Algebraic Linguistics; Analytical Models

SOLOMON MARCUS

Algebraic Linguistics; Analytical Models

SOLOMON MARCUS

UNIVERSITY OF BUCHAREST
AND MATHEMATICAL INSTITUTE
OF THE ACADEMY OF THE
SOCIALIST REPUBLIC ROMANIA
BUCHAREST, ROMANIA

ACADEMIC PRESS New York and London

ACADEMIC PRESS INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 65-28625

PRINTED IN THE UNITED STATES OF AMERICA

Preface

There are two fundamental types of models which are studied in algebraic linguistics: generative and analytic. Simplifying, we might say that within the framework of a generative model, the starting point is a certain grammar, while the object we study is the language generated by this grammar. An analytic model presents an inverse situation; here the starting point is a certain language, i.e., a certain collection of sentences, whereas the purpose of the study is to establish the structure of these sentences, their constitutive elements, and the relations among them within the framework of sentences.

As shown by the title, the present book is devoted to analytic models. These models cover to a great extent the area of descriptive linguistics and therefore present a great interest for linguists.

Special attention has been given to the axiomatic-deductive structure of analytic models. At the same time we have tried to explain the linguistic origin of the notions, the linguistic meaning of the theorems and the manner in which the models studied are used to investigate natural languages.

Most of the examples belonging to natural languages have a hypothetical and explanatory character; here we must take into account that the model is only an approximation of the reality. Hence there exists a certain lack of fit between a phenomenon and its model.

In view of the close connection between analytic and generative models and of the fact that some models have a mixed, generative-analytic character, we have also discussed some questions currently considered as belonging to generative models. An example of this sort is the calculus of syntactic types, discussed in the second part of Chapter III. We have also given those notions and results concerning generative models which permit us to understand the links between the two types of models; these links are pointed out in various paragraphs of the book.

The book is primarily directed to those mathematicians who desire to become acquainted with the mathematical aspects of linguistic struc-

tures and to those linguists who wish to know (and to use) one of the most powerful tools for investigating the structure of language: mathematical modeling. The book can also be useful to all those who are interested in the problems of linguistic information processing (automatic translation, informational languages, programming languages, etc.). Thus, the notion of configuration, dealt with in Chapter V, has already been used in construction of some algorithms of automatic translation (see the corresponding references in Chapter V).

In view of the rapid progress of algebraic linguistics, we made a definite effort to take into account the most recent contributions in this field. Of course, we have not presented all analytic models existing in literature. We hope that the selection we have made enables us to confer on the book a certain unity of conception and treatment.

A good portion of the book relies on some of the author's papers, as specified in the references placed at the end of each chapter. On the other hand, the book contains many results published here for the first time (especially in Chapters II, III, IV, and V).

We are very indebted to Professors Miron Nicolescu, Grigore Moisil, and Alexandru Rosetti for their support and encouragement in pursuing the research in the field of mathematical linguistics.

In writing this book we have been stimulated by the proposal made to us by Richard Bellman in June 1964 to publish in his famous series "Mathematics in Science and Engineering" an English version of our previous book "Lingvistică matematică" (Editura didactică şi pedagogică, Bucureşti, 1963). We thought it more appropriate to write an entirely new book, which would reflect the general status of analytic models and our own most recent views. We are deeply grateful to Richard Bellman for the opportunity to publish this book.

Bucharest
November, 1966

SOLOMON MARCUS

Contents

Preface vii

Chapter I. Languages and Partitions

1. LANGUAGES AND GRAMMARS 1
2. ENRICHING THE STRUCTURE OF A LANGUAGE 2
3. THE NOTION OF NATURAL LANGUAGE 4
4. DISTRIBUTION 6
5. *P*-STRUCTURES; DERIVATIVE OF A PARTITION 8
6. *P*-DOMINATION AND SOME OF ITS PROPERTIES 11
7. COMPARABLE PARTITIONS WITH THE SAME DERIVATIVE 16
8. PARTITIONS WITH THE SAME DERIVATIVE 18
9. CONDITIONS THAT A PARTITION BE A DERIVATIVE 20
10. MIXED CELLS; UNION AND INTERSECTION OF TWO PARTITIONS 23
11. CLASSES AND THEIR STRUCTURE 25
12. PARTITIONS OF THE FREE SEMIGROUP GENERATED BY Γ 27
13. BIBLIOGRAPHIC REMARKS 33

