I

THE LATTICE PROPERTIES OF SYNTACTIC RELATIONS IN AN OPEN LANGUAGE

A. F. PARKER-RHODES

See A General Procedure for Syntactic Description by
A.F. Parker-Rhodes
Cambridge Language Research Unit, M.L. 143

ABSTRACT

In this part, the underlying concepts of the type of grammatical analysis used in the work are defined; they are those of the "immediate constituent" model of the linguists, modified so as to allow of non-contiguous word-sequences being admitted as recognisable units. The basic terms used are "substituent" (replacing the linguists' term "constituent", which has a slightly different meaning), and "paradigm", by which we mean the overall set of uses of a given substituent. Using the definitions given, it is then shown that the set of all the paradigms possible in a language form a lattice; the ordering-relation in this lattice is that of set-inclusion, the paradigms having been defined, with this in view, as sets of occurrences of the substituents to which they refer.

PART I. THE LATTICE PROPERTIES OF SYNTACTIC RELATIONS IN AN OPEN LANGUAGE

Introduction

This section describes briefly a new model of grammatical description, devised originally with the purpose of providing a better tool for the machine processing of language material. Particular attention has been given to the advantages likely to accrue, for this purpose, from exploiting to the full whatever features could be found in common between all languages. The need to devise a new model became apparent when it was found how little attention had been given in the past to this point.

It seems that previous models of grammatical description fall into four main classes. The oldest of these, which has bean called by Hockett (1) the "Word-and-Paradigm" or WP model, originated in antiquity, and is well adapted to the description of inflected languages like Sanskrit, Greek and Latin. It is however, despite Robins' (2) recent reconsideration, far too limited in scope for our purposes. The next, the "Item-and-Process" or IP model in Hockett's terminology, works with the notion of items (words or short phrases) being modified by various processes (suffixation, vowel-change, root-replacement, &c) to produce all the various forms of the language. This model was first clearly systematized by Sapir (3); it is more adaptable than the WP model, but still not sufficiently general. The "Item-and-Arrangement" or IA model was evolved by descriptive linguists; it aims to describe the whole grammar of a language in terms of lists of items and of the ways in which they can be arranged (i.e. constructions). This model lends itself better to expressing the basic hierarchical structure of sentences, first recognised clearly by Husserl (4), than the previous models, and is somewhat easier to formulate mathematically; but it runs into numerous difficulties which have led to the formulation of yet another type of model.

This is the one originated by Harris (5) and greatly strengthened by

Chomsky (6); we may call it the Kernel-and-Transformation or KT model. It takes as its starting point, a number of simple standard sentence forms, called "kernels", and seeks to derive every possible correct sentence in the language by developing these kernels through a mechanism of substitution of their components by other kernels. This model has a number of advantages, notably in the description of what I here call interrupted substituenta, but it is very refractory to mathematical formulation. This model has received a more extensive application to problems of handling language material and mechanization of language processes than the others. This work is especially associated with the University of Pennsylvania, where it has been Ingeniously used by Hiž (7) and by Kaufman (8). Unfortunately the great complexity produced by these efforts, even though they have been confined to the description of a single language (English) casts some doubt on the effectiveness of the KT model for our purposes.

The new model which I propose here, for the purpose of meeting the needs of machine translation better than those previously have done, will set out so far as possible in an axiomatic manner, in order to emphasize its internal structure. The task of demonstrating in detail its application to the description of actual languages lies outside the scope of this paper. Evidence that it is so applicable comes from two sources: first, the operation of machine programs embodying ideas drawn from the model for the syntactic analysis of texts; and second, descriptions of various particular languages capable of being compared with each other and with more conventional descriptions. Evidence of both sorts is planned for publication in due course; here, I shall confine myself to exposition alone.

First, I shall define an operation called "replacement" by which parts of utterances may be substituted by other parts: this does no more than restate familiar ideas. Second, I shall use this operation to derive a rigorous definition of grammatical function (in a partly mathematical context this term, unfortunately, is too liable to be misunderstood, and must

be replaced; I use the term "paradigm" in an analogically extended sense for this purpose). Third, I show that the set of all possible paradigms (functions) constitutes a wall-defined mathematical system, namely, a lattice; this makes possible major simplifications in the description of syntactic phenomena.

