BSPS

Boston Studies in the Philosophy of Science

Volume 104

Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, T. N. Harris and S. Harris

The Form of Information in Science

Analysis of an Immunology Sublanguage

With a Preface by Hilary Putnam

Kluwer Academic Publishers

THE FORM OF INFORMATION IN SCIENCE

Analysis of an Immunology Sublanguage

ZELLIG HARRIS
MICHAEL GOTTFRIED, THOMAS RYCKMAN
PAUL MATTICK, Jr.
ANNE DALADIER
T. N. HARRIS and S. HARRIS

With a Preface by Hilary Putnam

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

Library of Congress Cataloging-in-Publication Data

CIP

The Form of information in science.

(Boston studies in the philosophy of science ;
v. 104)
Includes bibliographies.
1. Immunology--Language. 2. Linguistic
analysis (Linguistics) I. Harris, Zellig Sabbettai,
1909- . II. Series.
Q174.B67 vol. 104 001'.01 s 88-31513
[QR182] [415]

ISBN-13: 978-94-010-7777-4 e-ISBN-13: 978-94-009-2837-4
DOI: 10.1007/978-94-009-2837-4

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands

Kluwer Academic Publishers incorporates
the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AA Dordrecht, The Netherlands.

TABLE OF CONTENTS

PREFACE

DOES DISCOURSE HAVE A 'STRUCTURE'? HARRIS'S REVOLUTION IN LINGUISTICS

As a freshman back in 1947 I discovered that within the various academic divisions and subdivisions of the University of Pennsylvania there existed a something (it was not a Department, but a piece of the Anthropology Department) called 'Linguistic Analysis'. I was an untalented but enthusiastic student of Greek and a slightly more talented student of German, as well as the son of a translator, so the idea of 'Linguistic Analysis' attracted me, sight unseen, and I signed up for a course. It turned out that 'Linguistic Analysis' was essentially a graduate program – I and another undergraduate called Noam Chomsky were the only two undergraduates who took courses in Linguistic Analysis – and also that it was essentially a one-man show: a professor named Zellig Harris taught all the courses with the aid of graduate Teaching Fellows (and possibly – I am not sure – one Assistant Professor). The technicalities of Linguistic Analysis were formidable, and I never did master them all. But the powerful intellect and personality of Zellig Harris drew me like a lodestone, and, although I majored in Philosophy, I took every course there was to take in Linguistic Analysis from then until my graduation.

What 'Linguistics' was like before Zellig Harris is something not many people care to remember today. (The 'bible' of the subject – except at Penn! – was Bloomfield's *Language*, a knowledgable, scholarly, but deeply operationalist view of linguistics.) All of Harris's ideas were different from those that were being studied elsewhere: the idea of a 'transformation', later modified and made famous by Noam Chomsky, the idea of the autonomy of syntax, and the condensed mathematical notation which made it possible to represent a grammar in a few pages of what looked like equations. Since these three ideas have been taken over by 'generative grammar', it is important to be aware that Harris's view differs in important respects from that of the generative grammarians: For example, like the philosopher Nelson Goodman, Harris is deeply aware that any set of scientific phenomena admits of more than one description. He is not one to insist that a particular 'description' of the grammar of, say, English, describes *the* grammar coded in the brains of English speakers – fashionable as that sort of utterly unsupported speculation is today. And like the late Roman Jakobson, Harris is interested in the syntax of whole discourses, and not just of individual sentences. Indeed, the major part of Harris's long and incredibly productive scholarly life has been devoted to the development of tools for the responsible study of what so many 'literary theorists', 'structuralists', etc., talk about *ir*responsibly – the structures that characterize different types of discourse.

The great aim of Harrisian 'discourse analysis' is to do this purely *syntactically*. But, like Chomsky, Harris of course hopes that syntactic regularities will be associated with semantic ones: not, however, with 'innate' semantic structures, nor yet with 'universal' ones, but precisely with structures which grow and change as the discourse studied grows and changes.

Paul Mattick has drawn an interesting parallel between Harris's work and the ambitions of the Logical Positivists.[1] The Logical Positivists thought that by rewriting scientific theories in the artificial language of Symbolic Logic they would be able to discover what their structure really was. (This enterprise flourishes today under the direction of Wolfgang Stegmüller at the University of Munich, for example). Harris believes that one can give a precise description of scientific discourse (and, of course, non-scientific discourse as well) *without* first having to rewrite it in an artificial language. The importance of avoiding such rewriting should be clear: even if the ancient Italian proverb that 'to translate is to betray' is an exaggeration, it is clear that every translation expresses the translator as much as it does the original; and philosophical 'reconstructions' of scientific theories have richly illustrated this fact. When we are given a description of the 'structure' of a physical theory by a philosopher the one thing we can be sure of is that some other philosopher will give a totally different description of that structure. And those philosophers who, unlike the Logical Positivists, have not even attempted any kind of precision in their descriptions of 'structures' of discourse have been even more unconstrained than the Positivists – and have, not surprisingly, disagreed with one another even more.

In this state of affairs, one possible response is to adopt the sort of subjectivistic ideology that is today, lamentably, sweeping Parisian intellectual life: to conclude not only that accounts of linguistic and conceptual structures must be subjective, but also, by some kind of incredible extension, that *everything* human beings can think is subjective, is just a play with styles and with texts. Ultimately this view extends to a view of the human being: we are, we are solemnly told, just a play with mirrors, or (in another fashionable figure) 'centerless webs'.

This sort of pessimism mixed with irresponsibility is not new (a hundred years ago it was called the *fin de siècle* mood), and it is not destined to last. In the meantime, with unflagging brilliance and with unflagging energy, Harris has continued to pursue the task that even the Positivists thought impossible: to describe the structures of conceptual thought in a particular area with rigor and without first wholly rewriting those structures in the alien language of Symbolic Logic. The present volume, substantial as it is, is only a large 'pilot study' in this huge project. The idea that informs it is, like all of Harris's central ideas, breathtaking in its combination of simplicity and daring: to take a subfield of a particular scientific discipline and compare the structures of the texts in that subfield before and after a particular 'scientific revolution' in the subfield. The details look formidable at first blush – just as the papers on the grammars of various languages that I encountered as a freshman in Harris's notation looked formidable at first blush. But the reader who is serious about wanting to know if such a thing as 'structural analysis of discourse' is possible and who is willing to do a little work will find that the presentation is not as hard to follow as it looks, and that the payoff is large. This is a book every serious student of discourse, whether he comes from linguistics, from philosophy, from cognitive science, or whatever, will have to become aware of, and will have to learn to understand, at least in its central outline and key ideas. Bravo, Zellig! You have done it again.

HILARY PUTNAM

NOTE

[1] P. Mattick, Jr., "The Constitution of Domains in Science: A Linguistic Approach", in A. Fine and P. Machamer (eds.), *PSA 1986,* vol. I (East Lansing, 1986), pp. 333–341.

FOREWORD

This book presents a formal method for analyzing the word combinations in articles of a subscience, in a manner that gives the information in the science a more precise form, and may tell a good deal about the structure of the science itself. The method arises from the analysis of language as a mathematical system, and from the inherent correspondence between the form of language – under this analysis – and its information. The specific results obtained here arise from applying this method within the confines of a single area of science.

The field investigated here is in immunology: the search (c. 1940-65) to determine which cell produces antibody. Two different cells were claimed by different scientists; these were ultimately found to be different stages of the same cell-line. In analyzing the research articles, words were collected into classes on the basis of their occurring with each other, in regular ways in respect to other word combinations, in the sentences of the articles. The regular word combinations in these sentences were then found to fit into a closed set of word-class sequences. These word-class sequences are the formulas of the subscience, with each item of information having a stated form and location in the formula structures. The different views expressed in the articles, and the resolution of the controversy, were found to be represented by appropriately different formulas.

These formulas are thus shown to carry the information of the science. The structure ("grammar") of the formulas accords with the fact-structures of the science, i.e. with its objects and the relevant relations among them. The form indicates the content.

The purpose of the work presented here is to develop a formal tool for the analysis of science, and more generally of information. In respect to information, it has been found that a maximally unredundant description of the structure of language yields an approximate description of the information transmitted by language, and that this is all the more so within a science. In respect to the history of science, the formulaic representation of research done over a period shows, for example, changes in the way words for the objects of the science co-occur with words for the processes, changes which exhibit the actual development of the science. The analysis presented here for the statements of a science suggests new approaches for the resolution of some of the recognized problems in philosophy of science and in philosophy of language. And the results of analyzing a sample of articles show that one can discover for each science a specific grammar adequate for it. In particular, one can distinguish by purely formal procedures certain contributing linguistic systems: one for the results and theory of the given science, an ancillary one for the procedures of investigation, a meta-science system, and also material from prior sciences which is included in various statements of the given science. This isolating of various special and well-structured languages of science makes possible new investigation into the structure and information of science, with obvious relevance to Carnap's search for a language of science, but with more complex inter-sentence connectives than Carnap envisioned. This work also establishes the existence of sublanguages within natural language, and raises the question of what relations can be stated among sublanguages.

The initial application of these methods, in the present volume, has been carried out on a small field, the early years of immunology, and primarily on its central research problem as noted above. Enough work was done, beyond analyzing the articles excerpted in the Appendices, to make it clear that the sublanguage arrived at here would, with occasional additions, be adequate for other articles of the period. Hence the special grammar developed in this volume appears to be appropriate not merely for the articles listed but for the whole field at that period. This early period, in a field which was still small at the time, was selected in order to test the method in a reasonably simple case. However, nothing in this method would make it less applicable to larger and more complicated sciences; it will only require much more work, and the development of computer support.

The method applied here does not claim to yield a complete picture or an interpretation of the course of a science: that may require knowledge from outside its research articles, and even outside the science itself. The method also does not claim to yield a full analysis of the conclusions and theories of the science: that would require, in addition to the present analysis of the individual sentences, also the analysis of long sequences of sentences in the articles (argumentation and "proof"). However, any such further investigations require first of all that we establish the specific sentence structures, i.e. the formula types of the science, and it is this that the method as used so far has yielded.

Viewed step by step, the processing shown here seems informationless and unimportant. It is indeed informationless, in that it moves only from the sentences of the articles to paraphrases of those sentences. But it follows a systematic path through the maze of paraphrases (including stylistically cumbersome ones). It thus reaches, in an objective and non-semantic manner, a maximal similarity among the sentences, as is seen in the Appendix tables. The alignment of the similarities exhibits what is constant among the sentences.

This constancy has a meaning: it yields those categories of information in which the sentences are jointly dealing – the categories of the given research problem and of the subscience.

A guide to the chapters of the present volume:

The immediate results are presented in the tables of Appendices 1 and 2, in which one can see how the sentences of the articles are rearranged onto the formulas. Since the methods used in this analysis are novel, they ar presented here in great, perhaps unreadable, detail. However, an introduction to the work can be obtained from Chapter 1, sections *1* and *2*, Chapter 2, sections *5-7*, and Chapter 3. A summary of the findings is given at the beginning of Chapter 1. Chapter 2 shows how the sublanguage of the immunological material was obtained, and Chapter 3 contains a brief discussion of how such analyses open the way for characterizing the structure of sciences and their interrelations. In Chapter 4, ways of using the sublanguage formulas to clarify and specify various informational relations are presented. Chapter 5 gives details of the transformations used in obtaining the sublanguage formulas; an overview of the transformations is given in section *1*. Chapter 6 presents the special sublanguages of laboratory procedures and of measurement. A slightly different form of analysis is employed in Chapter 7 to obtain substantially the same informational units from papers written in French; the French material had not been included in the analysis presented in Chapters 1-6. Finally, Chapter 8 presents a historical sketch of the search for the cellular source of antibody, by two workers in the field.

The investigation was carried out with the support of a grant from the Division of Information Science and Technology in the National Science Foundation.

