[From: A.Makkai and D.G.Lockwood (eds.) *Readings in Stratificational Linguistics*, Alabama U.P.1973]

*Stratificational Linguistics as a Basis for Machine Translation**

SYDNEY M. LAMB

THE BASIC THEME of this paper is that simplicity is a good thing and that it can be achieved by isolating recurrent partial similarities, which involves separating various things from one another. It is especially the case that a machine translation procedure should be separated into stages in accordance with the stratification of linguistic structure. But I shall also consider briefly a couple of other types of separation whose importance for machine translation has not yet been recognized by all research groups.

Consider the expression

$$abc + abd + abe + ab(f + g)$$

from ninth grade algebra, and compare it with the expression

$$ab(c + d + e + f + g),$$

which conveys exactly the same information. Or at least it conveys exactly the same EFFECTIVE INFORMATION. But I would like to distinguish two kinds of information, namely SURFACE INFORMATION and EFFECTIVE INFORMATION. We may say that these two expressions have exactly the same effective information but that the second has less surface information than the first. This is another way of saying that the second is simpler than the first. In fact, the simplest and most effective way of defining simplicity is in terms of surface information. Given any two linguistic descriptions or partial descriptions which have the same effective information we will prefer that which has less surface information, and we will prefer it because

*Originally presented at the U.S.–Japan seminar on Mechanical translation, Tokyo, April 20–28, 1964, this paper appears in print for the first time in the present volume. This work was supported in part by the National Science Foundation.

of its simplicity. Its greater simplicity resides in the very fact of its smaller amount of surface information.

There is nothing abstruse about the concept of surface information. On the contrary it is a disarmingly simple concept. It can be precisely measured in terms of the number of symbol tokens and the amount of information per symbol token, which can be stated in terms of binary digits and depends on the number of symbols from which it must be distinguished. But it usually is not necessary to consider binary digits because the alternatives which the linguist ordinarily has to consider differ so grossly in simplicity that the very roughest kind of count of surface information clearly reveals a decisive difference. In the example above, it is not necessary to know how many bits of information each of the letters is worth, as long as they are all equal. We may consider the information content of the parentheses and the plus sign to be considerably less than that of letters, so that they may be ignored when comparing the surface information of the two expressions. The number of tokens of them is the same in both expressions anyway.

Now the first expression above has thirteen letter tokens while the second has only seven. It is precisely this difference which makes the second expression simpler than the first. Notice also that the difference between 13 and 7, namely 6, is exactly accounted for by the three repetitions of *ab* in the first expression. Six is the amount of the EXCESS SURFACE INFORMATION. There are six units of excess surface information because *ab* was written a total of four times (for eight units) whereas it was needed only once.

But notice also that the two expressions above do not differ just in simplicity. They also differ in generality and hence in the insight which they provide. The second expression above is more general as well as simpler. It has generalized the fact of the occurrence of *ab* with each of *c, d, e, f,* and *g,* whereas the first expression leaves these as isolated phenomena. Simplicity, then, is desirable not just in itself but because true simplicity goes hand in hand with generality. Simplicity and generality are like the two sides of the same coin. When you achieve one you also automatically get the other.

Now I would like to make the rather extravagant assertion that this same basic type of simplification operation lies at the very heart of all effective linguistic analysis, of all effective work on the design of machine translation systems—even of all science. (And I might also mention that this same operation of simplification is one of the most important aspects of the designing of efficient programs for a computer). To take an example from science, consider the Ptolemaic and Copernican descriptions of the solar system. It would be a mistake to say that the difference between these two descriptions is that one is correct while the other is incorrect, if by correct we mean something about whether or not the facts are correctly accounted for. All of the facts of the movements and positions of the planets can be accounted for with the earth taken as the center of the solar system as in the Ptolemaic model. These two accounts of the solar system are (or can be) equal in effective information. What makes the Copernican description much more attractive, insightful, and informative is simply that

