[*Automatic Translation of Languages*: NATO Summer School…July 1962]

SYNTAX †

A. F. PARKER-RHODES

Cambridge Language Research Unit, Cambridge, England

Summary—After a historical introduction, a notation is described by means of which the syntactic structure of any sentence in any language in terms of immediate constituents can be expressed. The notation is designed so as to give complete information as to the componency and equipollence relations of the units recognizable within the sentence. These terms are defined: equipollence is the relation between words or groups which have the same syntactic function. Rules for this notation are given in tabular form.

Since the notation is adapted for application to all languages, it provides a set of interlingual forms, the syntax of which forms can be ascertained from the rule of the notation, and which in its turn implies a great deal about the syntax of the various source languages which are reduced to this common form. We first apply the rules of the notation to extract a mathematically well-defined system of syntactic functions. These functions are identified with "total paradigms" which term is carefully defined. The system of total paradigms is found to be a lattice; the particulars of the lattice are given and sufficient of it is diagrammed to enable the whole, which is too large to show at once, to be constructed if desired.

A classification of substituent types ("substituent" being the term which we use in place of the "constituent" of the immediate constituent model) is now derived; this is based on the function and constitution of the substituents classified. Both properties are defined and classified on the basis of properties of the syntax lattice already discussed. It is shown that the total number of substituent types which any language might require is fairly certainly less than thirty; actual languages do not seem to require the recognition of more than about a dozen of these. With the use of this modest repertory of basic types, a sufficient classification of the words of any language can be undertaken.

The words of a language are classified on the basis of whether and how they participate in each of the substituent types which the given language possesses. This classification is encoded in a system which assigns one digit to each substituent type in the dictionary reading of each word; one of these digits may contain from one to perhaps seven or eight bits, but probably five or even four may be sufficient. Thus coded, the information is readily utilized in an algorithm designed to ascertain the syntactic structure of a given sentence. The possible variety of strategies for use in such algorithms is discussed briefly.

Examples of many of the preceding points, drawn from the English language, are given in full in three Appendices.

1. THE BASIC CONCEPTS

1.1. Historical introduction

THE PURPOSE of this paper is to maintain that there is a considerable area within the field of syntax which is common to all languages, and that an approach to the syntactic description of particular languages which emphasizes this interlingual element is likely to be more useful than one which ignores or conceals it. The purpose for which these ideas were first developed was that of automatic translation: it is perhaps obvious that, if there is an interlingual element in syntax, its exploitation could make possible the construction of at least part of a translation program which could be incorporated in every two-language routine, and so save effort and complication in the overall procedure. But it is also likely that explicit recognition of the interlingual part of syntax will help in the general study of language, that is, at the pure linguistic level, and even perhaps in certain levels of language teaching.

† Presented at the NATO Advanced Study Institute on Automatic Translation of Languages, Venice, 15-31 July 1962.

My enterprise in fact reduces to finding a model of grammatical description which will give us as much detailed information as possible *before* applying it to any particular language, subject, of course, to the overriding requirement that we can still attain a correct and sufficient description of the particular language. A survey of the literature soon shows that few if any of the models which have been proposed satisfy this aim very well.

The subject of grammatical models was well reviewed by Hockett [1]; though since the appearance of this paper at least one important class of grammatical model has been developed, named below as the KT type. Historically, the oldest model for grammatical description was the Word-and-Paradigm or WP model. This originated with Apollonius Dyscolus [2], and held the field until modern times, when in the face of wider knowledge of language types ill-suited to its limited categories it fell out of favour; but it has strong points still, as Robins for instance [3] has shown. The first model constructed in modern times was the Item-and-Process or IP model, definitively formulated by Sapir [4]. This was followed by the now more popular Item-and-Arrangement or IA model, often called in the terminology of Wells [5] the 'immediate constituent' model. This model was the first explicitly to recognize the fact (known from long before but not systematically exploited) that every sentence in every language can be represented as a hierarchy of constituents. However, to make this description logically watertight one must allow that a constituent may be discontinuous, or be a zero form: these complications make the principle far from easy to apply in practice, and are the main reason why the model has been less explored than it deserves to be. With slight modifications this model forms the basis of the description developed in these lectures.

The most recent addition to the repertory of grammatical models originated in the work of Harris [6], but is now mainly associated with the name of Chomsky [7] who has developed it much further. This is the Kernel-and-Transformation or KT model. Though very useful for various specialized applications, this model seems less suitable for machine translation purposes than the IA type, mainly because it depends on a complete listing and encoding of the transformations by which a given kernel, or prototype sentence, can be expanded and developed into all the possible types of sentence in the language. Moreover it is wholly unilingual in conception, and cannot be adapted for the interlingual description of syntactic phenomena without being modified out of recognition.

Current work on the general theory of syntax is mostly aimed towards formulating grammatical descriptions as mathematical deductive systems. This work cannot be adequately reviewed here; most of it has been done in America and Israel, and is exemplified, among others, by the work of Mooers [8], Lambek [9], and Bar-Hillel and Shamir [10]. From the linguistic point of view much of this work has a rather spurious air, since it applies rather to the artificial “languages” used in programming computers, and similar codes, than to actual spoken language; the guiding principle has come more from pure mathematics than from the actual handling of language material, whether by the linguist or by the librarian, and this is reflected in its somewhat academic flavour. Nevertheless, there will no doubt eventually be a meeting between this school and the kind of descriptive work, more prosaic and utilitarian, which is described here.

Another recent development is the methodological sophistication of the *process* of grammatical description. This trend is associated especially with the work of Halliday [11]. The discipline which this writer develops for the elimination of redundancies and omissions is perhaps more immediately useful to the pure linguist than to the machine translation researcher, but it has imported standards of rigour into the subject which make much

earlier work seem naive, and also shows up how much we mathematicians allow ourselves in the way of simplifications. But Halliday's methods do not tell us, yet, whether and where we have over-simplified; and if we have not, these complexities are of little practical value to us.

1.2. *Basic notions: substituent, equipollence, componency*

The syntactic model which has been developed by the Cambridge Language Research Unit may be characterized as a mathematization of the immediate constituent model. It rests on two basic notions, those of equipollence and componency; upon these is constructed a third, more valuable than either, namely the concept of a substituent.

Componency. It is assumed that any sentence in any language can be represented as made up of a limited number, most often two, of immediate constituents, each of which is said to be a 'component' of the sentence. In special cases one of the components may be zero: thus, in English, in common with some other European languages, the imperative sentence usually has a zero subject. It may also happen that one of the components is discontinuous: thus, in English, we can say 'it is hard to do that' in which the word 'it' serves as an introductory subject, the significance of which is filled out by the complementary subject 'to do that'. Since the presence of the 'it' is required by rule when the subject is an infinitive clause it is preferable to regard it as part of the subject, separated from the rest in order to reduce the expectation-load (which Yngve [12] has discussed under the name of 'depth'), which can become heavy in complex sentences involving infinitive clauses. When, however, cases of zero and split components are allowed for, the notion of componency presents no difficulty. It is necessary to remember, however, that componency is *not* a transitive relation. It is incorrect to say that 'that' is a component of the sentence cited above: for it is a component of the subject which itself is a component of the sentence. A component is always a component of *one thing only.*

Equipollence. Any two components, either of which can replace the other, in any context where both can occur, without the change involving any other alteration in the syntactic structure of the sentence, are said to be 'equipollent'; that is, equipollent components have the same syntactic function. Equipollence is an equivalence.

Substituent. In order to avoid the use of the terms constituent and component in other than their original senses, which may perhaps lead to confusion, I have preferred to coin a new term for entities belonging to the domain of the two relations of componency and equipollence. Such entities I call 'substituents'. A substituent which contains within itself smaller substituents is said to be 'compound'. A substituent which is not compound is 'simple'. Very roughly, a simple substituent is a word, and a compound substituent is a word-group. It is logically possible to extend the notion of compound substituent so as to include inflected words, making stem and endings their components; but in an elementary exposition this serves only to make the ideas harder to follow, and adds nothing to the scope of the descriptive technique as applied to English, which will be my main source of illustrative examples.

There is, however, one type of substituent which we have to recognize, which cannot be described as a 'constituent': this is what I call a syntagmatic substituent, or 'syntagm'. A syntagmatic substituent is an incident in the sequence of words in a sentence which according to the rules of the given language carries some definite syntactic function. Thus, by analysing the sequence of words in the English 'I think he's done it' one can infer that 'he's done it' functions as the object of the verb 'think', that is, as a substituent replacing

a noun; this situation can otherwise be indicated in English by the use of the word 'that' before the affected clause; thus, the syntagm consisting in a verb of appropriate type being followed by a complete clause, is found to be syntactically equivalent to a certain word. It need hardly be pointed out that, in translating from one language to another, it is a common thing to interchange syntagms for words and vice versa.

1.3. *The use of a syntactic notation*

The way in which a generalized description of syntactic structure is derived, from the three concepts described above, will now be explained. The link between the basic concepts and the description is provided by an appropriate *notation.* Since the relations of equipollence and componency, applied to substituents, are so defined as to be applicable in any language whatever, a system of notation can be devised which represents how these notions apply to particular words in a particular sentence and which is applicable to any language. Given such a notation, any sentence in any language can be transformed into the form required by this notation. Thus, any syntactic description which applies to the notation can be applied, following rules which can be explicitly formulated for each language, to the original language from which the sentences were drawn.

The first problem is therefore to devise an appropriate notation. Given that it is sufficient to indicate the boundaries, componencies, and equipollences of every substituent in every sentence, any notation which does this will serve our purpose; in practice, of course, we must also take pains to avoid its being redundant as well as to ensure sufficiency. I shall now give, as briefly as possible, the conventions which I have adopted to represent these basic relations; they are all, I think, fairly simple, and not excessively at variance with common usage.

Boundary of substituent. This is indicated by enclosing the words forming each substituent in brackets (...)• By this means the hierarchy of substituents appears in an obvious way as a nesting of brackets. It should be mentioned that the notation does not seek to conserve word-order: not only is word order one of the most obvious fields in which languages differ, showing it to be no part of any strictly interlingual syntactic system, but the rules applying in any given language are necessarily deducible from the rules which we shall eventually draw up for the reduction of each language into the form provided by the notation, and the generation of grammatical sentences from the notation. Thus, since we do not have to trouble with word-order within the notation, the existence of discontinuous substituents offers no special difficulty at this level.

We do, however, need a convention for dealing with zero forms and syntagms. I propose to write the sign ϕ wherever a zero form has been recognized and inserted, and \$, followed if desired by an index number to distinguish one from another in a given language, for a syntagmatic substituent.

Equipollence. We require a convention by which we can correctly infer, of any two substituents, recognized by their enclosing brackets in the notation, whether they are or are not equipollent. Three cases exist which need separate treatment: equipollence between components of one substituent, equipollence between substituents which are themselves in a componency relation (one being component of the other), and between substituents which are in an indirect componency relation (one being part of a component of the other).