The Concept of Replacement

Replacement in a Closed Language

We consider a closed language as being a closed corpus consisting of a set of utterances; each utterance is a sequence of signs having a beginning and an end. The signs in any such sequence are understood to have a unique simple ordering. Each sign may be a written letter or Ideograph, or a sound; there are thus various possibilities for the realization of the signs, and in some realizations it may be necessary to resort to special conventions in order that they may be unambiguously assigned a simple ordering; this however is a matter which at the present level of discourse need not be pursued in detail.

Any subset of the signs constituting an utterance, presented in the same order in which they occur in this utterance, is called a segment. If S is a segment of an utterance U, and if between the first and the last sign included in S, every sign in U is also a sign in S, then S is said to be an uninterrupted segment; otherwise, S would be interrupted. We shall have occasion to use the notion of a zero segment, that is, one consisting of no signs; just as the empty set, in set theory, is understood to be a subset of every set, so we shall admit the presence of an empty sub-segment in every other segment. In all the statements which we shall make about segments, the possibility that a zero segment may be referred to should be borne in mind.

If an interrupted segment consists of n sub-segments, each of which is Itself uninterrupted, the latter will be called fragments to distinguish them from general sub-segments, which may be themselves interrupted. A fragment.

being itself a segment, may also on occasion be a zero segment. We shall use, as a general form for denoting a segment, $.F_1...F_2...$, where F_1 and F_2 are fragments of an interrupted segment. Whenever such a form is used, it must be understood that though two fragments are shown, more than two fragments may in fact be present.

A segment $...F_1...F_2...$ is said to be replaceable by another segment $...F_1...F_2...$ if the following two postulates are fulfilled:

(a) for any X,Y,Z such that $XF_1\ YF_2\ Z$ is an utterance in the language, $XF'_1YF'_2Z$ is also an utterance in the language;

(b) for any $...G_1...G_2...$ in the language, of which $...F_1...F_2...$ is a sub-segment, there is at least one utterance of the form XF_1YF_2Z in the language, which does not contain . . . G_1. . . G_2. . .

The second condition is required to avoid saying that one segment is replaceable by another when they are only so as parts of larger ones. A closed language, as defined above, is a rather unsatisfactory model of actual speech. At the very least it needs to contain an enormous amount of material if it is to provide examples of all possible constructions. Furthermore, in a strict sense, the set of "possible constructions" in any actual language is an open one in that any speaker may coin a new construction without thereby ceasing to speak the given language. We therefore need to pass over from consideration of closed languages, to take account of open languages.

An open language is, like a closed language, considered as a set of utterances. But whereas in a closed language these utterances form an ostensibly given corpus, which can be examined to determine whether a given sequence is or is not an utterance, in an open language the criterion is, whether or not a given sequence is accepted by a competent speaker as a correct utterance in the given language. The definition of replaceability given above, needs modification in three particulars, in order to adapt it for use in an open language. We have to re-define the term 'segment'; we

have to consider carefully what is implied by a sequence being an utterance; and we have to re-phrase the definition of replaceability.

Segments in an Open Language

In effect, we are trying to substitute, for the closed corpus of a closed language, the behavioural response of a competent speaker, to define the compass of an open language. This being so, we cannot simply regard a segment as a sequence of signs, unless we admit as "signs" not only written marks and spoken sounds, but any sensory clue available to the competent speaker during the act of communication. We therefore regard all such clues as imaginary diacritics which could be added to the manifest signs composing a given utterance or segment. In other words, we allow our competent speaker to annotate any text before we subject it to further analysis.

The scope of such annotations may be illustrated by the example of the English phrases "you and not me" and "shorthand notes". Both, as they stand, are sequences of written letters, both can be parts of utterances in English, and both contain the uninterrupted sequence "and not".
By the definition above, this sequence is certainly a segment, of which both phrases contain exponents. We rely on the annotations or diacritics which a competent speaker might add, to recognise that the two letter-sequences are effectively different. This might, for example, be done by underlining the first and last letters of every word, in which case the two sequences would be "and not" and "and not". The particular device adopted does not matter, provided (a) it can be non-contentiously performed, and (b) it leaves the annotated text capable of complete analysis on the assumption that, if a segment S is replaceable by a segment T, S and T are sufficiently identifiable by the sequences of signs (including the diacritics) which they contain.