it is more simple. It takes less surface information to present the Copernican description. Here we may also note the generality that goes hand in hand with the simplification. The Copernican system is more general in that it ascribes the same basic patterns of motion to all of the planets, including the earth (which of course was entirely different from all of the other planets in the Ptolemaic model). We may also note one other highly significant fact about the simpler, more general description; namely, that the basic concepts used in it turn out to also be capable of accounting for phenomena observed outside of the solar system itself. The rotational and orbital motion, gravitation and centrifugal force and so forth are found also in other astronomical systems. In just the same way, one of the nice things about a simple linguistic analysis of a given corpus as opposed to a less simple one is that the simple one stands a much better chance of being able to account for new, previously unexamined material without the addition of extra rules.

Table I.

Sg.	N	dama	baba
	G	damy	baby
	D	dame	babe
	A	damu	babu
	I	damoj	baboj
	L	dame	babe
Pl.	N	damy	baby
	G	dam	bab
	D	damam	babam
	A	dam	bab
	I	damami	babami
	L	damax	babax

Let us now consider a linguistic example, shown in Table I. Here I show the inflectional paradigms of two Russian nouns. An alternative description of exactly the same material is given in Table II. Here we see what I hope is obviously a startling difference in surface information (but with the same effective information).

If we examine the method by which this simplification (i.e., reduction of excess surface information) was achieved, we see that it involved SEGMENTATION BASED ON THE FINDING OF RECURRENT PARTIAL SIMILARITIES. This is really the same process as that used in the algebraic illustration given above. The elementary algebra student is able to achieve the factorization by discovering a RECURRING PARTIAL SIMILARITY, namely the occurrence of *ab* in the four separate terms. The second expression shows the result of segmenting out this similar part and expressing it in the description only once instead of four separate times. In the same way,

the second linguistic description above achieved its simplicity and generality by the segmentation of the inflectional suffixes from the stems, again since they were recurrent partial similarities. All of the forms in the first column of Table I are partially similar, as are all the forms in the second column. Similarly the two forms in each row of Table I are partially similar. Table II shows the result of isolating the partial similarities and generalizing on the basis of them. The repetition found in the first description is excess surface information. (The process is the same, again, as that which Copernicus used when he found that all of the planets are partially similar to one another in their motion).

Table II.

Stems	*Endings*	Sg.	Pl.
dam	N	a	y
bab	G	y	—
	D	e	am
	A	u	—
	I	oj	ami
	L	e	ax

An efficient machine translation system, like a good linguistic analysis, is achieved by separating various things from one another on the basis of recurrent partial similarities. I shall describe now three important types of separation which enable one to get rid of excess baggage in a machine translation system, namely (1) separation of the program from the linguistic information, (2) segmentation of words, and (3) separation of the translation process as a whole into stages in accordance with linguistic stratification. It is this third type of separation, that based on stratification, that most of this paper is devoted to.

First let us consider the separation of the program from the linguistic information. Such separation is advocated by the University of California project as well as those of the University of Grenoble, the Rand Corporation, and the University of Texas. In an unseparated system, the linguistic information is built right into the program. I shall illustrate with an over-simplified illustration of a sequence of program steps with linguistic information built in that might form part of a syntactic decoder:

Is current item coded *adjective?*
  if YES, go to A
  if NO,
Is current item coded *verb?*
  if YES, go to V
  if NO,
Is current item . . .

Then beginning at A there might be instructions something like the following:

Is following item coded *noun?*
if NO, go to AA
if YES, form constitute and label it *noun.*

Ignoring the irrelevant fact that this illustration is oversimplified, we see that the linguistic information, which the ordinary linguist would be more accustomed to putting into rules or statements of relationships, has been stuck right into instructions telling the computer to follow a particular procedure, and that each part of the procedure is *ad hoc* for the particular information incorporated, rather than general.