For the indication of equipollence between components of one substituent, we make use of the fact that of the components of a given substituent at most one can differ from the others in its syntactic function. This fact can, indeed, be deduced in an *a priori* manner

from logical considerations: but I prefer to regard it as empirical since this avoids the introduction of a load of conceptual apparatus having little other expository value. I shall illustrate this as usual by an example in English. Can we construct, in this language, a phrase containing words of three different syntactic functions, or at least two words each of two different functions, without producing a form which requires further analysis to exhibit its structure? A little thought shows that any phrase involving three different kinds of words always contains at least one compound substituent. Thus, 'very tall trees' brackets as ((very tall) trees), since the phrase 'very tall' is obviously equipollent with 'tall' by itself. It will soon become apparent, on searching for further examples, that no phrase in English which does not require internal bracketing to exhibit its structure contains more than two different syntactic functions. The same conclusion can readily be verified in any other language.

If, therefore, a counter-example to the principle stated is to be found, it must take the form of a phrase with four words (or more), divided two and two between different functions. Again, a search for examples soon fails. Most phrases of this kind are obviously impossible: other languages than English, notably Chinese, are more tolerant in this respect, but I will do my best with a slightly forced example in English: consider the phrase 'tall dark houses trees'. Imagine that this is possible—it could pass in poetry—and consider how it is built up. Evidently, it is ambiguous: but there is a well-defined set of possible structures which it might have, and all of these are expressible, quite simply, by means of different bracketings. Thus, the phrase could be equivalent to *(a)* (tall houses) (dark trees)—two noun groups in conjunction; or *(b)* (tall dark houses) trees—again two noun groups but differently divided; or again *(c)* tall dark (houses trees)—where a pair of conjunct nouns is collectively modified by the two adjectives. Each of these three forms satisfies the condition that at most one component of any recognized substituent differs in function from the rest. The ambiguity of the given form is thus an ambiguity between alternatives all of which satisfy the proposed rule; it should need no saying that ambiguity as such is *not* a thing which we can hope to avoid.

We can therefore without difficulty make the following rule: that all the components of any one substituent shall be equipollent except the last. If, as in the (houses trees) of the above example, *all* the components are equipollent, we can introduce a zero form as the last, and represent the substituent as (houses trees ϕ). It is convenient, in such cases, to have a term for the two kinds of components, the repeatable and the unrepeatable: I propose to call the former 'dependents' and the latter, of which we have seen there can only be one and, if we adopt the zero-form convention, must be at least one, I shall call the 'governor'. We can later extend the range of these terms to cover cases where the criterion of repeatability cannot be directly applied.

To indicate equipollence between a substituent and one of its own components I shall use a comma placed after the affected component. Thus, to indicate that (tall trees) is equipollent with 'trees' I shall write it (tall trees,). Similarly, I shall write (tall dark (houses, trees, ϕ),) to indicate more precisely the equipollence relations in the particular analysis of the four-word phrase indicated by the brackets; 'tall' and 'dark' are shown to be equipollent by having no bracket between them, but to be non-equipollent with the whole phrase, by having no commas; but 'houses' and 'trees' are both equipollent with (houses, trees, *<f>)* and this in turn with the whole substituent.

A substituent which is equipollent with one or more of its own substituents is called 'endocentric'. One which is not so is called 'exocentric'. Because of the occurrence of

exocentric substituents, we cannot rely on the above two devices to indicate all equipollence relations in a complicated sentence. For example, if we observe the rules established above, the sentence 'I thought that you had it.' can be represented in the form (I (that (you it had)) thought). Because all the substituents are here exocentric, there are no commas, and there is thus nothing to indicate that (you it had) is equipollent with the whole sentence: which it clearly is, because 'you had it' is itself a sentence. We can get over this difficulty by placing a period (.) after any substituent which is itself equipollent with a complete sentence. In this way, the notation for the above sentence becomes (I ((you it had.) that) thought.). This completes the notational apparatus necessary to indicate equipollence-relations.

1.4. *The notational rules*

I have now given a discursive account of the rules necessary to enable the notation to represent correctly and simply the boundaries of substituents, their componency-relations, and the equipollence-relation between them in every case. It may help at this point if I now give the rules in a more formal manner for subsequent reference. After giving these rules, I shall give examples of each rule which seems to require any such illustration, avoiding those which have been already used in the text.

1 Rules for the Writing of Formulae

10 A 'syntactic formula' is a sequence of signs forming either a substituent formula or a sentence formula, in which each sign is *either*

101 a 'term', which is any lexeme in the given language (in conventional spelling or an assigned abbreviation); or any variable taking such as its values; *or*

102 a 'punctuation', which is any one of the signs ")", "(", ".", or "," whose use is defined below.

11 A 'substituent formula' is any term, or any sequence of substituent formulae (called its 'components'), preceded by "(", separated or not by "," or ".", and followed by ")".

12 Any component of a substituent formula

121 is equipollent with the substituent formula if and only if it is followed by the punctuation ",";

122 if not last in the substituent formula, is equipollent with the first component;

123 is not equipollent with any other substituent formula separated from it only by brackets.

13 Any sequence of substituent formulae followed by the punctuation "." is a 'sentence formula'.

Examples

101 'dogs' as in (dogs bark.) is a term;
so also is 'D' as in (D B.);
so also is *'x'* when we say that *x* is equipollent with y in the substituent *(x y z).*

11 'dogs' is a substituent formula;
so also is (barking dogs,) of which 'barking' is one of the components.

121 (barking dogs,) is equipollent with its component 'dogs' but not with 'barking'.

122 in (noisy barking dogs,) 'noisy' is equipollent with 'barking' but not with 'dogs';

123 in (tomorrow (never comes,).) 'tomorrow' and 'never' are not equipollent; the formula (noisy (barking dogs,),) contravenes this rule and is accordingly incorrect.

13 'dogs bark.' is a sentence formula;
so also is '(noisy dogs,) (often bark,).'

2. THE CLASSIFICATION OF SYNTACTIC FUNCTIONS

2.1. *The concept of 'paradigm'*

In ordinary linguistic usage, the paradigm of a given word, or rather of a given stem, is the set of all words which can be formed by regular and predictable processes from this stem. In some cases, forms which are conventionally treated as containing more than one word are admitted as members of a paradigm: thus, in Latin, it is customary to admit forms such as *'functus est'* into the paradigm of *'fungor'*; and in English, many older grammars were content to go through the whole of a verb in this style: 'I beat, I have beaten, I have been being beaten', and so on through the hundreds of possible combinations. I intend here to use the same term in an even wider sense, to include all the substituents in a language which contain the prepositus. Of course, one no longer writes out in full such a paradigm, which is an open set; but, not being limited to the arbitrary though useful hounds of a single word or even word-group, this extended sense of paradigm provides us with a *logically simpler* and more amenable concept than the traditional one.

A paradigm in this sense is essentially a set: a set of substituents, or 'contexts' in a syntactically-defined sense. It therefore provides us with an opening to use the powerful mathematical apparatus of set theory. It is this possibility which chiefly prompts me to extend the meaning of the old term in this way; for it means that if we can once delimit the set of paradigms existing in a language, we shall have before us a system of sets. If we arrange that this system shall include a null set and an all-inclusive set, it will be representable as a lattice, and enable us to bring in lattice-theory, an even more powerful tool than set-theory itself. The concept of paradigm thus enables us to approach the problem of mathematizing the process of syntactic description with greatly enhanced resources.

I shall bring in a formal definition of the paradigm of a given substituent in three stages. First of all, I must make a clear distinction between a *substituent,* which I shall normally regard as a type, and the *occurrence* of a substituent (in a particular sentence in a particular utterance on a particular occasion), by which phrase I shall designate a token of the type. With this in mind, I define the 'chain of determination' of an occurrence S_a of a substituent S as the set of all occurrences-of-substituents of which S_a forms a part. I shall explain this by an example. The substituent (word) 'an' has just occurred in this text; the chain of determination of this occurrence of 'an' contains *(a)* the occurrence itself, *(b)* the last occurrence before the last full-stop of the substituent 'an example', *(c)* the last occurrence of 'by an example', *(d)* the last occurrence of 'shall explain this by an example', *(e)* the whole of the sentence before the last full-stop. Ideally, the chain of determination does not stop here, but goes on to the paragraph, section, lecture and course. But to limit the scope of our enquiry I shall not pursue it beyond the sentence.

2.2. *Occasional paradigms*

The chain of determination thus deals entirely in *occurrences,* whereas a paradigm is clearly a matter of types, that is of the *substituents* themselves. We can make the transition in an obvious manner. Let us look at the chain of determination of the *latest* occurrence of 'an'. It consists of the latest occurrences of the following substituents: *(a)* 'an', *(b)* 'an obvious manner', *(c)* 'in an obvious manner', *(d)* 'can make the transition in an obvious manner', *(e)* the whole of the last sentence but one. Compare this with the chain of determination of the previous occurrence of 'an': it will be noted that corresponding links in the two chains are obviously equipollent substituents. Thus, they are successively, in conventional terms, *(a)* an article, *(b)* a noun group, *(c)* a prepositional phrase, *(d)* a predicate, *(e)* a sentence. We are thus led to define the 'occasional paradigm' of an occurrence S_a of a substituent S as the union of all chains of substituents link-by-link equipollent with the chain of determination of S_a.

Each chain of determination is a set, and their union is also a set. We can also regard the occasional paradigm of an occurrence in another light. For it is what we get if we express the chain of determination of the occurrence without citation of specific substituents, but replace each by some symbol expressive only of their syntactic function as determined by equipollence. (It will be obvious that the relation of equipollence, introduced into the definition of occasional paradigm, being an equivalence-relation, divides its domain into a set of equivalence-classes, which gives us one way of defining what we mean by the syntactic function of a substituent.) Such a chain of functions can be regarded as denoting each and all of the specific chains which could be formed by inserting specific substituents of the functions named, that is, all the chains of determination link-by-link equipollent with the prepositus chain.

The value of the occasional paradigm of an occurrence as it has now been defined is that we can ascertain a great deal about the system of all possible occasional paradigms directly from the rules of the syntactic notation which I have previously described. I shall regard the prepositus occurrence as the bottom element of the occasional paradigm, as of the chain of determination, and conversely the sentence as its top element. Now all sentences can be regarded, for the present purpose, as equipollent with each other. Thus the top element of every occasional paradigm is the same, namely the function of a sentence. I shall denote this function by Z. If we proceed step by step down the chain of determination, we are faced at each step with an often unlimited choice of substituents, but only a small range of possible functions. This follows at once from the notational rule which requires every sentence to be so analysed that it contains components of at most two different functions. There are in fact only four kinds of step which we can meet with as we go down such a chain, each allowing a different pair of functions to the components treated of: these are:

1. Recursive: governor new, dependent equipollent with sentence.
2. Conjunct: governor new, dependents equipollent with compound.
3. Endocentric: governor equipollent with compound, dependents new.
4. Exocentric: both components new.

Where the compound, that is the substituent whose components are represented by the elements in the occasional paradigm to which we are now stepping, is itself a sentence, cases 1 and 2 are identical. Where there are only two components, cases 2 and 3 are identical,

since we have not yet given an unambiguous definition of the governor in such cases. *Equipollent* substituents have *identical* OPs.

2.3. *The system of occasional paradigms*

In a chain of *one* element, we have only one possible OP, namely that which I have already designated as Z for a sentence. Note that I write symbols for OPs underlined: those representing the total paradigms discussed in Section 2.5, which will be of more extensive use when we come to define them, will be distinguished by being not underlined.