If this principle is applied to actual texts in actual languages, it is possible to find cases where it breaks down. These are cases of

irreducible ambiguity. An example is the sentence "Iceland fish catch drops": it is more than a competent speaker can do to annotate this text so as to distinguish non-contentiously all the meaningful segments in it. For it can bear two distinct meanings, which only a fuller context could disengage : either it concerns animal behaviour, or the fishing industry, according as "catch" or "drops" is taken as the verb. It is therefore necessary to prescind such cases of irreducible ambiguity in the rigorous analysis of open languages.

Recognition of Utterances

Whereas in a closed language, every sequence of signs either is or is not an utterance, there are four cases which may have to be considered in regard to open language.
These are exemplified by the following phrases :

1. "It's a nice morning" ; This is an utterance in English.
2. "I'se hungry" ; Not an utterance : the correct form is "I'm hungry".
3. "Lake three stand" ; Not an utterance, no comments occur.
4. "Verns hollip" ; Undecidable.

There is no novelty about either (1) or (3). The new cases not paralleled in a closed language are (2) and (4). The last is in fact peculiarly tiresome, in that there are in real life speech situations in which this phrase could be accepted as an utterance, and meaning could be attached to the words "vern" and "hollip". But in the context of any mechanical language processing we have to regard it as not an utterance, because it must always remain unrecognisable, until the words it contains get into the dictionary. The case (2) can be more constructively treated. We shall formulate the following definition :

Defn. 1. a sequence S in an open language L which differs from some utterance S' in L, if at all, in such a way that in the given context, a competent speaker of L will unambiguously identify S with S', is said to be corrigible to S', which is called its correction. Two different sequences

both corrigible to the same utterance are said to be not distinct.

This definition has been so formulated that it applies to the oases (1) and (2) of the above list, but not to (3) and (4). Its effect is, that in open languages the class of corrigible sequences will take the place occupied by utterances in closed languages.

Redefinition of Replaceability

The definition given for replaceability in a closed language was based on two postulates. The first of these, when its terms are interpreted in the light of what has been said above about segments and utterances, can stand. The second, aimed to exclude recognition of replacement between segments which are 'really' parts of larger segments, between which the replaceability relation is more usefully posited, requires amendment. For in an open language it is no longer sufficient, in order to exclude this situation, to find one instance to the contrary, or even a closed set of instances. Thus, in English, we could say that "ga" is replaceable by "ra", adducing instances in which "gain" is replaceable by "rain" ; this is not any the less silly because we can add a few other instances of the same replacement, such as "gate" being replaceable by "rate". Only if there is an open set of such cases can we count the replaceability as genuine.
We are therefore led to the following revised definition:

Defn. 2. : a segment . . . F_1 .. . F_2 ... in an open language L is replaceable by another segment ... , G'_1.. . G'_2... if and only if:

(a) for any X,Y,Z in L such that $XF_1\ YF_2\ Z$ is an utterance in L, $XG'_1YG'_2Z$ is a corrigible sequence in L;
(b) for any two distinct utterances $XF_1\ YF_2\ Z$ the corresponding $XG'_1YG'_2Z$ are also distinct, and
(c) for any segment ...G_1 ...G_2 ... containing ...F_1 ...F_2 ... as a proper sub-segment, there is an open set of utterances $XF_1\ YF_2\ Z$ not containing . . . G_1...G_2 . . .

Total Paradigms

Equipollence

As defined above, replaceability is an asymmetrical relation. It can happen that a segment S' can replace another segment S while S cannot replace S'. For instance, we can readily show that in English "them" is replaceable by "gypsies". But we cannot replace "gypsies" by "them". For if we make this replacement in the utterance "the gypsies came", we get "the them came". If this is accepted as corrigible, its correction can only be "they came". But, "gypsies came" is also an utterance, distinct from "the gypsies came". If we make the proposed replacement we get "them came" which is corrigible, but again corrects to "they came". It is not therefore distinct from "the them came" according to Defn. 1. The replacement therefore fails to satisfy postulate (b) of Defn. 2.
Nevertheless, it is easy to define a symmetrical relation, based on the replacement idea, as follows :

Defn. 5. two segments S,T in L are said to be equipollent if S is replaceable by T, and T by S, in L.

This relationship of equipollence is analogous, at the syntactic level, to that of "replacement" as defined by Jones (12) in regard to semantics. Like the latter, equipollence is a similarity relation; for it is reflexive (every segment is equipollent with itself), symmetrical (by definition), and transitive (for if S is replaceable by T, and T by U, then S is replaceable by U; and conversely). It therefore divides the class of segments in a given language into classes, whose members share common syntactical properties, Just as Jones' "replacement" divides the class of lexemes into classes whose members share a common meaning.