REFERENCES 33

Chapter II. Linguistic Typology

1. ADEQUATE LANGUAGES 36
2. HOMOGENEOUS LANGUAGES 40
3. VARIOUS TYPES OF HOMOGENEOUS LANGUAGES 44
4. COMPLETELY ADEQUATE LANGUAGES 49
5. OTHER TYPES OF ADEQUATE LANGUAGES 57
6. VARIOUS TYPES OF LINGUISTIC ISOMORPHISM 59
7. SOME CHARACTERISTICS OF FINITE-STATE LANGUAGES 66
8. SOME APPLICATIONS TO NATURAL LANGUAGES 69
9. BIBLIOGRAPHIC REMARKS 78

REFERENCES 79

Chapter III. **Parts of Speech and Syntactic Types**

1. INTRODUCTION 81
2. PARTS OF SPEECH AS CELLS OF THE DERIVATIVE PARTITION P' 82
3. GRAMMATICALITY 83
4. LINGUISTIC EXPLANATION OF THE MODEL OF PART OF SPEECH 84
5. PARTS OF SPEECH IN ADEQUATE AND IN HOMOGENEOUS LANGUAGES 88
6. SYNTACTIC TYPES 92
7. ANALYSIS OF THE ENGLISH VERB PHRASE 98
8. THE ASSOCIATIVE SYNTACTIC CALCULUS 101
9. NONASSOCIATIVE SYNTACTIC CALCULUS 104
10. CATEGORIAL GRAMMARS 107
REFERENCES 112

Chapter IV. **Grammatical Gender**

1. INTRODUCTION 115
2. FROM THE NATURAL TO THE GRAMMATICAL GENDER 116
3. GRAMMATICAL GENDERS IN NATURAL LANGUAGES 118
4. MATHEMATICAL MODELS OF GRAMMATICAL GENDER 124
5. GRAMMATICAL GENDERS IN HOMOGENEOUS LANGUAGES 133
6. CATEGORIES IN THE SENSE OF REVZIN 139
7. SUBPARADIGMS AND SUBFAMILIES 142
8. A MEASURE OF THE DIFFERENCE BETWEEN GENDERS 145
9. PERSONAL GENDERS IN RUMANIAN 148
REFERENCES 153

Chapter V. **Configurations**

1. INTRODUCTION 156
2. P-CONFIGURATIONS AND P-STRUCTURES OF ORDER n 157
3. THE P-CONFIGURATION AND THE P-STRUCTURE TYPES OF LANGUAGE 160
4. EXAMPLES OF E-CONFIGURATIONS 162
5. REMOVAL OF PARASITIC ELEMENTS 168
6. SEMIREGULAR P-CONFIGURATIONS 169
7. NORMAL P-CONFIGURATIONS, DEPENDENCIES, AND CONSTITUENTS 172

8. *P*-CONFIGURATIONS, *P*-STRUCTURES, AND REGULARLY FINER PARTITIONS 175
9. CONFIGURATIONS IN THE SENSE OF GLADKIĬ 183
10. GLADKIĬ CONFIGURATIONS AND GENERATIVE GRAMMARS 188
11. QUASI-CONFIGURATIONS AND SYNTAGMS 193
12. FINAL REMARKS 197
REFERENCES 198

