STRING ANALYSIS OF SENTENCE STRUCTURE

by

ZELLIG S. HARRIS

UNIVERSITY OF PENNSYLVANIA

1962

MOUTON & CO . THE HAGUE

This is a revised version of Computable Syntactic Analysis, no. 15 (1959) of *Transformations and Discourse Analysis Papers* (TDAP), the mimeographed series of the Department of Linguistics, University of Pennsylvania, and of the National Science Foundation project in linguistic transformations.

Printed in The Netherlands by Mouton & Co., Printers, The Hague

CONTENTS

1. THREE CHARACTERIZATIONS OF SENTENCE STRUCTURE

This paper introduces a method of syntactic analysis which is intermediate between the traditional constituent analysis and transformational analysis. One version of string analysis has been carried out on a computer, producing a decomposition of English sentences into their string-analytic elementary sentences and appropriately placed adjuncts.

A prefatory remark about sentences may be in order here. There are many difficulties in describing empirically what is to be included in the set of sentences of a language. For one thing, there is the problem of deciding what are utterances of the language. Native speakers of it may differ as to what they accept as being an utterance of the language. They may differ because of dialectal differences, which in urban society are often intercrossed, or because of individual strong differences in experience, or because of the vagaries of the conditions of observation, etc. And this though every speaker will accept utterances that he has not heard before. Furthermore, speakers may accept certain other utterances only hesitantly, or only as being bizarre, or as being a linguistic joke, or as being an otherwise occasionally usable departure from ordinary sentencehood.

Given the decision as to what are utterances of the language, there are further problems as to what are the sentences. Sentences of particular types may be characterized as those segments of speech (or writing) over which certain intonations occur or within which certain structures occur. (A particular structure is a particular combination of classes of elements.) Some utterances are longer than a sentence; that is to say, they can be shown to be sequences of sentences characterized as above (e.g. *We hurried over. The door was locked.* or *He was just here. Did you see him?*). Some utterances are shorter than a sentence; that is to say, they can be shown to be segments of a sentence, the non-completion of the sentence being of some regular kind (e.g. *But why?* or *What the!*). And some utterances have neither of these relations to otherwise recognized sentences; these must be considered as independent sentences (e.g. *Yes.* or *Scat!*).

Each utterance can thus be described as a sequence or fragment of sentences. The sentences obtained by decomposing any set of utterances of the language do not, however, constitute the set of all the sentences of the language, though they may constitute the starting point for determining that set.

First, if $A_1 B_1$ and $A_2 B_2$ are sentences obtained from utterances, where each A, B is a morpheme or word or word sequence, it is often possible to assign A_1 and A_2 to a class $\{A\}$, and B_1 and B_2 to a class $\{B\}$, in such a way that $A_1 B_2$ and $A_2 B_1$, which are not found in the given set of utterances, are also accepted by native speakers as being sentences. We do not wait to find all the sentences in utterances heard or written. Instead, having grouped the words, etc., into classes (in order to find within each utterance those regularities of class combination, over portions of the utterance, which we call sentence-structure), we then say that the sentences found in the utterance are combinations of particular members of these classes, but that the same combination of other members of these classes will also be accepted as sentences.

Second, in every language there are, in particular sentence types, certain structural features (e.g. the addition of certain classes of elements) which are unboundedly iterable. That is to say, if we find (in any collection of sentences of the language) that no sentence has the given feature more than n times, we can in stated circumstances produce from the sentences in the collection a new sentence, accepted by speakers but not in the collection, which contains the given feature $n + 1$ times. E.g. in addition to *It is very very large* we can produce *It is very very very large*, and in addition to *He picked up the pins and the needles* we can produce *He picked up the pins and the needles and the double bass.*

The sentences of the language are therefore not simply those culled from any collection of utterances, but are an unbounded set expanded in the above manner from the culled sentences. A grammar of a language seeks to show how all the sentences which would be accepted (under one or another criterion of acceptance, as noted above), can be characterized as particular types of combinations of particular classes of elements (phonemes, morphemes, words, sequences of words, sentences).

1.1 STRING ANALYSIS

String analysis characterizes the sentences of a language as follows: Each sentence consists of one elementary sentence[1] (its center), plus zero or more elementary adjuncts, i.e. word-sequences of particular structure which are not themselves sentences and which are adjoined immediately to the right or to the left of an elementary sentence or adjunct, or of a stated segment of an elementary sentence or adjunct, or of any one of these with adjuncts adjoined to it. An elementary sentence or adjunct is a string of words, the words (or particular sequences of them) being its successive segments. Each word is assigned (on the basis of its affixes or its position in elementary sentences and adjuncts) to one or more word-categories (rarely, word-sequence categories). Hence, we can replace each word of a string by the symbol of its category, thus obtaining a string of category-symbols (called a string formula) as a representation of the word-string in question. The term "string" will be used both for word-strings and for string formulas, depending on the subject under discussion.

For example, in the sentence

[1] An elementary member of a set $\{A\}$ is one which contains no member of the set $\{A\}$ as a proper part of it. The requirements for elementary sentences will be somewhat modified in axiomatic string analysis, presented in sections 2 and 3.

Today, automatic trucks from the factory which we just visited carry coal up the sharp incline:

trucks carry coal is the elementary sentence
today, is an adjunct to the left of the elementary sentence
automatic is an adjunct to the left of *trucks*
from the factory is an adjunct to the right of *trucks*
which we visited is an adjunct to the right of *factory*
just is an adjunct to the left of *visited*[2]
up the incline is an adjunct to the right of *carry*[3]
sharp is an adjunct to the left of *incline.*

How each sentence-string is subdivided into elementary strings will be seen in section 2. Here it is enough to note that it is possible to decompose each sentence into elementary strings which combine (to form a sentence) in accordance with specified rules. If in a given sentence we find a sequence of words which cannot be assigned to any known string formula occurring in it in accordance with some known rule, then a new string or rule of occurrence has to be set up. The intention is that a few classes of strings, with simple rules describing how they occur in relation to each other, will suffice to characterize all sentences of the language.

1.2 COMPARISON WITH CONSTITUENT ANALYSIS

In contrast with the above, the main method of structural linguistics has been the analysis of a sentence into constituents at a lower descriptive level. Constituent analysis characterizes the sentences of a language as follows: Each sentence consists of a sequence of constituents (say, noun-phrase and verb-phrase in the case of English and many other languages), each of which in turn is a sequence of lower-level constituents (say, noun and its modifiers in the case of a noun-phrase), and so on down until the final constituents are single morphemes or syntactically unitary (but morphemically complex) words (e.g. *conceive, boyhood*). The constituents into which a sentence or constituent is decomposed are not in general of the same class as it itself. A constituent analysis is accounted satisfactory if only a few (and not very variegated) classes of constituents and of decomposition rules suffice to characterize all the sentences.

In the sample sentence of 1.1 (disregarding the constituent status of *the*):[4]

$S = S$ adj *today* $+ Np_1$ *automatic trucks from the factory which we just visited*
$+ Vp_1$ *carry coal up the sharp incline*

$Np_1 = N$ adj *automatic* $+ N$ *trucks* $+ N$ adj_2 *from the factory which we just visited*

N $\text{adj}_2 = P$ *from* $+ Np_2$ *the factory which we just visited*

$Np_2 = N$ *the factory* $+ N$ adj_3 *which we just visited*

N $\text{adj}_3 =$ adjectivizer *which* $+ S_2 - Np$ *we just visited*

$S_2 - Np = Np$ *we* $+ Vp_2 - Np$ *just visited*

$Vp_2 = V$ adj *just* $+ V$ *visited*

$Vp_1 = V$ *carry* $+$ object Np $(= N)$ *coal* $+ V$ adj_1 *up the sharp incline*

$Vadj_1 = P$ *up* $+ Np_3$ *the sharp incline*

$Np_3 = N$ adj *sharp* $+ N$ *the incline*

[2] The sentence is ambiguous here, between *just* "merely" and *just* "this moment". The ambiguity arises because *just* is a member of two subcategories of a_{v1} (see section 3.24); in the second meaning it is a member of a subcategory which occurs almost only to the left of the verb, and is restricted to particular members of the tense category. A corresponding ambiguity will appear in the transformational analysis of the same sentence. When, in a string analysis of a sentence, we come upon a word which could, in the given position, be a member of more than one subcategory, or a segment of more than one string, we have to carry out the analysis of the sentence separately for each formulaic representation of the word. In the present case the two analyses would be identical except that the line in question would read, in one case, "*just* is in a_{v1a} to left of *visited*", and in the other case "*just* is in a_{v1b} to left of *visited*" (where *a, b* indicate the appropriate subcategories of a_{v1}).

[3] Though this adjunct occurs only to the right of *carry coal* (verb plus object), it could occur immediately to the right of the verb if the object were longer: *carry up the incline the last remaining loads of coal* (see under a_v in section 3.24).

[4] Capital letters indicate the class of a constituent. *S:* sentence; adj: adjunct; *N:* noun; *V:* verb (here including tense); *p:* phrase; *P:* preposition; *S* adj: sentence adjunct, etc.; *S* adj *today:* sentence adjunct represented by the word *today*. Subscript numerals are for identification within the example. $S - N$ or "*S* excising *N*" indicates an *S* string from which one *N* has been cut out (similarly for $\Omega - N$, etc. in section 3). $+$ or "plus" indicates concatenation of segments in a string, which is generally indicated here by successive symbols with space between them.

Note that the S adj is a separate constituent at the level of Np and Vp.

The transition from constituent to string analysis is given, at least for English, by the observation that most constituents either consist of a single word (of some category, or of any one of several categories, which characterizes that constituent) or contain a single word of the characterizing category plus adjunct words or phrases adjoined to it.[5] We can thus consider such a pluri-word constituent in any sentence A as being endocentric, i.e. expanded from its characterizing category by the addition of adjuncts; and this in the sense that we can replace each constituent by its characterizing category alone, and obtain a sentence B which would be related to A as a constituent-expansion of A. That is, given a sentence or constituent C whose immediate constituents (i.e. the next-level constituents into which C is decomposable in a regular way) are a C_1-phrase, a C_2-phrase, etc., we find that most C_i-phrases consist of a word of the C_i category (which characterizes the C_i-phrase) plus zero or more adjuncts of C_i. Thus in the example above, $S = Np_1$ Vp_1; and Np_1 consists of N: *trucks* plus adjuncts of N; and Vp_1 consists of V: *carry* plus object plus adjuncts of V; and the characterizing N plus V plus object constitute a sentence *Trucks carry coal*, of which the example sentence can be considered the expansion.

There are in English, as in other languages, certain exceptions to this expansional structure of constituents. Some constituents are exocentric, i.e. they are word-sequences (phrases) such that we cannot replace them by any word of a characterizing category contained in them and obtain thereby another sentence of the language. (Or if we indeed obtain another sentence, it is not one which – when compared with the given sentence – would fit into a satisfactory scheme of constituent-expansion relations among sentences). Another type of exception is the case of verbs whose subject or object is derived from a sentence:[6] *Whether the foregoing experiments succeeded, as is claimed, certainly interested many observers; I don't know whether he came.* In any satisfactory analysis of these subjects and objects, they are not expansions of a word contained within them.[7] Other exocentric constituents in English are the occasional nominalizations of verb-plus-object, e.g. *catch-all*, and some other compound words.[8]

Another difficulty in the way of constituent analysis is the case of single words which occupy the syntactic position of expanded constituents and not of their characterizing category alone; e.g. pronouns in English may be considered as replacing the whole left section of the noun-phrase (from its beginning and up to and including the noun): compare *He, who now entered...* with *The old man, who now entered....*

A constituent analysis is replaced by a string analysis when, given a sentence or constituent C whose immediate constituents are endocentric C_1-phrase, C_2-phrase, etc., we define the word-sequence $C_1 + C_2 + \ldots + C_n$ as the elementary string of C; and the adjuncts included in the C_i-phrase as adjuncts into the elementary string to the right or left of C_i. An exocentric constituent C_x of C is defined as a string which is itself an elementary segment of the string C, rather than being an expansion of a segment of C.

In the case of verbs, in a string C, whose subject or object is a whole phrase, we merely say that the subject or object segment of C is itself a string. In the sentence above, the elementary sentence is *Whether the experiments succeeded interested observers*, so that the subject of *interested* is not a single word or a phrase expanded from a word, but the whole string *whether the experiments succeeded*, in

[5] The terms *endocentric* and *exocentric* are used in the work of Otto Jespersen. Constituent analysis was further developed particularly in Edward Sapir's *Language* and in Leonard Bloomfield's *Language*. An attempt to develop constituents systematically from single words is made in Z. Harris, From Morpheme to Utterance, *Language* 22 (1946) 161-183 (and *Methods in Structural Linguistics*, chapter 16).

[6] But the constituents of the subject or object phrase are again endocentric.

[7] Some exceptions are only apparent: Consider the *wh*-phrases which occupy the syntactic position of a noun-phrase. Instead of taking these as exocentric N-replacers (*I eat what she cooks. I eat food.*), we can define a zero variant of *that* and the like (with attendant changes in the *wh*-pronoun), so that, for example, *What she cooked tasted fine* is a free (morphophonemic) variant of *That which she cooked tasted fine*. In the latter form, the Np (subject) endocentric constituent is N: *that* + N adj: *which she cooked*.