CHAPTER 1

REDUCING TEXTS TO FORMULAS

1. SEEKING CANONICAL FORMS

This book attempts to show that certain analyses of how words combine, when applied to reports in a science, suffice to transform the reports into a sequence of formulas which represent the information contained in the reports. The methods do not depend upon the investigator's judging or classifying the meanings of words or sentences, or upon any specialized knowledge of the science. The words are identified not by their meanings but by the combinations into which they enter in respect to other words, within each sentence of the science material. At least in part, the methods could be carried out in computer programs applied to the articles as published, without pre- or post-editing.

The major results of the pilot investigation reported here, carried out on research articles in a particular research area of immunology, are:

– The science subfield has a reasonably small set of word-classes, and not many individual words per class (disregarding synonyms); the latter constitute the vocabulary which is sufficient for the science.

– The word-classes are combined into a few sentence-types, which are the fact-structures of the science.

– Each fact-sentence in the science can be written as a formula. Each formula consists of particular members of the word-classes in a particular sentence-type, possibly with modifiers (in stated classes) and under conjunctions and meta-science operators.

– The formulas can be used to codify, locate and process the information in a subscience. They can also be used for making a critique of the discussion in scientific articles, and in some cases of the course of the research.

– Preliminary results suggest that discussion in the science is constructed largely out of selections of fact-sentences, possibly with particular modifications, under particular hierarchies of conjunctions, and of course under various meta-science operators.

– The possibility of specifying all the structures above shows that such combinatorial methods suffice for discovering the special grammar of a science, which in important respects represents the structure of the science itself: its objects and their relations. In so doing one can also specify the structural relation of the given science to its prior sciences, its subsciences, its success or sciences, and the like.

– The sequences of formulas in the articles in a given science can be looked upon as constituting discourses in a sublanguage of natural language, or alternatively as a new linguistic system structurally intermediate between natural language and mathematics.

The formulas thus obtained can be used to summarize the specific information in the given article, and any change in information. This can be a step in computer processing of the specific information in scientific reports. When these methods are applied to a number of articles in a subfield of science, the types of formulas characterize the information in the subfield – the entities with which the field deals and such relations among them as are studied in that field. In particular, the present investigation covers a problem in immunology and shows how the formulaic representations which are obtained for the various articles yield an organization of the successive stages of experimental results and conculsions as the problem developed. The formulaic representation makes possible an analysis and critique of the work and of the information in the science.

As to the terms used: a sublanguage is a proper subset of the sentences of a language, closed under certain grammatical operations of the whole language. That is, the result of these operations (e.g. transformations or conjunctions) operating on a sentence of the sublanguage, or on a pair of them, is again a sentence of the sublanguage. The sublanguage is characterized by particular word classes and sentence classes (word-class sequences) which are not necessarily classes of the language, and possibly by grammatical operations that are not distinguished as such in the language. A subscience for the purpose of the present discussion, is an aggregation of science reports characterizable by a sublanguage.

Both the methods and the results have had to be presented in some detail, but a general picture of this work and its conclusions can be obtained from Chapter 1, 1–2, Chapter 2, 5–7, and Chapter 3.

The methods used are mentioned immediately below, and discussed in Chapter 5. The specific procedures of analysis are introduced in *2*; details and problems of the analysis are presented in *3*.

The basis for obtaining the formulas of a science by grammatical transformations of its reports lies in the fact that the constraints on word-combinations in a language create the sentences of the language and at the same time determine the information carried by each sentence (given the meanings of the separate words). Different sequences of different words and word-classes yield, in a regular way, correspondingly different information. In addition, in the writings within a restricted subject matter, it is found that there are additional constraints on the combinations of words. In the present book it is proposed to show how these subject-matter-specific constraints can be used to exhibit the objects and relations which are involved in the information of that subject matter. To do this we first establish word-classes in such a way that combinations of the classes, i.e. of one or another word of one class with one or another of another class, recur frequently in the material. Then, in each sentence of the material, we seek insofar as possible to divide the sentence into segments such that each segment contains one of the recurring word-class combinations and also is grammatically a component sentence of the original sentence. Finally, in each segment having a given recurring combination we seek the maximal alignment of the word-classes: we permute the word-classes, to the extent that is permitted by known grammatical transformations, so that the order of word-classes in the combination is the same in as many of the segments as possible. These ordered word-classes, which are transforms of sentence or component sentences in the orginal material, are the formulas of the subscience. Since all the segmentations and transformations mentioned above are paraphrastic, that is, do not change the meaning of their operand, the formulas are paraphrases of the original material, and can be considered as simply a canonical, inspectable, and processible form of the orginal material.

The work described here was done on the basis of formally established transformations, which are presented in Z. Harris, *A Grammar of Engish on Mathematical Principles*, Wiley-Interscience, New York, 1982 (hereafter GEMP). However, once it is seen that recurring formulas can be obtained via such paraphrases, it becomes possible to carry out a resonable approximation to this work on the basis of common-sense paraphrases, so long as the paraphrases are based on general English or science-writing practice, and not on any specific issues which are under investigation in the given articles. This informal approach is possible because it is controlled by an internal check, namely whether or not it can lead to a small set of formulas covering the articles.

2. Analysis of word combinations

If one wishes to find, in the writings of a science, a canonical form for its information, one could analyze a set of articles dealing with one problem or area during one period, and try to show by interpretation that their structure mirrors the information they contain. However, a clearer and more definitive test can be achieved if one takes a succession of articles in which is traced the known development of a research problem or field. In such an historical overview it would be known by hindsight what new methods, new results, and changes of understanding appeared at what time and in what articles. If a formal analysis of the articles, specifically a reduction for formulas, made independently of this developmental knowledge shows changes in the formulas at those points at which the understandings are known to have changed, it would be clear that a relation exists between the change in formulas and the change in understandings or information.

An adequate research problem of this kind was found in the work on the cell responsible for antibody formation. There had been a controversy, largely between American and European scientists, as to whether the lymphocytes or the plasma cells of the lymphatic system were responsible; it was finally resolved by finding that both cells produced antibodies, and by recognizing that these names were being used for different stages of the same cell-type. The problem of which cell produces antibodies does not today exercise the scientists in the field, having been largely resolved. Nor does it loom large in the recent history of the field, since the issues that have moved into central importance have been concerned rather with the process of antibody production. However, in the period of approximtely 1940-1965, the "which cell" problem was important in the field, and while the evidence that both plasma cells and lymphocytes produced antibodies was obtained directly by experimental techniques (e.g. plaque production, cf. paper 11 in the Appendix), the recognition that they could be stages of a single cell-development sequence was a by-product of increasing data on cell-morphology and its development (e.g. in papers 4, 10, 11, 12). The main reason for selecting this problem for the present investigation was the fact that it had a clear beginning (between paper 1 and the first lymphocyte and plasma-cell papers) and end (as summarized in the Yoffey and Bussard extracts, in Appendix 1, paper 14 and Appendix 2), with a controversy and resolution pinpointed as to time, so that one could hope for a clear connection between the differences in sentential formulas as among papers, and the difference in information contained in those papers.

Fourteen papers in the areas were selected, on grounds given in Chapter 8 below. These are among the important articles in the field, from a 1935 paper which showed that antibody formation is located in the lymphatic system, through papers naming different cells as the producers of antibodies, and finally to electron-microscope papers that showed ongoing antibody production in plasma cells on the one hand and in lymphocytes on the other, and that discussed how the apparent conflict could be adjusted. In the course of this research the major new methods that came in at various time were the recognition of DNA-**RNA** involvement in production of proteins (in this case the gamma-globulin antibodies), the increased power of light microscopy, and finally the electron microscope.

We will see (in *1.3* of Chapter 3) that the formulas found for the 14 papers showed changes at the points at which there appeared the new methods and results, and that new kinds of discussion were necessitated by these. These changes in the formulas are of a kind that seems reasonable as a reflection of the known changes in information at these points.

The French papers considered in Chapter 7 were not part of the central investigation reported here. They were not selected for their relevance to the "which cell" problem, nor was their analysis used in judging the details and development of the formulas. Rather, they were selected as examples of research and review papers in a language other than English, to see if they exhibited the same gross formula structure as did the English texts.

2.1. Grammatical analysis

The selected articles were analyzed in the order of publication, beginning with the 1935 paper. In each article, the sentences were analyzed in the order in which they appear in the paper. This means in effect carrying out: first, a gross grammatical analysis of the sentence, determining for example the main verb of a sentence and its subject and object together with any modifiers of each of these; second, an undoing of any of the major transformations and zeroings (*3.3*) which have taken place in the sentence.

The gross grammatical analysis is determined by the classification of words as being arguments or operators, for example the class **N** (argument, mostly simple nouns, e.g. *cell*), $\mathbf{O_n}$ (operator on one **N**, e.g. *grow*), $\mathbf{O_o}$ (operator on one operator, e.g. *continue*) $\mathbf{O_{no}}$ (operator on **N** and **O**, e.g. *know*), etc., and the classification of affixes by how they change a word of one class into another, e.g., **O** → **O** (tense operating on a verb to yield a verb), **N** → **N** (e.g. plural), **O** → **N** (e.g. *-ment* in *development)*, etc. A sentence consists of one or more operator-words, each of which requires as argument words of particular classes: in *The cell's growth continued, continue* is operator on *grow* (as its argument), *grow* is operator on *cell*, and *-th* is the **O** → **N** indicating that *grow* has become the argument of a further operator.

The major transformations, including zeroings, are defined on the gross structural components of a sentence *A*, and produce a transformed or reduced sentence *B* differing in word sequence from the given sentence *A*, but paraphrastic to *A* in meaning. On finding a transformed sentence *B*., e.g., *John buys and sells old books*, we must first recognize that it is the product of a transformation; in this case, the trace (i.e. evidence) is the lack of an object (i.e. second argument) for the first operator (e.g verb), *buys*, and the lack of a subject (i.e. first argument) for second operator, *sells.* Then we undo the transformation; that is, we reconstruct the "source" sentence *A* as it was before the transformation had taken place, in this case *John buys old books and John sells old books*, where *buys* and *sells* are operators and *John*, *old books* are the arguments of each of these.

There are two reasons for recognizing the transformations which a sentence has undergone. One reason is the simplification of the structural analysis of sentences. By finding in a sentence the trace of a transformation, we characterize the sentence no longer as a sequence of words belonging to particular classes, but as a transform of a source sentence which in turn is a simpler sequence of words belonging to particular classes. For example, *B* above need no longer be described as a verb (*buys*) appearing without its objects joined by *and* to a verb (*sells*) appearing without its subject, but rather as a transform (made by the zeroing of repeated subjects and objects) from a source sentence *A* which is two occurrences of verbs, each with its subject and object, joined by *and.* If we considered each complex sentential structure, e.g. *B* above, separately, we might find various convenient ways of stating its word-class sequence; for example, we could say that *B* consists of two verbs joined by *and*, with a common subject and object. But if we consider all the partially similar (and paraphrastic) complex structures, e.g. also (*C*) *John buys old books and sells old books*, or (*D*) *John buys old books and he sells them*, we find that the least redundant description for all of them together consists in saying that repeated words can be zeroed, or pronouned, in stated situations, thus producing from a single source-sentence *A* many reduced sentences such as *B*, *C*, *D*. In the last analysis, the word-class sequence characterizing the source sentence, such as *A*, is the operator-argument relation (e.g subject-verb-object as above); all other sentences, such as *B*, *C*, *D*, contain this relation plus transformations.