Chapter VI. **Subordination and Projectivity**

1. INTRODUCTION 200
2. SOME NOTIONS AND RESULTS CONCERNING GRAPH THEORY 200
3. SIMPLE STRINGS AND PROPER TREES 203
4. AN AXIOMATIC DESCRIPTION OF SIMPLE STRINGS 208
5. ELEMENTARY STRINGS AND OPERATIONS WITH SIMPLE STRINGS 212
6. SUBORDINATION IN THE SENSE OF NEBESKÝ 215
7. VARIOUS TYPES OF PROJECTIVITY. PROPERTIES OF MONTONICALLY PROJECTIVE STRINGS 219
8. RELATIONS BETWEEN VARIOUS TYPES OF PROJECTIVITY 224
9. PROJECTIVITY IN NATURAL LANGUAGES 229
10. SIMPLE PROJECTIVE STRINGS 233
11. BIBLIOGRAPHIC REMARKS 240
REFERENCES 243

Author Index 247
Subject Index 250

Chapter I

Languages and Partitions

1. Languages and Grammars

Let Γ be a finite set called the *vocabulary*. The elements of Γ are *words*. Consider the free semigroup T generated by Γ, namely, the set of all finite strings of words endowed with an associative and noncommutative binary operation of concatenation. Since we are considering only finite strings, we shall say *strings* instead of finite strings. A string of words will also be called a string over Γ. The *zero string*, denoted by θ, is a string such that $\theta x = x\theta = x$ for each string x. Without contrary assumption, θ does not belong to Γ.

A subset Φ of T is a *language* over Γ. The semigroup T is the *total* or the *universal language* over Γ.

A *generative grammar* of Φ is a finite set of rules (called *grammatical rules*) specifying all strings of Φ (and only these strings) and assigning to each string of Φ a structural description that specifies the elements of which the string is constructed, their order, arrangement, interrelations, and whatever other grammatical information is needed to determine how the string is used and understood. ([5], p. 285). It is to be noted that in such a grammar the structural description is made with the aid of grammatical rules.

Such a point of view is closely related to the theory of formal systems and to other fundamental chapters of contemporary mathematical logic (such as Turing machines and recursive functions). But we shall consider in this book a quite different point of view: that of an *analytic grammar*.

An analytic grammar of Φ considers Φ given, and its purpose is to obtain an intrinsic description of the strings belonging to Φ, that is, a description of the relations between the words and between the substrings with respect to their position in the strings of Φ. Such a point of view is very closely related to the traditional structural linguistic theory, especially

to the so-called descriptive linguistics developed by Bloomfield [2, 3], Harris [13], Hockett [15], Wells [38], and others.

To provide a clearer distinction between a generative grammar and an analytic grammar, let us consider the following example. It is known that a finite-state language may be generated in several ways. If an ambiguous grammar is used, we may detect the so-called constructional homonymy that arises when a sentence has several representing sequences, that is, several different "constructions" ([1], pp. 93–94). Note, for instance, the ambiguous English sentence: *They are flying planes*, which is really two different sentences: (1) *They (are (flying planes))* and (2) *They ((are flying) planes)*. The grammatical structures, or the meanings of these two sentences are different ([5], p. 274); an ambiguous finite-state grammar or a nondeterministic finite automaton may detect this difference ([1], pp. 93–94). Such a situation is the basic concern of generative grammar.

Let us now consider another situation. We shall say that two strings x and y are Φ equivalent if, for each pair of strings u, v, we have either $uxv \in \Phi$, $uyv \in \Phi$, or $uxv \in T-\Phi$, $uyv \in T-\Phi$. A fundamental result of Rabin and Scott ([29], Theorem 1) and a theorem of Bar-Hillel and Shamir [1] imply that Φ is a finite-state language if and only if there are only finitely many Φ-equivalence classes. Such a characterization of the finite-state languages, which involves only the intrinsic structure of these languages, is at the basis of an analytic grammar.

The above example shows not only the difference, but also the close connection between the two types of grammars. Each completes the description given by the other.

The utility of an analytic study of the languages follows also from another fact. Since Γ is finite, the universal language T is denumerable, and, consequently, the set of all languages over Γ is not denumerable. On the other hand, as is noted in [4], the set of all generative grammars over Γ (more precisely, the set of all constituent-structure grammars over Γ) is denumerable. Therefore, there exists a nondenumerable set $\mathscr{L}$ of languages over Γ, such that, for $L \in \mathscr{L}$, there is no generative grammar of L. For such languages, the analytic study of their structure is the only method of grammatical investigation. An analytic study is applicable to every language.

