Generalized Phrase Structure Grammar

Gazdar, Klein, Pullum and Sag

Generalized Phrase Structure Grammar

GERALD GAZDAR, EWAN KLEIN,
GEOFFREY PULLUM, IVAN SAG

Basil Blackwell

First published 1985

Basil Blackwell Publisher Ltd
108 Cowley Road, Oxford OX4 1JF, UK

British Library Cataloguing in Publication Data

Generalized phrase structure grammar.
1. Generative grammar
I. Gazdar, Gerald
415 P158

ISBN 0-631-13206-6
ISBN 0-631-13459-X Pbk

Typeset by Stephen Austin and Sons Ltd, Hertford
Printed in Great Britain by T. J. Press Ltd, Padstow, Cornwall

Preface

This book contains a fairly complete exposition of a general theory of grammar that we have worked out in detail over the past four years. Unlike much theoretical linguistics, it lays considerable stress on detailed specifications both of the theory and of the descriptions of parts of English grammar that we use to illustrate the theory. We do not believe that the working out of such details can be dismissed as 'a matter of execution', to be left to lab assistants. In serious work, one cannot 'assume some version of the X-bar theory' or conjecture that a 'suitable' set of interpretive rules will do something as desired, any more than one can evade the entire enterprise of generative grammar by announcing: 'We assume some recursive function that assigns to each grammatical and meaningful sentence of English an appropriate structure and interpretation.' One must set about constructing such a function, or one is not in the business of theoretical linguistics.

This book will not be an easy read. Parts of it – the parts most crucial to the functioning of the general theory – are formalized rather precisely. Those parts – generally the final sections of the chapters – are naturally somewhat demanding. We hope that none of the formalism is gratuitous, and although it is doubtless not as clean as it should be, we have struggled to render it conceptually clearer than when it started out, and in every case, we explain in prose what each formalized principle says and what its role is in the general scheme of things. In the semantics chapters, in particular, we have attempted to allow for the reader who has little specialized knowledge of the subject. In general, we hope that graduate or advanced undergraduate students who have had courses in some version of generative grammar, and who have at least some introductory acquaintance with logic, mathematics, or computer science, will be able to understand generalized phrase structure grammar through studying this book, and will grasp something of what a rigorous syntactic and semantic theory of language might be like.

We have numerous people to thank for their assistance in the writing of this book. Some of them will be quite surprised that we regard them as having helped; but to us, virtually every colleague who curled a lip at some ugly patch of badly worded definitions, every student who refused to shut up when told to and pressed home an embarrassing question, and every linguist who paid us the compliment of addressing our work in published or presented papers has contributed something to this book. We cannot possibly identify every one of them, but we want to mention the names of some, even at the risk of slighting the others.

Much of our work has been conducted in the intellectual community of Stanford, Palo Alto, and Menlo Park. Many people who have worked there between 1981 and 1984 have talked to us, argued with us, and contributed clever ideas. Among them are Mike Barlow, Herb Clark, Mark Cobler, Robin Cooper, Chris Culy, Murvet Enc, Elisabet Engdahl, Chuck Fillmore, Dan Flickinger, Mark Gawron, Lauri Karttunen, Martin Kay, Paul Kay, Susannah MacKaye, Geoffrey Nunberg, Kathy O'Connor, Anne Paulson, Fernando Pereira, Stanley Peters, Carl Pollard, Kelly Roach, Jane Robinson, Stuart Shieber, Susan Stucky, Henry Thompson, Hans Uszkoreit, Tom Wasow and Michael Wescoat.

The reception of our ideas have received in the British linguistics and cognitive science community has often been sympathetic and stimulating, and we have benefited from the queries and comments of Bob Borsley, Gill Brown, Keith Brown, Brian Butterworth, Ronnie Cann, Richard Coates, Grev Corbett, Connie Cullen, Anne Cutler, Roger Evans, John Foster, Alan Garnham, Steve Harlow, Geoff Horrocks, Steve Isard, Phil Johnson-Laird, Christopher Longuet-Higgins, John Lyons, Rose Maclaran, Peter Matthews, Steve Pulman, Barry Richards, Graham Russell, Aaron Sloman, Neil Smith, Larry Trask, Nigel Vincent, Anthony Warner, and Yorick Wilks.