In a chain of *two* elements, the link between them may be of any of the three kinds 1/2, 3 or 4 of the above list. In the recursive/conjunct type, the new OP introduced by the governor needs a new symbol; I shall write it ZC. In the endocentric type, the new OP of the dependents will be denoted by S (for substantive). In the exocentric type, in addition to S for the dependents we shall need a new symbol for the OP of the governor, and this I shall write O (for operative).

In a chain of *three* or more elements, each link after the first opens up the possibility of all four different types. In the recursive type 1, since the OP of the compound (which need not now be Z) is not indicated by the inevitable Z of the dependent, it must be indicated by the governor. If the OP of the compound is X, I shall write that of its recursive governor as ZC.X. In the conjunct type, since the OP of the compound is indicated by that of the dependents, we only need one symbol for that of the governor, for which I choose IC (standing for indeterminate conjunction). The last two types require further consideration.

An endocentric compound requires a new symbol for the OP of its dependent; if its own OP is X, I shall write that of its dependent as XA. Since X may be any OP at all, it may be itself of the form XA, and in that case the OP of the dependent will be XAA. (Here A stands for adjunct.) In principle, one could generate an open set of new OPs in this way; but in practice the series soon terminates. This termination is presumably connected with Yngve's principle of limited expectation. While the absolute length of the series cannot be predicted *a priori*, all languages which have so far been examined from this point of view (these are English, German, Russian, Latin, Italian, Chinese) make no distinction between adjunct substituents after the first. That is to say, whatever OP X may stand for, XAA and XAAA are represented by equipollent substituents; thus, if X is S (dependents of a sentence, i.e. nouns), we can distinguish adjectives SA and subadjectives (such as 'very') SAA but at this point the series stops. All subadjectives can be used with other subadjectives as well as with adjectives (as in 'very nearly perfect'). This limitation of the series we can conveniently express in the notation by writing B for every sequence of two or more A's, so that the dependent of an endocentric substituent of OP XA will have the OP XB. The next letter C will then be free for use in connection with conjunctions.

An exocentnc compound whose OP is other than Z requires two new symbols for the OPs of its components. In the same manner that I use ZC.X side by side with ZC for a recursive compound, I propose to denote the governor of an exocentric compound whose OP is X by O.X and its dependent's OP by S.X. This convention, like the last, generates a potentially open set of new OPs; for if X stands for say S.SA (the OP of the dependent

of an exocentric substituent whose OP is SA) then O.X will stand for O.S.SA, and so on. The same principles which dictate the early termination of the adjunct series generated by endocentric substituents also apply here to cause a termination of this series. In practice, there are languages, such as English, in which such ternary OPs can be exemplified (in 'those for whom I speak' the word 'for' has the OP cited above as O.S.SA, being the governor of 'for whom' (equipollent with neither of its components and so exocentric) which is in turn a dependent of 'for whom I speak' which is equally exocentric, and which is the dependent of the whole phrase which is an endocentric noun-group). But if any cases occur of languages which allow of a greater complexity than this, I have not found any examples; many languages, such as Chinese, do not allow even ternary OPs to appear.

We have now the means for constructing systematic symbols to represent every possible OP. The constructions depend on the rules of the notation and through them on the basic regularities which these rules convey. We have therefore an interlingual system of OPs, valid according to the theory for every possible language; since each OP is by definition a set of substituents, we can usefully ask about the inclusion-relations between the various different OPs which we now have provided ourselves with symbols for. Confining attention at first to primary OPs (those *not* of the form X.Y), we observe that:

Z is included in all other OPs (because a sentence is included in every chain of determination).

S is included in SA (because in every chain of determination a noun-substituent stands above an adjective-substituent, but itself need not have an adjective-substituent below it).

SA is included in SB (for an exactly analogous reason).

O is included in OA, and OA in OB.

IC includes all other OPs (because any substituent can be replaced by a conjunct group, so that any chain of determination can be part of the chain of determination of a conjunction).

These relations can be summarized in the form of the following graph, which, it will be observed, satisfies the conditions for being a lattice, the relation of inclusion being represented by a descending line:

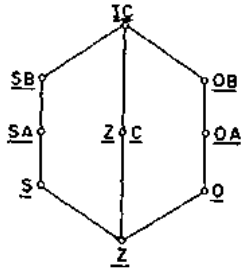


FIG. 1.

I shall refer to this figure as the 'lattice of primary OPs'.

As regards the secondary OPs, it is evident on the same principles that, for any X, both O.X and S.X include X, as does also ZC.X. When a substituent with such an OP as O.X

is itself an endocentric compound, it's dependents will have the OP OA.X, which therefore includes O.X just as OA includes O. It follows that to account for secondary OPs as well as primary ones, we must erect on each point of Fig. 1 another lattice isomorphic with the lattice of primary OPs; and to bring in tertiary OPs we must likewise erect a copy of Fig. 1 on every point of this secondary lattice. We shall thus generate the direct square of Fig. 1, then its direct cube, and so forth. But for the termination of the series already remarked upon the system would be an open one.

2.4. *Rules for assigning OPs*

In the preceding section, what I have done is to go down an imaginary chain of determination, and consider at each step what possibilities are open for choosing the next element, and to assign names to all which the syntactic notation given in Section 1 allows us to recognize as distinct. This, of course, implies a rigorous procedure for assigning an OP to every word or other substituent in a bracketed formula constructed (for any sentence in any language) according to the rules of Section 1.4.1 have shown that the system of OPs generated by the procedure is a lattice, and explained the form of this lattice; but I have not given the rules for assigning the OPs in a form in which they could be applied to an actual bracket-formula. This I shall now do:

2 Rules for assigning OPs.

20 An OP is assigned to every term and to every substituent formula, in the latter case being attached to the "("and")" which bound the formula.

21 A substituent formula followed by the punctuation ".",
- 211 if it is initial in a syntactic formula, has the OP Z,
- 212 but otherwise has the OP O.

22 A substituent formula followed by another having the OP Y, which is in turn followed
- 221 by".", has the OP S:
- 222 by a third substituent formula, has the OP Y:
- 223 by ")" with the OP X, has the OP S.X:
- 224 by ",", has the OP YA.

23 A substituent formula followed by ")" with the OP X and preceded
- 231 by ".", has the OP ZC.X:
- 232 by",", has the OP IC:
- 233 by any other sign, has the OP O.X.

24 A substituent formula followed by "," has the same OP as the substituent formula of which it is a component.

Examples

20 the formula: 'I (come shall,) .'
takes the OPs: S O OA O O

211 in 'Yes.' the term 'yes' has the OP Z.

212 in 'I came.' the term 'came' has the OP O.

22 in: 'I (it doing) (like shall,) .'
S S S.S O.S S O OA O O

221 '(it doing)' is followed by '(like shall,).' [note the "."]

222 'I' is followed by 'it doing' and *this* by '(like shall,).'

223 'it' is followed by 'doing)' with the OP S.

224 'like' is followed by 'shall,' and 'shall' by ",".

23 in: 'He ((I came . when) up got,) .'
S O OA S O ZC.OA OA OA O O

231 'when' followed by ')' with OA and preceded by "." has ZC.OA.

232 in '(men, women, and)' 'and' has the OP IC.

233 in example to 22, 'doing' followed by ')' with S has O.S.

24 in '(men, women, and)' which as a whole has the OP S, the term 'men' has also the OP S.

2.5. *The total paradigms*

In Section 2.2,1 defined the occasional paradigm of an occurrence of a given substituent, essentially as the set union of all chains link-by-link equipollent with the chain of determination of the prepositus. It is, however, more to the point to consider properties of substituents than of their occurrences; I must now therefore introduce a further definition. I shall define the 'total paradigm' of a substituent S as the set union of all the occasional paradigms of its occurrences in the given language.

This definition enables us to derive the system of total paradigms from that of the occasional paradigms already considered. As regards the primary occasional paradigms, whose set-inclusion relations are exhibited in Fig. 1, the procedure is very simple. If a given substituent is used invariably in chains of determination where it has a given OP X, its TP will be defined as being X (note that I distinguish TPs from OPs by underlining the symbols of the latter only). But if a substituent is used sometimes in chains of determination giving it an OP X, and sometimes where is has an OP Y, then, from the definition, its TP must be the set union of X and Y.

It follows at once that the system of primary TPs can be obtained from the lattice of Fig. 1 by adding to it points representing the set-unions of every pair of points whose union is not already shown there. For example, in Fig. 1 the set-union of S and SA is that point which includes both, but does not include any other point which includes both; this point is unique, by the fundamental theorem of lattice theory. In fact, it is SA itself, which, being already present in the system does not have to be added. But what of the union of S and O? In Fig. 1, this union is the point IC, which includes every other point in the system; in this case it is possible, as it was not in the case of S and SA, to insert a new point which includes *only* S and O. Obviously, the number of new union points which can be inserted into the lattice is finite; when the process is completed, the system is found to contain twenty-one points, which still of course form a lattice under the set-inclusion relation. With the same graphical conventions as were used in Fig. 1, the new graph is:

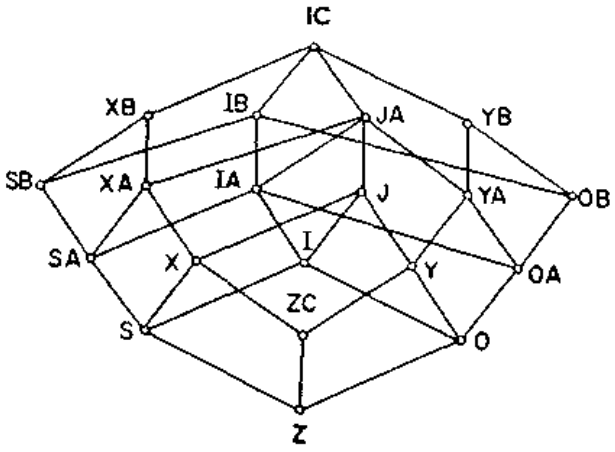


FIG. 2.

This I shall refer to as the lattice of primary total paradigms.

2.6. *The secondary total paradigms*

The above figure represents the primary TPs only, leaving out of account the TPs corresponding to OPs of the type of S.O, O.S.SA, etc. These last I call secondary TPs, tertiary TPs, and so on. Their derivation from the OPs is not quite so straightforward as is that of the primary TPs, for a reason which I shall now explain.