2. Enriching the Structure of a Language

There are many problems concerning a language Φ which can be successfully studied without enriching the structure of Φ, that is, by knowing

only that Φ is a determined subset of the free semigroup generated by Γ and being able to say, for each string over Γ, whether it belongs to Φ. An example of such a problem is that of morphologic homonymy. We shall say that the morphologic homonymy of the word x is not greater than the morphologic homonymy of the word y, if for each pair of strings u and v such that $uxv \in \Phi$, we have $uyv \in \Phi$. Moreover, if the converse is not true, that is, if there are two strings u and v such that $uyv \in \Phi$ but $uxv \in T - \Phi$, we shall say that the morphologic homonymy of x is less than the morphologic homonymy of y. Thus, if Γ is the French vocabulary and Φ is the set of all well-formed French sentences, the morphologic homonymy of *beau* is less than the morphologic homonymy of *mince*. Indeed, in each well-formed sentence containing the word *beau* the replacement of *beau* by *mince* also gives a well-formed sentence; but there exists a well-formed sentence containing the word *mince*, such that the replacement of *mince* by *beau* gives no well-formed sentence (compare *je possède une feuille mince* and *je possède une feuille beau*). A systematic development of this idea—which originates with Dobrušin [7, 8] and Sestier [34]—was given in [21–23]. For further developments, see [6, 24, 31, 32].

Another problem which may be studied without enriching the basic structure of the language is that of the morphemic segmentation. If Γ is the set of phonemes of a natural language and Φ is the set of all well-formed sequences of phonemes in this language, then, by counting the possible successors of each initial segment, one can obtain the morphemic boundaries in the considered sequence. Such a procedure was discovered by Harris [12].

We have discussed so far two problems of a pure distributional and syntagmatic character. Other such problems are considered in [24]. But there are many problems which also involve a paradigmatic structure of the considered language, that is, a partition of Γ. Such problems will be considered in Chapters I through IV. The customary linguistic interpretation of the partition of Γ is the decomposition of the set of words in paradigms, the paradigm of a word being the set of its flectional forms. For instance, the paradigm of book is {*book, books*} and the paradigm of great is {*great, greater, greatest*}. In fact, the paradigms do not form a partition of Γ, since there exist distinct paradigms which are not disjoint. Such nonconcordances are unavoidable in all modeling processes.

A triple $\{\Gamma, P, \Phi\}$, where Γ is a finite vocabulary, P is a partition of Γ, and Φ is a subset of the free semigroup generated by Γ will be called a *language with paradigmatic structure*. Since we are considering

especially such languages, we shall say, briefly, that $\{\Gamma, P, \Phi\}$ is a language.

The linguistic analysis needed in machine translation requires a richer structure of the considered languages. Here, a language must be considered a system $\{\Gamma, P, \Phi, K, \phi\}$, where Γ, P and Φ are the objects already defined, K is a class of subsets of Γ called grammatical categories (such as the set of words in nominative or the set of words in the past tense), and ϕ is a function which associates to each word x the intersection of all grammatical categories containing x. For a further discussion of this point of view, see [33], pp. 42–43.

3. The Notion of Natural Language

The notion of a language over the vocabulary Γ includes both natural languages and the artificial languages of logic and of computer-programing theory. The notion of a natural language is much more complicated, since its structure is very rich. Kalmár has proposed a definition of the concept of language, especially concerning the natural languages, which was intended to cover all parts of linguistics [16]. He defines a language as an 11-tuple $\{P, R, F, W, C, A, S, M_W, M_S, A_W, A_S\}$, with the symbols as follows:

P is an arbitrary set called the set of *protosemata* (in the case of a spoken language the set of physical sounds used as representatives of phonemes; in the case of a written language the set of geometrical figures used as representatives of letters).

R is an equivalence relation defined on the set of occurrences of the protosemata in the strings of the free semigroup generated by P. The classes of R equivalence are called *semata* (phonemes or graphemes, respectively).