Invitations from Harry Whitaker at the Department of Speech and Hearing, University of Maryland, in 1982 and Robert P. Stockwell at the Department of Linguistics, UCLA, in 1983 allowed us to present some of the results of our research to classes at two successive Linguistic Institutes. This was a valuable experience which taught us that we hadn't got everything as right as we thought we had. We learned a lot from Carol Anderson, Paul Chapin, Rob Chametzsky, Mary Dalrymple, Dominique Estival, Aryeh Faltz, Donka Farkas, Erhard Hinrichs, Chu-Ren Huang, Geoff Huck, Carolyn Jenkins, Mark Johnson, Ed Keenan, Michael Moortgat, Susan Mordechay, Young-Hee Na, Almerindo Ojeda, Jessie Pinkham, John Richardson, Jerry Sadock, Paul Schachter, Peter Sells, Mary Tait, and Shelley Waksler.

At longer distance, we have regularly had encouragement, support, and valuable feedback from Jan Anward, Emmon Bach, Greg Carlson, Ken

Church, Östen Dahl, David Dowty, Eva Ejerhed, Janet Fodor, Georgia Green, Takao Gunji, Lars Hellan, Pauline Jacobson, Aravind Joshi, Hans Kamp, Bill Ladusaw, Jim McCawley, Jim McCloskey, Joan Maling, Jerry Morgan, Dick Oehrle, Barbara Partee, Jeff Pelletier, Len Schubert, Frieda Steurs, Greg Stump, Rich Thomason, Mariko Udo, Annie Zaenen, and Arnold Zwicky.

Ideas are not like houseplants. They do not grow best in uniform conditions of bland comfort. They thrive when faced with opposition and counterattack. We have been greatly aided by our disputes with people who have argued that generalized phrase structure grammar is the wrong path to take in syntactic and semantic theory. We cherish particularly the challenging and stimulating critical attention our ideas have received at the hands of people like Joan Bresnan, Jean Gibson, Kris Halvorsen, Jorge Hankamer, Frank Heny, Dick Hudson, Ron Kaplan, Alec Marantz, Fritz Newmeyer, David Perlmutter, David Pesetsky, Paul Postal, Mark Steedman, Tim Stowell, and Edwin Williams. Our thanks to them are entirely sincere. We hope that if our ideas still seem wrong to them, this book will at least be valuable in making it clearer to them why they are right.

We owe a special debt to Elisabet Engdahl and Mark Gawron who made detailed comments on preliminary drafts of many of the chapters in this book (as well as several that have not been included), and to the people who have given their time to provide valuable research assistance: Mike Barlow, Mark Cobler, Dan Flickinger, Jerry Kelly, Susannah MacKaye, Karen Wallace, and Michael Wescoat.

There are several agencies and institutions without whose financial and general support the process of finishing a book with authors located in Brighton, Edinburgh, Palo Alto, and Santa Cruz would not have been possible. At Stanford University during several summers, support from the Sloan Foundation, two National Science Foundation grants (BNS-8102404, BNS-8309780), and a gift from the System Development Foundation to the Center for Study of Language and Information (CSLI) have helped support us, as has travel assistance from the British Academy (Small Grants Research Fund in the Humanities) and facilities provided by the Syntax Research Center at the University of California, Santa Cruz. Grants from the UK Social Science Research Council supported Gazdar's work at the University of Sussex. An Advanced Research Fellowship from the UK Science and Engineering Research Council (SERC) has supported Klein's work at the University of Edinburgh. A Summer Stipend from the National Endowment for the Humanities and a Faculty Research grant from the University of California, Santa Cruz, have assisted Pullum's research.