Substituents

However, not all sequences in a given language are either utterances or segments of utterances; likewise, not all segments are recognisable, either by a "competent speaker" or by a trained linguist, as meaningful

units of speech. In order to be able to isolate those segments which can be profitably used as units in the syntactic analysis of a text, we need to define a certain subclass of the domain of equipollence which shall contain only those segments which are useful for this purpose.

Defn. 4. a segment S, interrupted or not, is said to be a substituent in a language L if there is at least one segment T in L, distinct from S, such that :

(a) T is equipollent with S;

(b) there is no sequence U of segments U_1 U_2,... such that:

(b1) for every U_i there is at least one segment V_j in L distinct from and equipollent with U_i, and

(b2) the sequence U is corrigible to T.

The effect of this definition is to recognise as a substituent only segments which are equipollent with simple substituents, i.e. those which are unable to be further divided into substituents. Roughly speaking, this allows any meaningful unit up to a sentence to be a substituent, since sentences are in general equipollent with single units like "yes" or "no", and in all languages there exist sentences of so formal and stereotyped a character as to be admissible as simple lexemes. For instance, we do not get a true picture of the meaning of "How do you do?" if we analyse it into its component parts; such a sentence, while certainly equipollent with genuine sentences like "How is your stomach?" is a perfectly good candidate for inclusion as a whole in a dictionary.

It is convenient for some purposes, also, to recognise any two or more sentences as equipollent with a single sentence; if this is done, the restriction (b) in Defn. 4 is hardly needed. However, we aim eventually to consider the syntactic relations between the sentences in a paragraph or conversation, and for this purpose we must make a fairly clear distinction between "sentences" and higher units which Defn. 4 succeeds in doing.

The reason for introducing corrigibility into the postulate (b2) is to

allow for words like the French "au" which while apparently simple substituents (in that they cannot be analysed as they stand into smaller substituents) are inexpedient to admit as such, because in reality they are compounded of units having separate and definable functions in the sentence. But of course there exists the sequence "a le" which, though not a segment in French, is certainly corrigible to "au", and which is a sequence of segments each equipollent with at least one other ("a" with "dans"; "le" with "un"). The reason why we do not want to have to treat "au" as a single substituent, is that in an expression such as "au fond" we would like to recognise as substituents not "au" and "fond" but the more logical pair "a" and "le fond". In bracket notation we would wish to analyse "au" into "a (le ...)".

The following supplementary definition therefore suggests itself for use in connection with substituents:

Defn. 5. a substituent of S in L is said to be compound if it is the correction of * a sequence U of segments U_1 U_2,... such that:

(a) each U_i is a substituent in L, and

(b) the sequence left on replacing any one U_i by the zero segment is also a substituent in L.

In such a case, the segments U_1 ,U_2,... are the components of S.

* Note here, that by Defn. 1 every segment is its own correction.

Paradigms

We have already mentioned that equipollence is a similarity relation dividing any subclass of its domain, and in particular the class of substituents, into equivalence classes. Members of any one of these classes would be said by linguists to have the same syntactic function. However, the following definition proves to be more amenable to our purposes :

Defn. 6. the total paradigm of a substituent S in a language L is the set of all substituents in L which contain either S, or another substituent equipollent with S, as sub-segments.

It is part of the method of this work to replace the unsatisfactory unit of the "word", already abandoned by most linguistic schools, by the carefully defined concept of "substituent". It is this replacement which justifies the use of the term "paradigm" in this sense. It will shortly appear, that those members of the total paradigm of a "stem" (which, in an inflected language, is in general a simple substituent) which are "words" in the conventional sense form a set almost identical with the "paradigm" in the traditional linguists' sense.

It is evident that if two substituents S, T are equipollent, then according to Defn. 6 they must belong to the same total paradigm. Moreover, if T is not equipollent with S, then either (a) T contains S as a proper subsegment; in which case S which is contained in the paradigm of S, is not in the paradigm of T; or (b) S contains T, with the complementary effect; or (c) neither S nor T contains the other : on which case both paradigms contain substituents not in the other. Therefore, if S, T are not equipollent, they belong to different total paradigms. Thus, the total paradigms defined in Defn. 6 are Indeed equivalence classes under equipollence.