The other reason for recognizing the transformations which a sentence has undergone is that the transformations are paraphrastic – more so or less so depending on how they are defined. The transformations alter, in a sentence, the position and form of words (even down to zero) without altering the information that the sentence carries, i.e. the informational relations among the words. It has been found that in the sentences of a discourse, and of a sublanguage, words (or word-classes) often repeat in a given grammatical relation in respect to other words. That is, each sublanguage has certain operator-argument structures of special word-classes – call it certain sentence-types – that repeat. We can try to maximize this by taking those sentences which contain a particular constellation of word-classes and seeking such transformation in each sentence as would put the words of the constellation in the same information (ultimately grammatical) relation to each other in each of the sentences.

The importance of these grammatical methods for an informational analysis of language material is two-fold.

First, these methods apply to all sentences. The major word-classes of the grammar are fixed for the language as a whole, so that the classification of a particular word in a sentence does not depend on the sentence in question. The structural analysis of a language applies to all its sentences. And the transformational reconstructions apply to all sentences which contain the traces of the transformation in question, the traces – and, in general, the domain – being specified a priori in the definition of the transformation. Thus these methods are not ad hoc to particular sentences, and cannot be adjusted to the particular interests of the investigator analyzing the articles.

Second, the analysis of each sentence is made of the basis of the relative positions of the classified words. No semantic criteria or subjective judgment is involved, so that the work can in principle be carried out by a computer program, even though considerable complexity is involved in so doing. (Grounds for all statements in *2.1* are given in GEMP.)

2.2. Sublanguage classes and sentence structures

After the successive sentences of an article have been analyzed grammatically, they are subjected to a further analysis, in respect to sublanguage word-combinations. This work consists chiefly of forming classes of words which have the same grammatical relation to particular other words, e.g. the words **A** which appear (in Appendix I) as subject of *found in the lymph nodes after injection of an antigen*. These words **A** include *antibodies, agglutinin*, etc. Starting with this, we may then form a class **V** of operators *is found in, is contained in, is produced by*, whose subjects are **A** (e.g. *antibodies, agglutinin*). In this way we find that certain word-classes recur in a particular grammatical relation to certain others, e.g. **A** as subject of **V** with object or "complement" **T** (*lymph nodes, lymph, serum*). This creates a sentence type (structure), **AVT**, which is obtained at the same time as we set up (extensionally) the word classes **A, V, T,** since these are defined as occurring in respect to each other in the operator-argument relation which constitutes a sentence structure. The grammatical transformations applied previously will have strengthened this result, since for example a sentence *Lymph nodes produce antibodies* will have been recognized as a transform of *Antibodies are produced by lymph nodes*, hence as a case of **AVT**.

This method of setting up word classes in respect to their grammatical combinations is in principle the same as that used in finding the grammar of a whole language. When applied to the material within certain subject-matter language-uses it produces not the general word classes of whole-language grammar (**N**, $\mathbf{O_n}$, $\mathbf{O_{no}}$, etc.) but specific subclasses of these such as the **A, V, T** here. The restrictions on what words combine with each other in the material of a sublanguage are so strong that the major subclasses can be discovered readily. Aside from local problems discussed below, only a few nouns occur in these texts as subject of *is found, is produced by*, etc., and only a few operators (verbs, adjectives) occur between *antibody, agglutinin*, etc. and *lymph node, plasma cell*, etc. In all cases the criteria for classification were purely combinatorial and not semantic. The subclass **A** is the class of subjects of *is produced by the plasma cell* and the like, and not the class of words semantically close to *antibody*. Indeed, it will be seen below that *A* includes some words, such as *protein, gamma globulin*, which would not readily have been included on semantic grounds; and the subclass **V** of *is produced* includes, for example, *is secreted*.

2.3. Sublanguage subclasses

It would have been too much to expect that a few word-subclasses in various grammatical combinations would have sufficed without difficulty for the sentences of these articles. Difficulties were indeed met, but it was possible to overcome many of them by defining further subclasses of the major word-subclasses. For example, there is a subclass **G** of words (*antigen*, *diptheria toxin*, etc.) which are subject of *is injected*. In a few places we find the word *dye* in the position of **G**, as in *when vital dyes are placed in superficial cuts they are drained by the lymphatics to the regional nodes* (paper 1, p. 800). Here *is placed in superficial cuts* is very similar to members of **J** (the subclass of *is injected*), while *is drained by the lymphatics to the regional nodes* is a member of the subclass sequence **UTT** (below). The subject of **J**, and of **UTT**, is **G**. However, *dye* is not found in other combinations into which G enters, in particular the main combination, **GJ:AVT** (e.g. *following injection of antigen antibodies are found in the lymph nodes*). We therefore put dye into a "no-antibody" subclass $\mathbf{G}_{\mathbf{a}\sim}$ of **G**, and we have to say that the sentence structures **GJ** and **GUTT** hold for $\mathbf{G}_{\mathbf{a}\sim}$ as well as for other **G**, but not when these sentence structures combine with a following **:AVT**. The advantage in this formulation is that words which occur in some but not all positions of a major subclass can be thus fitted in as a subclass of it. The cost is that the formulas for word-subclass combinations now hold not for all words in each subclass but only for (at least) some of those words. (For convenience, the major subclasses of a sublanguage will henceforth be called its classes.)

Whereas the classes listed in Chapter 2, 2, are inescapable, as being the most efficient for the sentences of these articles, some of the subclasses proposed are less well established, so that alternative subclassifications are not excluded.

It will be seen that while the word-class formulas (i.e. the sentence structures in terms of word-classes) change little as one goes from article to article in order of publication, the word-subclass formulas change appreciably. The former express the general types of information dealt with in these successive articles; the latter express the specific information presented in each article.

In the present investigation, the articles were first analyzed in terms of word-classes, with notations where some members of the word-classes were restricted in respect to members of the other classes with which they combined. A second pass through the articles was then undertaken, after we had some record of the kinds and amounts of such restrictions, and of the subclasses that would be required. In this second survey the formulas for the successive sentences were rewritten with subscripts indicating subclass.

2.4. The tables

In the Appendices the sentences of the articles are presented in a transformed shape, the transformations being those that would grammatically align the words of the sentence with words of the same class sequence in other sentences. Each transformed sentence has been obtained from the original sentence of the article by a priori established transformations (Chapter 5), or, in a few cases, by a special transformation which is discussed in the notes to that sentence (Appendix 3). In any case, each transformed sentence can be seen to be a paraphrase of the original, so that the meaning of the article has not been changed.

In addition, each sentence in the Appendices is represented by a formula, which is a sequence of class symbols, one for each successive segment in the transformed sentence. The formula thus merely maps the ordered words (with their modifiers) in the transformed sentence into the symbols for the class to which each of the words belongs. Where the word has been specified as belonging to a particular subclass of its class, the class symbol for the word is provided with a subscript indicating the subclass.

The major modifiers of a word (including those of quantity, time, negation) are indicated by superscripts on the class symbol of that word in the formula.

The tables published in the Appendices contain only a portion of each article. The portion was selected as follows: After the word-classes of an article were determined, all sentences which contained only one of these word-classes were dropped from further consideration, since they could not be used to show sentential relations among the classes. This applied to almost all sentences of the Materials and Procedures sections of these articles, and to a very few sentences in the other sections of the articles. The remaining sentences were analyzed in detail, down to their formulaic representation (except for the internal structure of the meta-science segments). Since the full publication of the fourteen articles and the three French papers with their sentence-analyses, would have imposed too great a burden on the present book, a selection from each article had to be chosen. What was selected was those sentences which were essential to understanding the experiment reported in the article and the conclusions drawn therefrom by the article. The selection was made purely on the basis of the content of the original sentences, without regard to their formulaic representations. Because the sentences that were favored were those dealing with the main subject of each article, it was found that the main formulas of the article were relatively more frequent among these sentences than they were in the article as a whole. Other than that, the tables printed in the Appendix are similar to the tables obtained for the articles as a whole.

2.5. Validity of the procedures

The procedures used in moving from the original texts to the sequence of formulas differ from those used in many sciences, and their validity does not rest on statistical control as it does in much experimental work. Instead, the following considerations are relevant: The structural analysis (*2.1*) which ultimately describes a sentence as a partial ordering of particular operators and arguments is applicable to every sentence of the language. The reductions or other changes which describe a particular sentence as a transform of a particular other one are applicable to all sentences which contain the trace of that transformation. The paraphrastic effect of each reduction is a matter of interpretation, which can be verified once and for all in the case of each reduction separately. These properties thus hold for all applicable sentences, and therefore for the particular sentence in question.

There remains a difficulty, which in most cases can be overcome. Quite a few transformations are degenerate. That is, a given sentence may possibly be a transform of either of two different source sentences. Source *A* with reduction R_1, and source *B* with reduction R_2, may both yield the same word-sequence *C*. Given *C*, one can then reconstruct either *A* or *B* as source. When *C* occurs in a discourse, it becomes important to know which source was intended by the speaker or writer, for *A* and *B* may well differ in their informational effect in the discourse. If *A* would have greater word-repetitional similarities to the sentences around *C* than *B* would have, it may be presumed that *A* is the source for *C*.

Nevertheless, there may be some sentences in a text for which it is difficult to decide the "correct" source, i.e. the one meant by the author. In addition, there may be sentences which can be fitted into the neighboring word-repetitions only at the cost of ad hoc transformations which the reader may not accept. In all such cases the sentence can remain unanalyzed, or can be represented by formulas that do not fit in with the neighboring sentential formulas. It is important to understand, first, that this does not detract from the formulaic representations of the other sentences, and second that it does not destroy the formula-repetitions seen in the other sentences. The complete and partial similarities among successive sentence-formulas are so great that the intrusion of unanalyzed sentences, or of sentences largely unrelated to these formulas, does not vitiate the overall result, e.g. that presented in Chapter 2. Furthermore, the result that has been obtained – so far, at least – is not a tight sequencing of formulas, based on some criteria for succession, such as could be affected by the intrusion of unrelated material. Rather, this result is simply the existence of formula-repetition and of partial similarities among neighboring formulas.

3. Details of the Analysis

3.1. Word combination within segments

The analysis begins with the successive sentences as they appear in each article. For each sentence we have to know the operator-argument relations among its words. If the work is to be done by hand, as in the present investigation, it suffices if we recognize these grammatical relations by virtue of knowing the language. In knowing the operator-argument relations in a sentence, we would know what is the subject and object of each verb or adjective, and what are the ordered secondary sentences (modifiers) on each word. This is tantamount to knowing all the source grammatical relations among the words of the sentence.

We then investigate, over the whole corpus of articles, how words cluster in respect to each other within these relations, e.g. which verbs have the same nouns as subject, or which adjectives or relative clauses modify the same nouns. No useful result will be obtained if we ask this question without the framework of specified grammatical relations, for example if we ask what words occur in the same sentence as *antibody* or what words occur next to it. But if we ask what words occur as operators with *antibody* as their first argument, we find (a) *is in, is found in, is contained in, appears in, is produced by, is formed in*, etc. And if we check the relation among these operators, we find that some papers have *is found in, appears in*, but also *is not produced by* or *is not formed in* as operators on the argument-pair *antibody, lymphocytes*; but no paper has both *appears in* and *is not found in* (or *is found in* and *is not found in*) on the same argument-pair (unless one operator reports the author's work and the other reports someone else's work). We then say that all of (a) above are in the class of operators on the pair *antibody, lymphocytes*, but that the last two members (*is produced by, is formed in*) are in a different subclass from the first four, because only the last two of these can be (for the same arguments) under the further operator *not* in an article-section in which the first four are not under *not*. Such classification and subclassification has to be done for each word in respect to all the words which are arguments under it or operators over it, or co-arguments (i.e. joint arguments under the same operator); more rarely we may find that one word differs from another only in respect to words more distant in the partial order of operators in the given sentence.

We can now write each sentence not as a sequence of words but as a sequence of word-class symbols, with subscripts to indicate in which subclass the word belongs, and with superscripts to indicate the modifiers of the words. Any classification made at this point may be adjusted later, as further co-occurrence dependencies or regularities come to light; but at this stage we already have a good approximation to the final result.