[8] But most compound nouns and verbs are endocentric.

which in turn the subject of *succeeded* is *experiments;* with *foregoing* an adjunct of *experiments*, and , *as is claimed*, an adjunct of the whole subject string, and *certainly* an adjunct of *interested*, and *many* an adjunct of *observers*. In the case of a single word (or category) K which replaces the whole (or a portion) of a constituent C whose characterizing category is C_i, it suffices to include K as a sub-category within the category C_i, with the restriction that K does not take adjuncts (or does not take the adjuncts included in the portion). Hence in *He, who now entered, addressed the speaker*, the elementary sentence is *He addressed the speaker;* but the adjunction possibilities in this elementary sentence are more restricted than in *The man addressed the speaker:* in this way we provide that left adjuncts are excluded from *he*. As to the occasional exocentric constituents such as *catch-all*, these have to be similarly included as members of the word-categories whose position they occupy in the strings in which they occur: e.g. *catch-all* is a member of the N category.

We thus see that the relations formulated in constituent analysis can be included in string analysis, the latter being in various respects more general than the former. String analysis is the stronger of the two, in making the claim that for each class (or for many classes) $\{C\}$ of constituents there exist elementary members of $\{C\}$, and in particular that if the sequence of phrases $C_1p + C_2p + \ldots + C_np$ is a member of $\{C\}$, so is the sequence of word-categories $C_1 + C_2 + \ldots + C_n$. In contrast, constituent analysis makes the claim, lacking in string analysis, that the members of each class of sentences have identical segmentation into constituents, i.e. that the expansion and replacement of sentence segments are encapsulated within a fixed structural segmentation of all sentences of the class. This claim presents occasional difficulties. In English, for example, there are rare cases of adjuncts which are not contiguous with the other parts of the constituent to which they belong: e.g. certain subject adjuncts at the end of a sentence (*Finally the day arrived which we had so eagerly awaited.*); more frequently certain verb adjuncts occur at the beginning of a sentence (*Softly she tiptoed out.*). There are also various segments which are hard to assign to otherwise known classes of constituents, e.g. the *to go* in *I want you to go*. And the problem of sentence adjuncts (e.g. *in general, today*), which occur in various sentence positions, is uncomfortable for constituent analysis: At the least one has to say that an English sentence is characteristically not merely $Np + Vp$, but also provides for sentence adjuncts (at various points) which are not included in these constituents.

1.3 COMPARISON WITH TRANSFORMATIONAL ANALYSIS

Transformational analysis decomposes each sentence, without residue, into elementary sentences (not necessarily those of string analysis; and occasionally carrying primitive adjuncts, i.e. adjuncts not derived from sentences) that are operated upon by particular transformations. More specifically, it characterizes the sentences of a language as follows: Each sentence consists of one such elementary sentence (possibly with primitive adjuncts) under unary transformations (these include the identity transformation), plus zero or more elementary sentences (with unary transformations and, possibly, primitive adjuncts on each) each under a binary transformation which relates it to a particular other elementary sentence in the given sentence.

In the sample sentence of 1.1, we have (disregarding the transformational status of *the*):

elementary sentence	primitive adjuncts	unary transformations	binary transformations
1. *truck carries coal*	1 sentence operator: plural 2 sentence adjunct: *today,* 3 verb adjunct: *up the incline*	identity	—
2. *incline is sharp*	—	identity	sharing into 1.3

3. *truck is automatic*	—	identity	sharing into 1
4. *truck is from factory*	—	identity	sharing into 1
5. *we visited factory*[9]	verb adjunct: *just*	identity	*wh-* into 4

The transition from string to transformational analysis is given by the observation that most of the adjuncts are regular transformations of elementary sentences. In English, almost all the noun-adjuncts J are transformations of N *is* J sentences, as in 2-5 above (see under λ_{n_2} and r_n in 3.23). Almost all the conjunctional adjuncts and sentence adjuncts (r_x and a_c and r_c in section 3) are also binary transformations of sentences. E.g. in *He buys and sells books* and in *Entering the house, he was stopped by the neighbors*, we have:

elementary sentence	primitive adjuncts	unary transformations	binary transformations
he buys books	—	identity	—
he sells books	—	identity	*and* (with zeroing and permutation)
neighbors stopped him	—	passive	—
he entered the house	—	identity	*-ing* (with zeroing)

Not all the elementary sentences and primitive adjuncts are in one-to-one correspondence with the elementary sentences of string analysis. The D (adverb) and P N (preposition plus noun) adjuncts of verb, adjective, or sentence (as in column 2 of the table above) are apparently not ultimately derivable from separate sentences: there is no independent sentence containing *just* which combines with *We visited the factory* to form *We just visited the factory*. And so for the other adjuncts: we would not derive *five trucks* from *The trucks are five*.

Furthermore, certain sentences are segmented transformationally into elementary sentence plus unary constants;[10] whereas string analysis may make a different segmentation, i.e. a different choice of words (out of the given sentence) to constitute the elementary sentence, plus other segments (than the above constants) as the adjuncts. Thus in *Entering the house, he was stopped by the neighbors* string analysis gives, in contrast with the above:

elementary sentence: *He was stopped*
adjunct of Ω_{33} (see 3.22): *by the neighbors*
sentence adjunct: *entering the house*

Also, transformational analysis describes *He is slow in learning* as merely a unary transformation of *He learns slowly*. But string analysis gives for the first:

elementary sentence: *He is slow*
adjunct of Ω_{33} (see 3.22): *in learning*

and for the second:

elementary sentence: *He learns*
verb adjunct: *slowly*

Similarly, in *There was a man standing here*, which is a unary transformation of *A man stood* (with primitive adjunct *here*), string analysis gives:

elementary sentence: *There was a man*
sentence (?) adjunct: *standing*
verb adjunct on the sentence adjunct: *here*

Transformational analysis avoids some of the difficulties of string analysis (and of constituent analysis). For example, it is uncomfortable to list *by the neighbors* or *standing* as adjuncts, though they

[9] More exactly, sentence 5 is operated on by the unary transformation called "extraction", yielding *Factory is what we visited*, and by the binary transformation of sharing into sentence 4. The various transformations and operations mentioned here will be presented in detail in a later paper of this series.

[10] The constants of a transformation are the segments which are added to a sentence in the course of transforming it, and which serve to characterize that transformation. In *There was a man standing here*, the constants are *there*, *be*, *-ing*.

are clearly such by string analysis. It is not clear to what they are adjoined. Nor do they have the modificatory meaning that is seen in other adjuncts, including those other adjuncts in which the word categories are the same as in them (compare *stopped by the neighbors* with *stopped by the entrance* and *stopped near the entrance*).

1.4 COMPARISON OF THE THREE ANALYSES

If we consider all three types of analyses, we note first that string analysis is intermediate between the other two: It isolates one elementary sentence out of each sentence; constituent analysis isolates no sentence; while transformational analysis reduces the whole sentence to elementary sentences (with primitive adjuncts) and constants.

The differences among the three are not necessarily that one builds upon the other: It is possible to define the operations of each without using any results or concepts of the others[11] (though some results of morphology can be used, and may perhaps be required, in each). Nor does the difference lie in the power of the three to characterize different sets of sentences, or in that the maximal set characterized by one analysis is a proper part of the set characterized by another. For each of these types of analysis can describe all the sentences of a language (though at very different cost in complexity of the description). This is so because the complex detail of each language and, not to put too fine a point on it, the irregularities and the not-fully-carried-out analogies, force each type of analysis to provide in its statements for cases of special subsets of word-categories or structures; and statements of this form can be used to describe any special cases that diverge from the main rules and elements, or even any entirely different classes of sentences.

The difference is rather in how the three analyses interrelate the sentences and sentence-segments of the language: For each characterization of a sentence relates that sentence to its decomposition products and also to other sentences having a similar decomposition. Thus, constituent analysis shows to what extent the sentences can all be viewed as sequences of two constituents, subject and predicate, with sentence adjuncts deployed around them. String analysis relates all sentences having the same elementary sentence, the same adjuncts, etc. Transformational analysis goes far beyond either in bringing together the sentences which we feel should be brought together. Thus it relates *He is slow in learning* with *He learns slowly;* and *He began to speak* with *He spoke;* and *He seems young* with *He is young;* and *whom I saw*, adjoining *man*, with *I saw the man;* whereas neither constituent analysis nor string analysis shows direct relation between the members of each pair.

In addition, transformational analysis, in reconstructing the component sentences out of the transformed segments of the original sentence, tells much more about each component than do the other analyses. Thus it gives the sentential relation between the word-categories of a segment by transforming the segment into a sentence (even if part of the reconstructed sentence has to remain undetermined): e.g. it shows that *house* is the object of *enter*, that *neighbors* is the subject of *stop* in the active. And it reconstructs zeroed and shared elements, as in *We visited the factory* from *which we visited*, and in *He entered the house* from *entering the house.*

Nevertheless, though transformational analysis is the most refined, all three analyses are relevant, for language has the properties of all three. To see this even cursorily, we may consider what a language would be like if it lacked any one of these properties. To have no constituent structure, a language would have to distribute its adjuncts irregularly, or in any case not contiguously to the segment in respect to which they are adjoined. More generally, segments having the same or interrelated syntactic relations (in respect to specified other segments) would have to occupy, in different sentences, different positions in relations to the other segments. There would then be no reason to analyze both the elementary sentence and the sentences expanded from it by adjuncts as the same sequences of struc-

[11] This is clearly the case for constituent analysis, which was developed independently of the other two. For string analysis, see sections 2 and 3 below. For transformational analysis, the presentation will be given in a forthcoming paper of this series.

tures (that is to say, as having the same successive constituents).

To have no string structure, the segments of a sentence which is adjoined into a host sentence by a binary transformation would have to be distributed (irregularly intercalated) among the segments of the host sentence or of its other adjuncts. That is to say, material would be inserted into the sentence, but not all of it at one point. A stronger case would be if the contributions in a sentence which are due to the primitive adjuncts and to the binary sentence-adjoinings all took the form of alterations in the words, the morphophonemic shapes, or the word-order of the elementary sentence to which they are being added; so that an inserted adjunct or operation would disappear entirely from the segmentation, leaving its trace only in some modification of the words of the host sentence. Here we no longer have insertions at all, but only modifications or replacements (though these could preserve sameness of constituent structure for the elementary sentence and for those obtained from it).

To have no transformational structure, a language would have to have only one form to each elementary sentence (then it would have no unary transformations such as the passive); and all additions to the host sentence would have to be primitive adjuncts not derived from some independent sentence (then it would have no binary transformations for combining sentences).

The fact that sentences, to a large extent, have constituent regularity, i.e. that there is a correspondence between the successive C_i-word-categories of an elementary sentence and the C_i-phrases of other sentences, makes it possible to formulate transformations (to a large extent) in such a way that they can operate on an expanded (or even replaced) C_i-phrase in the (denumerably many) derived sentences of some sentence-class $\{A\}$ in the same way that they operate on the corresponding C_i-word in the (finitely many) elementary sentences of class $\{A\}$.

The result that sentences have a string structure is due to the fact that transformations are not merely arbitrary reshufflings and constant-addings and sentence-combinings, but (except for a few cases) are operations which send the sentences of one class of elementary sentences into the form of another class of elementary sentences.

Thus, the unary reformulation of a sentence sends the segments of the original elementary sentence and the added constants into the positions of the segments of some other elementary sentence. And the binary insertion of a second sentence into a host sentence sends the transformed second sentence into the position of some segment of the host elementary sentence (with primitive adjunct). If all (or certain classes) of the elementary sentences have certain properties (such as subject-predicate, with plural extending over both; verb-object; primitive adjuncts in particular places), then the new sentences created by transformations in the form of these elementary sentences will also have the same properties. Thus from *He walks slowly* we obtain transformationally *He is slow in walking* which however has string properties similar to those of *He is slow in the morning*. Hence the transformationally derived sentences are string-analyzable into apparent elementary sentences plus adjuncts, by use of much the same indications as mark the original elementary sentences and their primitive adjuncts.

2. HOW THE STRINGS ARE ESTABLISHED

To determine the elementary sentences and the adjoinable elementary word-sequences (adjuncts) of a language, we begin (at least in the present formulation) with the morphemes and words of the language, with their classification by morphological properties (nouns, verbs, etc.). The objectives of 1.1 will be satisfied if we then determine for each sentence, and for each adjunct isolated from the sentence (or from an adjunct), what is the elementary part of that sentence or adjunct. The elementary part A_0 of a sentence or adjunct A is that part of A which is an elementary member of the class $\{A\}$ to which A belongs. To be a member of $\{A\}$, A_0 must have as its segments a sequence of classes which is present in the other members of $\{A\}$, and A_0 must occur in the same positions relative to other sentences and adjuncts as do the other members of $\{A\}$: A_0 must have the same structure and the same properties of occurrence as the other members of $\{A\}$.