We are grateful to Hewlett-Packard Laboratories in Palo Alto for

allowing access to their computing facilities after hours while we worked on an electronic typescript of this book in the fall of 1983, and to Judea Pearl of UCLA for making his computing facilities available to us during the 1983 Linguistic Institute. We would also like to record our thanks to the many computer users and consultants at CSLI, Edinburgh, Sussex, UCLA, and UCSC who have uncomplainingly wasted their own time helping us with tapes, weekend machine crashes, shell scripts, nroff macros, printer drivers, terminal emulations, and the rest of the information technology that has become a *sine qua non* of producing a book in the 1980s. And we thank our publishers, in particular Stephen Ball, Philip Carpenter and John Davey: the existence of this book is more of a testament to their tolerant persistence than it is to ours.

Finally, we express our appreciation to the Digital Equipment Corporation (DEC), whose ubiquitous machines have enabled us to collaborate in eight different locations on two continents, and to the air traffic control staffs of London and San Francisco, who have so frequently had our lives in their hands.

Contents

1

Introduction

1 Generative grammar

This book is a contribution to the discipline known as generative grammar. This approach to linguistics is characterized by its goal of investigating natural language through the construction of fully explicit descriptions of particular languages and a formalized general framework for defining the space within which to locate such descriptions. The end to which this effort is directed is the development of a general theory of the structure of natural languages. If the formal framework is restrictive enough to make nontrivial, falsifiable claims about what may and may not be a natural language, or a grammar for a natural language, then we can identify the framework itself with the traditional notion 'universal grammar', and interpret it not merely as a formal language for representing grammars of particular languages but rather as a partial characterization of what natural languages and their grammars are like.

The basic assumption made in generative grammar is that languages can be regarded as collections whose membership is definitely and precisely specifiable. The elements of such a collection are the expressions in the language. Following Montague (1970) and Brame (1981), we assume that the grammars of natural languages should define not merely the expressions corresponding to sentences but also subsentential expressions of all categories. Clearly the set of compound linguistic expressions in a natural language is not finite, so we cannot list them. An interpreted formal system defining the membership of the collection of linguistic expressions, and assigning a structure and an interpretation to each member, is required. This is what we call a grammar. The study of language is intimately bound up with the study of grammars for those who accept this basic premise of the generative approach to linguistics, and parts of this book deal in considerable detail with the precise statement of particular rules and principles of grammar. Wherever possible, we do

more than just illustrate points of theory with suggestive examples or hint at what some components of the grammar might contain. There are three crucial methodological assumptions that lead us to proceed in this way:

I A necessary precondition to 'explaining' some aspect of the organization of natural languages is a description of the relevant phenomena which is thorough enough and precise enough to make it plausible to suppose that the language under analysis really is organized in the postulated way.

II A grammatical framework can and should be construed as a formal language for specifying grammars of a particular kind. The syntax and, more importantly, the semantics of that formal language constitute the substance of the theory or theories embodied in the framework.

III The most interesting contribution that generative grammar can make to the search for universals of language is to specify formal systems that have putative universals as *consequences*, as opposed to merely providing a technical vocabulary in terms of which autonomously stipulated universals can be expressed.

These three points merit some further discussion. Consider first I, i.e. the issue of whether an explanatory account of some grammatical phenomenon can be provided without the descriptive detail having been worked out. It has regrettably become more and more common of late to find linguists suggesting that broad hypotheses about grammatical theory can be discussed in the absence both of formal work that demonstrates that certain implications follow from those hypotheses and of descriptive work showing that the putative implications are well confirmed. Our experience is that even after quite a significant amount of work has been done on a proposal for the description of some fragment of a language, it is quite difficult to see in full detail what its consequences are.