Next, we check each successive sentence of each article, or a segment of the sentence (for example, up to a conjunction), or rarely a sequence of sentences, seeking repeating sequences ("formulas") of word-class symbols. Thus in paper 1, p. 783.1.7, (hereafter 'p.' will be omitted) we have *pathogenic bacteria carried on the lymph stream* and in 783.2.1 we have *antigens arriving by the lymph stream*, both of which we can represent as $\mathbf{GU^{y}T_{\ell}}$, where **G** is the class of antigen terms, $\mathbf{T_{\ell}}$ is used for *lymph* as a subclass of tissue words (**T**), **U** is for verbs whose subject is **G** with a second-argument **T** (or **C** for cell words), and superscripts **f, t, y** are respectively the prepositions *from, to, by* (with *on, along* as variants of *by*) introducing this second argument. When in the same paper we find (801.2.1) *the rapid lymphatic distribution of antigen* we have **G** and $\mathbf{T_{\ell'}}$, and we accept *distribution* as a **U** operator on the grounds given above; the reconstructed sentence might be *antigen is distributed rapidly lymphtically* (or: *along the lymphatic system*), represented by $\mathbf{GU^{iy}T_{\ell'}}$ (with superscript **i** for *rapid*).

In some cases, the decision as to the formulaic representation is more complicated. In paper 1, 796.4.3, we find *The nodes were equally inflamed.* Since the two preceding sentences distinguish *the lymph nodes from the side injected with that antigen* from *the nodes from the other side*, the referentially reconstructed *the* in 796.4.3 refers to these two, so that 796.4.3 can be referentially reconstructed to *The lymph nodes from the side injected with that antigen and the nodes from the other side were equally inflamed.* We have two ways of fitting this sentence into the elsewhere-established formulas. One is to use the reciprocal (reflexive) status of *equal* to transform 796.4.3 into *The nodes from the side injected with that antigen were inflamed equally with* *the nodes from the other side* (GEMP 6.71); this could be represented as $\mathbf{T_nYT_n}$, where **Y** is the set of operators whose two arguments are necessarily of the same class, and $\mathbf{T_n}$ is *lymph nodes* (the references to *side* being superscript **B**, discussed below). The other possible formulaic representation is in two sentences connected by a conjunction: *The nodes from the sides injected with that antigen were inflamed, equally as* (or: *to an equal extent as*) *the nodes from the other side were inflamed.* This is represented as two $\mathbf{T_nW_f}$ sentences conjoined by *equally as* between them; the $\mathbf{W_f}$ is a subclass of **W**, which is the set of operators whose first (and usually only) argument is **T** or **C**. Since *equally inflamed* is quite different from the established members of **Y**, and since **Y** occurs virtually only with the pair **C**, **C** as its arguments (and never otherwise with **T**, **T**), it is better to use the two-sentence analysis here.

3.2. Obtaining repeating types of sentences

The goal of finding the greatest amount of regularities of word-combination in this material makes us seek, in each sentence of the articles, the largest repeating sequence of word classes, even though in some cases the longest formula has to be rejected, as in the case of $\mathbf{T_nYT_n}$ above. The sentences as printed in the articles repeat only rarely. Even if we represent the words by the symbols for word classes and subclasses, we do not get many repeating sequences. However, we can segment the sentences of the articles in such a way that many segments are class-symbol sequences which recur, as in the **GUT** sequences above. The recurrence is greatly enhanced if we permit each symbol to carry different superscripts, e.g. if the $\mathbf{GU^{iy}T_{\ell'}}$ above is considered a repetition of the $\mathbf{GU^{y}T_{\ell'}}$. Thus, we look for repetitions of particular operator classes or subclasses, with their arguments, allowing each of these words to carry various ordered modifiers. The modifiers are, grammatically, secondary-sentence operators on the given word. For example, many occurrences of *cell*, especially of *lymphocyte*, carry the modifier *large*. In many of these occurrences the *large* is clearly used to indicate a type of the cell and can even be taken as part of the cell name: $\mathbf{C_y^g}$ for what is called large lymphocytes as against lymphocytes in general. In a few cases the *large* is rather just a property of the cell, and could be given in a relative clause of *cell*: paper 7,11.3.3 has *Both the large cells and the smaller ones, the lymphocytes, do contain antibody*, where *the large cells contain antibody* could be derived from *the cells which are large contain antibody* from *the cells contain antibody; the cells are large.* A modifier (*large*) is simply the operator in a secondary sentence, as here (after the semicolon above), with the host of the modifier (*cell*, here) being repeated as subject in the secondary sentence. The modifier may also appear in the primary (non-secondary) sentence, as in *A few of these...lymphocytes... were large* (paper 13, 453.1.2). When *large* is taken as a modifier it is written as a superscript, as in $\mathbf{C}_{\mathbf{y}}^{\mathbf{g}}$ (*large lymphocyte*); when it is taken as an operator it is written as $\mathbf{W}_{\mathbf{g}}$, a subclass of the word-class **W**, as in $\mathbf{CW}_{\mathbf{g}}$ (*the cell is large*).

In addition to the modifiers, there are certain other operators which are written as superscripts rather than in a separate formula. These are the local ("aspectual") operators on a verb, i.e. operators whose subject is the same as the one of the arguments of the verb on which it is operating (or is a classifier of that argument): for example **b** for *begin, start* as in *Antibody production starts at this stage*, derivable from *Antibody starts being produced...*, from *Antibody starts its being produced...* (paper 13, 470.3.3). We represent this sentence by $\mathbf{AV}_{\mathbf{p}}^{\mathbf{b}}$, with **A** for *antibody*, **V** for *produce, form*, and **b** as supercript on $\mathbf{V}_{\mathbf{p}}$ rather than as a verb in a new sentence. There is a particular set of operators which functions in these articles much as the aspectual operators do in English: this includes *have a role in, participate in*, etc., as in *The lymphocytes constitute a factor in antibody production* (paper 3, 122.1.1). These operators appear on $\mathbf{V}_{\mathbf{p}}$ (*produce, form, synthesize*), and if we compare the sentences containing $\mathbf{V}_{\mathbf{p}}$ under these operators with the sentences containing $\mathbf{V}_{\mathbf{p}}$ alone, we see that the arguments are the same in the two cases, but that there is a difference in respect to the metalinguistic segment over the $\mathbf{V}_{\mathbf{p}}$ (*2.2* in Chapter 2) and in respect to neighboring operators: we may find such two-sentence sequences as

$$\mathbf{AV}_{\mathbf{p}}^{\mathbf{r}}\mathbf{C}_{\mathbf{y}} \textit{ but } \mathbf{AV}_{\mathbf{p}}^{\sim}\mathbf{C}_{\mathbf{y}}$$

(*Lymphocytes have a role in the production of antibody but lymphocytes do not produce antibody*). Therefore, rather than treat these words as an independent operator-class, we treat them as local operators on $\mathbf{V}_{\mathbf{p}}$ and write then with a superscript **r**.

Another situation in which additional material can be included in a single repeating sentence-type is that of the repeating sentence-pair. For example, we find very many occurrences of sentences such as *Following injection of antigen, antibody was found in the lymph nodes.* The two component sentences are occasionally found one without the other, and in fact there is a certain background presence of antibody in the lymphatic system and blood which is not in response to infection or injection of antigen. However, in these articles the great bulk of occurrences of the two component sentences are paired, connected by *following, thereafter*, etc., as above (although the *injection* component may be zeroed). Therefore, rather then consider each component a separate sentence, with a conjunction (*after*, etc.) between them, we write a single double-sentence formula **GJB:AVT** (e.g. *Antigen injection into the footpad is followed by antibody appearance in the lymph node*), while those cases where the components appear separately are written as **GJB** alone or as **AVT** alone.

The conjunctions between sentences (more precisely, between rows in Appendix 1) have much less repetitive regularity in respect to the sentences which are their arguments than do the various word-classes inside a sentence-type formula. In saying this, we take the colon which represents *following, after*, etc., as an intra-formula word-class. Therefore, to the extent that we are able to represent a sentence or sentence-segment of an article by a single sentence-formula rather than by a sequence of smaller sentence-formulas connected by period, *wh-* (which introduces a relative clause, i.e. a secondary sentence), or conjunctions, we obtain a better record of the co-occurrence regularities of the words in the articles. In addition to this, there are related informational reasons for maximizing, in particular ways, what is to be included in a sentence formula. The main objective is to get maximal information into the confines of a single formula because within a formula the information is explicitly organized by its sublanguage "grammar." A related objective is not to leave out of a formula (together with any conjunction on it) anything which would seriously alter the information in the rest of the formula.

By the main objective, we would favor keeping modifiers in a sentence segment X as superscripts on a word in the formula of X rather than reconstructing them grammatically into a relative clause, i.e. into a secondary sentence connected to X by the *wh-* conjunction. For example, in *Suspensions of various killed organisms were employed* (paper 1,794.5.2), we would write the whole sequence before *were employed* as **G** (with *suspensions* and *killed*, which are not in special word-classes of the sublanguage, as modifiers), rather then transform the sentence first into *Various organisms which were killed, which were in suspensions, were employed* and then into three conjoined sentences *Various organisms were employed; the organisms were killed; the organisms were in supensions*. (In the last form, *in suspension* and *killed* are the operators in their respective formulas, and would have to be put into sublanguage classes.)

However, if the modifier contains members of the formulaic word-classes, it is reconstructed into a separate conjoined sentence. For example, in *They (the large cells) synthesize antibody specific for the antigen which stimulated their development* (paper 9,66.4.3), we treat *wh-* as the conjunction which introduces secondary sentences (the semicolon in the above examples); hence we transform the sentence into *They (the large cells) synthesize antibody specific for the antigen; the antigen stimulated their development*, which would be represented by the two formulas $\mathbf{G^{w}J{:}A^{G}V_{p}C^{g}}$ and $\mathbf{G{:}C^{g}W_{p}}$ connected by *wh-*. (Here the $\mathbf{w}$ superscript indicates which word is carrying the following *wh-* sentence as modifier; the $\mathbf{G^{w}J{:}}$ – read *after injection of antigen* – is reconstructed from the *specific for the antigen*, which is a modifier (superscript $\mathbf{G}$) on $\mathbf{A}$ and refers to the injected antigen; $\mathbf{V_{p}}$ is *synthesize*; $\mathbf{W_{p}}$ is *develop*; $\mathbf{C^{g}}$ is *large cells*; the colon, which usually represents *thereafter* or the like, but also various causal verbs, here represents *stimulate*. Some words may appear in one sentence as a modifier written as a superscript, and in another as a full operator. For example, *mature* (written as superscript $\mathbf{m}$) appears frequently as modifier on *cells*, especially on plasma cells; but we also find *when the plasma cells reached maturity* and *when the cells were fully mature*, which are $\mathbf{C_{z}W^{b}_{m}}$ and $\mathbf{C_{z}W^{+}_{m}}$ (paper 6,154.3.3,4).

A second situation in which we can maximize the information carried in a single formula is seen in the case of the comparative. Grammatical analysis decomposes comparative sentences into two sentences neither of which contain a comparative, plus a sentence which contains a comparative operator (*is more than*) and which can be reduced to the comparative *-er*: *I am taller than John* from (a) *I am tall to a degree which is more than the degree to which John is tall.* However, it is possible to get the comparative word or suffix into the first component sentence by making an artificial, ad hoc, transformation form (a) to *I am taller* and *John is tall* with *than* conjoining them. This has been done in many of the comparatives in Appendix 1.