As in the case of primary TPs, I shall define the TP of a substituent which *always* has an OP of the form X.Y as being X.Y. Thus, to each of the secondary OPs denoted by S.O, S.SA, S.SB there corresponds in this direct way a secondary TP, viz. S.O, S.SA, S.SB. Now all these are modifications or special cases of the one primary TP: S. Therefore, their union must be included in S (while, as union, it must include each separate one of the TPs named). But, in this example, the second components of the OPs, which are O, SA, SB, come from both sides of the lattice in Fig. 1, so that their set union is identical with the union of *all* the sets represented in that lattice. This being so, the union of S.O, S.SA, S.SB is identical with the union of *all* the secondary TPs beginning with S, which is simply S itself. Thus we can assert that, in the system of secondary TPs,

$$SO \cup S.S = S \tag{1}$$

We have already seen that the OP of a compound substituent, being just that part of the chains of determination of its several components which is common to all of them, is the intersection, in the system of OPs, of the OPs of the components. In particular, the OP of a substituent whose components are O.X and S.X is S.X $\wedge$ O.X which is X. Since (from their definitions) the system of OPs must be contained as a subset in the system of TPs, it follows that

$$O.X \wedge S.X = X \tag{2}$$

But, as a particular case of this, we have from Fig. 1 that O $\wedge$ S = Z. We may subsume this under the same formula as equation (2) by writing, for the TP corresponding to O, not O but O.Z; and likewise for every primary OP. Thus, side by side with (2) we may write

$$O.Z \wedge S.Z = Z.Z \tag{3}$$

Equally, we can rewrite (1) in the form

$$S.O \cup S.S = S.Z \tag{4}$$

On comparing (3) and (4) we see that when S and O appear as the first components of secondary TPs their intersection is Z; but that when the same points are named as the *second* components of TPs, their *union* appears to be Z. The former relation is that which is given by the lattice of Fig. 2 representing the system of primary TPs alone. The latter would be given by the *dual* of this lattice (i.e. the lattice formed from it by reversing every one of its inclusion-relations: which, keeping to the same graphical conventions, means turning it upside down). Thus, while we found the system of secondary OPs to be the direct product of the lattice of Fig. 1 with *itself*, we now find that the system of secondary TPs must be the direct product of the lattice of Fig. 2 with its *dual*. This product is called the self-dual product or SDP lattice, and contains 441 points. It is too large to be given in full; but, as we shall see, it can be drastically simplified without loss of useful information.

One important property of the SDP lattice must, however, be pointed out here. In this lattice, it is clear without any elaborate calculation that

$$O.X \wedge S.X = Z.X \tag{5}$$

which shows that the point Z.X must stand for the same TP as the point X.Z, which we have already found reason to substitute for the X appearing in (2). A similar relation will clearly be derivable, by the same argument, for every primary TP; thus, every primary TP has not one but *two* distinct representations in the lattice of secondary TPs, one of the form X.Z and the other Z.X. Of these two, it is convenient to adopt the convention that the one with the Z in the second place is standard, and the other deviant. Thus, we have the following rule for finding the TP of a compound substituent: find the intersection of the TPs of its components in the SDP lattice, and if the result is a point of the form Z.X replace this at once by X.Z, but in every other case the result is the TP of the compound.

The lattice of secondary TPs is a finite system, with a definite upper and lower bound. If, therefore, we are to find the TPs of compounds, as the definitions of the paradigms require us to, by taking always the *intersection* of those of the components, we shall find ourselves, as we pass *up* the chain of determination, from the words of a sentence through the word groups of increasing size up to the sentence itself, continually passing from points higher in the lattice to points *lower down;* to this process there must be a limit, if there is no provision in the algorithm for any other operation. It would therefore follow that sentences must have an upper limit of complexity. But the device, which we have seen to follow inevitably from the principles of construction of the lattice of TPs, of providing a dual representation for each primary TP, makes possible a form of the algorithm which does not lead inexorably downwards, but enables us from time to time to pass from a lower to a higher point. This occurs whenever we reach one of the points *included in Z.Z:* this point itself represents a completed sentence; therefore the points included in it represent substituents containing a complete sentence as part of themselves. We thus see that the means by which indefinite complexity of sentence structure is made possible in the lattice is exactly the means which we commonly find applied to this end in the languages familiar to us: the use of subordinate clauses. This is only one of a very large number of fundamental facts about language which are succinctly expressed in the lattice of 441 points: but even so this lattice is quite needlessly large, as I shall next try to show.

3. THE CLASSIFICATION OF SUBSTITUENTS

3.1. *The nature of the problem*

In Section 1 I considered the problem of the classification of syntactic functions. I put forward a solution of this problem based on the mathematical and formal exploitation of a demonstrably sufficient syntactic notation. The resulting classification represented each syntactic function, now identified more clearly as what I call a total paradigm, by a point on a certain lattice. Up to the point at which the discussion was left, this lattice had been defined as the self-dual-product of a certain lattice of twenty-one points (shown in Fig. 2).

For the purposes of machine translation, in relation to which this theory was originally propounded, we are not directly concerned with the classification of syntactic functions, however these are defined. What we need is a means of *recognizing* the function of a given occurrence of a word in a given text. When the function of every word has been recognized, the syntactic structure of the text has been fully discovered. Our aim is therefore the discovery of syntactic structure: that is to say, syntactic analysis. This also is an important aim in general linguistics. Both the ordinary linguist and the language-technologist have to effect this analysis on the basis of suitable given information; this information in the case of machine translation *must*, and in all cases may with advantage, take the form of dictionary-readings upon the smallest conveniently recognizable units or lexemes of which texts in the given language are composed. The practical question, therefore, is not the classification of syntactic functions, but the classification of *words*.

One way of doing this would be to write down, using say the encoding suggested by the above lattice system, *all* the functions which each given lexeme can have in the given language. This list would certainly be sufficient for the purpose. Very roughly, in fact, this is what grammarians in the past have generally done when making dictionaries. The procedure suggested, however, suffers from two serious disadvantages: first, the dictionary readings would be of unequal length, which somewhat complicates look-up procedures and makes the algorithms for effecting the analysis of texts somewhat harder to program; and second, the information is 'in the longer entries at least' very redundant, as well as being in an inconvenient form. We are therefore led to seek a better way of coding it.

The method which has been developed by the Cambridge Language Research Unit for this purpose consists in interposing an extra stage between the classification of syntactic functions and that of words. First, we devise a classification of *substituents* based directly on the properties of the lattice; we then classify the words according to their capacities for participating in the various types of substituent.

The value of this method stems from the fact that the number of types of substituent which the theory provides for us can be cut down without loss of essential information to a figure which is usefully low. Whereas we find in most languages, probably in all, that the number of points in the lattice of secondary TPs which are occupied at least by one word in one expression in the language runs to 100 or more, the number of substituent types is in all cases well under thirty, generally it seems about a dozen. I shall, therefore, devote this section to explaining how we arrive at this classification of substituents.

Briefly, substituents are classified in respect of two properties: function, defined as before by their TPs; and constitution, which is a category based on the distinction already remarked between exocentric and endocentric compounds, but refined in accordance with the concepts afforded us by the lattice representation.

3.2. *The functions of compound substituents*

I have already explained that we take all sentences to be equipollent with each other; that is, they are taken to have the same syntactic function. And I have also pointed out that individual words are capable of a large number of different functions, of the order of 100. These are but the end-points of a general trend, for which there is a sound linguistic basis. The larger the units we consider, the smaller the variety of functions which they need to distinguish. The basic reason for this is the necessity to reduce the increasingly large numbers of possible forms to a limited number of basic types, so that the work of extracting them from the memory as required may be accomplished in the very brief time which normal speaking-rates allow us. An average man may dispose of a vocabulary of say 10,000 words; from these he could make perhaps $10^{10,000}$ sentences, a number impossible to contemplate without a great deal of simplifying devices in the way of recognizable structural procedures.

The way to exploit this diminution in the number of necessary distinctions as we pass up the scale of complexity from word to sentence is to apply to compound substituents intermediate between word and sentence a simplified classification of functions. Statistically, the optimum, if we are to introduce (as we would like to) only one intermediate stage, is to assign to this level the square root of the number of functions required at the word level, or as near to this as the symmetries of the system permit. Thus, if we have 100 word-functions, we should aim to distinguish about ten functions among compound substituents; if we can get away with fewer word functions (as we can) we may expect correspondingly to reduce the number of compound substituent functions also.

This square-root norm fits in admirably with the lattices: we have seen that the number of secondary TPs is the square of the number of primary TPs [only roughly—the actual relation is $n(n\text{-}1)$]. This suggests that we adopt the TPs of the primary series, without the secondary ones, as defining the range of functions to be distinguished among compound substituents. And if we can seriously simplify the 441-point lattice as a representation of the system of secondary TPs, we should be able to use well under twenty-one different functions for classifying compound substituents.

A closer inspection of Fig. 2 shows that this is indeed the case. Many of the distinctions which it allows us to make are clearly redundant. For instance, the point XA represents the total paradigm of a substituent able to function either as a subordinating conjunction (I anticipate the result that this is what the point ZC refers to) or as an adjective, SA. Now, though it is quite likely that some such word exists in some language, it is safe to predict that no language has compound substituents having this function: for since compound substituents are built up out of recognizable and recombinable units the existence of one with a given function implies the existence of a class of them—and there is clearly no utilitarian basis for a whole class of substituents combining these particular functions. If, therefore, we are content to consider only classes of substituents having more than one or two members, we will certainly be able to manage with a smaller lattice than that of Fig. 2; words with unusual combinations of functions will then be regarded as individual exceptions: such exceptions do, of course, occur whatever system of functions we work with, short of one which allows all the combinations of the primary OPs of Fig. 1, of which there are 511.

We obtain, therefore, a more realistic picture of the range of syntactic functions actually occurring in language if we use a simplified version of Fig. 2. The most conspicuously otiose of the points in this lattice are those representing combinations of the function ZC with others: we shall therefore simplify the system by confounding these in groups. The following

is the simplest way of eliminating all these points while leaving the rest undisturbed: confound X, XA, XB with SB; confound Y, YA, YB with OB; confound J, JA with IC. The result is the lattice:

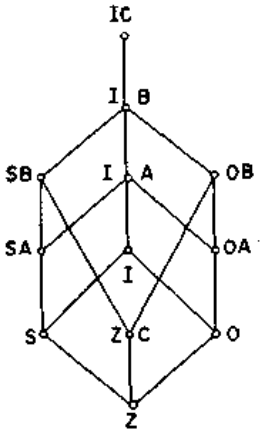


FIG. 3.

If we confine our attention to compound substituents alone, we may expect to find a still more simplified lattice sufficient. As indicated by the argument above, the degree of such simplification which can be accepted ought ideally to depend on how many steps of the hierarchy we are removed from the complete sentence. In practice, it is a needless complication to consider these different degrees of compounding separately; a single system of syntactic functions for application to compound substituents less than a sentence is what we want. How far we take this simplification beyond the stage represented by Fig. 3 is ultimately a matter for empirical test; I am currently working on the assumption that we can classify compound substituents sufficiently by assigning them to one or another of only five basic functions, provided we make provision in each language for a short list of 'anomalous' substituents, of which I shall say more below. These five functions are obtained from those of Fig. 3 as follows: first, we eliminate the 'conjunction' points, IC and ZC, which represent functions which seem to be carried by compound substituents only of rather exceptional kinds if at all; next, we confound SB, OB with IB; and SA, OA, I with IA. The justification for this step is that there is a strong tendency in a wide range of languages for compound substituents having adjunct or subadjunct functions to operate indifferently as components of substantive and operative parts of a clause; it is purely empirical, and the discovery of a language which makes this distinction systematically (or of a language which freely forms compound conjunctions not otherwise analysable) would be sufficient to necessitate a more complex system of functions at this stage. The five-point lattice has the form:

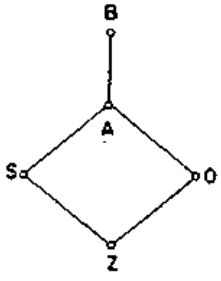


FIG. 4.