There may not even be algorithmic ways of confirming the consequences of some theories of grammar, of course: if the theory allows grammars for nonrecursive sets, then we run the risk that the claim that some string is not generated by some grammar cannot be verified in principle. Familiar statements of the type 'Thus our grammar excludes examples like (158)', in other words, may simply be untestable conjectures.[1]

This observation leads naturally to our second assumption, II, that a grammatical framework is best construed as a formal language having itself both a syntax and a semantics. A grammar characterizes a language, say Japanese, DEC10 Prolog, or the Polish postfix notation for arithmetic. A family of grammars characterize a family of languages, say the indexed languages, the natural languages, or the assembly languages for 8-bit

microprocessors. Each of these object languages has a syntax and a semantics. In formal language theory, the syntaxes of a given family of object languages are themselves specified in a formal language (e.g. the Backus–Naur notation for the grammars of context-free languages). This metalanguage itself has both a syntax and a semantics. Since such a metalanguage has the grammars of the object languages as its topic, it follows that the semantics of the metalanguage has as its domain syntactic entities (strings, trees, categories, etc.) in the object languages. The syntax of the metalanguage is in many respects arbitrary, but it needs to stand in some fairly perspicuous relation to the intended semantics, and it needs to be explicit enough for one to see what can and cannot be expressed by means of it. Much of the technicality of the present work stems from an attempt, not uniformly successful, to be as explicit as possible about what our grammar formalism is, and what it means.[2]

Assumption III, that universals are most interesting when embedded as integral parts of a formal system that has some nontrivial structure, involves just as clear a break with the approaches adopted in much current work. It goes without saying that the process of searching for grammatical universals initially involves attempting to discover facts about language (as opposed to facts about some particular language or set of languages). But there is a sense in which even a precise formulation of a successful discovery of this sort will not constitute a truly interesting result in theoretical linguistics. If the fact needs a special statement, as opposed to following from the very form in which the theoretical reconstruction of the notion 'natural language' has been cast, the job is not done.

Thus, for example, one might propose that natural language grammars never exhibit direct grammatical dependencies between elements separated by more than two phrasal categories of a certain sort, or that they never permit a full category in a certain position in the clause, or whatever. But these proposed universals are not accounted for by the mere fact of their having been written down in some uninterpreted algebraic formalism. The explanatory task has not even begun when a constraint or generalization is merely stated. Only when it can be shown to be a nontrivial consequence of the definition of the notion 'possible grammar' can it be regarded as explained, because while it resides in the form of an autonomous statement it can be modified, enhanced, weakened, or even discarded with no consequences for the rest of the theory (cf. Dowty 1982b, pp. 107–8, on this important point). The penalty for failure of such a universal is effectively zero; a new universal saying something carefully hedged to avoid the last known counterexample can be constructed in a moment. Ironically, in view of the fact that such universals are often presented with a considerable fanfare of rhetoric about explanation, they have much the same status as the descriptive universals we find in the

typological work that takes its lead from Greenberg (1963) – only these claims, being better researched, generally have a much longer half-life.

Our goal in the work that has led to GPSG has been to arrive at a constrained metalanguage capable of defining the grammars of natural languages, but not the grammar of, say, the set of prime numbers. (The phrase 'capable of' indicates a rather ambitious program; a somewhat less ambitious one, under which we need not require that the set of prime numbers be literally indescribable within the terms of our theory, is obtained if we replace this phrase by 'suited to'.) The universalism is, ultimately, intended to be entirely embodied in the formal system, not expressed by statements made in it. Consider, by way of illustration, the statement in (1).