Just as we try to maximize the information carried by a single formula, we try to minimize the occurrence of formulas which carry almost no information. For example, in paper 3,128.3.1, we have *Lymphocytes act an antibody producers*. Here *as* could be taken as a conjunction between the two sentences: *Lymphocytes act*, and *Lymphocytes produce antibody*. However, *lymphocytes act* carries virtually no information, even though *act* operating on *produce* affects the meaning. Hence we treat *act*, written as superscript $\mathbf{r}$, as a local-operator modifier of *produce* in a single formula $\mathbf{AV^{r}_{p}C_{y}}$.

We turn now to the subsidiary objective of not leaving out of a formula anything which would alter its information. A major example of this is the restrictive relative clause: the case when a sentence is said about a given argument only when that argument is under a particular modifier. For example, given (a) *This raises the question whether the "primary response" exists as such on a cellular level* (paper 9,67.3.5), where we know from other material that *primary response* is the same as *response to primary injection*, we obtain a formula **GJ^1:AVC**, representing (b) *To | a primary injection, | response exists as such | on a cellular level*, where the superscript **1** represents *primary* as modifier on *injection*. If this modifier were taken out, as being the operator in a secondary sentence (a relative clause), we would have two sentences: (c) *To an injection, response exists as such on a cellular level* (**GJ:AVC**) plus a conjoined (d) *The injection is primary*. But (c), without (d), is certainly not being posed as the question of (a) or of (b). True, we could say that in any sentence we don't know what is being said or asked until we add any conjoined sentences which may be present. But it is preferable if, in the course of connecting the formulas through their conjunctions, we do not present in one formula wrong or unintended information which has to be corrected in later formulas. It would be better if the formulas were such as to be only additive informationally in respect to preceding ones.

To achieve an informationally additive (and not correction-requiring) character for the formulas is not always easy. To do this, we would have to tie to each formula the degree of assertedness stated about the sentence – e.g. whether it is being asserted, or said to be possible, or questioned, or negated, etc. In the present set of tables we have usually done this in the case of negation, where the tilde ~ appears after the operator (or elsewhere in the formula, if needed). But many indications of assertedness are stated in the meta-science (**M**) portion of a sentence (as in the word *question* above), and the mechanisms for seperating these indicators from the **M** have not yet been fully worked out. This does not detract from the formulas as records of what kinds of information are presented in these articles; but the assertion-markers of a formula will have to be derived from the **M** and from the relation to neighboring formulas, if we are to use the present formulas as a record of the specific information given in the articles.

One other consideration should be mentioned as to how much should be represented by a single formula. When a word in one formula refers to a word in another, the apparatus to indicate the reference is complex and is not indicated in the present set of tables. (The development of such an apparatus depends upon further investigation into the textual distance, and other sublanguage restrictions, between the referent and its antecedent.) However, if a word in a formula refers to another in the same formula, it is easy to indicate this because the words which are part of a formula have a priori fixed positions. The chief example of this situation in the present material appears in such terms as *regional lymph nodes* and *lymph nodes on the injected side*, which refer to the region or side of the injection. When the injection sentence GJB and the response sentence $\mathbf{AVT_n}$ are included in the same formula, we place a superscript **B** on $\mathbf{T_n}$ to refer to the **B** (body-part, body) of the injection. This gives an additional reason for including these two sentences within one formula $\mathbf{GJB{:}AVT_n^B}$.

3.3. How much transformation?

Grammatical transformations in the set of sentences are mappings from one subset of sentences onto another, which preserve the original operator-argument relations (even if in derived manner), and hence the meanings, within each sentence. In order to achieve repeating sentence-types, written as formulas, it is necessary not only to segment many sentences into two or more, but also to transform certain sentences in such a way that their words will appear in the position that their classes occupy in the formula. For example, by the side of *antibody is found in lymphocytes*, $\mathbf{AV_iC_y}$, we would transform (a) *lymphocytes contain antibody* to *antibody is contained in lymphocytes*, again $\mathbf{AV_iC_y}$. In some cases, the work of transforming can be replaced by simply writing the word-class sequence of the sentence, or part of it, backward (indicated by an arrow): thus (a) could be written directly as *antibody* | *contain* | *lymphocytes*←, which has the order of $\mathbf{AV_iC_y}$. An example of this was seen in the $\mathbf{GJ^1{:}AVC}$ above (cf. also the use of arrows in Chapter 4).

Most transformations consist of reductions of those words which contribute little or no information to the sentence in which they occur. The most frequent are the reductions, to pronouns or to zero, of repeated words which have occurred elsewhere in the sentence or in preceding sentences. In the present work such pronouns and zeros have been replaced in some sentences by the antecedents which they are repeating. In other sentences the pronouns or zeroes (zero being absence of the expected word) are left standing, in order not to burden the tables with too much reconstruction. It is hoped that further work will establish more precise criteria for replacing a zero by the word which was zeroed.

In the tables, the transformed sentences which accompany each formula have been so presented as to differ as little as possible from the original sentence, just enough to make the rows in the table conform to one or another of the formulas. The result of minimizing the use of transformations is that we are left with a larger number of formulas. For example, there are rows such as *Antibodies are found in large number* represented by $\mathbf{AV_i^+}$, and other rows such as *Antibodies are found in large number in plasma cells*, or *Plasma cells contain many antibodies* represented by $\mathbf{AV_i^+C_z}$. These two can then be considered as variants within a family of partially-similar sentence-types. We could also have defined a transformation-like reconstruction among the variants, which would fill out the $\mathbf{AV_i^+}$ type to $\mathbf{AV_i^+C}$ (or to $\mathbf{AV_i^+C_y}$ or $\mathbf{AV_i^+C_z}$ according to what the neighboring rows show to be the antecedent of the zero after $\mathbf{V_i^+}$). It is easier to justify reconstructing this **C**, or to determine whether it is $\mathbf{C_y}$ or $\mathbf{C_z}$ in a given occurrence, when we compare a formula with neighboring formulas, than when we are calculating the grammatical structure of a sentence. Hence for many situations of zeroing, it is best to leave the reconstruction of what has been zeroed until after the given sentence has been represented by a particular formula, and after comparison with neighboring formulas is possible.

In some cases it is easy to see that several sentence forms are variants of one another. For example, we find *They (the cells) had basophilic cytoplasm* (paper 7, 3.5.5), which is represented by

$$\mathbf{C}\ \textit{have}\ \ \mathbf{S_cW_s}$$

(abbreviated to $\mathbf{CS_cW_s}$, where $\mathbf{S_c}$ is *cytoplasm*); but also *The slightly large nucleus of these cells showed a loosening of the central chromatin* (paper 13, 454.1.3), which is represented by

$$\mathbf{S_n^g}\ \textit{of}\ \mathbf{CW}$$

(abbreviated to $\mathbf{S_n^gCW}$, where $\mathbf{S_r}$ is *nucleus*). It is easy to consider **C** (*has*) **SW** and **S** (*of*) **CW** as variants of a single formula, the more so as a transformation betweeen *N_1 is of N_2* and *N_2 has N_1* is known in English (the subscript numerals here identify the nouns, written *N*). In other cases the relation between two partially similar sentence types is unclear, as for the many sentences written $\mathbf{GJ{:}AV_i}$, e.g. (a) *Antibody appears after injection of antigen* as against (b) *The antibody is specific to the antigen*, which is written **G:A**. Such a (b) appears usually together with an (a), as in paper 12,109.2.3. This combination of (a) and (b) can also be written $\mathbf{GJ{:}A^GV_i}$, as in paper 5,205.1.1, where the superscript **G** represents *homologous* (rather than *to the antigen*). If (b) can occur independently of (a), its operator (*specific to*, or the like) would have to be a new verb-like word-class (not conjunction-like, as is the colon), even though the sublanguage meaning of that operator is related to that of the colon conjunction.

The main transformations which have been used in segmenting and aligning the original sentences into the repeated word-class sequences represented by the formulas are listed below. A fuller discussion is given in Chapter 5.

(1) Zeroed arguments and secondary sentences: Since each operator requires stated word-classes as its arguments, the appearance of an operator without its argument permits us to reconstruct that argument, as in our occasional inserting of *of antigen* after *injection*. In many cases the papers report antibody appearance and other cellular responses without saying the implicit *after antigen was injected*. This last can be inserted in the rows of the table (i.e. in the transforms of the original sentences) and in the formulas, although such insertions have been made only when some word in the row referred to some part of the absent **GJB:** segment.

(2) Pronouns: As noted, pronouns and zeros (word-absences) have in some cases, but not always, been replaced by the antecedent word whose repetition they indicate.

(3) Nominalizations: When a sentence (or its operator) occurs as the argument of a further operator, it is in many situations "nominalized," i.e. it carries a "noun-like" suffix showing that it is being used as an argument; and when the verb is nominalized, its adverbs become adjectives. If we return the sentence to its free-standing form, these adjectives are returned to adverb form. Thus, *after peritoneal injection of antigen* is reconstructed to *after antigen is injected peritoneally.*

(4) Passive: If both a sentence and its passive, or other permuted form, occur in the text, one can choose either the active or the passive order of symbols for the formula, e.g. $\mathbf{AV_pC}$ for both (a) *Antibodies are produced by the cell* and (b) *The cell produces antibodies.* To fit (b) into $\mathbf{AV_pC}$ is tantamount to transforming it into the passive (a). In certain cases the passive presents ambiguities which can be resolved by appeal to the known argument-classes of the given operator. For example, we have (a) *Mice were injected intradermally in the right ear with 0.03 cc. of the paratyphoid bacterin* and *after intradermal injection of antigens* (ibid, 4.1, nominalized from *after antigens were injected intradermally*), and (b) *0.03 cc. of the paratyphoid bacterin was injected intradermally in the right ears of mice.* In conformity with many **GJB** (*Antigen was injected into animals*) sentences, we transform (a) into (b) (via *We intradermally injected the right ears of mice with 0.03 cc. of paratyphoid bacterin*), and represent it by the **GJB** formula.

(5) Secondary sentences: Single text-sentences which contain residues of conjunctional material can be expanded into two sentences, e.g. by filling

out the secondary sentence of a comparative from the primary, or by reconstructing a modifier (adjective, relative clause, etc.) into a secondary sentence. Some of the considerations as to when to do this have been mentioned above.

(6) Shifting modifiers: Modifiers of operators or of sentences can be moved from their position in a sentence to certain other positions, in a way that aligns the word-classes of their sentence with those of other sentences; e.g. in paper 1, 783.1.1 and 792.4.1. Less generally, even certain modifiers of an argument (a noun) can be shifted into the status of modifiers on the operator on that argument. For example, we can transform *The cytoplasm in active cells is basophilic* to *In active cells, the cytoplasm is basophilic*, and vice versa. These possibilities of transformation can be used to locate in similar position all modifiers which are similar in informational character. For example, most modifiers referring to time (*immediately, on the 6th day*, etc.) occur on the colon which represents *after*, etc.; we can then transform others, such as *early*, from the word on which they occur to the colon in their row. With greater difficulty we may be able to move quantifiers (e.g. *first*) from nouns (e.g. *antigen*) to the verbs which operate on those nouns: e.g. (a) *the first antigen was deposited* derivable from *the antigen which was first deposited*, from *the antigen which was first injected was deposited* (by "appropriate" zeroing of *injected*, cf. Chapter 5), which is represented by

$$\mathbf{G^{w}U|||wh|||GJ^{1}};$$

this analysis is supported by the fact that (a) is attached to a subordinate sentence *if a second injection is given a month after the first*, which involves $\mathbf{J^2}$. In particular *no* on nouns can be moved to *none* on verbs, as in *No antibody was found* transformed to *of antibody, none was found* (GEMP 7.13).