3.3. *The constitution of a substituent*

The adoption of the lattices shown in Fig. 3 and Fig. 4, as containing a sufficiently detailed classification of syntactic functions for simple and compound substituents respectively, entails that for the complete system of secondary TPs we shall require not the self-dual product of Fig. 2, but the product of Fig. 3 with the dual of Fig. 4. This lattice, which I call the simplified syntax lattice, contains $12 \times 5 = 60$ points. It is the lattice from which any general classification of substituents should be derived.

In Section 1.2 we saw that in proceeding down the chain of determination of a substituent, there were four possible kinds of step, presenting four different choices for the function of the components of the substituent being subdivided. These four choices were called recursive, conjunct, endocentric, and exocentric. These terms can now be redefined in terms of the syntax lattice. In any product lattice especial importance attaches to *(a)* certain points, called the 'vertices' of the lattice,† defined as the points of the product-lattice which involve only upper or lower bounds of its factors: there are 2^n vertices, where n is the number of prime factors of which the given lattice is the direct product; *(b)* the sets of points which include, or are included by, each of the vertices: these sets are called the 'upper' and 'lower ideals' of the vertex concerned. We can use the vertices of the syntax lattice and their ideals for defining the constitution of a substituent. The great advantage of presenting the matter in this way is that the definitions will hold irrespective of the degree to which either factor of the syntax lattice is simplified.

The 12×5 syntax lattice, having two factors, has four vertices, denoted by IC.Z, IC.B, Z.B, and Z.Z; of these four, Z.B refers by definition to the same syntactic function as IC.Z and need not be considered further, and IC.B would characterize a conjunction able to join together only other conjunctions or subadjuncts, which is a type of word which is unlikely to be of importance in any language. We shall then consider only the vertices IC.Z and Z.Z. The first of these is the upper bound of the whole lattice, and its ideals are therefore trivial; but the UI and LI of Z.Z are non-trivial and will be used. The following classification emerges:

Both components represented by vertices	'Conjunct Sentence'
One component represented by IC.Z	'Conjunct Group'
One component represented by Z.Z	
Second component in UI of Z.Z	'Endocentric Sentence'
Second component not so	'Subjunct'
Neither component at a vertex	
Intersection of components at Z.Z	'Simple Sentence'
Intersection in UI of Z.Z	'Word-group'
Intersection in LI of Z. Z	'Clause'
Intersection elsewhere	'Anomalous Substituent'

If this classification is compared with the linguistic one given in Section 1.2, we find that *(a)* recursive steps correspond to subjuncts; *(b)* conjunct steps correspond to conjunct sentences or groups; *(c)* endocentric steps correspond to endocentric sentences or groups; *(d)* exocentric steps correspond to simple sentences, clauses, and anomalous substituents. Theoretically, endocentric anomalous substituents and exocentric groups could occur, but

† In some textbooks on lattice theory the vertices of a lattice are called its 'centre': this misleading term should not be used in applications of the theory.

seem to be wanting in the languages studied. The class of clauses can be subdivided further as follows.

In order to have an intersection in the lower ideal of Z.Z, the two components of a clause must have TPs of the form O.X and S.Y respectively; the former component I call the 'operative' one, and the latter the 'substantive' component. The meaning of these terms in relation to ordinary grammar can be seen from the fact that, in Section 1.3, we assigned O as the OP of the governor of a sentence. Now in general a sentence can be made up of a verb and a number of noun groups; thus, the verb, as having the function which can occur only once in a sentence, is the governor, so that the TP O.Z is that which, in our notation, characterizes a verb-substituent of any kind. Now if *both* components of a clause lie in the UL of Z.Z it follows that the clause is simply a sentence; if either component lies outside the UI, whereas the other is an ordinary verb or noun substituent, this will be a special kind of word determining the clause as having a particular function. Thus, a clause whose components have the TPs S.Z (which lies in the UI) and O.B (which lies outside) consists in plain language of a noun-substituent and a word which marks it as having the function, denoted by Z.B ($\rightarrow$ B.Z), which is that of a general qualifier; in English one example of a clause of this kind is a prepositional clause, in which the preposition carries the function O.B. Again, a clause consisting of an ordinary predicate O.Z and a 'marked' subject S.SA will exhibit the function of a noun-qualifier; such a clause in English is a relative clause. Thus, we can classify clauses in the following manner:

Both components in the UI of Z.Z	O.Z+S.Z	Z-clause
Operative component *not* in this UI	O.X+S.Z	O-clause
Substantive component not in this UI	O.Z+S.X	S-clause
Neither component in the UI of Z.Z	O.X+S.Y	I-clause

It will be noticed that what is here called a Z-clause is simply another way of defining a simple sentence. I-clauses are rare, and perhaps are always open to interpretation in terms of concord rules operating on O- or S-clauses: a typical example is afforded by the accusative-infinite construction in Latin. Thus, in 'ilium abiisse dixerunt', it could be argued that *both* 'ilium' and 'abiisse' were forms specialized for use in infinitive clauses, and as such not in the UI of Z.Z which contains only unspecialized and unmarked forms; but equally one could say that 'ilium' is the inflection of 'ille' required by concord when it is the subject of an infinitive clause, in which case we should have simply an O-clause.

Omitting the duplication of simple sentences, and setting aside the case of anomalous substituents for separate consideration, we have thus nine different constitutions definable in terms of those properties of the syntax lattice which are invariant under changes in the number of points admitted in its factors.

3.4. *Elimination of inconsistent combinations*

With nine constitutions and five functions one could distinguish forty-five different types of substituent. Not all of these, however, are logically possible or distinct. Thus, a simple, endocentric, or conjunct sentence must have the function Z; but no other constitution can be combined with this function. Conjunct groups present special difficulties: first, most, if not all, 'conjunctions' can be used to connect substituents of any function at all, so that the function of a conjunct group is an intrinsically less useful idea than the function of other types of substituent; second, they present peculiar problems of recognition, since the limits of the substituents which figure as their dependents often depend on semantic

rather than on strictly syntactic considerations, even in those languages which put in the conjunctions at all (not all do). For these reasons, it is convenient to treat conjunct groups as a special case, not subdivided further in regard to function. With this understood, there remain of the forty-five possibilities only twenty-four which need be recognized in a general classification of substituents.

These twenty-four substituent types are:

1.	Adverbial Group	abbr. Bg	function: B	
2.	Adverbial O-clause	Bo	„	
3.	Adverbial S-clause	Bs	„	
4.	Adverbial I-clause	Bi	„	not in English
5.	Adverbial Subjunct	Bx	„	
6.	Adjunctive Group	Ag	function: A	Ao and Bo are not distinguished in English
7.	Adjunctive O-clause	Ao	„	
8.	Adjunctive S-clause	As	„	
9.	Adjunctive I-clause	Ai	„	not in English
10.	Adjunctive Subjunct	Ax	„	not in English
11.	Nominal Group	Sg	function: S	
12.	Nominal O-clause	So	„	Bs and Ss are not distinguished in English
13.	Nominal S-clause	Ss	„	
14.	Nominal I-clause	Si	„	not in English
15.	Nominal Subjunct	Sx	„	
16.	Operative Group	Og	function: O	possibly not in any language
17.	Operative O-clause	Oo	„	
18.	Operative S-clause	Os	„	possibly not in any language
19.	Operative I-clause	Oi	„	
20.	Operative Subjunct	Ox	„	ditto
21.	Z-clause or Simple Sentence	Zz	function: Z	
22.	Endocentric Sentence	Zg	„	rare in English
23.	Conjunct Sentence	Zc	„	
24.	Conjunct Group	Ic	function indeterminate	

This, then, is the outcome of our attempt to apply the principles of the theory to the classification of substituents. We find we can draw up this list of twenty-four substituent types, and claim that every substituent in every language must belong to one or other of them. It is necessary to be very clear on the status of this list. How definite are its limits? What if anything is excluded from it?

As regards the limits of the list, they are clearly not very final; being based on the empirical observation that, as it seems after a very inadequate survey of the world's languages, we shall be able to get by on a five-function classification of compound substituents, it may well turn out in the end that we need more. If, for instance, we should need, for complete interlinguality, to divide the adjunctive substituents into two function-classes, corresponding to the primary TPs SA and OA, then we should have to divide each of Nos. 6-10 into two, and so increase the list by five. On the other hand, should it turn out that all I-clauses can be plausibly explained away, we could reduce the list by four. It is very likely, though not as far as I can see provable *a priori*, that the types Oo and Oi also do not occur,

and this would again reduce the list. Thus we see that though the list of types is not final, its possible variations are rather limited and unlikely to be catastrophic.

This is, of course, still at the interlingual level. For any *particular* language by no means all of the otherwise permitted substituent types will occur. I have indicated above which ones appear to be missing in English; I have also pointed out that two pairs, which in general have to be kept apart, Ao/Bo and Bs/Ss, are in English not distinguished. That means that every substituent in English which belongs to the interlingual type Ao (exemplified by the participal clause in 'women *wearing high heels* look silly.') Bo (exemplified by the prepositional clause in 'women walking *in high heels* look silly.') can each take on the other's functions (wearing high heels, she looked silly: women in high heels look silly.). In general, such confusions of what are potentially distinct types are as common among languages as mere omissions of types. In fact it is rare to find any substituent type *totally* ruled out in any language; what usually happens is that any possible example of a certain type has a more obvious or easier explanation in other terms. Thus, the noun-group 'the fact that people make love' can be construed as containing an adjunctive subjunct as one of its dependents, viz.: 'that people make love.' But it is easier in this case to regard this substituent as a nominal subjunct preceded by a zero preposition, replacing here the 'of seen in 'the fact of making love' or 'the City of London'; and every other alleged example of adjunctive subjuncts in European languages can similarly be explained away. But in Chinese, such alternative explanations are not available, or unacceptably forced, and in that language we admit Ax as a valid substituent type.

3.5. *Anomalous substituents*

The list of constitutions given in Section 3.3 ended with 'anomalous substituents'; this constitution was expressly set aside in drawing up the list of twenty-four interlingual substituent types, and the time has come to revert to it.

The class includes all substituents whose overall function, represented by the intersection in the syntax lattice of the functions of its components, is not in either of the ideals of Z.Z. The points in these ideals all have a direct mapping on to the lattice of primary TPs (in the form shown in Fig. 3). The points excluded are those of the form X.Y where neither X nor Y is Z. It is probable that in some languages, Chinese for example, all compound substituents fall within the ideals of Z.Z; this is not generally the case in European languages. However, these substituents are precisely those whose components have to be attributed *tertiary* OPs, and we saw in Section 2.3. that there can never be very many of these because of the difficulties of analysis which they can cause, and that further complications of this kind are likely to be ruled out, if only because of the need to limit the expectation-load in the comprehension of an utterance. We must thus expect that anomalous substituents will be few in numbers, although there may be in any one language several different types of them. Because of this statistical aspect, it is unlikely that it will ever pay, for the machine translation programmer, to modify the system formally so as to admit of them: it will always work better to treat them as special cases relegated to the care of the unilingual part of the overall procedure. But for the general linguist this obviously will not do: if anomalous substituents exist in a given language, they must be described and classified. I shall now try to do this for English.