(1) [VFORM FIN] ⊃ [−N, +V]

This states, in the terms we introduce in chapter 2 below, that having the value FIN, i.e. finite, for the feature VFORM implies being verbal and non-nominal; in other words, only a verb can have tense. We could state this in our theory of grammar as a universal feature co-occurrence restriction. But this would only amount to an admission of – hopefully temporary – defeat (an admission we may have to make at the present state of our knowledge, of course). If (1) is universal, then it should not need saying. It ought to be a consequence of the grammatical metalanguage itself – for example, by virtue of a theory of features which (unlike ours) ties tense securely to the semantic notion it expresses and simultaneously restricts its syntactic realization to verbal categories in the theory of grammar. If this were done effectively, the discovery of a language with tensed adjectives would severely compromise the theory of features as a whole and force revisions that would alter the consequences of the theory in other domains. If we simply rest content with the universal stipulation '[VFORM FIN] ⊃ [−N, +V]', we can drop it, or modify it to say '[VFORM FIN] ⊃ [+V]', at no real cost. There are, of course, languages with constructions that have been held to exhibit tensed adjectives (the so-called non-nominal adjectives in Japanese, for instance). The fact that it would be so easy to modify (1) to take account of them is precisely what we are drawing attention to.

We therefore regard universals stated within the metalanguage as inherently less interesting than those which are built into it. We exhibit in this book some claims, for example the Exhaustive Constant Partial Ordering claim about linear precedence in grammars (see chapter 3, section 2), which follow as consequences of our overall formal system. It is this sort of result that is an important goal of the GPSG approach to linguistics: the construction of theories of the structure of sentences under which significant properties of grammars and languages fall out as

theorems as opposed to being stipulated as axioms.

In view of the fact that the packaging and public relations of much recent linguistic theory involves constant reference to questions of psychology, particularly in association with language acquisition, it is appropriate for us to make a few remarks about the connections between the claims we make and issues in the psychology of language. We make no claims, naturally enough, that our grammatical theory is *eo ipso* a psychological theory. Our grammar of English is not a theory of how speakers think up things to say and put them into words. Our general linguistic theory is not a theory of how a child abstracts from the surrounding hubbub of linguistic and nonlinguistic noises enough evidence to gain a mental grasp of the structure of a natural language. Nor is it a biological theory of the structure of an as-yet-unidentified mental organ. It is irresponsible to claim otherwise for theories of this general sort. It may even be incoherent, as Katz (1981) and Soames (1984) have argued.

Thus we feel it is possible, and arguably proper, for a linguist (*qua* linguist) to ignore matters of psychology. But it is hardly possible for a psycholinguist to ignore language. And since a given linguistic theory will make specific claims about the nature of languages, it may well in turn suggest specific kinds of psycholinguistic hypothesis. Stephen Crain and Janet Fodor, in a series of papers (Crain and Fodor, in press; Fodor 1980, 1983a, 1983b) have argued that GPSG does have implications for psycholinguistic concerns. Nonetheless, it seems to us that virtually all the work needed to redeem the promissory notes linguistics has issued to psychology over the past 25 years remains to be done. If linguistics is truly a branch of psychology (or even biology), as is often unilaterally asserted by linguists, it is so far the branch with the greatest pretensions and the fewest reliable results. The most useful course of action in this circumstance is probably not to engage in further programmatic posturing and self-congratulatory rhetoric of the sort that has characterized much linguistic work in recent years, but rather to attempt to fulfill some of the commitments made by generative grammar in respect of the provision of fully specified and precise theories of the nature of the languages that humans employ. Even when that is done, the psychology of language will doubtless have a vast amount of work to do before we have a scientific understanding of how the human species acquires and uses language. After all, geometrical optics long ago provided us with a fairly clear and stable means of characterizing the objects of visual perception, but the psychology of visual perception still has many problems to solve. So far, linguistics has not fulfilled its own side of the interdisciplinary bargain.