(7) Conjunction: More problems are met with in the transformations that enable us to include in the colon all the sentence material which we want to include there. For example, consider the hidden *wh-* conjunction in *The nodes on the side injected with paratyphoid bacterin became slightly larger* (paper 1, 792.1.2, derivable from ...*on the side which was injected*...), which we transform into *The nodes became slightly larger on a side; paratyphoid bacterin had been injected on that side*, represented in inverse order by $\mathbf{GJB{:}T^{B}_{n}W_{g}}$. The *which* is decomposable into the *wh-* conjunction (written as semicolon) and the pronoun *-ich* (here replacing *that side*); the *wh-* occupies here the position of *and then, causing*, etc., as though we had *paratyphoid bacterin was injected in a side and then the nodes on that side became slightly larger*. Although the details of the transformation have to be

specified, the motivation for including this occurrence of the *wh-* conjunction in the colon conjunction is that this occurrence joins **GJB** to **TW**, and whatever does that has the status of the colon conjunction – indeed, joing **GJB** to **TW** (or **CW**) or **AVC** is the definition of the colon in this sublanguage.

(8) Special-domain transformations: There are a number of transformations involving particular subsets of operators, which have been used in the tables. One is between N_1 *has* N_2 and N_2 *is of* N_1 (above). Another expands sentences with reciprocal verbs into two sentences, as in deriving N_1 *and* N_2V from N_1VN_2 *and* N_2VN_1 (with V for verb, GEMP 6.71: e.g. *X and Y met* from *X met Y and Y met X*). Yet another decomposes certain transitive verbs into *cause* operating on the corresponding intransitive verb (N_1VN_2 into N_1 *cause that* N_2V, GEMP 6.8). We use this, for example, when we find *agglutinin-forming antigen* (paper 1, 792.1.1), which seems to come from *Antigen forms agglutinin*; but we would like to avoid a formula $\mathbf{GV_pA}$ which does not otherwise occur. We then transform *Antigen forms agglutinin* to *Antigen causes agglutinin to form*, which is a case of $\mathbf{G{:}AV_p}$ and is close to the existing $\mathbf{GJ{:}AV_p}$.

3.4. Summary of procedures of analysis

The word-classes of articles listed in Appendix I were established by observing how the words combined with each other within the framework of operator-argument grammatical relations. Sentence-type formulas of these word classes were found by seeking repeating sentence-making sequences of the word classes, aided by paraphrastic transformations which aligned certain word-class sequences with others. Once the formulas are obtained, some of them could be transformed into others, by tranformations which are more readily justified when we know what word-class combinations are common in this corpus than when we are simply recognizing the structure of an English sentence.

To a first approximation, this work can be done with very little grammatical specialization. It would be enough to state explicity what words in a sentence are the subjects and objects of what verbs (or of predicate adjectives or of predicate nouns), and what words in it are the modifiers (GEMP 5.3, 6.6) and local operators (GEMP 6.5) on what words. Within these relations one could seek the repeating word-combinations that would justify setting up word-classes, and the repeating word-class sequences that would justify setting up sentence-type formulas. The test of the analysis would lie in finding a small number of formulas that repeat many times over. The reason that one can obtain good results even with a rather rough grammatical formulation is that the repetition of just a few formulas is so great that they are bound to be discovered even if some sentences are misanalyzed or left unanalyzed.

The precise grammatical analysis is needed if we wish to avoid having many variegated formulas in addition to the few repeating ones, and if we wish to see in detail what are the patterns of recurrence of formulas and how they make up the whole article and the whole area of research.

It should be mentioned as an aside that precise grammatical analysis is sometimes not possible because the sentences of the text are not in all cases perfectly grammatical. Slips of grammar enter into some long sentences, and the analysis then has to be made on the evident intent of the writer rather than on the actual form of the text. (an example is *its* for *their* in paper 1, 789.4.1).

3.5. Output

The output of the analysis of an article is a sequence of formulas. Each formula is readable as a sentence (in a language whose words are class symbols); it is a sequence of word-class symbols, with subscripts to indicate subclasses and ordered superscripts to indicate modifiers or local operators. Each formula represents all of the specific words (other than meta-science) in a text sentence, or in a segment (or sequence) thereof, and is a paraphrastic transform of that piece of the text. The sequence of formulas together with the conjunctions and meta-science segments on them, cover the sequence of text sentences in the article.

In the work done so far, and in the tables of Appendix I, certain kinds of meanings are not specified in the formulas: e.g. The specific time and quantity modifiers, such informationally complex words as *ratio*, the distinctions among semantically different negative words (e.g. in *deplete*, inadequately written $\mathbf{W}_{\mathbf{i}}^{\sim}$ and *restore*, inadequately written $\mathbf{AV}^{\sim\sim}$ (as in *The antibody response can be restored*, paper 10, 303.1.1 and 2.2). This would have to be amended in further work.

Given **M**, we find that the subjects of **M** verbs are a particular set of nouns, **N′**, which includes *workers, students, investigators* (these being derived from **M** words), *we*, and capitalized words not usually listed in dictionaries: the names of scientists. Given **N′**, we then find that its members appear also as subjects of another set of verbs, **M′**, whose second argument (object) is a noun of the science-language rather than a science sentence. **M′** includes *use, examine, obtain, extract, excise, separate... from*: e.g. *We excised small pieces of red pulp*. Here we should include *use this technique* (or *method*), and the like. Problematic members of **M′** have **N′** as subject but usually no object, as in *work* (*on*), *experiment* (*on*). The words *table, Fig., article, paper* may be assigned to **M′**, if we reconstruct their occurrences as being from **N′** *made a table* (*of antibody titers*, or the like), and **N′** *wrote a paper about*.... There are also whole sentences which may contain **N′**, **M′**, or science-language nouns, but not science-language sentences, e.g. *The sampling problem for electron microscopy becomes very great* (in paper 12, 113.5.5). All these segments have been marked **M** in the tables, although the term "meta-science" may not be precisely appropriate for them.

To return to the operators on science-language sentences: There are many verbs, adjectives, and nouns which have the grammatical status of operators whose first and only argument is a science-language sentence. Such verbs are: *emerge, result, appear, may be* (as in *It may be that*...). Such adjectives (with *is*): *possible, probable, likely, significant, clear, evident, logical, true*. Such nouns (with *is*): *fact, thesis, theory, problem, case, not the case, data, evidence, factor, difficulty, development, subject of confusions, point at issue, matter of semantics*. All of these may be assigned to a new class **M″**. Some of them may be thought to be part of the science-language sentences, since to say *S is a fact*, or *S is not the case*, is the same as the assertion or denial of *S* in the paper. On the other hand, one can say that each science-language sentence in the paper carries a meta-science operator of the writer's asserting (or denying, or stating the improbability, etc., of) that sentence.

Meta-science operators on a sentence can also appear as modifiers of it, the latter being a transformation of the former: e.g. *In the present study, cells have shown pleomorphism* can be derived from *Cells have shown pleomorphism; that cells show pleomorphism is (found) in the present study* (where *is (found) in the present study* would be **M**).

There are some occurrences of **M** verbs where both subject and object are science-language sentences. Such are *demonstrate, show, indicate, suggest, confirm, point to*. These occurrences are similar to conjunctional verbs between science-language sentences such as *cause, accord with, support,* *speak in favor of, represent, mean that, by means of, is a result of, is a condition for, is consistent with, is corrected by, is borne out by.* The purely conjunctional verbs do not occur with **N′** as subject. Other verbs, such as *demonstrate*, can also occur with **N′** as subject, as in *Rich demonstrated that S_1* above. There is a transformational relation to the conjunctional status (as if one said *Rich demonstrated S_1 on the basis of some S_2*, whence *S_2 demonstrated that S_1*), but this is not always the case.

The procedures sketched above suffice to separate out, within the sentences of the articles, grammatically characterizable meta-science segments from a residue which is the science-language and is grammatically characterizable by itself. Separating these may involve complex transformations, which can be avoided if we allow some occurrences of meta-science words to remain within the science-language sentences. For instance, in *Peripheral lymph flow is far more rapid than is generally supposed* (paper 1, 783.1.2) we have a comparative with **M** in the second part: roughly *Peripheral lymph flows with a rapidity which is more than the rapidity of lymph flow which is generally supposed.* However, we can consider this occurrence of *supposed* as a word for quantity rather than **M**, and leave *is generally supposed* in the science sentence as though it meant *a moderate degree* or the like; this if the environment shows that *supposed* is not being used here to refer to actual supposing by scientists. Somewhat similarly, in *Some endoplasmic reticulum was demonstrable* (paper 13, 453.3.1) we can derive *demonstrable* from an underlying sentence such as *It was possible to demonstrate that some endoplasmic reticulum was present*; alternatively we can consider that *demonstrable* here did not refer to actual demonstration but was a rough synonym for *present* in *Some endoplasmic reticulum was present.* And *found* appears here both as **M** and as a synonym of *present* in the science sentences (e.g. in paper 1, 798.3.4).

In particular, science sentences can be filled out to conform to the sentence types worked out in *6* below by transforming certain kinds of modifiers from the **M** segment into the science sentence under that **M**. Thus we find (1) *Workers who examined the primary response were at first led to believe that the lymphocyte was responsible* (paper 9, 62.2.3). In terms of the word classes of *4*, *responsible* is merely a superscript on a word of the $\mathbf{V_p}$ class; hence *the lymphocyte was responsible* does not suffice for any sentence type of *6*. However, (1) could be derived from (2) *Workers who examined the primary response... believed that the lymphocyte was responsible for the primary response*, where *for the primary response* would have been zeroable as a repetition, yielding (1). Here, the *lymphocyte was responsible for the* *primary response* is a case of a $\mathbf{GJ^1{:}AV^r_pC_y}$ sentence type (section *6*). The reconstruction (2) of the zeroed segment is supported by the continuation of (1) in the article, which is *because of its very great predominance in antibody-containing suspensions made from once stimulated lymph nodes.* The sentence-types in this continuation are:

$$\mathbf{C_yW_i^{+}\,{}^{+}T_n^{wsw}}$$

$$\mathbf{AV_iT^s}$$

$$\mathbf{GU^1T_n}$$

representing

$\mathbf{C_yW_i^{+}\,{}^{+}T_n^{s}}$: *The lymphocyte had very great predominance in lymph node suspensions.*

$\mathbf{AV_iT^s}$: *Antibody is contained in suspensions.*

$\mathbf{GU^1T_n}$: *(Antigens) stimulated lymph nodes once.*

The conjunction *because* is understandable here only if *once-stimulated* is matched by *primary* in the first argument of *because.*

Every **M** segment is a grammatically constructed (i.e. argument-requirement-satisfying) chain of **M**, or **M′**, or **M**-type conjunctions (above), either operating on one or more science sentences or else occurring as a separate sentence. The sentences of the articles are composed entirely of the following: **M** segments, conjunctions, and science-language sentences. There are differences among the **M** segments, depending on the kind of science sentence on which they operate (e.g. observation sentences or conclusion sentences). However, those differences, as also the properties of conjunctions, relate to the structure of sentence sequences, and fall beyond the scope of the present book.

2. Word classes

In principle, word classes in a closed corpus of texts are established by characterizing each word-occurrence by its "co-occurrents," i.e. the words to which it has a grammatical relation in a sentence, and then putting into one class those word-occurrences which have the same co-occurrents, or nearly the same. The possibility of forming classes depends on how the word-occurrences cluster with respect to their co-occurrents. In the present corpus of articles, it was found that the subject-verb-object (or subject-predicate) relations sufficed to partition the word-occurrences into a few classes. The noun classes were easy to distinguish on the basis of their occurrences with other nouns and with verbs or adjectives; they are given in detail below. The operator classes (chiefly verbs) were defined chiefly in respect to the noun classes which appeared as their arguments, i.e. their subjects and objects. Since they are more complicated they are only introduced below, with the detailed membership given in *5*. The classes listed below are drawn only from the sentences presented in the tables of Appendix 1. In the articles, the sections on Materials and Procedures contained words of a few additional classes, which are not included here. Words are listed in order of appearance; parenthesized numbers indicate the article in which the word first appears in this corpus.