We now consider each sentence S as a sequence of morphological word-categories (or sub-categories, or disjunctions of categories, or rarely sequences of categories) s_i. When we are given an arbitrary sentence S, we isolate out of it the elementary part S_0 by asking what contiguous sequences of the s_i can be excised, one sequence at a time, by operations of general or nearly general applicability, the residue of S after each excision being still a sentence of the language.[12] For example, in the sample sentence of 1.1 we can

[12] The nature of this analysis as a sentence-completion process (corresponding to the excision process presented here) was stressed by Henry Hiż in Steps Toward Grammatical Recognition, *Advances in Documentation and Library Science*, vol. III, part 2 (1961).

excise	leaving a sentence
1. *today,*	*automatic trucks from the factory which we just visited carry coal up the sharp incline*
2. *automatic*	*trucks from the factory which we just visited carry coal up the sharp incline*
3. *from the factory which we just visited*	*trucks carry coal up the sharp incline*
4. *up the sharp incline*	*trucks carry coal* $= S_0$

No further excisions are possible preserving sentencehood (though we could have excised first 3a. *which we just visited*, and then 3b. *from the factory*). The order of excisions above is irrelevant (except for 3a, 3b), though in some situations order, and even non-contiguity of the excised sequence, may have to be included in the operations determining S_0. For a given S, there are often several ways of extracting an S_0 – different S_0. However, if we wish the isolation of S_0 to be such that the excised sections can each be further analyzed into elementary adjuncts (as below), and all in a regular way, we may find that only one way of isolating S_0 (i.e. only one choice of S_0) is satisfactory, for each reading (formulaic representation) of the sentence (see footnote 2 and section 6.7). The isolated S_0 has the same properties of occurrence as S, namely those of a sentence.

We now turn to the adjoined word-sequences Z which were excised in the course of isolating S_0 (namely, the excised items 1-4 above). From each of these, Z_i, we again seek to isolate the elementary part Z_{i0}. We do this in the same way as for S_0, namely by excising successively various word-sequences included in Z_i, the residue of Z_i after each excision having the same properties of occurrence that Z_i had. Thus from excised item (adjunct) 3 above we can omit 5. *which we just visited*, leaving *from the factory*. This process can be repeated for each of the adjuncts excised out of Z_i in the course of isolating Z_{i0}, until every word-sequence is in elementary form. Thus from excised item 5 above we can excise 6. *just*, leaving *which we visited*.

Any such analysis of a single sentence or a small group of sentences would not be considered as valid for the language unless the (class of) elementary strings which are produced by it combine in the same general way in all sentences of the language, with restrictions that are not artificial for a reasonable theory of the language and for a reasonable interpretation of the theory.[13]

The informant problem and the repeated testing of tentative segmentations for their general validity are much the same in string analysis as in the usual descriptive linguistics. E.g., given a very short sentence, we check (via an informant) if any proper part of it (not necessarily connected) is also a sentence. If so, we list the residue as a tentative adjunct (possibly a combination of adjuncts), and note what are the elements or sequences to which we may tentatively say the adjunct is adjoined (and on which side). When we excise the tentative adjuncts of another sentence, we test them to see if they contain one or more of the previously recognized adjunct structures, appropriately adjoined. In addition to known and new adjunct structures, we may find previously recognized structures adjoined in a somewhat different way, or structures that are similar but not identical to ones previously recognized. In some cases we may have to redefine our word-categories, or at least to establish new subcategories of them, as for example if it turns out that a particular adjunct structure occurs not after all members of some word-category but only after an identifiable subset of it.

The excisability of adjuncts is improved, if we correct for automatic differences which may appear in the host string (sentence or adjunct) when a particular adjunct is adjoined to it. For example, we may hesitate to excise *not* from *does not contain*, since the residual *does contain* occurs only in a different stress. However, we can take

[13] Various difficulties may arise which can only be discussed in a fuller treatment. E.g. there may be words, or even whole subcategories, which occur (at least in a particular neighborhood of other categories) only when certain other categories are adjoined to them. Thus in *The meters are of a dubious quality*, or *They produced meters of a dubious quality*, we can hardly excise *dubious* as an adjunct of *quality* (since *The meters are of a quality* is doubtful). In this case, we would accord both with the transformational analysis and with the semantic interpretation, if we said that sequences which would be considered adjuncts when adjoined to other nouns are not such when the noun is of the classificatory subcategory (as *quality* is).

does before a verb to be a variant of *-s* after a verb, the variant appearing (inter alia) when certain *D* adjuncts are adjoined to the left of the verb. Then excising *not* from *does not contain* leaves *contains*, which is eminently acceptable, just as excising *not* from *may not contain* leaves *may contain.*

Various sentence-structures may be found which do not admit directly of a string analysis. In particular this is the case with sentences which include another (undeformed) sentence as a required part of their structure. For example in *I said* [*that*] *the control is excessive* or *That he's wrong is certain* we can find elementary sentences *The control is excessive* and *He's wrong*, and each residue has a structure which occurs in all sentences of its type: $N\ V'$ [*that*]... (where V' includes *say*, *claim*, etc., and the sentence occurs both with and also without the bracketed material), and *That*... is A_s (where A_s is a subcategory of situation-describing adjectives including *certain*, *clear*, *odd*, etc.). These residues are not themselves independent sentences, and would be considered by the present analysis to be adjuncts. However, they differ from adjuncts in certain respects (e.g. they don't really modify the elementary sentence); and, differently from adjuncts, these residues would themselves become sentences, if we replace *that* (if present) plus the elementary sentence by a sentential pronoun (a pro-nominalized-sentence) such as *this:* e.g. *I said this. This is certain.* We can describe these structures, then, either as containing an elementary sentence plus an adjunct which is a fragment of an elementary sentence; or else, modifying the definition in footnote 1, as consisting of an elementary sentence that includes an elementary sentence (considering, e.g. $N\ V'$ *that* plus pro-nominalized-sentence as itself an elementary sentence). (See $\Sigma_{5,6}$ and $\Omega_{8\text{-}13,\,15\text{-}18}$ in section 3.11.)

Another special problem is that of conjunctions. In English we may find after almost any constituent X_1 of a sentence a conjunction K followed by a sequence X_2 structurally identical (or grammatically equivalent) to that constituent: $X_1\ K\ X_2$. We may say that the $K\ X_2$ is an adjunct of X_1. This can also be said when X is a whole sentence structure S. However, it is also possible to say in the case of $S\ K\ S$ or $K\ S,\ S$ that we have conjoined elementary sentences rather than that the conjunctional $K\ S$ is the adjunct of the other.

The process of determining what is the elementary sentence and what are adjuncts and to what these adjoin, first for very short sentences and then for longer ones, is not hard to grasp. The difficulty, as in much of linguistics, lies in the complexity of the material, in the fact that there will be many sub-types of each elementary sentence structure or of the rules about adjoining of various adjuncts, and so on. Some of the special conditions that may be met have been mentioned here. There can be many more special conditions: for example, some elementary sentence structures may be unexpandable (may take no adjuncts).

3. AXIOMATIC STRING THEORY

It may happen, as in English, that the elementary strings obtained from any general method of decomposition are not quite the most convenient ones for characterizing the string structure of all sentences. We may, after obtaining the elementary strings as in section 2, see that some modification of our results would yield a more satisfactory characterization. For example, in *He runs the farm profitably* we have to say that the elementary sentence is *He runs*, with *the farm* and *profitably* each adjoined to *runs.* In such cases, the adjoining of *the farm* incurs a difference in meaning of the verb *runs* and a difference in its adverbs (e.g. *profitably* as against *quickly*). This is not the case in, e.g., *He read slowly* as against *He read the letter slowly.* We may therefore wish to say that in *He runs the farm profitably* the elementary sentence is *He runs the farm*, and is not directly related to the elementary sentence *He runs* of *He runs quickly;* while in *He read the letter slowly* the elementary sentence is *He read.* This would be expressed by setting up various subcategories of verbs: V_a, which does not accept object-adjoinings; and V_b, which does. In addition, there would be V_c, which has an object N in its elementary string (not as an adjoining), and so on. If we adopt a particular categorization of words, designed to yield the above results (*sleep, run* in V_a, *read* in V_b, *wear* in V_c), we would obtain:

	elementary sentence
He sleeps quietly (verb V_a)	*He sleeps*
He runs quickly (V_a)	*He runs*

He read slowly (V_b)	*He read*
He read the letter (V_b)	*He read*, adjoined object *the letter*
He wears a hat (V_c)	*He wears a hat*
He runs the farm profitably (V_c)	*He runs the farm*

Since *run* is listed as a member of V_a, it cannot have an object adjoined to it. Hence when we see an object following *runs* it can only be as part of the elementary string of *runs;* thus *runs* is here a member of V_c.

The characterization of sentences by such a modified set of elementary strings can be expressed in an axiomatic theory, which presents, in terms of a particular syntactic categorization of the words of the language, a particular body of elementary string-formulas (but departing only in some simple way from some method of determining strings as in section 2), together with rules for combining them. Each string-formula (including a sentence-formula) is a sequence of segments, each of which consists of a stated word-category (or subcategory or disjunction of categories) or of a stated string-formula (or disjunction of them). And each has particular properties of occurrence: it occurs independently; or it occurs to the right or left of a particular string-formula, or of a stated segment of a particular string-formula, or of a particular category in any string in which that category occurs. When each segment of a formula F is replaced by a word which is a member of the category occupying that segment, the result is a word-sequence which occurs in sentences of the language precisely as F occurs in the string-formulas of the grammar.

For English we set up the following sets of axiomatic string formulas:[14]

[14] The following list is incomplete and not stated in detail, but it includes the great bulk of reasonably common string formulas of English. The strings and subcategories can obviously be classified by various properties which would make the list more coherent. For example, $\Omega_{6\text{-}29}$ (except *14*) are cases in which the verb has an independent sentence as its object; in $\Omega_{14,\ 30\text{-}32,\ 33.6\text{-}8}$, the range of Σ for the verbs that take these are the same as the range of Σ for the first verb within the Ω (e.g. *I thought to go. I go.; He is good at running. He runs.*). The subcategories of the word-categories which are required for the formulation of all these strings are all the subcategories which a string analysis has to distinguish. Thus we recognize the various V_{ij} subcategories of verb, such subcategories of N as N_s (sentence-names, such as *result, idea, plan*) or those which may be required under Σ_1, Ω_2, and possibly the detailed pair-categories of $\Omega_{28,\ 29}$. The subcategory D_b in $\Omega_{33.3}$ includes a few locative adverbs which are objects of *be* (*nearby, there*). The strings will be given in greater detail in later papers of this series.

3.1 CENTER STRINGS

(see 3.4 for the term "center")

c = center strings, which occur as sentences.

3.11 The major sentence type

$c_1 = \Sigma_i\, t\, V_{ij}\, \Omega_j$ (The major assertion-sentence formula in English), where[15]

t = tense word or morpheme: *-ed* past, zero (and *-s*) present, after the V; the words *will, can, may*, etc., which occur before the V may be considered members of t or else adjuncts to t.

V_{ij} = that subcategory of V which occurs with Σ_i and Ω_j. Many verbs are members of more than one subcategory, especially for closely related Ω: e.g. *ask* is $V_{1,\ 17}$ (*ask him whether he finished*) and also $V_{1,\ 18}$ (*ask of him whether he finished*); this may be abbreviated as $V_{1,\ 17\text{-}18}$.

Σ = the subject of the verb.

$\Sigma_1 = N$ (with various subcategories): *The food tasted fine.*

$\Sigma_2 = \Omega_6$: [*His*] *leaving home surprised everyone.* Also $= \Omega_8$: *The fact that he left home.* Also, rarely, $\Sigma_i\, V_{ij}$ *ing* Ω_j: *Children fleeing napalm was a familiar sight in Angola.*

$\Sigma_3 = \Omega_7$: [*The*] *barking of dogs was loud.*

$\Sigma_4 =$ [*for* N_k][16] *to* $V_{k\lambda}\, \Omega_\lambda$: [*For him*] *to go there is foolish.*

[15] Each symbol indicates a single word except as otherwise stated. V can be taken to include the cases of V plus immediately following adverbial-prepositional complement, unless the latter is considered an adjunct of class $r_{v_{ij}\ k}$ (*take over*, *break up*, etc.); P includes rare $P\ P$ (*over against*, *near to*, *out from*, etc.); N includes *the A* (*the good*, *the smaller*, *the large*, etc.) which may be called an N-replacer, and also certain V *ing* (and, of course, V with nominalizing suffixes, e.g. V *ation*). Certain N (called count-N) are always preceded by a word of the *a* category ($\lambda_{n\ 4}$).

[16] Brackets around symbols indicate that the string occurs (without change of its properties of occurrence) both with and also without the bracketed material.

$\Sigma_5 = \mathit{that}\ \Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *That he came here surprises me.*

$\Sigma_6 = \mathit{whether}$[17] $\Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *Whether he will do it [or not] doesn't interest me. Whether* (or: *If*) *he will do it is unclear. Who will do it is unclear. What he will do is a question.*

Note that the verb following Σ_6 is either a V_{6j} (like *interest*) or $V_{33}\ A_s$ or $V_{33}\ N_s$ (i.e. a verb of the *be* category plus an appropriate subcategory of A_s or N_s as in *is unclear, is a question*). And so for $\Sigma_{4,5}$.