In a system which keeps the information separate from the program, it is expressed in rules or formulaic statements, whose form has to be precisely defined, and a program is written which will have the capability of operating with any statement which is in that specified form. The advantages of the separation are: (1) it allows the linguist to write his rules as rules rather than in a flow chart or in some programming language: (2) when the linguist wants to revise some of his statements he can do so very easily, without any need for reprogramming (which is usually very time-consuming); (3) the various basic operations which must be carried out by the machine to perform the process have to be written out only once in the separated program, whereas in the integrated one they must be repeated over and over again with the different units of linguistic information which are subjected to the same basic operations; (4) the program, since it is written to operate with statements of a specified form, can operate on such statements not just for one language but for any language, so that new programs do not have to be written when we decide to translate from a new source language.

Next we may consider the desirability of segmenting words. (For a more detailed treatment see Lamb, 1961a). This type of separation is illustrated above in Tables I and II. It brings about a reduction not only in the number of dictionary entries needed but also in the amount of linguistic information needed per dictionary entry. In the examples above, a dictionary based on the analysis of Table I would require eighteen entries for the material shown, whereas a dictionary based on Table II requires only eleven. The difference between eighteen and eleven does not seem very great, but if we think in terms of 100 nouns of this declension type, then the numbers are 900 and 109. And in general, for b bases and s suffixes each of which can occur with any of the bases, the figures are $b \times s$ and $b + s$.

Let us consider an abstract example, shown in Tables III and IV. Here I show only four bases and two suffixes. The rectangles in the figures enclose the dictionary information for each entry. There is a certain amount of information needed for each base as well as for each suffix, so in the unsegmented dictionary of Table III each entry must be large enough to contain the information for both the base and the suffix. (The translation of, say, the genitive singular suffix depends on factors which are independent of

Table III.

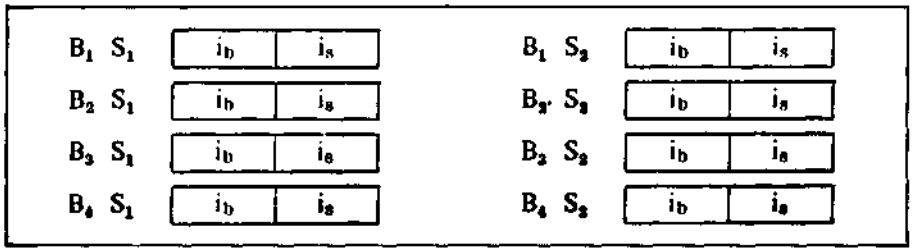
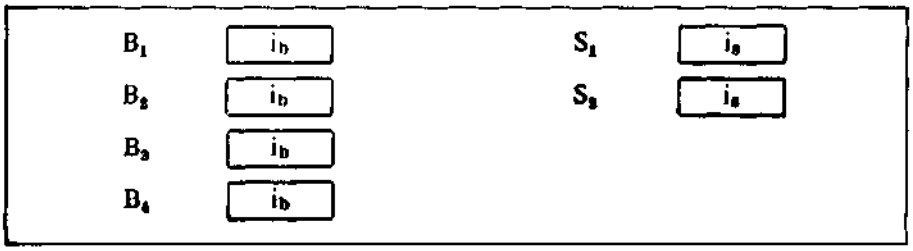


Table IV.



the stem with which it occurs). But in the segmented dictionary of Table IV each entry for a base requires only the information for that base, so each entry is smaller than in the case of Table III. By isolating the recurrent partial similarities of Table III, we achieve a reduction in number of entries from 8 to 6 (4 x 2 reduced to 4 + 2) but a reduction in total VOLUME OF RULES from 16 units to 6 units (considering the dictionary information for each base and suffix to be the same, namely one unit). The formula for volume of information in an unsegmented dictionary, considering only bases and suffixes and allowing only one suffix per word, where b is the number of bases, s is the number of suffixes, i_b is the average amount of information per base, and i_s is the average amount of information per suffix, is