First, two classes, defined in respect to each other, can be set up for a set of nouns which occur as object of any of a particular set of verbs: The noun set is **G** (*antigen*), including (1) *antigen, bacteria, diptheria toxin, paratyphoid organisms, B. enteritidis, B. prodigiosus, ch. spirilla, typhoid vaccine, staphylococcus, bacilli;* (2) *sheep erythrocytes;* (3) *pneumococcus, sheep blood cells;* (4) *horse serum, s. typhi;* (5) *influenza virus, viral protein, cellular agents, agent;* (9) *antigenic material, organisms, diphtheria toxoid;* (10) *tetanus toxoid*; (12) *horseradish peroxidase;* (13) *antigen bearing red blood cells, SRBC.* The verb set is **J** (*inject*), including (1) *inject, incision, utilized, introduce, employ, vaccinate;* (3) *immunized*; (4) *sensitized, administered, deposited;* (5) *received injection*; (6) *received*, (9) *stimulation*; (10) *challenged with.* In most cases **G** is the subject of the passive of the **J**, as in *Paratyphoid bacteria was injected on one side.* For a few inverse members of **J, G** is the "object" – with *by* or *with* – of the passive **J**, as in *These animals were challenged with tetanus toxoid* (paper 10, 306.5.2) There are also a few nouns which can, on the grounds of their larger sentence-environment, be put into **J** unaccompanied by **G**: such as *scratch, puncture wound* in paper 1, 783.1.1 In many sentences, **GJ** is followed by a preposition plus noun (or an equivalent single word) such as *in these animals, in rabbits, on one side, intravenously, subcutaneously, intradermally.* These have been marked **B** ("body-part").

Words of **G** are also found, though much less frequently, as subjects of certain verbs marked **U** (or of the passive of certain inverse members of **U**). In **U**, whose general meaning is "move," are included (1) *travels, there exists a ready route for, has a path*, etc. There are certain preposition-plus-noun combinations which follow **U** whether **U** is active or passive. The prepositions are in most cases *from, to, along, by*, and the nouns are (1) *lymph nodes, blood stream, lymph stream, ear, blood*; (4) *red pulp, follicles, white pulp*. These nouns can be put into a class **T** (tissue); these and many additional members of **T** appear in other combinations too (below). Before **U** the class **G** includes a new member $\mathbf{G_f}$ (section *3*): *infection* (paper 1). Examples of **GUT** and **GU** are: *Antigen arrives by the lymph stream* (ibid. 1.7), where no preposition-plus-noun is added. **U** differs from **J** in that it may be followed by up to three **T**, each with a different preposition (*from, to along* and their synonyms), as in the transformed sentence (ibid. 1.5): *The infection has a path between the lymphatic capillaries of the skin and the entrance of the larger channels into the blood stream, along which path stand the regional lymph nodes* (where *between... and* is equivalent to *from... to*).

We next consider the co-occurrents of the word *antibody*. This word is the subject of a large set, marked **V**, of verbs, such as *appear in, are formed by*. Since they fall into several subclasses, these verbs will be discussed in the listing of subclasses (*3*). The subjects of **V** are marked **A**, and include (1) *antibody, agglutinin, bacteriolysin, antibody protein*; (2) *hemolysis*; (7) *immune globulins*; (13) *anti-ferritin, anti-peroxidase*. Many **V** are followed by a preposition (usually *in*) plus a noun of the class **T** (especially in paper 1) or of the class **C** (*cell*, in later papers). The main **T** words after **AV** are (1) *lymph nodes, serum*, but also e.g. *the ear tissue*, (2) *lymph*; (3) *adipose tissue*; etc. **C** words after **AV** are (1) *collections of lymphoid cells*; (2) *lymphocytes*; (3) *plasma cells*; etc. The **T** and **C** words also occur in other combinations (below), and will be listed in their subclasses (*3*).

As to the other combinations into which **T** and **C** words enter: There are rare constructions in which two words of **T** are the two arguments of an operator, e.g. *The lymph stream passes through the glands* (paper 1, 783.1.7) and the more common construction seen for example in *the lymph follicles in the spleen* (paper 4, 12.4.2). Much more common, and different, are the constructions in which the first argument is one of a specified set of words which are names of cell types, as in *lymphocytes present in the fat of the renal sinus* (paper 3, 128.8.2), *Cells of characteristic appearance occurred in the reaction centers* (paper 4, 1.3.4), *Lymphocytic hyperplasia becomes organized into the characteristic follicular structure* (paper 5, 204.2.2), *the chronic drainage of cells from a thoracic duct fistula* (paper 10, 303.2.1). The subject position here is occupied by **C** words, but not by **T** words; and the second argument is always **T** and not **C** as it is in respect to **Y** verbs, below.

The verbs in the **C–T** (and rare **T–T**) sentences above are marked **W**. These are two-argument members of **W**. There are also sentences in which **T** or **C** appears as subject of a one-argument operator such as *develops,* For example, *after←* indicates that *after* precedes **GJ** in *After the reinjection,... it was possible to observe the occurrence of cells of characteristic appearance* (paper 4, 1.3.4); and *to←* precedes **GJ** in *the response to a second intravenous injection of toxoid* (paper 10, 306.4.2). In contrast, *produce* without arrow indicates e.g. *Diptheria toxin was utilized to produce local inflammation* (paper 1, 792.1.1). The colon class includes (1) *wh*-conjunctions and pronouns ←, *after←, with, to←, following←, produce, call forth, induce, result in, upon←, in←*; (2) *yielded*; (4) *conditioned*; (5) *prior to←, specific to ←*; (9) *outcome of, detonates, results after←*; (10) *gave*; (12) *to trace*; (14) *give rise to, is stimulus to.* Although the colon words are grammatically conjunctions and sentence-connecting verbs, they differ in these articles from the other conjunctions in that they connect the two members of a very frequent sentence-pair: **GJB** and **CW** (or **TW**, or **AVC**). There are in addition many other conjunctions, which connect various sentences (including the above pair-sequence as a unit, e.g. **GJB:CW**) to others. While these other conjunctions are noted in the tables, they are not represented in the formulas, because their subclassification depends on an analysis of long sentence-sequences, which is not part of the present study.

These, then are the gross word-classes that can be distinguished by their co-occurrents in the material here investigated. In certain gross classes (**G, J, U, A, Y**), many members can co-occur with almost any member of the co-occurring classes (e.g. *antibody* in **A** can occur before any **V**), while other members (such as *plaque* in **A**) can occur only with particular members of the co-occurring classes, or with particular members of the co-occurring classes, or with particular grammatically-farther words. We put such words into a subclass: e.g. *plaque* in $\mathbf{A_q}$. There are other gross classes (**V, T, W, C, S**) in which virtually all the words are restricted as to co-occurrents, and thus are members of one subclass or another. In such classes, any word that is not thus restricted has the meaning of a classifier or a pronoun for the restricted subclass words.

3. Word subclasses

The subclasses are marked by a subscript after the class symbol. The different words in a subclass are in effect synonymous in respect to the given articles; that is, the semantic differences between them are immaterial to the research discussed in these articles.

G has a subclass $\mathbf{G}_{\mathbf{a}\sim}$ for foreign substances that do not call forth antibody response. Differently from **GJ**, the $\mathbf{G}_{\mathbf{a}\sim}\mathbf{J}$ is not followed by **:AV** (but by $\mathbf{:AV}_{\mathbf{i}}^{\sim}$). Members are, e.g., (1) *dye substances, diphtheria toxin. Infection*, *disease*, and some related words form a subclass $\mathbf{G}_{\mathbf{f}}$, noted in section *2*.

U has a subclass $\mathbf{U}_{\mathbf{i}}$ expressing the antigen's stopping at a tissue or its presence in the cell: (1) *is arrested in*, *is held by*; (4) *accumulates at*, *is found in*; (9) *presence of*. There is a subclass $\mathbf{U}_{\mathbf{d}}$: (4) *perish in*. We find a $\mathbf{GU}_{\mathbf{d}}^{\mathbf{k}}{-}\mathbf{T}_{\mathbf{t}}$ sentence-type in the transformed *Thymus has an insignificant phagocytizing capacity toward antigen* (paper 4, 12.4.2). There is also $\mathbf{U}_{\mathbf{p}}$: (5) *multiply*, in *eliminates the question of multiplication of the agent* (paper 5, 204.2.7); and $\mathbf{U}_{\mathbf{s}}$ *sensitize*.

The class **T** has a few non-specific members, which have in most occurrences the status of a classifier (see below): (1) *tissue, site, organ*; (4) *places*. The other words that appear in **T** position are assignable to distinct subclasses on the basis either of the **W** subclass with which they occur or of the neighboring sentences to which they are conjoined. One subclass is:

$\mathbf{T}_{\mathbf{b}}$ (1) *blood, serum, vascular, circulating*; (12) *humoral.*

Subclasses referring to lymph tissue are:

$\mathbf{T}_{\mathbf{n}}$ (1) *lymph nodes, lymph glands*;
$\mathbf{T}_{\ell}$ (1) *lymph, lymph stream.*
$\mathbf{T}_{\ell'}$ (1) *lymphatic plexus, lymphatic capillaries, lymphatics, lymphatic tissues;* (2) *lymphoid tissue*;
$\mathbf{T}_{\ell''}$ (5) *interstitial fluid, lymph supernatant.*

Subclasses of other tissues containing antibody-forming cells are:

$\mathbf{T}_{\mathbf{t}}$ (3) *thymus*;
$\mathbf{T}_{\mathbf{k}}$ (3) *adipose tissue of the renal sinus, fat of the renal sinus, pelvic fat*;
$\mathbf{T}_{\mathbf{p}}$ (3) *retroperitoneal adipose tissue, retroperitoneal fat.*

Subclasses naming tissue structures are:

$\mathbf{T}_{\mathbf{s}}$ (1) *spleen*;
$\mathbf{T}_{\mathbf{d}}$ (4) *red pulp of the spleen*;
$\mathbf{T}_{\mathbf{f}}$ (4) *white pulp of the spleen, lymphatic follicles, follicular tissue*;
$\mathbf{T}_{\mathbf{m}}$ (4) *Malpighian bodies, periphery of the lymph follicles*;
$\mathbf{T}_{\mathbf{x}}$ (9) *cortex*;
$\mathbf{T}_{\mathbf{u}}$ (9) *medulla*;

$\mathbf{T}_r$ (1) *germinal centers*; (4) *reaction centers*;
$\mathbf{T}_h$ (10) *thoracic duct fistula.*

Subclasses of tissue not containing antibody-forming cells:

$\mathbf{T}_v$ (1) *liver*;
$\mathbf{T}_c$ (2) *muscle.*

In **W**, whose one or two arguments are **T, C**, or **S**, a few general properties are named: (1) *size, appearance*, (3) *weight, have histological features*, (11) *have morphological features.* Otherwise, there is a host of subclasses, each characterized by particular arguments. The major ones (for the others, see *5*) are:

$\mathbf{W}_a$, whose argument is **T** or **C**, contains (1) *reaction, are affected, involved*; (3) *active*; (9) *response, biological event*;

$\mathbf{W}_f$ has only **T** as argument: (1) *painful, inflamed, hemorrhagic*; in this context *normal* is $\mathbf{W}_{f\sim}$ (where the tilde means *not*).

$\mathbf{W}_g$ after **T** is (1) *enlarged*; after **C** or **S** it is (4) *large*, with $\mathbf{W}_{g\sim}$ standing for *small*; after S it is also (12) *extensive.*

$\mathbf{W}_{c'}$, only after **T**, contains (1) *rupture, open*; it occurs in particular sentence-sequences.