Ω = the object of the verb.

Ω_1 = zero: *The furies slept.*

$\Omega_2 = N$ (with various subcategories)[18]: *He invented the calendar.*

$\Omega_3 = P\ N$ (subcategorized according to the particular P): *He relies on luck. He trusts to luck.*

$\Omega_4 = N\ P\ N$ (subcategorized according to the particular P): *He attributes everything to her.* For a certain subcategory of $V_{k\,4}$ we have as Ω_4 not only $N_i\ P\ N_j$, but also $N_j\ N_i$: *He gave the book to me. He gave me the book.*

$\Omega_5 = A$ (many small subcategories: most verbs that have A object have only a few particular A as object; there may also be some categories of particular D objects): *It loomed large. The moon shone bright.*

$\Omega_6 = [N_k\text{'}s]\ V_{k\lambda}\ \mathit{ing}\ \Omega_\lambda$ (and also, for $\lambda = 2{,}4$: $[N_k\text{'}s]\ V_{k\lambda}\ \mathit{ing\ of}$ (or: *by*) Ω_λ); also with nominalizing suffixes in place of *-ing*: *She opposed his leaving home. I reviewed his description of the experiment.*

$\Omega_7 = [\mathit{the}]\ V_{k\lambda}\ \mathit{ing\ of}\ \Omega_\lambda\ \mathit{by}\ N_k$ (and also, for $\lambda = 1$: $[\mathit{the}]\ V_k\ \mathit{ing\ of}\ N_k$); also with nominalizing suffixes in place of *-ing*: *They oppose the testing of bombs by anyone. We heard the barking of dogs.*

$\Omega_8 = N_s\ \mathit{that}\ \Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *We derive the result that n = 3.* Verbs $V_{k\,8}$ usually occur also as $V_{k\,2}$ with members of the subcategory N_s of N, and could be derived from these $V_{k\,2}$ by the adjunct $r_{n\,12}$.

$\Omega_9 = [\mathit{that}]\ \Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *I know [that] he is here.* Many but (apparently) not all members of V_{k9} are also in V_{k8}.

$\Omega_{10} = \mathit{it\ that}\ \Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *I doubt it that he is here.* Acceptability varies.

$\Omega_{11} = [\mathit{that}]\ \Sigma_k\ \mathit{should}\ V_{k\lambda}\ \Omega_\lambda$: *They require that this should be done.*

$\Omega_{12} = [\mathit{that}]\ \Sigma_k\ V_{k\lambda}\ \Omega_\lambda$: *He insisted that this be done.*

$\Omega_{13} = \mathit{whether}\ \Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$: *I wonder whether he will come.*

$\Omega_{14} = \mathit{whether\ to}\ V_\lambda\ \Omega_\lambda$: *I wonder whether to come.*

$\Omega_{15} = N\ [\mathit{that}]\ c_1$:[19] *I told him [that] they came.*

$\Omega_{16} = P\ N\ \mathit{that}\ c_1$: *I reported to him that they had come.* (Subcategorized according to the particular P, as is also Ω_{18}.)

$\Omega_{17} = N\ \mathit{whether}\ c_1$: *I asked him whether it was true.*

$\Omega_{18} = P\ N\ \mathit{whether}\ c_1$: *I inquired of him whether it was true.*

$\Omega_{19} = \mathit{for}\ N_k\ \mathit{to}\ V_{k\lambda}\ \Omega_\lambda$: *She prefers for him to stay.* The acceptability of Ω_{19} varies.

$\Omega_{20} = \Sigma_k\ \mathit{to}\ V_{k\lambda}\ \Omega_\lambda$: *I want him to take it. They ordered him to take it.*[20]

[17] The position of *whether* in all Σ, Ω can be occupied by *if*. It can also be occupied by *wh* plus a pronoun of a subcategory of N or $P\ N$, or N*'s* or A, in which case a N or $P\ N$ or A of that subcategory is excised from the Σ_k or Ω_λ following the *whether*. (If Ω_λ itself contains some other Ω_λ, the excision may be from the included Ω.) The excision may also be from certain adjuncts of $V_{k\lambda}$. From a string point of view, there is no act of excision such as there is in transformations, or (differently) in the discovery procedure of section 2. If we say that a string $z = x - y$, i.e. x with excision of y, we mean merely that the segment-sequence $z = x$ except that a segment y present in x is not present in z. Compare *whether he will do it* with *who will do it* (lacking $N = \Sigma_1$), or *what he will do* (lacking $N = \Omega_2$); compare *whether he will do it with elan* with *how he will do it* (lacking $P\ N = a_{v\,2}$).

[18] All N in Ω, and after *for*, *of*, *by*, and P in general, are accusative if pronouns (*him*, *them*, etc.) except in $\Omega_{33\cdot1}$ optionally, and in Σ following [*that*], *whether: I saw him. I saw him go. This is he. I know* [*that*] *he went. I wonder whether he went.*

[19] We write c_1 for $\Sigma_k\ t\ V_{k\lambda}\ \Omega_\lambda$. The symbol was not used heretofore in order to bring out the differences among Ω_9, Ω_{11}, and Ω_{12}. When c (or $\Sigma\ t\ V\ \Omega$) is given as part of a formula, it is understood that c or its segments can have added to them all the adjuncts which are permitted to them in this theory. Thus if we state here that $\Sigma_i\ t\ V_{i\,15}\ N\ \mathit{that}\ c$ is a string of the class c_1 (and this is what is stated in 3.11 under c_1 and Ω_{15}) it follows that the c which is inside Ω_{15} may have any permitted adjuncts (as in *I told him that they of course came immediately*), and also the newly resulting c_1 may have any permitted adjuncts (as in *I of course told him immediately that they came*).

[20] In $\Omega_{20\text{-}29}$, the Σ (if long, e.g. with many adjuncts) and the N (also in Ω_4) also occur after the rest of the Ω instead of before it: *He threw open the door. He viewed as the end what was only a later stage of development.* Compare a_v and $r_{v_{ij}\,k}$.

$\Omega_{21} = \Sigma_k V_{k\lambda}$ *ing* Ω_λ: *I felt it coming to a head. They set it going.*

$\Omega_{22} = \Sigma_k$ *to* $V_{k\lambda}$ *ing* Ω_λ: *They set the bell to ringing.*

$\Omega_{23} = \Sigma_k P V_{k\lambda}$ *ing* Ω_λ: *They restrained him from going.*

$\Omega_{24} = \Sigma_k V_{k\lambda} \Omega_\lambda$: *They made him go.* There is also *He made do with it.* And in addition to *He let it go* also *He let go of it* and *He let go* (the Ω constiting of particular V).

$\Omega_{25} = \Sigma_k$ *as* Ω_{33}: *I view it as the end.*

$\Omega_{26} = \Sigma_k \Omega_{33}$: *This left him stranded.*

$\Omega_{27} = \Sigma_k \Omega_{33.1}$: *They elected him president.*

$\Omega_{28} = N A$: *He threw the door open. He drinks it black.* There are many small subcategories here, including D as the second part, somewhat as in Ω_5.

$\Omega_{29} = N D$: *He left the door ajar.* Here, too, small subcategories.

$\Omega_{30} =$ *to* $V_k \Omega_k$: *I thought to go.*

$\Omega_{31} = V_k$ *ing* Ω_k: *He began slipping.* Also for $k = 2,4$: [*the*] V_k *ing of* Ω_k (also with nominalizing suffixes in place of *-ing*): *He began the painting of frescoes.*

$\Omega_{32} = P V_k$ *ing* Ω_k: *He refrained from painting frescoes.*

$\Omega_{33} =$ is a set of sequences all of which occur as object of the verb *be*. In the category V_{33} of *be* there are subcategories of verbs, called $V_{33.i}$, which occur with particular objects $\Omega_{33.i}$ (e.g. *seem, turn out*).

$\Omega_{33.1} = N$: This N generally has the same plural and gender affix as does the Σ in $\Sigma V_{33} N$ (where $\Sigma_{4\text{-}6}$ are taken as being singular): *They are actors. She was an actress. Whether he came is a question.* Many count-N which require a (footnote 15) occur here without words of the a category: *He was president from 1932 to 1945.*

$\Omega_{33.2} = P N$: *He is in class.* Some count-N occur here without the a category.

$\Omega_{33.3} = D_b$: *He is here.* Other D and $P N$ occur here when *that* c_1 follows them (with $\Sigma = It$): *It was quickly that he ran.* The transformational source for these is different than for $\Omega_{33.2,3}$.

$\Omega_{33.4} = A$: *They were bitter.*

$\Omega_{33.5} = A P N$: *They are fresh from the field.* This applies to particular $A P N$ sequences where the $P N$ is not an adjunct of A as it is in r_a below (compare *They are freshly from the field*).

$\Omega_{33.6} = A$ *to* $V_k \Omega_k$: *We are ready to go.*

$\Omega_{33.7} = A P V_k$ *ing* Ω_k: *He is good at running.*

$\Omega_{33.8} = V_k$ *ing* Ω_k: *We are coming.* (Or consider *be* here a subcategory of λ_v.)

$\Omega_{33.9} = V_k$ *en* Ω_k, excising the first N or Σ of Ω_k, except that for $k = 6\text{-}9$, 11-14, 19 (and rarely *31*) the whole Ω is excised.[21] For $k = 3$; 16, 18, 32, the N (or V *ing* Ω) is not excised after certain verbs. V_k in whose Ω_k there is no excision do not occur in $\Omega_{33.9}$. That is to say that there is no passive of verbs whose object is $\Omega_{1,5,30}$, and in certain verbs whose object is Ω_3, etc. *The calendar was invented* [*by him*]. *Luck is unhesitatingly relied on* [*by many people*]. *Everything was attributed to her. His leaving home was opposed. That he is here is known.* We cannot readily obtain in string analysis the fact that the excised part of Ω_k appears as the Σ_λ of the formula Σ_λ *is* V_k *en* Ω_k with excision; nor the fact that V_{ik} *en* Ω_k with excision is often followed by an a_{v2} which consists of *by* Σ_i (*attributed to her by him*, related to *he attributed*...); for these are transformational.[22]

3.12 **Other sentence types**

$c_2 = t \Sigma_i V_{ij} \Omega_j$?[23] But when $t =$ *-ed* or zero (and *-s*) alone, the V does not include *be*: instead we have t *be* $\Sigma \Omega_{33}$ (and also t *have* $\Sigma \Omega$); also when *wh-* plus pronoun of a category of N or $P N$ or N's or A (or P *wh-* plus pronoun of N) occurs before the t, an N or $P N$ or A of that category is excised from the Σ or Ω or from certain adjuncts. *Will he come? Did he come? Does he have time? Will he be here? Is he here? Has he the time? Who will come? What fell? Where will he sit? On what is he standing? How did he do it? Why was he elected president?*

[21] There are various additional conditions: e.g. certain V with Ω_{20} do not occur in $\Omega_{33.9}$: There is no direct passive of *I want him to take it.*

[22] We could also define $\Omega_{33.10} =$ *to* $V_k \Omega_k$ as in *He is to go soon. He was to go;* but the *is, was* is peculiar in not accepting *will, can*, etc.

[23] And $t \lambda_v \Sigma_i V_{ij}$ *en* Ω_j: *Has he brought the books?* When *-ed* and zero (and *-s*) tense occur before V (or if the V before them is excised as in 3.25) they have the morphophonemic forms *did, do, does: He walked. Did he walk? He has it. Has he come? Does he have some?*

$c_3 =$ zero morphophonemic form of *you* plus $V_{ij}\ \Omega_j$*!: Go home! Wash yourself!* The imperative has no tense t. That the subject Σ is an (optionally) excised *you* is seen in the form *yourself* in Ω: this form occurs elsewhere in Ω only when Σ is *you*.

$c_4 = \Sigma_i\ t$ *not* $V_{ij}\ \Omega_j$; $\Sigma_i\ t'\ V_{ij}\ \Omega_j$. ($t'$ indicates t with contrastive stress.) This is like c_1 except that the rules for t and V are as in c_2 and footnote 23, yielding: *did not V, do not have, have not, will not be*, etc., and *I did see some, I do have some.*

$c_5 =$ *It* $t\ V_{33.i}$ *that* c_1; *It* $t\ V_{ij}\ \Omega_j\ \Sigma_i$ $(i = 4, 5, 6)$. *It seems that he did it. It interests me whether he did it* [*or not*]. *It is clear that he did it.*

$c_6 =$ *There* t *be* Σ_1; *There* t *be* $\Sigma_1\ \Omega_{33.i}$. *There's a man. There's a man coming.*

$c_7 = D_b\ t\ V_{1,1}\ \Sigma_1$: *Nearby sat a sailor.*

$c_8 =$ either of the two sections of $\Omega_j + \Sigma_i\ t\ V_{ij} +$ the other section of Ω_j $(j = 4, 15\text{-}18, 20\text{-}29, 33.5\text{-}7)$; $\Omega_j\ \Sigma_i\ t\ V_{ij}$ ($j =$ all other values, including all $33.i$, except $j = 10, 11$): *Him we restrained from going, From going we restrained him* $(j = 23)$; *This I like* $(j = 2)$; *That he is here I know* $(j = 9)$.

