efficient word-recognition processes for which psycholinguists such as Marslen-Wilson and Tyler (1980) have presented experimental evidence. What they have shown is that words are recognized either at or prior to their “uniqueness point,” i.e. that point at which a word becomes uniquely different from all the other words in the language that share the same initial sound segments in the incoming speech stream. Given the rapidity of the speech stream, the recognition of the intended word at or prior to its uniqueness point means that there must be an equally rapid and efficient process mapping speech sounds into a mental lexicon in which words are encoded and grouped according to their left-to-right phonological structure, as in a dictionary (cf. further Hawkins and Cutler 1988).

At a more abstract level, we are evidently equipped with mechanisms for recognizing on-line syntactic structure, such as the groupings of words into phrases and constituents. This higher sentence structure is derived crucially from the lexical information provided by word recognition, which in turn presupposes phonological identification. Some proposals for universal syntactic parsing mechanisms can be found in Frazier (1979a), Frazier and Fodor (1978), and Inoue (1991). Again, there are plausibly innate neural mechanisms mediating between word recognition and the construction of syntactic representations. There are also plausibly innate mechanisms for interrelating all the levels of linguistic representation that need to be postulated in an explicit model of grammar and that contribute to the real-time understanding of any utterance. Crucially important, of course, is the mapping from word recognition and syntactic representation onto some representation of meaning, which in turn serves as input to a pragmatically enriched discourse representation.

For all these processes, there is evidence for performance mechanisms that operate obligatorily and immediately in response to the sense data within their respective domains. Psycholinguists disagree, in fact, over the extent to which these domains are “encapsulated” or autonomous from the
Some proposed processing explanations for grammar

“central processes” of natural language understanding, i.e. specialized for the task at hand (cf. Marslen-Wilson and Tyler’s 1987 arguments against Jerry Fodor’s 1983 modularity thesis). But they are agreed about their fundamental speed and efficiency, and also presumably about the fact that these processing mechanisms are biologically pre-programmed in some form.

Some attempts to explain grammars in terms of processing mechanisms are the following. In phonology there are quite striking regularities in the composition of vowel inventories across languages, as Crothers (1978) has shown. Languages exhibit vowel systems of varying size, comprising three vowels, four, five, six, and so on, and they expand these inventories in largely predictable ways, starting from a basic three-way distinction between [i], [u], and [a]. Lindblom et al. (1984) have provided a processing explanation for this structuring, arguing that these expanding inventories are those that achieve “sufficient perceptual differences at acceptable articulatory costs” within the available phonetic space, and they have corroborated their predictions using computer simulations of the relevant phonetic variables. For a similar explanation for expanding consonant inventories, cf. Lindblom and Maddieson (1988). We have here a production and perception explanation for phonological structuring across languages.

In morphology, Hawkins and Gilligan (1988) document a clear preference for suffixing versus prefixing and infixing across languages, following early work by Greenberg (1957) in this area; and Hawkins and Cutler (1988) offer an explanation for this preference in terms of principles of morphological processing. Cf. Hall (1991) for a revised and more detailed explanation.

In syntax, Kuno (1973a, 1974) and Dryer (1980) have shown that there is a correlation between the perceptual difficulty of center embeddings and their frequent ungrammaticality in English and many other languages, cf. example (1.1) from English and the cross-linguistic summary of data provided in Hawkins (1988b):

(1.1) a. *Did it [that John failed his exam] surprise Mary?
b. Did it surprise Mary [that John failed his exam]?

The even more systematic avoidance of self-embedded relative clauses across languages, as in (1.2), has been given a processing explanation in Frazier (1985) in terms of the excessive number of non-terminal syntactic nodes that need to be postulated within a local viewing window of three adjacent words:

(1.2) *The farmer [that the cow [that gave bad milk] kicked] died.
6 Introduction

Frazier (1979b, 1985) has examined the word-order universals of Greenberg (1966) from the perspective of the two-stage parsing model developed in Frazier and Fodor (1978), in which the first stage has a limited viewing window of five to six adjacent words. She argues that certain syntactic sequences, such as (1.3) (a postnominal relative clause within a noun phrase co-occurring with a postposition within a postpositional phrase) would regularly exceed this limit, and she predicts that they will be non-occurring:

(1.3) \text{RefNoun[Head Noun \ Rel. Clause] Postposition]}
Y \quad \text{Complement} \quad X
\text{of Y}

Janet Fodor (1978, 1984) has proposed a parsing explanation for the Nested Dependency Constraint, illustrated in (1.4):

(1.4) \text{What are boxes easy to store in \_\_\_?}

The nested reading store boxes in what is grammatical, but the intersecting store what in boxes is not, and Fodor argues that this constraint has been grammaticalized in response to pressure from the parser. The cross-linguistic generality of this constraint has yet to be established, and for at least one language (Dutch) there are systematic exceptions, cf. Steedman (1984).