The most frequent operator on the pair **C, T** is $\mathbf{W}_i$, whose first and second arguments are mostly **C** and **T** respectively (or **T** and **C**, in the case of words marked "inv," for "inverse"): chiefly (3) *infiltrate, found in, present in, present in*, $\mathbf{W}_{i\sim}$ *free from* (inv, *the lymphatic tissue investigated was free from plasma cells*, 128.3.3), $\mathbf{W}_i^+$ *predominant*; (4) *met with, localized in*, $\mathbf{W}_i^+$ *abundant, contain* (inv), *abundant in* (inv, in *the pieces of red pulp were abundant in plasma cells*, (5.1.1), *in, have number* (11.1.5); *content*; (5) $\mathbf{W}_i^+$ *hyperplasia*; (7) *consist of* (inv); (9) *scattered*; (10) $\mathbf{W}_{i\sim}$ *depleted of* (inv), $\mathbf{W}_{i\sim}$ *are lacking*; (12) $\mathbf{W}_i^-$ *few, scanty*, etc.; (13) *occupy.*

$\mathbf{W}_p$ also has **C** as subject, in many cases with a second argument **T**: (3) *proliferate, -poietic*; (4) *formation, development* (11.2.3), *production*; (5) *multiply*; (7) *output, -genesis.*

$\mathbf{W}_c$ has mostly **C** as subject, but with no second argument: (1) *change* (where the subject is still **T**); (4) *transition, develop*; (5) *become organized*; (9) *differentiation, changing character, course of events*; (11) *pleomorphism*; (13) *adaptations, different*; (14) *sequential changes.*

A related subclass with one argument, **C** or **S**, is $\mathbf{W}_m$: (6) *reach maturity*, (7) *well-developed.*

Somewhat less common is $\mathbf{W_u}$, whose first argument is **C** (rarely **T**) and **T**: (1) *flow, pass through* (with **T** subject); (6) *leave, separated from*; (14) *held up in, enter, settle in, migrating throughout, reach.* Inspection of the neighboring sentences shows that $\mathbf{W_i}$ deals with the presence or absence of cells in tissue, while $\mathbf{W_u}$ deals with their motion.

A very few words can be put in another subclass, $\mathbf{W_d}$, with **C** as subject: (4) *disintegrate*, or **T** as object: (10) *damage.* And $\mathbf{W_o}$: (7) *mitoses.*

In addition, there are several subclasses, each with a particular subset of S as subject. Chief among these are:

$\mathbf{W_e}$	(4) *eccentric*, and $\mathbf{W_{e\sim}}$ *round*, with subject $\mathbf{S_n}$ (*nucleus*);
$\mathbf{W_s}$	(4) *red*, (7) *basphilic, pyroninphilic, bright* with subject $\mathbf{S_c}$ (*cytoplasm*);
$\mathbf{W_r}$	(11) *rough* with subject $\mathbf{S_r}$ (*endoplasmic reticulum*);
$\mathbf{W_t}$	(13) *electron-opaque*, with subject $\mathbf{S_n}$.

Finally, there is a subclass $\mathbf{W_\ell}$ of laboratory prodedures, whose object is **C** or **T**, such as (7) *seperate by sedimentation*, (1) *excise, tease.*

As to C, the following major subclasses can be distinguished:

$\mathbf{C_\ell}$	(1) *lymphoid cells*;
$\mathbf{C_y}$	(1) *lymphocytes*;
$\mathbf{C_r}$	(2) *reticulo-endothelial cells*; (4) *reticulum cells, reticulo-endothelial elements*;
$\mathbf{C_z}$	(3) *plasma cells*, (13) *plasmacytic*;
$\mathbf{C_b}$	(9) *hemacytoblasts*, *blast forms*;
$\mathbf{C_h}$	(10) *pyroninophilic cells*;
$\mathbf{C_m}$	(13) *macrophages.*

In S, there are few classifier words serving for all subclasses: (11) *structural units.* The other words are in subclasses on the basis of the W subclasses whose subjects they are:

$\mathbf{S_c}$	(4) *cytoplasm*;
$\mathbf{S_n}$	(4) *nucleus*;
$\mathbf{S_u}$	(7) *nucleolus*;
$\mathbf{S_m}$	(11) *mitochondria*;
$\mathbf{S_g}$	(11) *Golgi apparatus, Golgi bodies*, (13) *Golgi area*;
$\mathbf{S_r}$	(11) *endoplasmic reticulum*, (12) *ergastoplasm*;
$\mathbf{S_b}$	(11) *ribosomes*;
$\mathbf{S_p}$	(13) *perinuclear space.*

S_rC—	**S—C**	**T_m—B**
well developed	*develops*	*were well developed in* *showed proliferation* *of* ←

T_r—T_f
proliferation in

W_o: **C—**
undergoing mitotic division
dividing
divides
mitotes were found

W_u: **T_ℓ—**
flows

W_u^f: **C—T_n**
transferred (from)
output from
output of ←
leave from

$W_u^{f\sim}$: **C_z^m—T_n**
would be held up in

W_u^t:	**C—T_n, C—T_b, C—T**	**C—T**
	entered *settled in*	*deposition* ←

$W_u^{t\sim}$: **T_ℓ^f—T_b**
was prevented from reaching

W_u^y: **C—T_n**
pass through
migrating throughout

referent is only likely (*3.1*) The tables of Appendix 1 do not reconstruct all instances of repetitional zeroing – relevant considerations are noted below.

3.1. Parallel-zeroing and end-zeroing

Parallel-zeroing is widespread under *and*, *or*, and other conjunctions, e.g., *but*, the comparative (7). In (a) *both lymphocytes and plasma cells produce antibodies* (from 3, 218.9.1; the conjunction here is *both...and*), parallel-zeroing in the source sentence: *both lymphocytes produce antibodies and plasma cells produce antibodies* results in *both lymphocytes produce antibodies, and plasma cells.* To obtain (a), the residue of the zeroing, *and plasma cells*, is requiredly transposed to after the last word which did not serve as an antecedent for the zeroing (here, *lymphocytes*). For a sentence involving a comparative form consider (b): *the total bacterial content had in most cases fallen considerably and at a greater rate in the red than in the white pulp* (4,9.1.1) *And at a greater rate* indicates a zeroing of the second sentence (under *and*) aside from its modifier: *the total bacterial content had in most cases fallen.* Under the comparative *-er... than*, which raises the likelihood of word-repetition, *the total bacterial content had in most cases fallen at a rate* is reconstructed following *than.*

In end-zeroing, the final sequence of words (usually in the second sentence) has been repetitionally zeroed. End-zeroing is recognized under many operators, e.g. *and*, *or*, comparative, and other conjunctions (chiefly O_{oo}, an operator whose first and second arguments are operators). In (c) *The lysed lymphocytes did not contain specific agglutinin, whereas the cultured lymphocytes did* (14,577.1.5), the sequence *contain specific agglutinin* is reconstructed under the contrastive conjunction *whereas.*

The reconstruction (expansion) of all the text-sentences in accord with zeroings just mentioned would entail considerable extension of the tables. To avoid this situation, conjunctions, principally *and* and *or*, have been left in the rows and are indicated in the formulas by a comma. For instance, (d) *the antibody production in vitro of red and white splenic pulp* (from 7,3.5.1) is not expanded in the tables; its formula is abbreviated as $\mathbf{AV^{v}_{p}T_{d}, T_{f}}$.

3.2. Subject-zeroing

Under various prepositions, and subordinate conjunctions, the subject of the second sentence, if it is the same as an argument of S_1, is zeroable, along with *is*. In (e), *When present, it occurs chiefly in the interior of some or all of* *the large flattened sacs*... (from 12,113.2.2), *it* is a pronominal reduction of an antecedent *antibody* (the second sentence in this example has been moved to before the primary sentence, S_1). The sentence is then expanded to *When antibody is present*....

Another, infrequent, case of subject-zeroing arises where the subject of a lower sentence has the same referent as the subject of a higher operator. This zeroing is reconstructed in example (f): *if agglutinins had seeped through the permeable vessels on the inflamed ear for agglutinins to be drained to the lymph nodes*. The text-sentence has ... *on the inflamed ear to be drained* (from 1,792.4.1), where the *for* (of the *for*... *to* argument indicator) preceding the zeroed lower subject is also zeroed.

4. Reconstruction of Low-Information Zeroing

This section examines the considerations according to which a text-sentence can be regularized by reconstructing occurrences of words present only in zero phonemic form. Word occurrences with high likelihood in a stated situation make little or no informational contribution to their sentence and are readily zeroable. In terms of the present analyses, it is often unnecessary to reconstruct all zeroed forms. In general, this has been done when some feature of the analysis depended upon, or was made clearer by, such reconstruction. In the tables of Appendix 1 reconstructions of zeroing are enclosed within parentheses.

4.1. Broad selection words

Certain words normally occur with an exceptionally large domain of operators over them or arguments under them. These words have only very general meanings and corresponding to their high likelihood, the informational contribution they make to their sentence is low. As such, they often occur in zero form but can be reconstructed, e.g., by noting that their presence is required in order to satisfy the argument requirement of a neighboring word. However, unlike reconstructions of repetitional zeroings, words which have been zeroed on grounds of low information are often not uniquely reconstructible. Rather the trace of the zeroing suggests only that some word or words from a small set of words, all of which have roughly the same favored likelihood in the specified environment (and thus are locally synonymous), may be reconstructed. In the present material an important set of these broad selection words are the classifiers *amount, quantity, degree, number, period, time.* These, under a characteristic preposition, may occur as modifiers of many of the verbs (main operators) of sublanguage sentences, e.g., *antibody production was in a quantity, plasma cell proliferation was to a degree, antigen uptake by the cell occurred at a time, cell differentiation occurred throughout a period.* When occurring under their (selectionally) favored "appropriate" operators, these modifiers are often zeroed. For words like *quantity, number, amount, degree* this appropriate operator may be the comparative *more* (*than*), a specifying adjective e.g., *high, some, little* or this adjective under the comparative as in *higher, greater* (cf. *7*). As the zeroing of these broad selection classifiers is extremely widespread and of little significance in establishing the informational structures of the sublanguage, only rarely have reconstructions been performed in the projected sentences, and then only to preempt possible unclarity as to the choice of a word class or subclass. A case in point is *the changes in nucleic acids in lymph nodes* (from 6,158.2.1) which is reconstructed under conditions discussed below and which is represented formulaically as $\mathbf{DV}_{\mathbf{i}}^{\Delta}\mathbf{T}_{\mathbf{n}}$. In such cases, the reconstruction serves to illustrate that what may appear as a new sentence type or subclass can be accommodated within existing forms.[3] Similarly, in (6,164.4.2) *the rise in lymphocytes did not prevent the PNA from dropping* is reconstructed as *the rise in numbers of lymphocytes present did not prevent the PNA quantity* (or: *concentration*) *from dropping.*

4.2. Strong selection zeroing

A case related to the zeroing of broad selection words under an appropriate operator is that of strong selection, i.e., the zeroing of certain words with exceptionally high likelihood of having particular cooccurrents. For example, in GEMP (6.14) apparent O_{oo} (that is, bisentential) occurrences of the time-order prepositions *before, after, following* and the like are derived from base occurrences as O_{on}.[4] As O_{on} operators, these prepositions can have as first argument an aspectually modified sentence – S_1 *at a time/in a period* – and as second argument a duration noun such as *time, moment, period.*[5] Their apparent conjunctional occurrence stems from strong selection to the duration words which, by this fact, can occur only in zero form. A second sentence may then be appended as a relative clause via *when* or another relative pronoun. Schematically, the reductional path from O_{on} to O_{oo} is S_1 *in a period after the time when* $S_2 \rightarrow S_1$ *after* S_2. By application of a relinearization transformation of *2*, this becomes *after* S_2, S_1. Taking a