$$b \cdot s \cdot (i_b + i_s),$$

which, if we let $i_b = i_s = i$, is equal to

$$b \cdot s \cdot 2i.$$

But if we segment, then the amount of information is only

$$b \cdot i_b + s \cdot i_s$$

or roughly

$$b \cdot i + s \cdot i = (b + s) \cdot i.$$

Thus the EXCESS SURFACE INFORMATION RATIO is

$$\frac{b \cdot s \cdot 2i}{(b + s)i} = \frac{2bs}{b + s}$$

So if we are dealing with 10,000 bases each of which occurs with each of ten suffixes, the excess surface information ratio is

$$\frac{2 \cdot 10{,}000 \cdot 10}{10{,}000 + 10} = \frac{200{,}000}{10{,}010} \quad \frac{20}{1}$$

In other words, under these conditions the unsegmented dictionary has about twenty times as much surface information as the segmented one, for the same effective information. But even this ratio is not as high as it would actually be in the more realistic situation which would allow more than one suffix (e.g. both derivational and inflectional) per word.

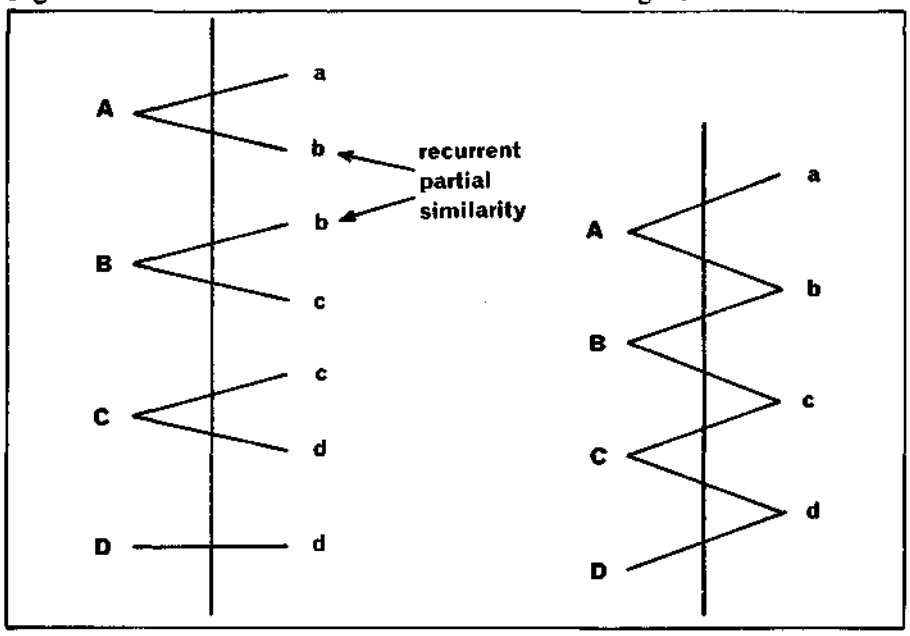
In the case of the Russian dictionary constructed at the University of California, words were segmented into bases, prefixes, derivational suffixes, and inflectional suffixes. The dictionary has around 20,000 entries (not counting those for the chemical nomenclature, which amount to an additional 2,000 to 3,000); but the number of words that are formable from these units (defining word as a sequence of graphemes that can occur between spaces or punctuation), using only grammatical constructions, is more than 2,000,000. This is a difference of more than one hundred to one in number of entries required. Obviously this difference, particularly when we add to it the consideration of amount of information needed per dictionary entry, is of enormous importance when we consider the question of the amount of storage space needed by a computer, and the amount of time required for obtaining the dictionary information for items in a text.

The third type of separation I would like to consider (and the one which relates to stratification) is that of the translation procedure as a whole into stages. As an illustration of the kind of economy that can be achieved by such separation, consider the bilingual dictionary—say a Russian-English dictionary—and the separated equivalent, which would consist of a Russian dictionary on the one hand and an English dictionary on the other, with addresses in the Russian dictionary identifying locations in the English dictionary. In the unseparated Russian-English dictionary, a given English lexeme might occur over and over again in different dictionary entries because it would be a suitable translation equivalent for several different Russian items. But in the separated version each English lexical item needs to be listed only once. (For further details see Lamb, 1965a).