F is a subset of the free semigroup generated by the set of semata (the elements of F are called *word forms*).

W is a subset of the power set of F, that is, a set the elements of which are subsets of F, or a decomposition of the set F into not necessarily disjoint subsets. (The elements of W, or the subsets of F into which it has been decomposed, are called *words*, every word being identified with the set of all its forms).

C is a partition of the set W into subsets called *word classes* or *parts of discourse.*

A is an application of the set C onto some set the elements of which are sets of functions such that if $c \in C$ (that is, if c is a word class) and G is the image of c under application A, then G is a set of functions f defined for all elements w of c (that is, for all words w belonging to the word class c) and for each such w, we have $f(w) \in w$ [that is, $f(w)$ is one of the forms of w]. For example, if c is the class of all nouns (suppose this to be a word class), the elements of the corresponding G are the functions "the nominative of . . . ," "the accusative of . . . ," etc.; if c is the class of all verbs (supposed to be a word class), the elements of the corresponding G are the functions "the indicative present tense singular second person of . . . ," etc. A is called the *morphologic application.*

S is a subset of the free semigroup generated by the set F. The elements of S are called *grammatically correct sentences.*

M_W is a set called the *set of word meanings.*

M_S is a set called the *set of sentence meanings.*

A_W is an application of the set W into the power set of M_W. For any word $w \in W$, we call the elements of the set onto which w is mapped by A_W, the (possible) *meanings* of w.

A_S is an application of the set S into the power set of M_S. For any sentence $s \in S$, we call the elements of the set onto which s is mapped by A_S, the (possible) *meanings* of s.

Tentatively, we can regard the sets M_W and M_S as arbitrary abstract sets; however, to have a better model of natural languages, we suppose them to be sets having some logical structures still to be determined. Approximately, M_W corresponds to the set of concepts and M_S to the set of propositions in the sense of traditional logic. The sets M_W and M_S are common for different natural languages, which makes translation from one to the other possible.

A theory based on this definition needs some structure axioms (the term "structure" being used in a sense similar to that of an algebraic structure). In such a theory, phonology, morphology, syntax, and semantics will appear as subtheories similar to those of the additive group of a ring in relation to ring theory. Thus, P, R, and F define the phonetics, the graphematics, and the phonology; W, C, and A define the morphology; S defines the syntax; M_W, M_S, A_W, and A_S define the semantics. In such a theory, a generative grammar may show how to generate the set F of word forms or the set S of grammatically correct sentences.

The customary nonconcordance between a phenomenon and its

logical model appears also in the above construction. So, in a natural language the parts of discourse are not disjoint, and the passage from physical sounds to phonemes is not simple enough to describe by an equivalence relation. See, in this respect, [17, 27, 28, 36].

The sets M_W and M_S are ambiguous, for we do not have a clear criterion for deciding when two word meanings or two sentence meanings can be regarded as identical. The definition of identity has to be the main part of the determination of the logical structure of the sets M_W and M_S. For the delicate questions of semantics and the possibility of using the methods of generative grammars here, see [18, 28, 39]. We also note the absence, in the above construction, of such a fundamental linguistic notion as morpheme. Finally, let us remark that, according to some recent papers [14, 33], the notion of grammatical correctness, attached to the set S, may be reduced to simpler notions.

By postulating appropriate axioms, the above model can probably be improved, so as to become more adequate to the nonbanal aspects of natural languages.

4. Distribution

Let us first consider the most simple notion of a language, given as a pair $\{\Gamma, \Phi\}$. The strings which belong to Φ are called *marked strings*. In many linguistic problems we are concerning with various partitions of Γ, that is, decompositions of Γ into nonvoid mutually disjoint sets.

The most important partition of Γ which arises in linguistics is the so-called partition in *distributional classes*, defined as follows. Two words a and b will be considered in the same distributional class if for each pair of strings x, y, the relation $xay \in \Phi$ implies $xby \in \Phi$, whereas the relation $xby \in \Phi$ implies $xay \in \Phi$.