Frazier (1985) discusses “garden-path” sentences, which are known to cause processing difficulty, and notes that they do nonetheless occur quite frequently in English. She argues, however, that structures that would garden-path the hearer on every occasion of use regardless of the particular lexical items chosen will be blocked by the grammar. An example would be (1.5b) which will always be misanalyzed as a main clause on-line, prior to reanalysis as a subordinate clause when the matrix verb is encountered (cf. Bever 1970):

(1.5) a. That Mary is sick surprised Bill.
   b. *Mary is sick surprised Bill.

Frazier labels this the Impermissible Ambiguity Constraint.

A rather different tradition of research has also contributed to parsing explanations for grammatical principles, that of computational linguistics, exemplified in works such as Marcus (1980) and Berwick and Weinberg
Some problems for processing explanations

(1984). One principle that has been much discussed is subjacency, the proposal that no rule can relate two constituents separated by more than one bounding node (cf. e.g. van Riemsdijk and Williams 1986). Berwick and Weinberg offer an explanation for this in terms of the bound on left context within their proposed parser.

1.3 Some problems for processing explanations

These explanations are very suggestive, but they raise more questions than they answer, in my opinion, and they do not yet enable us to conclude that processing has shaped grammar to any significant extent. Why not?

The first issue that has to be resolved involves the precise status of many of the sentences under discussion. Are they unacceptable in performance and also ungrammatical (i.e. not generated by the relevant grammar)? Or unacceptable, yet still grammatical? If the latter, then any processing difficulty associated with e.g. center embeddings (cf. (1.1a)) or self-embeddings (cf. (1.2)) will not be reflected in any grammatical rules or constraints, since grammars will freely generate such structures. If there are rules or constraints blocking these structures, then processing may be a possible explanation.

The Standard Theory of generative grammar (cf. Chomsky 1965) was founded on the premise that processing difficulty could determine only the acceptability status of sentences, and not their grammaticality. The rules of the grammar generate all and only the grammatical sentences of a given language, and these are then assigned different degrees of acceptability within a performance model that contains grammar as one of its components. I will call this the assumption of pure acceptability. Naturally, if we accept this assumption, then processing difficulty will not be reflected in grammars, and performance and competence will be quite orthogonal.

One possibility that is admitted by Chomsky involves processing explanations that hold at the level of the evolution of the species, cf. Chomsky and Lasnik (1977) and also Newmeyer (1990b, 1991). Considerations of performance might have shaped the human language faculty in its evolution, favoring some innate universals of grammar over others.2

The processing explanations summarized in section 1.2 are clearly predicated on a rejection of the assumption of pure acceptability. That is, they assume that there is some grammatical mechanism blocking sentences of the relevant type, and that processing explains why it operates the way it does. But some of these explanations raise a second set of problems: not all
8 Introduction

languages appear to have responded to the processing difficulty in question; and not all sources of processing difficulty seem to result in a regular grammatical response.

For example, if processing difficulty is supposed to explain the absence of center embeddings in English and in many other languages, why is it that exactly the same type of center embedding does nonetheless show up in certain languages, such as Japanese? The structure of (1.6a) is grammatical in Japanese, even though many center-embedded Ss are regularly preposed within their containing clauses in Japanese performance, as in (1.6b) (especially longer ones, cf. ch. 4.1.5), thereby confirming the processing difficulty they cause. Example (1.6a) is nonetheless both grammatical and acceptable:

(1.6) a. Mary-ga [kinoo John-ga kekkonsi-ta to] it-ta.
   “Mary yesterday John married that said,” i.e. Mary said that John got married yesterday.


The productivity of (1.6a) is an embarrassment for whatever explanation is proposed for the status of (1.1a) in English and many other languages. The logical problem here can be summarized as follows: if one is going to appeal to fundamental properties in the architecture of human processing mechanisms in order to explain why grammars block certain structures, then why are they not blocked in all grammars? The same problem arises for self-embedded relative clauses, which are systematically avoided in some languages, but tolerated in others (cf. example (1.9) below).

Conversely, some sources of processing difficulty seem to invite more of a grammatical response than others. Frazier (1985) points out that garden paths occur quite frequently in English (cf. the famous sentence The horse raced past the barn fell and many others), whereas center embeddings are regularly avoided. Both types of structures are known to cause processing difficulty, so why the differential response?