Since $c_{4,5}$ (and to a lesser extent $c_{6,7}$) have segments which can be assigned to the successive $\Sigma\ t\ V\ \Omega$ of c_1, most of the combinings that c_1 undergoes (e.g. in $\Omega_{6ff.}$, r_n, a_c, r_c), $c_{4,5}$ (and to a lesser extent $c_{6,7}$) also undergo.[24]

[24] Some rarer classes of c can be defined, e.g. $c_9 =$ *Would that* c_1; we would consider this as a case of c_1 with zero I and zero $V_{1,9}$ as though from *I would wish that* c_1. We may also define $c_{10} = X\ t$ *be wh* + *pro-X* $+ c'_1 - X$ (where $c'_1 = c_1$, plus possibly any of $r_{v_{i(j)}}$, a_v, $a_{c\,1\text{-}6}$, $r_{c\,1}$; and where X ranges over N, $P\ N$, Σ, $\Omega_{5\text{-}9,11\text{-}15}$, and over the first symbol or else the post-first-symbol-residue or the last symbol of $\Omega_{16\text{-}29}$ and over the whole or the last symbol of $\Omega_{30\text{-}33}$ and over the adjuncts listed at the beginning of this parenthesis). E.g. *The book is what fell. Agressively is how he talks. Whether it was at all true is what I enquired of him.* With this c_{10} we can also define $c_{11} =$ *wh* + *pro-X* $+ c'_1 - X\ t$ *be* X (*What fell is the book.*) and $c_{12} =$ *It* t *be* X *that* $c'_1 - X$ (*It is the book that fell. It is whether it was at all true that I enquired of him.*).

3.2 ADJUNCT STRINGS[25]

3.21 Adjuncts of P, D

λ_p: P-adjuncts, adjoined to the left of prepositions P in any string in which P occurs:[26] D: *almost at the entrance.*

λ_d: D-adjuncts, adjoined to the left of adverbs D in any string in which D occurs:[27] D (i.e. the string D occurs as left adjunct of a string D): *very nearly free, quite nearby, extremely sloppily dressed.*

3.22 Adjuncts of A

λ_a: left adjuncts of adjective A in any string in which A occurs, but rarely on any A except the first in a repeating sequence $AA..A$; $\lambda_{a\,2}$ occurs to left of $\lambda_{a\,1}$.

$\lambda_{a\,1}$: N-[28]: *stone-cold.* Not repeatable, except that N- itself may have $l_{n\,1}$ to $l_{n\,3}$.[29]

$\lambda_{a\,2}$: D including *very*: *very young, quietly happy.*[30]

r_a: right adjunct of A (and in general of the strings in $\lambda_{n\,2}$, $\Omega_{33.2\text{-}9}$)[31]: $P\ N$: *serious in intention;* repeatable only with considerable restric-

[25] The grouping and characterization of the adjunct strings has been changed here from the earlier form in TDAP 15 on the basis of TDAP 27, N. Sager, *Procedure for left-to-right recognition of Sentence Structure* (1960): Classes of strings are named by lower-case letters. In the adjunct name x_y, x indicates the adjunction point and y indicates the category or string to which the adjunct is adjoined: e.g. r_n strings are adjoined to the right of N.

[26] Adjunctions are repeatable without definite limit (in some cases with various restrictions and with increasing stylistic difficulty) unless marked as non-repeatable. Hence we can have *almost quite at the entrance*, etc. The adjoining of x to y is independent of any other adjuncts y may already have, unless restrictions are stated: adjoining x_1 and x_2 to y yields $x_1\,x_2\,y$ or $x_2\,x_1\,y$.

[27] The symbol D is used here for a variety of word-categories which differ considerably in their properties of occurrence. Statements here about D apply only to particular subcategories of D.

[28] N- indicates N attached with compound-word stress to the following word.

[29] N- occurs similarly as left adjunct of V*en*, V*ing*, N*'s* when these are in positions of A.

[30] D also occurs as left adjunct on members of $\lambda_{n\,3}$ (e.g. *fully two weeks*). And D occurs as left adjunct of V*en* (*very shaken*), but only before certain V*ing* (*very interesting*) and only certain D before N*'s* (*entirely his*); and this only when these are in positions of A (as $\Omega_{33.4}$ or as λ_n).

[31] An adjunct of a category will be understood as applying to that category in any string in which that category occurs.

tions. Rare when A is itself an λ_n string, and then only with compound-word stress: *aged-in-wood whiskey.*

3.23 **Adjuncts of N**

λ_n: left adjuncts of N excluding pronouns. $\lambda_{n\,(i+1)}$ occurs to the left of $\lambda_{n\,i}$.

$\lambda_{n\,1}$: *N-*: *labor-union.* Non-repeatable, except that *N-* may itself have $l_{n\,1}$ to $l_{n\,3}$. Hence *N-N-N* is not two occurrences of $l_{n\,1}$ on N, but one occurrence on N and one on *N-*: *jacket-design-exhibit.*

$\lambda_{n\,1a}$: N (particularly for certain subcategories of N): *iron railing, group effort.*

$\lambda_{n\,2}$: *A, V en, V ing, N's.* There are several subcategories of A, and each of these subcategories of $\lambda_{n\,2}$ has particular relative position, making a number of successive $\lambda_{n\,2}$ positions, all to the left of $\lambda_{n\,1}$. Within a sub-position there is no repetition (except with comma or comma-intonation, which is a case of $r_{x\,1}$).[32] The N of *N's* may itself have $\lambda_{n\,1}$ to $\lambda_{n\,3}$. The position of A is occasionally occupied by longer sequences which occur in Ω_{33}, such as *A to V* Ω in *hard-to-distinguish;* these have compound-word stress here. *Wild plan, broken hopes, changing ideas, an old man's thoughts, large white canvas.*

λ_{n2a}: *N of*, for a particular subcategory of N including *kind, type, sort: He is a sort of investigator.*

$\lambda_{n\,3}$: numbers and quantifier words, in several positional categories, all preceding $\lambda_{n\,2}$: *five men, few men.* Not repeatable within a position.

$\lambda_{n\,4}$: The *a* category (article) including *a, the, some, either, no: some man.* Not repeatable. *a* and *the* are the main subcategories.

$\lambda_{n\,5}$: pre-article qualifiers: *Scarcely a man came.* More correctly, this is a left adjunct of the article, $\lambda_{n\,4}$, and of the numbers, $\lambda_{n\,3}$.

r_n: right adjuncts of N. The subclasses of r_n are not explicitly ordered. $r_{n\,1\text{-}9}$ correspond to $\Omega_{33.1\text{-}9}$.

$r_{n\,1}$: N (noun in apposition, with only particular subcategories of N, e.g. names of occupations, occurring in r_{n1}: *my friend the cellist.*

$r_{n\,2}$: *P N: the colors in the painting.*

$r_{n\,3}$: D_b: *the people nearby.*

$r_{n\,4}$: A_d (a particular subcategory of A): *the people present.*

$r_{n\,5}$: *A P N: strawberries fresh from the woods.*

$r_{n\,6}$: *A to* $V_k\,\Omega_k$: *workers ready to strike.*

$r_{n\,7}$: *A P* V_k *ing* Ω_k: *experimenters good at adapting instruments.*

$r_{n\,8}$: V_k *ing* Ω_k: *refugees seeking a haven.*

$r_{n\,9}$: V_k *en* Ω_k with excision as in $\Omega_{33.9}$: *books given* [*to*] *her.*

$r_{n\,10}$: $N_i\ V_{ij}\ \Omega_j - N$ (excising one N from Ω_j, except as in $r_{n\,11}$): *the book he gave her.*

$r_{n\,11}$: *that* $c_1 - N$ (excising one N from Σ or Ω, provided the N is not preceded by a *that* inside c_1, or by *whether* (or *wh*-words), or *for* (of Ω_{19}, Σ_4); but the zero form of *that* is not a bar to the excision). *The painting that disappeared, the man that I saw, the book that I thought was missing.*

$r_{n\,12}$: *that* c_1 (without excision), only after N_s: *the fact that he saw the man.*

$r_{n\,13}$: *whether* c_1, only after N_s: *the question whether he saw the man.*

$r_{n\,14}$: *wh-* plus a pronoun of a category of *N, N's, A, P N* plus c_1 excising (as in $r_{n\,11}$) a member of the corresponding category from c_1 or from certain adjuncts of it. Certain members of $r_{n\,14}$ occur also as right adjuncts of $\Sigma_{2\text{-}6}$. *The man whom I saw, the book which I thought was missing; That he wrote me, whichwas quite surprising, became known to all.*

$r_{n\,15}$: *to* $V_j\,\Omega_j$: *the man to do it.*

$r_{n\,16}$: [*for N*] *to* $V_j\,\Omega_j - N$ (excising one N from Ω_j as above): *the man* [*for you*] *to see.*

3.24 **Adjuncts of V**

λ_v: left adjuncts on the verb;[33] this is not a modifier of the verb. *have ... en* (and for a very few verbs, *be ... en*): *has gone, has taken, is gone.* Not repeatable.

[32] This observation is based on the work of Zeno Vendler, to appear in a later paper.

[33] The relation of *have* and *be* (as in $\Omega_{33.8\text{-}9}$) to following verb could also be formulated differently. This *have* gets the *t* and *ing* which are assigned to the V in the various formulas. The verb forms with *be, have,* etc. are given by the descriptions of t, c_2, λ_v, $\Omega_{33.8\text{-}9}$.

r_v: right adjuncts of the verb.

$r_{v_{i(j)}}$: Ω_j. We use $V_{i(j)}$ to indicate those verbs which occur sometimes as V_{i1} (i.e. with zero object Ω_1) and sometimes as V_{ij} (i.e. with some other object Ω_j. When $V_{i(j)}$ occurs with Ω_j, the Ω_j has to be considered as an adjunct of some kind, since the definition of an elementary sentence in $\Sigma_t\ V_{i(j)}\ \Omega_j$ is satisfied by $\Sigma_t\ V_{i(j)}$: *He reads books. He reads.* This adjunct does not have the meaning of a modifier of the verb, and is not repeatable.

$r_{v_{ijk}}$: certain D_k associated with particular V_{ij} which we may therefore call V_{ijk}. The D_k occurs before Ω except when the object is Ω_2 and consists of a pronoun without adjuncts: *look up the telephone number, talk over about this.* The D_k occurs after Ω_j when $j = 2, 6, 7$: *look the telephone number up.* Hence, when the object is a pronoun without adjuncts the D_k occurs only after Ω: *look it up.*

a_v: all-position adjuncts of the verb, occurring (differently for different subcategories) to the left or right of it (or of $V\ \Omega$). These do not occur before an Ω which begins with N, unless the N has several or long adjuncts: *He pronounced the sounds clearly. He pronounced clearly some very difficult sounds. He relied heavily on me.*

a_{v1}: D (not including *very*): *I quite forgot. I forgot quite. I forgot completely.*

a_{v2}: $P\ N$ (several subcategories, including of time, place, manner, with different relative positions): *He spoke with fervor.* Occurrence to the left of the verb is quite restricted: *He may with some success obliterate all the traces.*

3.25 Adjuncts of c

If an adjunct to c itself contains $V\ \Omega$ (as $a_{c\,4ff.}$) the occurrence of $V\ \Omega$ which is second or is headed by K may be excised (if it contains the same words as the other $V\ \Omega$). Hence we obtain $\Sigma\ t$ fragments as strings: *I'll go if you will.*

a_c: all-position adjuncts of the center, occurring to its left or right, or to the right of Σ (with its right adjuncts), or in the positions stated for a_v; and more rarely between any other segments. They are usually separated from the center by commas.

a_{c1}: certain individual words and phrases, e.g. *in general, today.*

a_{c2}: certain D: *clearly* (with subcategories of time, etc.).

a_{c3}: certain $P\ N$: *at this time, of a certainty* (with subcategories as in a_{c2}).

a_{c4}: subordinate conjunctions K_s plus c_1; there are several subcategories of the subordinate conjunctions (*since, if, because, while, as*, etc.) and after various ones the c_1 excises the $\Sigma\ t$ and adds *ing* to the V, or excises the $V\ \Omega$, or excises $\Sigma\ t$ *be*. Hence a_{c4} includes $K_s\ c_1$, and $K_s\ V_j$ *ing* Ω_j, and $K_s\ \Omega\ t$, and $K_s\ \Sigma_{33}$. *Since he will return here, Since returning here, Because he will, While a boy.*

a_{c5}: $P\ V_j$ *ing* Ω_j (not for all P): *after returning here;* also $P\ N_i$'s V_{ij} *ing* Ω_j: *before his coming here.*

a_{c6}: $P\ \Sigma_t\ \Omega_{33}$, and $P\ \Sigma_t\ V_{ij}$ *ing* Ω_j; for P = *with, what with, without* (if Σ_t is a pronoun, it is in the accusative): *with his head bowed, what with my friends having gone.*

a_{c7}: *wh* plus pronoun of a category of N, N's, A (with following N), $P\ N$ plus *ever* plus $c_1 - N$ (with excision as in r_{n14}): *whatever he says.*

a_{c8}: [*in order*] [*for* Σ_t] *to* $V_{ij}\ \Omega_j$, and *so as to* $V_j\ \Omega_j$, and similar constructions.