The general case, as it exists between any two strata, is illustrated in Figures 1 and 2. If *A* realizes, or is realized by, *a* and *b* alternatively, but *b* is also one of the alternatives in the same relation to *B*, and so forth; then in the system which does not separate into stages on the basis of the stratification the elements *b*,*c*, etc. have to be repeated as in Figure 1. But if one isolates these recurrent similarities and states them only once instead of repeatedly, then one has the situation diagrammed in Figure 2.

The greater simplicity of Figure 2 is of larger proportions than would be apparent if the units shown as *a*, *b*, *c*, and *d* are taken to be in the midst of a series of strata (rather than at the end), since in that case there is further branching from each of them (to the right in the diagram, if it were extended). Suppose, for example, that we take them to be sememic units which are the realizates of the lexemes *A*, *B*, *C*, and *D* (for explanations of these concepts, see Lamb 1964a). In one type of translation system, these sememic units (*a*, *b*, *c*, *d*) would connect, with branching and merging, to lexemic realizations in the target language. This situation amounts to

Figure 1 Figure 2



an extension of Figures 1 and 2 two steps beyond what is shown in the diagram, and the extended Figure 1, with its compounding of excess surface information, represents the system which tries to convert directly from lexemes of the source language into lexemes of the target language, while the similarly extended and far simpler Figure 2 may be interpreted as the situation achieved by separating the translation process into stages in accordance with the stratification present. (Some of the statements that have been made about the impossibility of economically high-speed dictionary look-up using existing computers have been fallacious simply because, among other things, the calculations were made in terms of excessively bulky dictionaries in which each entry would have to contain all the information that would be needed if no stratificational separation were made).

A more concrete example and one which is contained within a single language is furnished by the genitive lexeme of Russian, which has several morphemic realizations and is in turn the realization of several different semons. To illustrate with simplified diagrams of the alternative decoding processes, let a, b, c be different realizations of the genitive lexeme G, and let S, T, U be three semons which are realized by G, according to the realization rules, which may be diagrammed as R within a circle.

Then Figure 3 shows the repetition for a system which fails to recognize the lexemic stratum but which tries to convert directly from the morphemic shapes to the sememic realizates; while Figure 4 shows the corresponding situation for a stratified system. In reality, the difference in complexity is much greater than that shown in the diagrams, since (1) there are more than three morphemic realizations of G, (2) all of them are portmanteaus

Figure 3

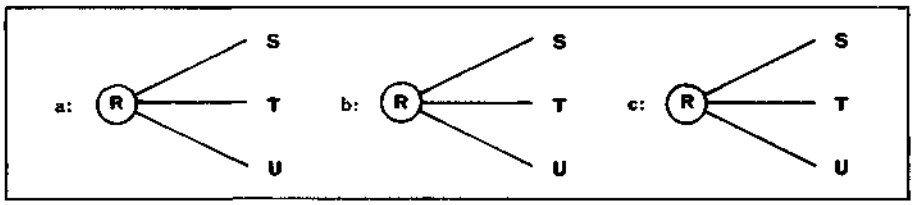
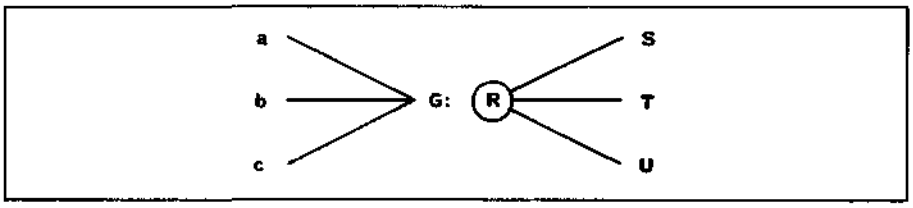


Figure 4



for G combined with singular or plural, and (3) there are not just three but perhaps a dozen or more sememic realizates of G.