One of the difficulties which soon appears, is that there is a rather wide margin of doubt as to what forms should and what should not be included; for almost all anomalous substituents, for the reasons already given, involve components belonging to closed and often

very short lists, and, as is well known, such components very easily become bound forms, whereupon they cease to be substituents in strictest sense at all. Thus, it *would* be possible to argue that the component '-ever' appearing in 'whoever', 'whatever', etc., is a substituent in the English language: but it would be simpler to regard these compounds as unanalysable and treat them as simple substituents. In the following list, I have taken a narrow view as to what forms should be admitted as compound substituents. The following is the list that emerges:

Conventional name of substituent type	Function	Component in UI of Z.Z	Example
Compound Subjunction	ZC.B	neither	in case
Relative Prepositional Clause	IB.A	neither	with whom
Relative Infinitive Clause	IB.A	neither	to lose which
Subject of Infinitive	S.S	Substantive	for me [to come. ..]
Infinitive Verbs:			
adverbial use	O.B	Operative	[I came] to see [him]
adjunctive use	O.A	Operative	[a man] to trust [her]
free use	O.S	Operative	to eat [a bun]

It will be noted that all these substituent types involve special and for the most part easily recognized word uses. Compound subjunctions are quite a short list and could very well be treated as single lexemes. Relative prepositional and relative infinitive clauses are easily recognized through their containing relative pronouns (the difficult 'that' is not used in these cases), and can be analysed by straightforward methods. The subject of the infinitive introduced by 'for' involves a special and fairly easily recognized use of this special word; but it could equally well be treated as an ordinary prepositional clause modifying the infinitive clause in its capacity as a noun-substituent. As to the infinitive with 'to' with its three distinct uses, there is much to be said for regarding this 'to' as an inflecting prefix rather than as an independent substituent. Thus, although the existence of these anomalous substituents cannot be denied, it is clear that we gain little in the practical problem of syntactic analysis, whether mechanically or even perhaps 'manually' performed by recognizing them on the same footing with other more regular types, whose functions are within the upper ideal of Z.Z.

5. THE CLASSIFICATION OF WORDS

4.1. *The participation class of a word*

As indicated at the outset of the previous lecture, the best way to arrive at a classification of words, from the point of view of their syntactic capabilities, is by way of the classification of substituents. I have already outlined how we arrive at a classification of substituents into their different types; I have given a list of twenty-four such substituent types which, while not necessarily final, is likely to serve as a useful starting-point. The subject of the present lecture is the way in which we can use these substituent types to classify the individual words, or rather lexemes, of a given language.

The principle on which we work is this. Any given word in any given language either can or cannot be used as a component in substituents of each of the twenty-four types in the list. The particular set of substituent types in which a word *can* appear constitutes a description of the syntactic capabilities of the word. We can obviously classify all the words

of a language by putting into one class all the words which can participate in any given set of substituent types. Such word-classes I shall call the 'basic participation-classes' of the language.

It is clear from this that the basic participation class of any word in any language can be very simply coded in twenty-four bits. Such a coding would, however, be redundant, in two ways. First, there are in every language certain substituent types which either never occur or are instanced so rarely that it is convenient to neglect them; such substituent types will be represented by a zero in the participation-coding for every word in the language, and this constitutes a redundancy. Second, there are certain predictable correlations between the substituent types which a given word can participate in; these are most in evidence when we enter more information than mere presence or absence into the participation-coding, but there is one instance of this valid for most European languages, namely that all words (with negligible exceptions) can be joined by conjunctions: this being so, every word can participate in conjunct groups, and so all have a 1 against this substituent type, which constitutes another redundancy.

It follows that the basic participation class of a word can in all languages be coded in less than twenty-four bits. In English, we need twelve bits only; and in general the figures will be nearer twelve than twenty. But even of the 4096 participation-classes that this coding allows for, the great majority will be empty. In English, less than 100 of these basic participation classes are filled. A very brief trial suffices to show that the information contained in the basic participation class of each word is not of itself sufficient to make a satisfactory syntactic analysis of even a simple sentence possible. We therefore require to encode more information than this allows for.

It is not difficult to see that the principle of the participation class can be readily extended. Instead of asking merely *whether* the prepositus word can participate in each substituent type, we can ask *how* it participates. Under this head we can bring in a diversity of information. First, we can indicate whether the word functions in the given substituent type as governor or as dependent; second, we can ask what restrictions may apply as to the position the word can occupy in the word sequence; third, we can ask whether it is subject to any concord restrictions; fourth, we can ask whether its appearance signals any oddity in the construction.

All these items of information can conveniently be encoded separately under each substituent type. In some cases, however, much of the information is redundant anyway: thus, in most languages which have concord, the phenomenon is restricted to a very limited number of substituent types. Noun groups account for the greater part of it in the Indo-european languages, though there is also concord of number in the sentence and certain correspondences between clauses; for such languages, therefore, it would be pointless to give the complete spectrum of substituent types, and the concord information would be encoded under two or three types only. In English the bare remnant of the concord system applies only for Z-clauses and certain types of predicate. On the other hand, governor-dependent information is about equally valuable in all the types (except perhaps conjunct groups), as also is word-information.

Information on governors and dependents is limited by definition to two bits per substituent type; since these two bits include the information, where applicable, that the given word does not participate, these two bits subsume the one required for the basic participation class. Word-order information is open to indefinite elaboration: two bits give useful data, three give all that can reasonably be desired. The extent of concord information varies

enormously from one language to another, and may amount to eight or ten bits altogether in some cases. Warnings about peculiar constructions, which may be useful especially in connection with interrupted substituents, will rarely require more than one bit in each of a minority of substituent types. Altogether, the participation-coding for each word in a language could be expanded from say twelve bits up to perhaps eighty if all these data were included.

On the basis of work done so far by the Cambridge Language Research Unit, it appears fairly certain that for English a total of four bits per substituent type per word will suffice to disentangle all but those exceptional constructions which complicate the procedure in every language. The following are two sets of questions which could be asked concerning each word W in each substituent type T:

Set A (two bits each to governor-dependent and word-order)

Bit 1: Can W be the governor in T?
Bit 2: Can W be the dependent in T?
Bit 3: Can W figure in an initial sequence in T?
Bit 4: Can W figure in a final sequence in T?

Set B (one bit for governor-dependent and three for word-order)

Bit 1: Can W be a dependent in T?
Bit 2: Can W be the last component in T?
Bit 3: Can W be an intermediate component of T ?
Bit 4: Can W be the first component in T?

Extended participation-classes based on the questions A have been drawn up for a select list of English words and are currently under test. Set B, which is likely to give better results, has not yet been put to the test.

4.2. *The making of dictionary entries*

The above procedure would be of little real use if the procedure for making the required dictionary entries were to be found unduly laborious. It may be presumed that a linguist, investigating a given language, will be prepared to take enough trouble to ascertain, if necessary by answering the questions one after the other all through his vocabulary, what participation classes are present; but it is otherwise when we are engaged on a mechanical translation project. In this case we have to get ready a number of dictionaries based on the above principles for various languages, and will expect that the work involved on each will be not too great, and above all we shall expect that it can be done by a relatively untrained person. That is to say, the skill of making correct entries must be quickly acquirable.

The stages in the procedure, when setting out to make a syntactic dictionary for a language not previously worked on, are as follows. First, one must decide upon the set of substituent types which have to be recognized. If our purpose is strictly linguistic, this list should be as complete as possible, and should include anomalous substituents; for machine translation, or other technological purposes, however, the aim is to minimize the number of types required, and anomalous substituents should certainly not be included. This assigning of the list of substituent types is the hardest part of the task; it constitutes in itself a schematic description of the grammar of the language, and can hardly be done

successfully except by one with a thorough mastery of the language. Conventional schooling in grammar is no help, rather the reverse, but a thorough understanding of the principles on which syntactic functions, substituents, and words are classified is essential.

Once the substituent types have been decided on, the next question to be discussed is the kind of information which it will be most profitable or convenient to record in the participation-coding. In practice, this question is usually answered before one starts by the properties of an existing analysis program; in any case, it depends largely on the overall strategy to be adopted, and this will be further considered in the next section. I shall therefore assume that a list of questions to be answered for each word and each substituent type has been prepared.

We next have to find the answers. It is obvious on general grounds that the number of participation classes in a given language will be finite. Even when we pass from the basic participation-classes, of which we have seen that there are only a limited number, to the extended participation-classes including supplementary information, we are still dealing with a finite set. In the case of English it appears that the questions of Set A in Section 4.2 distinguish between 100 and 200 participation-classes altogether. The fact that this number, though large, is small compared with the total number of words to be classified, means that we can considerably simplify our task by drawing up lists of common participation-classes beforehand. Thus, it is likely that in most languages there will be only a few common classes of nouns; if then we can say of a given word, 'this is a noun of class 4', we can at once write down its appropriate participation class, or some abbreviation for this, and go on to the next word.

The labour of compiling a syntactic dictionary is thus greatly alleviated. As the work proceeds, the number of pre-edited classes will grow; but provided it is not allowed to exceed, say, a few dozen it vastly reduces the time required to classify each new word. Only those words which do not fit into any of the pre-edited classes will then need to be exhaustively questioned. The chief difficulty is to be able to recognize *which* words are going to give trouble: it is here that a native speaker of a language has a great advantage. To be able to spot that the common auxiliary verb 'will' is also a noun may not be so difficult, but it is easy to pass over a word like 'accordance' as an ordinary abstract noun, when in fact it has the very unusual property of never taking an article and never being the object of a verb.

Another technique which saves a lot of time and trouble is the use of a certain quite small number of basic participation-classes, from which the participation-class of any word of multiple use can be derived by taking their union. Thus, 'cheat' which functions both as a verb and as a noun can do all that 'read' can do and all that 'chest' can do; its participation-class is therefore simply the union of those of the two latter words. A very large proportion of all the words in English can be assigned participation-classes by this operation upon about a dozen basic prototypes. In languages less rich than English in homographs, this expedient will be proportionately less rewarding, but is likely always to be needed often enough to be included in the procedure.

Once the necessary prototype participation-classes have been prepared, and when the rate of appearance of words which do not fall into any of the existing classes falls low enough, the operation of making the dictionary reduces to pigeon-holing each new word as it comes up; in practice, if we write the data on cards, what we get is a number of piles of cards, and a large scatter of twos and threes representing the less usual classes. Our experience of the operation is not yet great enough to enable us to pronounce upon the

overall rate of work which can be expected; but it is sufficient to say that the labour of compiling a syntax dictionary for a new language is not excessive, and a single worker can expect it to occupy at most a few months. Of course, it is essentially a clerical task, and one rather strongly subject to errors at a clerical level. It must therefore be followed by a phase of correction: this may well prove to be a slow and laborious process.

4.3. *Strategy of analysis*

Ultimately, what we want our dictionary for is to enable us to arrive at a correct syntactic analysis of the sentences of a given text. This is to be done by an algorithm operating on the participation-classes of the successive words of the text. I do not propose in these lectures to go into much detail regarding the structure of such algorithms, but it is necessary to indicate what sort of range of possibilities is open, since upon this depends to some extent the relative value of the different sorts of information which can be incorporated in the extended participation classes of the words.