The notion of distributional class becomes more intuitive if we introduce the notion of *context*. A context over Γ will be defined as an ordered pair of strings over Γ and will be denoted by $\langle x, y\rangle$, where $x \in T$ and $y \in T$. A word a is *allowed* by the context $\langle x, y\rangle$ if the string xay belongs to Φ. Denote by $\mathscr{S}(a)$ the set of all contexts with respect to which a is allowed. It follows immediately that two words a and b belong to the same distributional class if and only if $\mathscr{S}(a) = \mathscr{S}(b)$, that is, if and only

if a and b are allowed by the same contexts. This notion has its origin in descriptive linguistics (see, for instance, [9] and [13]).

If we interpret Γ as the English vocabulary and Φ as the set of well-formed English sentences, the words *book* and *chair* are in the same distributional class, whereas *book* and *books* are not. If we interpret Γ as the French vocabulary and Φ as the set of well-formed French sentences, the words *mince* and *maigre* are in the same distributional class, whereas *grand* and *mince* are not; indeed, the sentence *j'ai une feuille mince* is well-formed, whereas *j'ai une feuille grand* is not. One of the principal tasks in the study of a language is the establishment of its distributional classes.

It is easy to see that two different distributional classes are disjoint; thus these classes define a partition S of Γ, called the *distributional partition* of Γ. The first mathematical study of this notion was made in 1958 [19] and will be the point of departure in the following considerations. A distributional class is called, in [19], a *family*. We shall use these two denominations as equivalent.

The properties defined exclusively in terms of contexts and of distributional classes are the simplest and the most elegant in a linguistic description. We may consider the following situations concerning the reciprocal distribution of two words a and b: (1) $\mathscr{S}(a) \subset \mathscr{S}(b)$ (where $\subset$ means that the inclusion is strict); in this case we shall say that a and b are in *defective distribution*. If Γ is the French vocabulary and Φ is the set of well-formed French sentences, then $a =$ *grand* and $b =$ *mince* are in defective distribution. (2) $\mathscr{S}(a) \cap \mathscr{S}(b) \neq 0, \mathscr{S}(a) - \mathscr{S}(b) \neq 0 \neq \mathscr{S}(b) - \mathscr{S}(a)$; in this case we shall say that a and b are in *equipollent distribution*. If Γ is the English vocabulary and Φ is the set of well-formed English sentences, then $a = a$ and $b =$ *the* are in equipollent distribution. (3) $\mathscr{S}(a) \cap \mathscr{S}(b) = 0$; in this case we shall say that a and b are in *complementary distribution*. (4) $\mathscr{S}(a) = \mathscr{S}(b)$; in this case a and b are in *identical distribution* (that is, they belong to the same distributional class).

The most frequent type of distribution in a natural language is that of equipollent distribution. But the three other types are very significant from the linguistic point of view. Let us consider, for instance, the French word *grand*. It is an adjective with values singular and masculine. The words which belong to S(*grand*) are also singular, masculine adjectives, but there are singular, masculine adjectives which do not belong to S(*grand*); such adjectives are *mince*, *large*, *maigre*, and others. It is possible to find a formal procedure which detects all adjectives with the values singular and masculine? The answer is affirmative and involves

the consideration of defective distribution. Indeed, let us consider all adjectives a such that *grand* and a are in defective distribution. Denote by $\mathscr{A}(grand)$ the set of these adjectives. The union $S(grand) \cup \mathscr{A}(grand)$ contains all adjectives with the values singular and masculine for two reasons. First there exists no word a such that a and *grand* are in defective distribution; second, *grand* and a are in defective distribution if and only if a is a singular, masculine adjective and $a \notin S(grand)$ since a must have a greater morphologic homonymy than *grand*.

The above considerations may be generalized. Consider, in a natural language, a word b for which no word a exists such that a and b are in defective distribution. Then, the union $S(b) \cup \mathscr{A}(b)$, (where $\mathscr{A}(b) = \{a; b$ and a are in defective distribution$\}$) is exactly the set of words whose set of values contains those of b.