a_{c9}: *whether* c_1 *or* c_1*; whether* c_1 *or not: whether he goes or not, I will.*

λ_c: $D_c\ t\ \Sigma\ V\ \Omega$ *than: Scarcely had I returned than* c_1. D_c is largely the category in λ_{n5}. There are also other $D\ c_1\ K$ adjuncts, e.g. *I had but returned when* c_1.

r_c: right adjuncts of the center.

r_{c1}: Ω_{33}: *He died a failure, He walked off unrepentant, He lectures standing.*

r_{c2}: , *which* $c_1 - N$ (with excision as in r_{n14}): *He went away, which was a big help.*

3.26 Adjuncts of x

r_x: right adjuncts of x, where x = string or segment (i.e. subsequence, including word-category) of a string of any class except λ_{n4} (article).

r_{x1}: conjunction K (*and*, comma, *or*, *but*) plus a string or segment of the class (or category) x (or plus one listed as K-equivalent[34] to x): for $x = D$, *a quietly but intensely rebellious mass;* for $x = A$, *a deep and growing concern;* for $x = V$, *He will buy and sell books;* for $x = t\ V$, *He will buy and will sell books;* for $x = V\ \Omega$, *He will buy books and sell them;* for $x = c$, *He left, and she left too.* For $K =$ *and, or,* the x to which r_{x1} is adjoined may be preceded by *both, either,* respectively: *Either he or she will leave.* For $x = c_i$, the K-equivalent strings are most of the fragments (partial sequences, not necessarily contiguous) of c_i (such as do not leave a V without its Ω in the right adjunct of c_i): *He left and she too. He left and she will soon* (excising $V\ \Omega$, but keeping t). *He plays piano and she violin.*

r_{x2}: comparative conjunction consisting of a scope marker (indicating what is being contrasted) *rather, more, less, as* before or after a string x_0 or a segment x_1 of it (also *-er* after A), and a corresponding comparative conjunction *than, as* followed by a string or segment of class x_0 or x_1 or a fragment of it (as in r_{x1}). The string after *than, as* may have the same excisions in respect to the x preceding as were stated in r_{x1} (and also as in a_{c4}).[35] *He ran rather than walked. As much time as money will be lost. As much time will be lost as money. More men came than I had ever met. More men came than women. More men than women came. Men more than women came.*

[34] For example, λ_{ni} is not K-equivalent to λ_{nj}. But $r_{n14} = r_{n11}$, and certain subsets of a_{c2} (e.g. of time) are K-equivalent to the corresponding subsets of a_{c3}. A derived string or segment of class x, in the sense of the derivation rule below, is K-equivalent to an elementary member of x. A few details have to be added to the statement above, which would exclude *She will list and he will pack books*, but would admit *She will list and he will pack such of the books as we think worth keeping.* In addition, c_i is K-equivalent to c_j only with certain restrictions: *He studied it but how could he remember it?*

[35] Slight additions to this statement are required in order to make it apply also to the cases in which the scope marker is *too, enough* and the comparative conjunction is *for... to, so as to*, etc. Various restrictions have to be stated concerning the occurrence of the scope markers listed above.

3.3 PROPERTIES OF THE STRINGS

There are various restrictions and operations within the strings. The restrictions are mainly of two types: One segment of a string may in some cases be filled by members of a particular subcategory of the required category only if some other segment of the string is filled by members of a particular subcategory: e.g. certain verbs occur only with animate subjects. Or a string may be adjoined to a segment of another string only if that segment is filled by a particular subcategory. E.g. r_{n12} is adjoined to N only if N is N_s.

In English there are not many operations on the strings other than that of adjoining one to another. However, there are in addition some operations within single strings. Chief of these is the plural, which adds a plural suffix to the $\Sigma_{1\text{-}3}$ and removes the *-s* (present tense) of the t (the *-s* does not occur when $\Sigma_1 = I$, *we, you, they*). Some adjunctions to the center, which can be considered restricted members of a_c, are operators in that they do not simply adjoin one position of the string. Thus *not* can be viewed as an operator on c_1, inserted between t and V, requiring that the members of t normally suffixed after V appear as independent words before V: *He walked. He did not walk.* This would eliminate c_4.

In addition to this, the strings of a language have many properties which can be studied, and which help to characterize and classify them in a coherent string grammar of the language. In English, the adjuncts modify in meaning the strings or adjuncts to which they are adjoined (except for particular classes of adjuncts). The left adjuncts are mostly only one word long. The right adjuncts are mostly longer, and mostly have a characteristic marker at their head (to their left).

3.4 THE RULE OF DERIVATION

Let $X = X_1\, X_2 \ldots X_n$ be a string of class x, and Y a string of class y which has the property of occurring in a particular position i of strings of class x (i.e. to the right or left of X, or of X_j, or of any occurrence of a particular category A in X). Then the result of adjoining Y into X at i (written $Y X_{(i)}$) is again a string of class x: $Y X_{(i)}$ has the same properties of occurrence as X has. Hence if Y can adjoin X_k, and Z can adjoin X_k, Y can adjoin the result of adjoining Z to X_k and conversely, unless restrictions are stated (e.g. that Y or some subcategory of it can adjoin an occurrence of X_k only if Z or some subcategory of it is, or is not, adjoined to that occurrence of X_k); these restrictions are generally expressible by an ordering of the adjoinings. We may call the position i or (the right or left side of) the segment X_j the adjunction point for Y in X. The resultant $Y X_{(i)}$ may be called a derived string of the class x. And if Y is adjoined to a word-category A (occupying some segment X_k), we may call the sequence $Y\,A$ (or $A\,Y$) a derived segment (phrase) X_k of X belonging to the derived (phrase) category A. Within the derived (or the elementary) string (whether sentence or adjunct) or segment, the elementary string or segment from which it was derived by adjunction may be called the center of the derived string or segment. Thus an elementary string of class c is the center of the sentence in which it occurs.

Various statements follow from the above. The properties of occurrence of a center are those of the string or segment of which it is a center. Each sentence has only one center, for each reading of that sentence (see end of 6.7). If a string Y and a string Z both can be adjoined to X_k in a string X, then (aside from special restrictions) $Y\,Z\,X_k = Z\,Y\,X_k$, $X_k\,Y\,Z = X_k\,Z\,Y$. (The interpretation is that the meanings of the equivalent sequences are the same.) If a string Y can be adjoined either to Z at i or to X at j, and Z can be adjoined to X at j, then $(Y\,Z_{(i)})\,X_{(j)} \neq Y\,(Z\,X_{(j)})_{\,(i)}$. Hence, e.g. *The color of the book which I like* has two readings: *which I like* adjoined to *book*, and this (derived) *book* adjoined to *color;* and also *book* adjoined to *color*, and *which I like* adjoined to this (derived) *color*.

4. DECOMPOSITION OF SENTENCES

In terms of a list of axiomatic elementary strings (each having stated properties of occurrence), it is possible to decompose a sentence into strings which are present in it in accordance with their properties of occurrence. This will be called recognition of the string structure of the sentence in respect to the given list. No claim is made here that any list of strings can be complete for a language, or that all properties of a sentence can be given by its string decomposition. However, a great amount of information about the sentences of a language can be obtained by decomposing them in respect to a reasonably adequate string list. Recognition cannot be directly based on an unstructured scanning of the sequences of word-categories, since each category occurs before and after almost every other one, in one sentence or another. However, these categories are bound to strings: in an analyzed sentence, each category appears in a particular position in a small number of elementary (center or adjunct) string formulas, which in turn can be adjoined only to the left or right of particular other categories or string formulas.

4.1 EXPECTED WELL-FORMEDNESS

It follows that the recognition process proceeds in respect to an expected structure, that is to say, to the requirement of well-formedness in respect to the axioms and derivation rules:

1. from sentence beginning, it seeks a complete center sequence $\Sigma_i\ t\ V_{ij}\ \Omega_j$ (or, in the absence of this c_1, one of the other c_i) before reaching sentence end;

2. before or after each category in the sentence formula it seeks only such string formulas (either sequences or single categories) as would be permitted by the properties of occurrence of the axiomatic string formulas.

The problem, then, is to assign each word, as we come to it, to a given category (or a disjunction of them), and then to assign the category to a given position in a string which is permitted at that point. To do this, we may have to know the immediate (or more distant) neighbor, and (in some cases) at what point in the sentence-structure we are. To a large extent the recognizer can keep a record of its position, as it moves along the sentence, in the following way: when it enters a sentence *X*, or a string *X* permitted at that point, the entry into *X* requires that certain categories be met before exit; this is called the well-formedness requirement for the sentence or string *X*. If a category *Y* which is permitted but not required in *X* is met before well-formedness of *X* is satisfied, the well-formedness requirements incurred by the presence of *Y* must be satisfied before (or at the same point that) the well-formedness of *X* is satisfied; this is the requirement of nesting of well-formedness. In the Univac program the information obtained in a single scan (either left-to-right or right-to-left) is in many cases not sufficient to specify what point in what string has been reached.[36] In such cases a second scan, or a back-and-forth check is used. Where it is impossible to specify a unique analysis of a sentence, it is possible to indicate two or more analyses (readings), at least one of which must hold in the case under consideration (e.g. in footnote 2 above).

4.2 IDENTITIES

In computing the structural recognition, one can consider each string *I*, adjoined by the derivation rule, to be a left or right identity in respect to the computation of the well-formedness of the sentence, since for any string $A_i \varepsilon \{A\}$,[37] we have $A_i\ I = A_j \varepsilon \{A\}$. Our meeting each of Σ_i, t, V_{ij}, Ω_j, in order yields the decision that the sentence is well formed. Meeting any string permitted by the derivation rule constitutes adding an identity to the decision process at that point; it neither adds to the well-formedness nor detracts from it. The decision that a particular category within the sentence sequence is part of an identity depends on the conditions in the derivation rule, and can thus often be made without complete knowledge of the position of the category in the whole sentence: it is sufficient to see the position of the category within its string, and the relation of the string to its adjunction point within the including string. The fact that there are strings within a sentence whose internal composition and whose acceptability for sentence well-formedness is independent of their position in the sentence (but relates only to specific categories or adjunction points in other strings wherever these may occur) is what made possible the use of these strings for recognizing sentence structure.

4.3 INTERPRETATION

The string-recognition process reports for each sentence whether it is well-formed or not; and if not, what is lacking; and what sections of the sentence (or string) constitute the center of the sentence (or string), and what strings are adjoined to what parts of the including string (or sentence).[38] Such a report is particularly useful because, in all except certain types of cases, these relations have direct interpretation in what we may call grammatical meaning: most strings adjoined to *X* modify the meaning of *X*, the participants in a

[36] Such cases may arise: when a word can be assigned to more than one category; when a string can have more than one well-formedness requirement (e.g. *while* may be followed by a full sentence form or by just the objects of *be*); or when the well-formedness of a string, or the determination of its status as an identity or an *N*-replacer (i.e. adjunct of zero *N*), depend on its position within the sentence.

[37] $\{A\}$ indicating the set of strings having the same properties of occurrence that *A* has.

[38] The Univac program below, however, assumes that the sentence is well-formed, and decides only whether a particular reading of it is well-formed, i.e. whether a given structural characterization of it is possible (e.g. a given decision as to dictionary ambiguities, or as to the boundaries between the adjuncts of a preceding *N* and those of a following *V*).

well-formedness requirement have a relation to each other (subject, object) which they do not have to other material in the sentence; and words occupying different structural positions have different semantic properties.[39]

[39] For example, whereas all N may be followed by the string *that* c_1 excising N (*the piano that he bought, the report that they made*), a certain subcategory N_s of N may also be followed by the string *that* c_1 (without excising N: as in *the report that they made an H-bomb;* we cannot say *the piano that he bought it* or the like). Members of this subcategory of N have the general meaning of being descriptions or names of types of statements: N_s includes *idea, plan, suggestion, information*, etc.

5. COMPUTING THE RECOGNITION OF SENTENCES

We will speak of computing the structure of a sentence if, given a sentence and a set of string formulas and derivation rules, we can offer an effective procedure for deciding of which application of rules to which strings the sentence is a case.

The computing of sentence recognition requires more than a finite state device, but more only in a few specific ways. A finite state device suffices for the computing of sentence center (though, in the formulation of section 3, these can be unboundedly long); and when the structure of every sentence is described in terms of sentence centers, it is possible to say at what points more powerful devices become necessary.