Two additional terminological points are in order about this account of our theoretical orientation. First, notice that although Langendoen and

Postal (1984) take the term 'generative grammar' to be restricted to those theories that characterize recursively enumerable sets of sentences (and not, for instance, the proper classes of finite and infinite-length strings that they claim constitute natural languages), we do not regard ourselves as committed to any such limitation. It is straightforward to interpret our grammar as admitting a proper class of structures most of which are infinite in size if that is thought desirable.[3]

And second, note that the term 'generative grammar' is sometimes used as if it referred to (even solely to) contemporary work in Chomsky's 'Revised Extended Standard Theory' (REST) such as 'Government-Binding' (GB; Chomsky 1981) but not, for example, GPSG, Lexical-Functional Grammar (LFG; Bresnan 1982b), or Arc-Pair Grammar (APG; Johnson and Postal 1980). Van Riemsdijk (1982), for example, is particularly explicit in this usage, and even critics of REST sometimes adopt it (Comrie 1984). It will be clear that our use of the term 'generative grammar' covers GPSG, LFG, APG, Montague Grammar in all its varieties, the work presented in *Syntactic Structures* (Chomsky 1957), Stockwell et al. 1973, Lasnik and Kupin 1977, and other work, but includes little of the research done under the rubric of the 'Government Binding' framework, since there are few signs of any commitment to the explicit specification of grammars or theoretical principles in this genre of linguistics.

2 Syntax and semantics

While a purely syntactic approach to a language takes it to be simply a collection of expressions or other linguistic objects, natural languages have meanings associated with their expressions. Presumably, it is only because of the meanings carried by expressions in natural languages that they exist at all. That is, regardless of whether the natural languages employed by human beings function primarily as internal representation codes in which thinking can be carried out, or media for artistic expression, or systems for inter-organism communication, or have some other rationale for their existence, there would appear to be no value in knowing a natural language if no meanings were associated with its expressions. (Compare this with the case of phonetics: to say that there is no value in knowing a natural language if one cannot vocalize and process auditory data is simply not true. Human languages subsist in other modalities than the phonetic, e.g. in private thinking, in visual, tactile and electronic codes, in hand signing, and so on.) Thus it is uncontroversial (or should be) to assume that the specification of a relation between the expressions of a language and their meanings is a central goal of linguistic theory.

Furthermore, it is now widely accepted, though not entirely uncontroversial, that the theory of meaning for natural languages falls into two subcomponents, namely pragmatics and, under a narrow construal, semantics. Under the classical definitions of these components, semantics deals with the relation between expressions and what they denote, while pragmatics deals with the relation between expressions, their denotata, and their use. A related contemporary view maintains that semantics deals with the inherent meaning of expressions, while pragmatics deals with the meanings communicated by expressions on the occasions of their use.

This book will have almost nothing to say about pragmatics, but semantics plays a crucial role in a number of places (in the treatment of agreement, for example) and the two final chapters are devoted to the technical details of the semantic theory that we adopt, which is in essence that of Montague.

The distinction between semantics and pragmatics means that semantics, as we use the term, should not be thought of as the study of meaning *simpliciter*, because there may be aspects of linguistic meaning that are not treated by semantics at all, and are not supposed to be. This raises the issue of what a semantic theory for a natural language does have to do. In this book, we adopt, without extended justification, the view of semantics that has become dominant over the past decade. We assume that, minimally, a semantic theory for a natural language has to be able to provide a recursive definition of *denotation in a model* for the linguistic expressions of the language. Denotation is a general notion, applicable in principle to sentences of imperative, interrogative, and other types, but in the case of declarative sentences (which are canonically used to make statements), it amounts to the notion of truth in arbitrary states of the world. In order to know whether a given sentence is in fact true or not, we need to know two kinds of things: what the sentence means, and what the facts of the world are. This suggests that a theory of meaning should define a function which, given an arbitrary sentence of English and a possible state of affairs, tells us whether the sentence is true or false in that state of affairs. This is done by means of a *model*. A model is an abstract 'state of affairs' in which basic expressions of the language are assigned denotations. For example, proper names might be assigned individuals as denotations, and *n*-place predicates might be assigned *n*-ary relations on the domain of individuals. The recursive clauses of the semantics will then specify how complex expressions receive a denotation in the model on the basis of the denotations of their component expressions. This sounds quite a trivial enterprise when very simple cases are considered (for example, deciding that *All swans are white* is true in a state of affairs where all the swans there are (if any) are members of the set of white objects), but it