Very broadly there are four types of strategy, the product of the two categories, predictive versus exhaustive, and forwards versus backwards. A predictive strategy is one which operates by formulating expectations as to how the sentence under consideration is going to be continued, which expectations are compared with the next word in the text and cancelled, satisfied, or suspended as the case may be. An exhaustive strategy on the other hand takes the words in order and asks how the first *n* words *could* be bracketed, i.e. built up into a hierarchy of substituents; all such bracketings are accepted except those which cannot be successfully carried through to the end of the sentence, and it is hoped that not too many alternatives will survive. If alternatives are wanted from the predictive strategy, it is necessary to start again after having found one acceptable one to see whether there isn't another, and so on. The contrast of forwards and backwards is self-explanatory: either we take the words in the order they appear in the text, or in the reverse order. Very roughly, it seems that predictive strategies work best forwards, and exhaustive ones backwards, at least for European languages.

It is certainly the case that which strategy works best depends on the language under investigation. A language like English in which most of the position-specific words are initials (like articles or prepositions) will not respond like Chinese, which has mainly final particles. No one has yet got enough experience to be able to make any generalizations on this topic. Machine translation researchers have tried out most of the possible strategies, but mostly on a very limited range of languages. Russian and English have been the main test languages. The predictive strategy, associated with the name of Ida Rhodes [13], but developed mainly by Oettinger and his associates [14], is known to be applicable to Russian, though it presents certain difficulties which at the time of writing have not been fully resolved. There is good reason to expect that it will work better for English, which is a language in which word-order is much more important than it is in Russian. The exhaustive type of strategy has been more widely used, but never on a large scale with the type of word-classification which I am describing here. Our work at Cambridge tends to show that it can be used to good effect if the number of alternative bracketings which the dictionary information permits is not too great.

Obviously to operate a predictive strategy one needs primarily word-order information, on which governor-dependent information can operate only as a check, on the same footing as concord information. It does not follow that governor-dependent information is the

best to use with an exhaustive strategy, though this may be the case. It is possible, though not very likely, that in a language like Latin which relies very heavily on concord and very little on word-order, a strategy using concord information as primary would have some hope of success. On the whole, however, it is probable that of the three types of information mentioned, word-order is in general the most useful. As to the direction in which the analysis proceeds, there is a very strong *a priori* reason for expecting that the best strategy will work forwards rather than backwards—because, of course, that is how we talk and understand: but in this connection it must be borne in mind that human language operators dispose of information organized in a very different way from that which I have here been describing, and that given our type of dictionaries the advantage of forward analysis could be reduced or reversed. The presumption in favour of forward predictive strategy remains, however, until practical experience teaches us otherwise.

REFERENCES

[1] C. F. HOCKETT: Two Models of Grammatical Description Word, 1954,10.
[2] DYSCOLUS APOLLONIUS: *De Syntaxi seu constructione orationis libri iiij*, Frankfurt.
[3] R. H. ROBINS: In Defence of Word-and-Paradigm, *Trans. Philol. Soc.*, 1959,**1959.**
[4] E. SAPIR: *Language*, New York (1921).
[5] R. P. WELLS: Immediate Constituents, *Language*, 1947, 23, 81.
[6] Z. S. HARRIS: *Methods in Structural Linguistics*, Chicago (1951).
[7] N. CHOMSKY: *Syntactic Transformations*, The Hague (1957).
[8] C. N. MOOERS: The Mathematical Theory of Language Symbols in Retrieval. Int. Conf. Sci. Inf. (U.S. Nat. Acad. Sci., Washington) (1958).
[9] J. LAMBEK: The Mathematics of Sentence Structure, *Amer. Math.*, 1958, 65, 154.
[10] Y. BAR-HILLEL, E. SHAMIR and M. PERLES: On Formal Properties of Simple Phase-structure Grammars. Hebrew University, Jerusalem (1960).
[11] M. A. K. HALLIDAY: Categories in the Theory of Grammar Word, 1962,17, 241.
[12] V. H. YNGVE: A Model and a Hypothesis for Language Structure, *Proc. Amer. Phil. Soc., 1960,*104, No. 5.
[13] I. RHODES: (1959). A New Approach to the Mechanical Translation of Russian. U.S. Nat. Bureau Standards, Rep. 6295.
[14] A. G. OETTINGER: Automatic Syntactic Analysis and the Pushdown Store. Symposium on Structure of Language and its Mathematical Aspects, *Amer. Math. Soc.*, New York (1960).

PUBLICATION OF THE C.L.R.U.

The following publications are not items from journals or official reports, but they may be obtained (subject to availability) from Cambridge Language Research Unit, 20 Millington Road, Cambridge, England.

A New Model of Syntactic Description. M.L. 146.

C.L.R.U. Syntax Dictionary for English (Vocabulary No. 1). M.L. 159.

APPENDIX A. SUBSTITUENT TYPES IN ENGLISH

In this appendix I give a list of the substituent types which can be exemplified by reasonable examples in English. I number them by arbitrary serial numbers, not the same as those in the complete list of twenty-four substituent types in Section 3.4; the relation between the latter, interlingual, substituent types, and the selection of them occurring in English, is indicated by the use of the two-letter abbreviations given in the former list. Anomalous substituent types are not included in the list below. In addition to the systematic name of each type, I have also given a suggested conventional name.

1. *Adverbial group or compound adverb (Bg)*

Consists of two adverbs, one qualifying the other; either may be replaced by an adverbial substituent. The governor is taken as the component which is modified, the modifier being its dependent. Examples with more than one dependent are rare.

almost exactly
truly in my opinion
far but not too far out

[in all examples, the component underlined is the governor]

2. *Adverbial O-clause or prepositional or participal clause (Bo or Ad)*

Consists of a preposition or participle in -ing, which is the governor, and a nominal substituent which is the dependent. Limited to one dependent. Participial clauses are almost always identical in form with corresponding infinitive clauses (see No. 8) owing to the identity in English of the participle in -ing and the gerund. This makes it necessary to discount the latter interpretation in considering the following examples:

reaching the shore	participial clause
considering the circumstances	„ „
they having at last come	„ „
between the house and the street	prepositional clause
under duress	„ „

3. *Adverbial S-clause or absolute clause (Bs)*

Consists of one of a small class of words including certain uses of the interrogative pronouns, and forms with the suffix '-ever', which is the dependent, and an ordinary predicate which is the governor. It is convenient but not essential to split up the governor into a verb-group and a second dependent, so that we do not have to call the sequence 'I do' in 'whatever I do' a 'predicate', which offends against common usage. If this is not done there is never more than the one dependent.

whatever I say [he'll go]
[unpleasant] however you take it

Nominal S-clause (Ss)

All adverbial S-clauses also have nominal uses; in addition, there are clauses with 'what' which have only nominal uses; properly therefore we should count these two types as distinct, but since this use of 'what' seems to be unique, it is convenient to treat nominal S-clauses as another use of adverbial S-clauses, just as we treat adjunctive O-clauses as another use of adverbial O-clauses. Examples

what I say [goes]
[arrest] whoever you see

4. *Adverbial subjunct or subordinate clause (Bx)*

Consists of a simple sentence (very occasionally containing a special verb form) prefixed by a subordinating conjunction; the clause may be replaced by certain participles or adjectives, but cases where it appears to be replaced by a noun group ('if a man') are better explained as due to ellipsis, since they cannot stand alone without a preceding sentence

giving the missing parts. The clause or adjective is the dependent, and there is only one dependent.

if I were you	case of special verb form
in case it should rain	
though exhausted	
when wet	

5. Adjunctive group or adjectival group (Ag)

Consists of an adjective, which is the governor, qualified by one or more adverbs or adverbial substituents. There is often more than one dependent.

very funny
discredited only in parts
rather cold for the time of year

6. *Adjunct S-clause or relative clause (As)*

Consists of a relative pronoun and a predicate, the latter being the governor; as with absolute clauses, it is convenient to split the governing predicate into a verb group and a second dependent. A relative clause frequently ends with a preposition, which properly speaking forms the governor of the relative pronoun in an anomalous substituent. The relative pronoun can be zero.

[the man] who taught me
[the man] that I gave it to
[someone] ϕ others look up to

7. *Nominal group or noun group (Sg)*

Consists of a noun with qualifiers, which may be adjectives, articles, or any other adjunctive group. The noun is the governor, and there is no limit to the number of dependents.

the three white and slightly battered pigeons that arrived
a bigger one
such a man

8. *Nominal O-clause or infinitive clause (So, Ao, or Bo)*

Consists of an infinitive verb or gerund as the governor, and one or more nominal substituents as dependents; normally, these are objects of the verb. Every infinitive clause has also adjunctive and adverbial uses: but in the case of those where the verb is a gerund these latter uses are regarded in English as containing not a gerund but a participle. This is one of the traditional distinctions which our theory enables us to discard. There may on occasion be no dependent.

swimming in hot water [is tiring]	
[it is hard] to be sure	zero dependent
[a man] to be trusted with money	adjunctive use
[we did it] to gain time	adverbial use

9. *Nominal subjunct or nominal clause (Sx)*

Consists of a simple sentence prefixed by a subordinating conjunction which must be either 'that', 'whether', or zero. The clause which is the only dependent cannot be replaced by any single word except 'so' or 'not'.

[he thought] that it was too late
[I wonder] whether they'll come
[suppose] ϕ something goes wrong
[I think] ϕ not

10. *Operative group or verb group (Og)*

Consists of either (a) auxiliary verb as governor and main verb as dependent, *(b)* complete verb as governor, adverb(s) as dependent(s), (c) auxiliary verb or one of a special class of complete verbs as governor and an adjunct substituent as dependent. *(a)* and *(b)* may be combined in one substituent. There is no limit on the number of dependents. The governor may be itself a verbal group consisting of two auxiliary verb forms, such as 'will have' or 'was being'.

have come	type *(a)*
came recently	type *(b)*
have recently come	mixed type *(a)* and *(b)*
will be very cold	type (c) with compound governor
paint [it] green	type (c) interrupted by the object of the verb
paint [it] out	type *(b)* ditto, with postverb as dependent.

11. *Operative S-clause or predicate (Os)*

Consists properly of a verbal group accompanied by one or rarely more noun groups or prepositional clauses as its dependents; when the latter are certain pronouns, they take a special form. It is convenient to allow that there may be zero dependents, so that intransitive verbs may be counted as predicates as in the traditional terminology; but strictly speaking intransitive verbs constitute a special use of a verb group as governor of a sentence.

[I] hit him	dependent in special form
[they] put the matter up to him	interrupted governor, one dependent a prep. clause
[whom] we found	

12. *Sentence (Zz)*

Consists of a noun-group (which may be zero: e.g. in imperative sentences) as dependent and a predicate (or verb-group with intransitive verb). It may also have additional dependents in the shape of prepositional clauses and adverbial subjuncts.

I hit him
they put the matter up to him
a quite hard and in some cases brittle substance resulted

ϕ get out of here	
meanwhile the police arrived in case trouble should start	
	three dependents, one a noun-group, one an adverb, and one an adverbial subjunct.

Conjunct groups

Consist of two or more substituents sharing some common function joined by a conjunction. There are as many types of conjunct group as there are functions which the dependents can have. Here are a few:

quietly and steadily	
big, strong, and muscular	note that we may count the commas as part of a split governor
horses, sheep, cattle, etc.	
either go now or sit it out	another split governor
some clapped but others laughed	
[men] and/or [women]	conjunct conjunction.