The complementary distribution is very important in the phonological descriptions, where two individual sounds which differ only by their position (such as an initial a and a final a) are in complementary distribution [17, 36, 37].

5. *P*-Structures; Derivative of a Partition

A more complex concept considers a language to be a triple $\{\Gamma, P, \Phi\}$, where P is a partition of Γ other than into distributional classes. Formally, we may also admit the possibility that $S(x) = P(x)$ for each $x \in \Gamma$, but this situation is of no linguistic interest.

In a language with paradigmatic structure there are three species of properties: (1) properties of a purely distributional (syntagmatic) character, which involve only the sets Γ and Φ (such properties are, for instance, those discussed in the preceding section); (2) properties of a purely paradigmatic character, which involve only the set Γ and the partition P (such properties appear, for instance, in the description of flectional forms in Latin, Russian, and other flectional languages; see a model description of these phenomena in [25] and in Chapter III of [24]); (3) properties of a mixed character, which involve all three components Γ, P, and Φ. We are concerned in the first five chapters of this book especially with properties of the third species. Thus we need some preliminary notions and propositions.

If P is a partition of Γ, each set of P will be called a *cell* of P or a *P-cell*.

If the partition P is written

$$\Gamma = \bigcup_{i=1}^{n} P_i;$$

then each P_i denotes a cell of P and the number of cells is equal to n. Since the sets P_i are mutually disjoint, each word belongs to a single cell. We denote by $P(a)$ the cell of P containing the word a. It follows that, for two distinct words a and b, we have either $P(a) = P(b)$ or $P(a) \cap P(b) = 0$.

As we have remarked, the customary interpretation of the set $P(a)$ in a natural language is the consideration of $P(a)$ as the set of flectional forms of the word a. This situation suggests the introduction of the so-called *unit partition* of Γ, in which each cell is formed by a single word. With the interpretation just adopted for P, a language whose partition P is the unit partition is a language without morphology; following traditional terminology used in the classification of natural languages, such a language will be called an amorphic language (for instance, Chinese). This type of language will be studied in Chapter II.

Another simple partition of Γ is the *improper partition*, which has a single cell identical to Γ.

The starting point of linguistic analysis is the unit partition of Γ. Each process of abstraction involves an equivalence relation which leads to a partition with fewer cells. This situation makes the following definition natural.

Let us consider two partitions P and Q of Γ. We shall say that P is finer than Q if $P(a) \subseteq Q(a)$ for each $a \in \Gamma$.

The unit partition is finer than every other partition of Γ, and each partition of Γ is finer than the improper partition. If we interpret $P(a)$ as the set of all flectional forms of a, partition P seems to be finer than the partition of Γ into the parts of discourse. This idea will be expanded in Chapter III.

If $x_1x_2 \ldots x_n$ is a string over Γ, the sequence $P(x_1)P(x_2)\ldots P(x_n)$ is called the *P-structure* of the string $x_1x_2 \ldots x_n$. If $P_i \subseteq \Gamma$ for $1 \leq i \leq s$ and there exists a string $x_1x_2 \ldots x_s$ over Γ, such that $P_i = P(x_i)$ for $1 \leq i \leq s$, then the sequence $P_1P_2 \ldots P_s$ is called a *P-structure*. This *P-structure* is *marked* if the string $x_1x_2 \ldots x_n$ may be chosen so it belongs to Φ. In other words, the *P-structure* $P_1P_2 \ldots P_s$ is marked if there exists a marked string $x_1x_2 \ldots x_s$ such that $P_i = P(x_i)$ for $1 \leq i \leq s$.

The *P*-structures may be composed by concatenation. This operation leads to a new *P*-structure.

Let us consider two *P*-structures $\mathscr{P}_1 = P(x_1)P(x_2)\ldots P(x_n)$ and $\mathscr{P}_2 = P(y_1)P(y_2)\ldots P(y_m)$. We shall say that $\mathscr{P}_1$ and $\mathscr{P}_2$ are *P*-equivalent and we