The relation between total paradigms, and the syntactic functions of the linguist, is now clear. If any two substituents belong to the same paradigm, then they share a common function. If they belong to different paradigms, they have no common function, unless their paradigms have a non-trivial union, in which case the latter provides them with a common function. We therefore postulate a one-to-one correspondence between syntactic functions and total paradigms; the first are the properties which characterise the second as classes.

However, in formal statements I shall prefer the term paradigm to function, on the grounds that the latter word has too many other uses not entirely excluded by the context. I shall normally drop the epithet "total" before "paradigm", where no confusion is likely to follow.

Thus, while we have this simple relationship between our total paradigms

and the relation of equipollence, their structure under the relation of replaceability is somewhat more complex. It may be reduced to the following five Lemmas:

Lemma 1. If a substituent A replaces both B and C, where B, C are not equipollent with each other, then the paradigm of A is the set union of those of B, C.

Lemma 2. If a substituent A consists of two or more segments, each a substituent, B, *C,* ..., the paradigm of A is included in each of those of B, C, ...

Lemma 3. If there is in a language L a substituent Z such that any other substituent Z' containing Z is equipollent with Z, then the paradigm of Z is contained in that of every substituent.

Lemma 4. If there is in L a segment which can replace every other substituent in L, then this segment is a substituent in L, and has a paradigm including those of all other substituents in L.

Lemma 5. The paradigm of any substituent is unique (provided we take due account of the procedures mentioned in S 2. 1).

The substituent Z mentioned in Lemma 3 is exemplified by a complete sentence not forming part of any other sentence and associated only by concatenation with other segments in an utterance. Formally we may state the following:

Defn. 7. A substituent in a language L is a free sentence of L if it is a component of an utterance equipollent with the whole utterance.

The segment mentioned in Lemma 4 is exemplified by a sign of omission such as ..., or a word such as "thingummy" used to replace any word which a speaker will not trouble accurately to recall.

The Syntax Lattice

The above five lemmas are sufficient to prove that, if we assume the existence of the substituents postulated in (3) and (4), the system of all the total paradigms is a lattice under the set-inclusion relation. For if S, T are any two non-equipollent substituents, their respective paradigms have, potentially, a join defined by (1) and a meet defined by (2), while the bounds of the lattice are provided by (3) and (4); these points satisfy the definition of a lattice (see Birkhoff (9)).

The preceding argument has established that the syntactic structure of any language, expressed as the system of paradigms, can be represented by a lattice. We have not yet been shown what lattice. The actual form of the syntax lattice can be established either empirically, by applying the definitions given above to a corpus of texts in a given language; or a priori, in the way exemplified in Part II. An empirical construction for one language would not have evidential value for any other language; the fact that the lattice can be constructed from non-empirical considerations shows that in fact, at the least, the syntax lattices for different languages must have a great deal in common. This is so important a conclusion, especially for the development of M.T. procedures, that it seems proper at this point to present the argument leading us to the form of lattice which we shall use in the applications of the model at the programming level.

Whereas in the preceding part the term "paradigm" has been used, to obviate collision with the mathematical meaning of the term "function" more familiar to linguists, it seems best now to revert to the more normal usage; so that what we have up to now called "total paradigms" we shall henceforward call "syntactic functions" of the various substituents; and the word "function" will continue to have this meaning, even where the adjective "syntactic" is omitted.

References

1. C.F. Hockett : "Two Models of Grammatical Description", WORD, 10, 1954.

2. R.H. Robins : "In Defence of 'Word-and-Paradigm' ", TRANSACTIONS OF THE PHILOLOGICAL SOCIETY, 1959.

3. E. Sapir : "Language", New York, 1921.

4. E. Husserl : "Logische Untersuchungen", Halle, 1915.

5. Z. Harris : "Methods in Structural Linguistics", Chicago, 1955.

6. N. Chomsky : "Syntactic Transformations", The Hague, 1957.

7. H.I. Hiz : "The Intuitions of Grammatical Categories", University of Pennsylvania Transformations and Discourse Analysis Projects, No. 29.

8. B. Kaufman : "Iterative Computation of String Nesting", University of Pennsylvania Transformations and Discourse Analysis Projects, No. 20.

9. G. Birkhoff : "Lattice Theory", American Mathematical Society Colloquium Publication, 2nd Edn., 1948.