The computing of sentence structures may be aided by the recognition in each sentence (or adjunct) of certain stations, indicating satisfaction of the successive well-formedness requirements of the string formula. The recognition of these stations when the simple center is interrupted by various strings is not always directly or uniquely calculable. In some cases, the task of calculating a station can be replaced by calculating a particular amount and kind of separation between two marks (word-categories or later-inserted brackets) in the sentence.[40]

It is useful to consider just how the adjunct strings are recognized, since this is at the base of any simple recognition process for sentences. A convenient method in the case of English is to define two sets of strings: first-order, which do not contain the verb-plus-object

[40] These methods are used in TDAP 19.

sequences ($V\,\Omega$ and $V + r_v$), and second-order, which do. This distinction is useful in English computation because many verbs have several possible objects, so that determining for a given V what part of the following material satisfies its object in that sentence may be quite difficult. In the Univac program the class of first-order strings had the following members:

1. N with any λ_n or $r_{n\,1\text{-}5}$ adjoining it. These are the first-order N-strings;

2. any A, with any λ_a or r_a adjoining it, which is not inside a first-order N-string;

3. any D or $\lambda_d\,D$ which is not inside an A- or N- or V-string;

4. the sequence P plus first-order N-string (or A-string);

5. the sequences $t\,V$ and $t\,\lambda_v\,V$, together with a_v; this is a first-order V-string;

6. the conjunction-category K followed by a word-category or a first-order string.

Inspection of these shows that first-order strings are sequences each element of which is one of the following:

word-category marks (which may be considered zero-order strings as in the case of P, or else fill the position of first-order strings, as in the case of pronouns);

in some positions first-order strings.[41] In any given first-order string not all of these may occur, and some may occur more than once. Some of the elements in a first-order string may be computational identities, e.g. A in an N-string; or a $P\,N$-string in a first-order V-string.

In contrast, we define second-order strings as any string containing the verb-plus-object sequence. In the Univac program this class had the following members: 1. c, 2. $r_{n\,6ff.}$, 3. $a_{c\,4ff.}$, 4. the conjunction K followed by a second-order string. It is seen that these second-order strings are sequences each element of which is one of the following:

zero-order string-heads (e.g. *that*);[42]

first-order strings (and N-replacer second-order strings), those being the elements required for well-formedness of the second-order string;

plus, possibly, as identities: certain first-order strings (D, A, $P\,N$, K-strings);

and any second-order strings as identities (except N-replacers); it follows from this statement that these second-order strings can be nested without limit, also that a sentence is itself a second-order string (except that it has no special string-head), so that no higher-order strings exist (in English).

For computation, the following properties of strings (in English) are important. Any two strings whether elementary or derived (including the sentence-center) either are disjoint, or one includes the other as a proper part (is nested within the other). Each string (again, derived or elementary) can be said to be internally connected, there being no element within its boundaries that is not part of it; this is achieved by treating each included string as an element (if only a computational identity) within the including string. Each category and each included first-order string (if any) occurring in a given position of a first-order string is either an element required to occur in that position of the string (except for identities), or else an identity permitted by the rule of derivation to appear in that position; in second-order strings, the string-heads are zero-order words, affixes, etc. (see footnote 42), the first-order strings are elements required or identities permitted at the position in which they occur, the second-order strings are identities permitted (or rarely N-replacers required) at the positions in which they occur (in the same sense as in the case of first-order strings). Hence, every string can

[41] Zero-order strings are single categories (word, affix, space, or word sequence assigned to a single dictionary category, as $P\,P$ to category P) which have the property that their computational status is never filled by a sequence of categories. First-order strings are included in $P\,N$ and in K-strings; and the $t\,V$-string includes any D or $P\,N$-string which is between the t and the V. Strings headed by a conjunctional X will often be called X-strings: e.g. K-strings. Otherwise strings are named by their center, e.g. $P\,N$-strings.

[42] In the case of *-ing*, *-en*, the string-head (which is generally at the left of the string) is suffixed to the first V; and in the case of r_{n10} the string-head is simply the position of the N as a second (non-appositional) N within the including string (see 7.5). In the case of c (the sentence center) we might wish to say that its string-head is sentence-initial space.

be treated as having a recognized computational status in the position at which it occurs within the including string.

Since strings are connected, and the return from n^{th} nesting to the n-1^{th} is simple, all that remains is:

1. to recognize the entry into a string: this is simple because almost every right-adjoined string has a string-head on its left;[43]

2. to compute to the end of the string, given its head. Some heads always introduce the same string, e.g. *because* $\Sigma_i\, t\, V_{ij}\, \Omega_j$. Some heads introduce shorter or longer strings, often depending on what precedes the string-head, e.g. *to* $V_j\, \Omega_j - N$ occurs only after N, but *to* $V_j\, \Omega_j$ (with no excision) occurs anywhere (including after N).

3. It remains to decide what kind of element the n^{th} nested string is within the n-1^{th}; most strings are always identities, or always N-replacers; but *wh*-strings ($r_{n\,14}$) following a zero N are in effect replacers of N: *What she cooked tasted good* (zero-N equivalent of *That which she cooked tasted good*), compared with *The food tasted good.*

[43] Some string-heads are identical with words in other categories (e.g. *that* in *I've seen that*); this is treated as a dictionary alternative (see section 7.3).

6. COMPLEXITY OF THE RECOGNIZER

The nature and degree of the complexities in a recognition process vary according to the different word-categories and sequences. The types of devices which are sufficient for sentence-decomposition are noted in 6.3-5. The difficulties peculiar to language recognition are noted in 6.2, 6.6-7.

6.1 UNIQUE CATEGORY SEQUENCES

The simplest recognizer would be one which assigns to a given word-category a single value (a single contribution to sentence well-formedness) without regard to its position or neighbors. This is possible for word-categories which occur only in a single class of strings, and is very rare in English: e.g. whenever the recognizer meets the word *the*, it knows that this is an identity operating on a (not necessarily immediately) following N (or zero N in *the* A).

6.2 LOCAL ALTERNATIVE VALUES

Next simplest would be a finite state recognizer which could recognize what is the computational status of a particular occurrence of a word-category (if it has different statuses in different occurrences), on the basis of a finite (and, to be at all useful, small) number of different previously-examined word-category sequences. This can be done to most (but not all) first order N-strings: scanning backward, the recognizer can start with every N it meets, read through certain predecessors determined by a tree, and decide the boundary (beginning) of an N-string (except for certain stateable situations). This can also be done to most cases of the other first-order strings, and to those second-order strings whose heads specify the string structure independently of their position in the sentence.[44]

6.3 NETWORK OF TREES

Certain complexities that go beyond a finite state device can be handled by a succession of scannings, each expressed in a tree, in which each end-node of the tree determines a particular string (marked, say, by particular bracketings) which is to be considered as a single element in later scans. For example, many N-strings can be recognized by scanning backward; other N-strings (and the N-replacer *the A*) and most other strings can be recognized by scanning forward between the sections bounded off by the backward scan.[45]

6.4 AUTOMATON WITH ERASURE AND CYCLING

More important: nesting, which can be treated (below) by keeping count, can also be treated by repeated scannings without keeping count: In the course of a left-to-right scan, if we meet a string-head X, we drop the calculation thus far, compute the end of the string headed by X and exit, replacing the string by a single-element mark

[44] If the string structure depends on the immediate predecessor of the string-heads (or on a predecessor whose distance can be stated in terms of a set of finite sequences of marks), a finite state recognizer could decide which string structure to accept in each occurrence. However, in all cases an unbounded number of adjuncts could intervene between the determining predecessor and the string-head. E.g. after all N, *that* heads an $r_{n\,10}$ string, but after N_s, *that* can also head a c_1 ($r_{n\,12}$) string. However, between the N_s and its following *that* any number of $P\ N$ may intervene.

[45] This is done in TDAP 18.

which has the computational status of an identity and is of zero order;[46] if before we reach the end of the string we meet another string-head, we again drop the preceding calculation and do as above. In this way, the only computation that is completed in each scan is that of a second-order string which contains no second-order string. Since this second-order string, upon being computed, is replaced by a zero-order identity mark, its immediately including string no longer has this second-order string within it, and may thus become available for computation on the next scan. Finally, the sentence itself, as the most inclusive second-order string, can be thus scanned after all its second-order strings have been replaced by single elements.

6.5 COUNTER OF NESTINGS

The work done in a succession of finite-state scannings can be performed in a single scan if a record is kept from the point at which a decision is made (at which a requirement, i.e. restriction, is incurred) to the point at which the decision is carried out (the requirement is discharged). There is no limit to the distance (in number of intervening words) between these two points within a sentence. Where the incurring and discharging of requirements is separated only by the incurring and discharging of similar other requirements as is the case in nesting, we need only count the successive incurrings and successive dischargings.[47]

[46] If the head has different values in different positions, the mark will be a variable, whose value in this occurrence will be decided when it is met in the course of scanning the including string (that is to say, when it is met in its environment). If different lengths of string occur after the given head, we have to record the string as having two or more readings (i.e. formulaic representations), one for each length that could be read. We can also use the mark immediately preceding the string-head to help decide the value or length of the string in this occurrence.

[47] In *I saw some water-colors* (*which*) *that artist* (*whom*) *you met had painted*, we find that (*which*) *that artist* incurs the first obligation for a verb and (*whom*) *you* incurs the second obligation for a verb. Following upon the second incurring, *met* is the first discharging of a verb, and *had painted* the second. The dischargings are: *met* for *you*, and *had painted* for *that artist*.

6.6 RESTRICTIONS AT A DISTANCE

In considering all restrictions at a distance, whether discontinuous elements or grammatical agreement or other types of obligation, which originate with one element and must (or may) be discharged at some later point, the following should be noted. There may indeed be unboundedly many words between the incurring and the discharging of a restriction. But, by the considerations of section 3, any restriction at an unbounded distance must also occur at reasonably small distances; and the restriction at great and unbounded distances must be structurally the same (except for intervening computational identities) as the corresponding dependence at small distances, and can differ from the latter only in the number of repetitions of some recursive operations. A recognizer embodying the rules (restrictions) necessary for any large set of sentences, and providing for recursive iteration (possibly equinumerous iterations of separated parts of an operation, or of related operations, at related points) would suffice for any restriction at unbounded distance.

Examples of restrictions are:

(a) When the recognizer makes a decision at the particular point in the string formula that permits this decision, but cannot be sure that the decision will be satisfied until some later point which is in a fixed position in respect to the earlier point: e.g. discontinuous elements; or displaced members of a set, as when *not* requires the t suffixes to appear to its left instead of after the V. We can check from the earlier point, at which some requirement is incurred, to the later point, at which it is discharged, or vice versa; but in any case, some cross-check is necessary.

(b) If the later point is not stateable in a simple way with respect to the earlier point. E.g. the requirement for an article ($\lambda_{n\,4}$) is made as soon as a count-N is met (scanning backward); and the well-formedness requirement for $t\ V$ is incurred as soon as an initial N is met (scanning forward).

(c) When the recognizer meets the beginning of a string before it has computed to its end, a count must be kept for each entry into a nested string. The well-formedness decision for the n nested string-beginnings consists of n nested string-completions.

6.7 DEGENERACIES

In all the preceding cases, the recognizer matched the complexities of the axiomatic generator, i.e. of the rules for applying well-formedness requirements and operators. In addition, the recognizer may meet difficulties which are due to degeneracies (ambiguities) in the dictionary (if a given word is a member of more than one category)[48] or in the grammar (due to there being more than one way in which rules operating on strings may produce a given sequence of words or word-categories: rules A operating on strings B may produce the same sequence of words or categories as rules C operating on strings D).

Each of these different assignments of word to category or of category to a segment of a string incurs particular requirements later in the sentence. The recognizer has to hold in view enough other parts of the sentence to see which of these requirements are met by the rest of the sentence. If a later degeneracy in the sentence has multiple values each of which satisfies one of the requirements of the earlier degeneracy, the ambiguity of the sentence is unresolvable, and the sentence can be read grammatically in more than one way. In all these cases it is possible to indicate the choice of decisions at the point in which it arises, and then to follow the path for

[48] That is to say, if some particular choice from one class and some particular choice from another class yield the same word or the same spelling or phonemic sequence. A special case is that of words or morphemes whose phonemic content is zero but which belong, in particular environments within a sentence, to particular categories. The fact that there is a zero variant of *that* (or, alternatively stated, that the string $r_{n\,10}$ has no word as string-head) makes it necessary to count the number of successive free N or Σ (whose V-requirement has not yet been discharged) in a sentence (see 7.5). The zero *that* which occurs in a certain object type (*I know that he came, I know he came*) is easier to treat. Zero t (present tense in *We know*, etc.) is handled by treating V like $t\ V$ except in certain positions (e.g. except in the object of verbs requiring V, and in certain $V\ K\ V$ sequences).

each decision and to see which path matches one of the possible paths for a well-formed sentence (cf. TDAP 17, 27). Each computational path followed to the end of the sentence will be called a reading of that sentence, yielding a unique formulaic representation of it (see also footnote 2).