In conjunct groups systematic use is made of ellipsis, resulting in the appearance of dependents which are not, without supplying the missing components, substituents at all. Thus

for whom I but not you were working	either supply a second 'for whom' or a second 'was working'
they might have hit and ruined it	supply a second 'might have' and a first 'it'.

APPENDIX B. CONSTRUCTION OF A DICTIONARY READING

In this appendix I shall show in detail how one can ascertain the participation class of a given word. I shall assume that no previous participation classes are known, and thus I shall not make use of any of the many devices for shortening the procedure which in practice make the construction of dictionary-readings a far less laborious operation than this. I shall include for the word chosen under each substituent type two bits of governor-dependent information and three bits of word-order information; from this five-bit entry I shall then derive two four-bit readings, one corresponding to each of the two sets of questions listed in Section 4.1.

As to coding, I shall use the following conventions:

Question 1: can the word be the governor?	bit 1	digit 1
Question 2: can the word be the dependent?	bit 2	
Question 3: can the word be initial?	bit 5	digit 2
Question 4: can the word be intermediate?	bit 4	
Question 5: can the word be final?	bit 3	

The test word will be 'these'. Under each of the twelve substituent types listed in Appendix A I shall subject this word to the above five questions, which I shall represent by their last words alone. The answer given to each question, if YES, will be followed by an example. If NO, in some cases explanations may be added.

Dictionary reading for 'these'

1. Adverbial group

Governor?	NO		
Dependent?	NO		digit 1 = 0
			∴ digit 2 = 0

2. Adverbial O-clause

Governor?	NO		
Dependent?	YES	'in these'	digit 1 = 1
Initial?	NO	but possibly in 'these notwithstanding'	
Intermediate ?	NO	the type has only two components	
Final?	YES		digit 2 = 4 (5)

3. Adverbial S-clause

Governor?	NO	
Dependent? *	same as in S.T. 6 below	digit 1 digit 2 = 6

4. Adverbial subjunct

Governor?	NO		
Dependent?	NO	'if these' would be regarded as an ellipsis	digit 1=0 ∴ digit 2 = 0

5. Adjunctive group

Governor?	YES	'just these'	
Dependent?	NO		digit 1 = 2
Final?	YES		
Intermediate?	NO	it is hardly possible to attach two dependents to this word	
Initial	NO		digit 2 = 4

6. Adjunctive S-clause

Governor?	NO		
Dependent?	YES	if we split the predicate: '[people] whom these fit'	digit 1 = 1
Final?	YES	'[the people] who fit these'	
Intermediate ?	YES	(as in previous example)	
Initial?	NO		digit 2 = 6

7. Nominal group

Governor?	YES	'all these'	
Dependent?	YES	'these people'	digit 1 = 3
Final?	YES		
Intermediate?	YES	'all these people'	
Initial?	YES		digit 2 = 7

8. Nominal O-clause

Governor?	NO		
Dependent?	YES	'to buy these'	digit 1 = 1
Final?	YES		
Intermediate?	YES	'to give these food [would be wrong]'	
Initial?	NO		digit 2 = 6

9. Nominal subjunct

Governor?	NO	
Dependent?	NO	digit 1=0
		∴ digit 2 = 0

10. Verbal group

Governor?	NO	
Dependent?	NO	digit 1 = 0
		∴ digit 2 = 0

11. Predicate

Governor?	NO		
Dependent?	YES	'take these'	digit 1 = 1
Final?	YES		
Intermediate?	YES	'give these food'	
Initial?	NO	I regard 'I these bestow' as not in the relevant dialect	digit 2 = 6

12. Sentence

Governor?	NO		
Dependent?	YES	'these are the best'	digit 1 = 1
Final?	NO	exceptions explained by ellipsis	
Intermediate?	YES	'are these the best?'	
Initial?	YES		digit 2 = 3

Final result: the dictionary reading which we have constructed can be exhibited best in two lines, thus:

these	0110	21	310	011	dependency information
	0460	46	760	063	word-order information

We can derive from this the reading for the reduced set of questions A of Section 4.2 by the following mapping of digit 2: 0 → 0; 1, 3 → 1; 4, 6 → 2; 2, 5, 7 → 3 (in the last case we map 2 on to 3 only because it will fit nowhere else). This yields for the second line of the above reading the revised form:

these	0220	22	320	021	word-order information

For set B the mapping is for digit 1: 0, 2 → 0; 1, 3 →1, giving for the first line

these	0110	01	110	011	dependency information

APPENDIX C. WORKED EXAMPLE

I shall now give an example of analysing the syntactic structure of a simple sentence by the methods described. I shall use for this purpose a simple version of the predictive strategy, operating on dictionary-readings in which each word is coded with three bits of word-order information per substituent type. The substituent types will be the twelve described for English in Appendix A (omitting the conjunct group). To work the example we need to have *(a)* a sufficient outline of the steps of the procedure, *(b)* the participation-classes of the words occurring in the test sentence, *(c)* the participation-classes of each substituent type, using the code understood in the procedure.

Summary of procedure

1. Read next word, 2.
2. Test word: *(a)* it can begin or continue the 'target' substituent type as required, 1.
   (b) it can only end the 'target', 6.
   (c) any other case, 3.
3. Find a target substituent (which the word in hand can begin); or if a target is already in hand, find another; in either case the target must itself be able to fit in the required position. If no target can be found, 4; otherwise, 1.
4. Delete last word, with any completed bracket-groups which may include it. If only one component left in bracket-group, 3; otherwise, 5.
5. Can what is now the last word or bracket end the target substituent? Yes, 6; No, 4.
6. Record the current target as a completed bracket. If this is substituent-type 12, 8; otherwise, 7.
7. Does the newly completed bracket-group end the previous target? Yes, 6; No, 1.
8. Is the sentence complete? Yes, STOP; No, 3.

Participation-classes of substituent types

1.	Adverbial group	0000	50	000	363
2.	Adverbial O-clause	6004	50	600	063
3.	Adverbial S-clause	0400	00	040	045
4.	Adverbial subjunct	0000	50	000	045
5.	Adjunctive group	0004	00	700	040
6.	Adjunctive S-clause	0000	00	400	040
7.	Nominal group	0464	06	060	043
8.	Nominal O-clause	0404	40	400	045
9.	Nominal subjunct	0000	00	040	045
10.	Operative group	0060	06	000	015
11.	Predicate	0040	04	040	004
12.	Sentence	0004	00	004	007

Participation-Classes of Words in Test Sentence

Test sentence: I shall try to explain it.

I	0000	00	100	003
shall	0040	04	000	334
try	0000	04	000	714
to	4100	04	010	440
explain	0000	04	000	714
it	0400	00	140	041

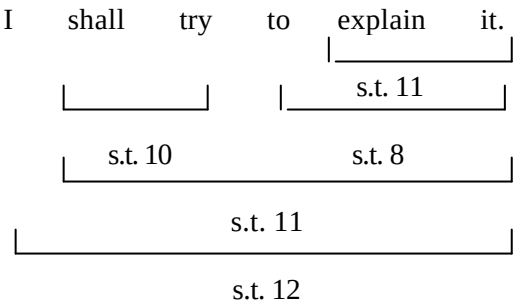
Sequence of steps in the work

1. read 'I'
2. can 'I' start the s.t. 12 (sentence)? YES
1. read 'shall'
2. can 'shall' continue or end s.t. 12? END
6. record bracket 'I shall' as s.t. 12
8. is the sentence complete? NO
3. what s.t. that 'shall' can begin will continue or end s.t. 12 ? first take s.t. 10 which ends it
1. read 'try'
2. can 'try' continue s.t. 10? YES
1. read 'to'
2. can 'to' continue or end s.t. 10? END
6. record bracket 'shall try to' as s.t. 10
7. does s.t. 10 end s.t. 12? YES
6. record bracket 'I shall try to' as s.t. 12
8. is the sentence complete? NO
3. what s.t. which 'to' can begin continues or ends s.t. 10? NONE
4. delete 'to': leaving two words in current target group
5. can 'try' end s.t. 10? YES
6. record 'shall try' as s.t. 10
7. does s.t. 10 end s.t. 12? YES
6. record bracket 'I shall try' as s.t. 12
8. is sentence complete? NO
3. what s.t. which 'try' begins continues or ends s.t. 10? NONE
4. delete 'try': leaving only one word
3. what s.t. after 10, which 'shall' can begin, continues or ends s.t. 12? s.t. 11
1. read 'try'
2. can 'try' continue or end s.t. 11 ? NO
3. what s.t. which 'try' can begin continues or ends s.t. 11? NONE
4. delete 'try' leaving only one word
3. what s.t. that 'shall' can begin begins s.t. 11 ? s.t. 10
1. read 'try'
2. can 'try' continue or end s.t. 10? EITHER
1. read 'to'
2. can 'to' continue or end s.t. 10? END

6. record bracket 'shall try to' as s.t. 10
7. does this end previous target? NO
1. read 'explain'
2. can 'explain' continue or end s.t. 11 ? NO
3. what s.t. which 'explain' begins will continue or end s.t. 11 ? NONE
4. delete 'explain': leaving only one component 'shall try to'
3. what s.t., which 'shall try to' (s.t. 10) can begin, will begin s.t. 11 ? NONE
4. delete 'to' and bracket group labelled s.t. 10 before: leaving two components 'shall try' in the group
5. can 'try' end s.t. 10? YES
6. record 'shall try' as s.t. 10
7. does s.t. 10 end s.t. 11 ? NO
1. read 'to'
2. can 'to' continue or end s.t. 11? END
6. record bracket 'shall try to' as s.t. 11
7. does s.t. 11 end s.t. 12? YES
8. is sentence complete? NO
3. what s.t. which 'to' can begin continues or ends s.t. 11? first take s.t. 2
1. read 'explain'
2. can 'explain' continue or end s.t. 2 ? NO
3. what s.t. which 'explain' can begin continues or ends s.t. 2 ? NONE
4. delete 'explain', leaving only one component 'to' in current group
3. what s.t. after 2, which 'to' can begin, continues or ends s.t. 11? next take s.t. 8
1. read 'explain'
2. can 'explain' continue or end s.t. 8 ? NO
3. what s.t., which 'explain' can begin, continues or ends s.t. 8? s.t. 11
1. read 'it'
2. can 'it' continue or end s.t. 11 ? END
6. record bracket 'explain it' as s.t. 11
7. does s.t. 11 end s.t. 8? YES
6. record bracket 'to explain it' as s.t. 8
7. does s.t. 8 end s.t. 11 ? YES
6. record bracket 'shall try to explain it' as s.t. 11
7. does bracket s.t. 11 end s.t. 12? YES
6. record bracket 'I shall try to explain it' as s.t. 12
8. is sentence complete ? YES

STOP.

The resulting analysis of this sentence is therefore:



Note: it is not possible, without the use of a complete algorithmic notation which would be out of place here, to make the description of the procedure entirely unambiguous. It is, therefore, not strictly possible to check the correctness of the steps given above against the procedure outlined; but, of course, the steps may be taken as illustrating the procedure. Nevertheless, it will probably be agreed that the step taken is, if not unique, at least the most reasonable interpretation of the form of words given under the corresponding number in the outline of the procedure.