In order to avoid these kinds of insufficiencies, another set of processing explanations makes (at least implicitly) what I will call the assumption of absolute unprocessability. Some absolute limit on human processing resources is proposed within a given domain, and a grammatical response is then explained on the grounds that this limit would be exceeded by structures of a certain type. The word order combination of (1.3) above would regularly exceed the five-to-six-word viewing window of the first stage of Frazier and
Some problems for processing explanations

Fodor's (1978) parser. Subjacency violations would exceed the bound on left context within Berwick and Weinberg's (1984) parser. Hence the relevant structures are necessarily non-occurring.

This assumption leads to a third set of problems, however, because these proposals construct strong claims on rather weak foundations. We just don’t know, in the current state of the art, what the absolute limits on the human processing system are. Nor has there been enough cross-linguistic research to establish with confidence that certain types of grammatical phenomena are not attested. Even if we grant that the parser operates in two stages along the lines of Frazier and Fodor (1978), does it necessarily follow that immediate constituents whose recognition exceeds the five to six words of the first-stage viewing window are unprocessable? How good is the evidence for a five-to-six-word limit? Can’t other parsing or memory systems interact in these cases and help out? I shall argue in ch. 5.1.3 that there are, in fact, languages whose constituent orders will regularly exceed Frazier’s (1985) proposed limit. Similarly, how good is the evidence for Berwick and Weinberg’s (1984) parser? And is subjacency a genuine universal? Many surface structures of English are prima facie counterexamples to both subjacency and the Berwick and Weinberg parser, since the WH-word can be separated by two or more bounding nodes from its trace, as in (1.7), and they are real counter-examples to the extent that successive Comp-to-Comp movements (preserving subjacency) are either linguistically or psycholinguistically unmotivated:

(1.7) Who, did you say that you imagined Fred would consider asking Susan to bring it, to the party?

A further problem that will plague all proposals based on the assumption of absolute unprocessability is the directionality-of-explanation issue. Does the grammar respond to the parser, or vice versa? Smith (1986) draws attention to this in his review of Berwick and Weinberg’s parsing explanation for subjacency: “one could equally well argue that the (arbitrary) existence of subjacency as a principle of grammar explained the bound on left-context in the parser, as opposed to vice versa” (p. 225).

To summarize, the first set of problems for any processing explanation for grammars involves disentangling the unacceptability-versus-ungrammaticality status of sentences. Secondly, grammars do not respond uniformly to processing difficulty: not all grammars block a structure that is difficult to process; and not all sources of difficulty seem to result in a regular grammatical response. And thirdly, assumptions about the absolute unprocessability of certain structures are largely premature at this point.
10  Introduction

1.4  The approach of the present book

I believe that there are solutions to these problems. I believe, moreover, that the impact of processing mechanisms on grammars has been absolutely fundamental, and that by solving these problems we will be able to see the extent of this impact more clearly.

First, I reject the assumption of pure acceptability. Despite the problems alluded to above, there is simply too much anecdotal evidence suggesting that grammars can respond, and have responded, to processing, by incorporating rules, constraints or principles that owe their raison d’être to the need for language users to produce and comprehend language in a rapid and efficient manner in real time. Once we admit that processing can determine grammatical properties in any one area or for just one structural type, we have of course disproved the assumption, and it becomes important to establish to what extent performance is a significant determinant of grammars, and how performance interacts with those aspects of grammar that are independent of it and perhaps derivable from an innate UG. Among the processing explanations enumerated in section 1.2, at least some of them will, it seems to me, be able to survive critical scrutiny in some form, and thereby disprove Chomsky’s (1965) assumption of the relationship between grammar and performance.

With the benefit of hindsight, I would suggest that the assumption of pure acceptability was not well argued to begin with. It seems reasonable enough for verb-particle constructions in English such as *I called my long lost uncle from America who had been trying to reach me for several weeks without success up*, in which increasing distance between verb and particle results in declining acceptability, and there is no clear point at which, and no obvious grammatical mechanism by which, one could stipulate that sentences of this sort have become ungrammatical. But this assumption is, at the very least, questionable for structures such as center embeddings of $\hat{S}$ in (1.1a), which are also difficult to process, but whose unacceptability/ungrammaticality is now definable in terms of grammatical structure: e.g. star an $\hat{S}$ with non-null material to both left and right within a containing $S$.

It is not sufficient to argue that sentences like (1.1a) are grammatical because they can be generated by independently motivated rules or principles of English. For there are countless ungrammatical strings that are blocked by a grammatical constraint on otherwise productive generative mechanisms, especially in the highly interactive, principle-based models of current generative theory. Nor can the “grammatical but unacceptable” status of (1.1a) be