6.8 FIXED SYNTACTIC VALUES FOR CATEGORIES; INVERSES

There are various possibilities for basing the computation directly on some property of each successive word-category in a sentence. This means giving each word-category the value of all its possible contributions (in one environment or another) to the well-formedness of the sentence (or to the denial of well-formedness). Since almost every word-category makes different contributions in different environments, we would have to indicate for each category what environmental information determines which of its contributions. In the more individually peculiar cases, this can only be done by assigning a variable value to the word or category. However, certain general types of contribution can be indicated in a general way; and one might try to represent the contribution of each category by particular *n*-tuples of integers. To devise such a representation, the following considerations are necessary (for convenience, the symbols of addition are used here):

1. If *a* requires *B*, and *A B* together constitute *C*, then the value of *A* = the value of *C* – the value of *B;* e.g. a string-head requires its completion (usually a center-structure) in order to constitute with it an identity in respect to the computation. Thus, string-heads are inverses, in respect to the computation, of their strings. It is not desirable that the center of the string itself (not including the string-head) have value zero, since the string-head would then be the inverse of zero. Hence, since many identity strings have the form of a string-head plus c_1, it will not be desirable to let the c_1 have the total value zero: so a sentence which contains c_1 as center will not have the total value of zero. In addition to string-heads, a free (subject) *N* requires a (free) *V*, and each *V* requires its object.

2. In addition to the relation of being required, there is another kind of boundedness: being permitted. *N* permits string-heads on its right, and the λ_n on its left; *A* permits *D* or $\lambda_{n\,4}$ on its left (in *the A* as *N*-replacer); any constituent permits *K* (with its string) on its right; and so on. A permitted string should add zero to its permitter. But since the permitted string is locally restricted, it has to be represented by at least a number pair, one member of which is zero while the other member is used to cancel a corresponding member of its permitter's number. String-heads are thus binary operators: they are inverses of their following strings, but also operate upon their preceding permitter.

3. There is an ordering of requirements. For example, if V_2 follows V_1 before the object of V_1 has been completed, the object of V_2 is required before (or at the same time as) the object of V_1. This may be expressed by saying that the value of the sentence computation must never fall outside a certain range in the course of the computation: the value of $V_1 + V_2 +$ object$_1$ would fall outside the range, while the value of $V_1 + V_2 +$ object$_2$ would be acceptable.

7. SUMMARY OF THE UNIVAC PROGRAM

The program which operated on the Univac (in 1959), and for which descriptions and flowcharts are presented by the authors of TDAP 16–20, covers all English sentences of the major type c_1, although in less detail than the string list given in section 3 here. Some strings were omitted because of limitations of space in the Univac. There are also many more or less «idiomatic» strings, mostly involving individual words (e.g. *no doubt*, *try as he might*) that are not analyzed by this program. Such strings can, however, be fitted into the present program: for in every case they consist of a specific and short sequence, either of individual words or of word-categories or of both, which operates as an identity, or a replacer for one (or more) of the well-formedness requirements.

As to the other sentence types of English, they are structurally related to the major type, and require only rearrangements of parts of the Univac program plus specific changes in the well-formedness requirements and in certain strings. (Compare the various c_i in section 3.1.)

In the Univac program, a particular arrangement of the work was selected, partly to fit the particular computer.

1. The successive words of each sentence are compared with the entries in a category-assigning dictionary, and each word is replaced by its dictionary equivalent – namely the category and subcategory to which the word belongs.

2. The sequence of category-marks which represents the sentence is now scanned for dictionary alternatives, i.e. for cases where there are two or more category assignments for a given word. Each such

indecision calls in a program (TDAP 17) which tries to decide which value of the alternative is the correct one in the given occurrence. Each resolved indecision can be replaced by a single category-mark.

3. The sequence of category-marks (zero-order and first-order symbols), with most dictionary indecisions now decided, is scanned several times, once each for the various types of first-order strings (TDAP 18). Each scan looks for a class which starts (leftward or rightward) a first-order string; upon finding the start, it computes the finish-point of that string. Hence each first-order string can be replaced by a single first-order symbol.

4. The sequence of first and zero-order symbols is scanned for satisfaction of sentence well-formedness (TDAP 19, 20). Each time we meet a symbol which, in its position, satisfies a well-formedness or an identity status, it is marked as such. Each time we meet a string-head, we turn aside to compute the end of the string, replace the string by a symbol indicating its status (identity or N-replacer), and resume the scanning of the sequence as before we met the string.

5. If, when we reach the end of the sentence, the well-formedness requirement is not satisfied, we check back to see what other possible paths of computation could have been taken that would have avoided the unsatisfactory result. If we satisfy the well-formedness at the end of the sentence, we check back to see what other combinations of paths could have been taken for this sentence that would also have led to the satisfaction of well-formedness.

In each of these programs, particular methods were used:

2. in the dictionary alternatives: a battery of tests to decide each indecision;

3. in the first-order strings: a tree which starts only at the start of such a string, takes a fixed preferred path whenever it meets an alternative, and exits at the boundary of the string;

4. in the second-order strings (which includes the whole sentence): a tree which starts at the beginning of the sentence, admits side-trees to compute every included second-order string by going over (virtually) the same tree itself, and then resumes and completes its computation at the end of the sentence.

7.1 WORD CATEGORIES[49]

The word categories are important initial elements in the computation. They represent that classification of words in respect to which it is possible to define operators, mostly recursive, which produce corresponding derived categories, such that these corresponding string categories can fill within the formation rules of the sentence a position analogous to that filled by the (corresponding) word category within the formation rules of the center *c*. In practice, there are almost no problems of choice in selecting the major categories. Only a particular selection satisfies in a simple way the considerations stated above. The word categories which appear in the Univac program are roughly those of section 3 above.

7.2 DICTIONARY MATCHING[50]

This requires only that each word in the sentence be looked up in the dictionary, and replaced by the classification given there. The classification gives not only the category and subcategory, but also any underlying classification from which the given word is grammatically (morphologically) derived – this to the extent that may be useful for transformations, later constituent analysis, or later applications: e.g. *formalization* would be classified *N V A*, indicating that it is an *N* returnable under certain transformations to *V* and to *A;* but we would not classify it *N V A N* (i.e. we would not mark *formal* as derived from *form*) because *formal* is probably not returnable to *form* under any transformation. The classification also gives other information for later handling of the word: e.g. a number to indicate each individual affix which may be relevant in transformations (e.g. *-ation*) or in string analysis (e.g. plural, for grammatical restrictions); and a number to indicate each stem to the extent that words of that stem are interrelated in transformations or in informational applications (e.g. the same stem number would be given to *purity* and *purefy*, but not to *quality* and *qualify*).

If a word is a member of more than one category, the dictionary indicates the disjunction of categories to which it belongs.

The Univac dictionary lists each word as it appears in print, from space to space. It is also possible to list the more common affixes as separate entries, and to have the computer discover the fact that various words of the sentence which are not in the dictionary are combinations of listed affixes and words (or stems). The presence of morphophonemic irregularities makes this cumbersome.

In some cases, it is necessary to become free of word-space conventions. If the value (in the including string) of a sequence of words is not equal to the sequence of the (dictionary) values of those words, the program will misstate the contribution of the sequence to the formation rules of the string. E.g. *in general* is *P A*, but its computational status is that of an identity rather than that of the beginning of a *P N*-string. If this condition applies to a sequence of category marks, the program that scans the string (as a succession of category marks) will have a special output for this sequence. However, if this condition applies to sequences involving individual words, it is best handled near the dictionary level. The first activity which evaluates words in relation to their neighborhood, therefore, is the word-complex dictionary. All individual words which participate in some special sequence of this type are so marked in the dictionary. The word-complex dictionary then scans the sentence for this mark. When it finds the mark at a given word, it checks the neighboring words (immediate neighbors or at stated possible distances) to see if they match the sequence which, with the given word, produces (always or sometimes) the special value; if so, the sequence of values is replaced by a single value for the sequence. We thus obtain a single dictionary category for a sequence of words.

At a later rewriting of the present program, it will also be desirable to classify certain words as sequences of dictionary categories. The words may be composed of two or more morphemes, each of which is assignable to a category: e.g. the *wh-* words are K_s plus

[49] TDAP 21 and the paper cited in footnote 12.
[50] TDAP 16.

pronoun (*who*, for instance, can even be replaced by actual English sequences such as *such that he*). Some words may not be phonemically divisible into morphemic elements, yet may syntactically equal a sequence: pronouns = article plus noun or adjective (or *P N*) (*he* replaces *the man*, *the boy*, etc.). We would then obtain a sequence of dictionary categories for certain single words, thus not only reducing the number of categories but also simplifying the formation rules.

7.3 ALTERNATIVE CLASSIFICATIONS[51]

Another scan is now made through the sentence, in order to decide the assignment of words that can belong to any of a disjunction of categories, *X Y*... *Z*. We can say that such a word is classified by a category variable, which can take as values the various categories of the disjunction: e.g. *study* is classified *N*/*V*, which can take as values both *N* (as in *a study*) and *V* (as in *to study*). When the scan meets a category-variable, it calls in an appropriate battery of tests. Each test matches the neighborhood of the variably-classified word with a particular type of neighborhood in which one of the values *X* of the variable can not occur. If the neighborhood in the given sentence matches the test, then the given word cannot have the value *X* in this occurrence; it can only have one of the remaining values of its category-variable. If the test neighborhood includes some category *Z* in a certain position, and in checking the sentence neighborhood we find that the test is satisfied except that in that position there is a category-variable, one of whose values is *Z*, the test cannot be completed. If subsequently the second category-variable is resolved, either as *Z* or as not *Z*, we can return and complete the original test: to success if the second variable was decided as *Z*, to failure otherwise.

There are various considerations in planning this section: linguistic considerations, in determining, on the basis of the sentence-environments in which each category cannot occur, what are the most convenient test-neighborhoods (convenient for the simplicity and ordering of the tests, and for frequency of usefulness); combinational considerations, in determining, on the basis of the relevant tests and the particular constellations of values which appear in various category-variables, how the diagnostic (test) neighborhoods can best be pitted against each other; and programming considerations, in deciding how to arrange the tests, their calling in by the variable, the network of success, failure, and non-completion and later check of non-completed tests.

[51] TDAP 17.

7.4 FIRST-ORDER STRINGS[52]

The sentence is now scanned several times in order to bracket those sequences of category-marks which, in their sentence environment, satisfy the definitions of the various first-order strings. First, it is scanned leftward for *N*; at each *N* the leftward neighbors are matched with the branches of a tree that describes the leftward identities on *N*, exiting at the left boundary of the first-order *N*-string, which is placed into square brackets [ ]. Then *A* not within [ ] is bracketed in its own [ ], if preceded by *T* (article: $\lambda_{n\,4}$) or if it is a verb-object; otherwise it – with its left identities – is placed in parentheses ( ). *D* not otherwise bracketed, and the sequence of (one or more) *P* plus [*N*], are also placed in parentheses ( ). *V′* (i.e. *V* accompanied by *t* or *to* or *-ing* or *-en* or zero, with any included parentheses) is placed in braces { }.[53]

If we meet an unresolved category-variable, we can in most cases choose one of its values as preferred, and determine the bracketing by following the branch containing that value, while indicate what the alternative path would be at that point. The preferred value is in general the one that makes for the longer *N*-string, at least in the

[52] TDAP 18 and Aravind K. Joshi, Computation of Syntactic Structure, *Advances in Documentation and Library Science*, vol. III, part 2 (1961).

[53] In the Univac program, objects of *V* which contain only *V* are also included in the braces, i.e. in the first-order *V′*.

case of scientific writing. If there is no preferred value for a given variable, both possible paths are indicated by appropriate alternative bracketings (path-selector variables).

7.5 SECOND-ORDER STRINGS AND SENTENCE WELL-FORMEDNESS[54]

Each bracket is now replaced by a single (first-order) symbol, and the sequence of these is scanned to satisfy the well-formedness of a sentence S or of included second-order strings. The first-order symbols each have specific status in respect to this. An N-string is called *free* with respect to a given second-order string (including S itself), if it is not required for the object of that string and is not part of any string included within that string. The problem of the string which has no special string-head (e.g. *my friend met*, in *the man my friend met*) is handled by counting the free N-strings and considering a second free N-string to be the head of a string (produced by $r_{n\,10}$). Other strings can be recognized by their heads, although there are various cases of string-heads which are identical with words in other categories. P plus N-string (for particular P), and certain other first-order sections which had been placed in parentheses can be required for the object after particular V; otherwise sections placed within parentheses ( ) are identities within S or within one of its second-order substrings. Wherever a V' is reached, whether within S or within any string, the program computes what following neighborhood within the string satisfies (as "least required string") one of the possible object-requirements of (the last V of) that V'. Many V are members of several subcategories, e.g. can have various types of object; and the program marks the set of all fits between the sentence neighborhood and the neighborhoods required for each object-subcategory of that V. Each second-order string, after its head is recognized and its length computed, is bracketed and replaced by a (second-order) symbol indicating its status as identity or N-replacement within the including string. Second-order identities were bracketed by $< >$.

Finally, the material remaining after all strings have been accounted for is inspected to see if it is the well-formed sequence c_1.

[54] TDAP 19, 20. The discussion here accords with TDAP 19. TDAP 20 gives a revised presentation with a programming worked out in Fortran coding. The plan in section 6.4 above is related but not identical to this.

INDEX