To take another example, consider the English lexemes L/go/, L/go crazy/, L/go in for/, L/go through with/, L/undergo/. Each of these lexemes must be treated as a separate unit for economical handling of its sememic and syntactic relations (which is why they are to be analyzed as lexemes), yet if each is treated separately with regard to the past tense, not to mention the past participle, then the rule which provides the correct realizations *went, went crazy, went in for, went through with, underwent* must be given repeatedly. But if the strata are kept separate this statement is needed only once, as a morphological realization rule relating to the single lexon L/go/, a component of each of these five lexemes.

Thus, in short, the identification of recurrent partial similarities in linguistic structure leads to recognition of, among other things, the stratification of language, and in machine translation research the recognition of this stratification leads to the separation of the translation procedure as a whole into several stages. Since proper treatment of stratificational phenomena is so vital to machine translation, and since they have been dealt with so unsystematically in the past, I shall now go into some detail with regard to stratification and the simplicity to be achieved by recognizing it.

A. The stratification of language

The stratification of language has been recognized in varying degrees by many people who have worked with language, both amateurs and professional scientists. Hjelmslev (1943, 1954) and Hockett (1954, 1958, 1961) are prominent among the linguistic scientists who have recognized it, and I (1964a) have attempted to extend their observations to an explicit recognition of four structural strata, the *phonemic*, *morphemic*, *lexemic*, and *sememic*, which are apparently present in at least most natural spoken languages.

A language, by its nature, relates sounds (or graphs) to meanings. The relationship is a very complex one which turns out to be analyzable in terms of a series of systems each of which connects two neighboring *strata*. The sememic stratum has units directly related to meaning. These sememes may be thought of as encodable into units of the next lower stratum, which in turn are themselves encodable, and so on, until one comes out with units directly related to speech or writing (i.e. with phonemes or graphemes), which may now be spoken or written as the case may be. In understanding the importance of stratification to machine translation, it is helpful to look at the structure of spoken languages even though machine translation research is currently concerned with written languages, since the latter are based upon spoken languages and derive much of their structural patterning from them.

A few examples will give an indication of the various types of situations dealt with between different pairs of neighboring strata. If we consider the *t* in *eighth* in relation to the *t* of *water* in spoken English, we can see that from one point of view they are quite different. The former is dental, tense, and voiceless; the latter is postalveolar, lax, and voiced. These are phonetically two different entities, but phonemically the same, since the phonetic differences are non-distinctive. A rule can be given to account for the various features that are present in different environments, and those features, thus accounted for, no longer need be considered at the higher stratum. Similarly, but one level higher, if we compare *sane* and *sanity*, or *vain*, and *vanity*, or *nation* and *national*, we see a recurrent variation between two entities which are phonemically different. But in some other sense there is a single unit *sane* underlying both of the units *sane* and *sanity*, and the recurrence of the alternation for *vain* and *nation* indicates that it is not directly a property of *sane* as a whole, but rather a property of one of its components. Similarly, but one level higher, the forms which we represent orthographically as *good* and *better* are altogether different at one level, but at another they are partially the same. That is, the latter, from the lexemic point of view, consists of the former plus the comparative lexeme. Finally, *can*, as in *he can go*, is lexemically different from *be able to*, as in *he will be able to go*, but at another level these are one and the same unit. These examples are all concerned with DIVERSIFICATION, one of the several phenomena which characterize linguistic realization.

The reason for this complexity in linguistic structure is, in part, that sounds and meanings are, by their natures, patterned differently from each other. They each have their own set of structural relationships. Phonemic systems must be adapted to transmittability of speech through the air and to the articulatory and auditory organs, while sememic systems must be adapted to thought and memory patterns and to cultures and the world about which people talk. Speech takes place in time, which is linear; but the brain and the world are three-dimensional. Moreover, the processes of linguistic change affect phonological and semological systems in different ways. Thus a close correspondence between them would be impossible.

The same is true for written languages because writing systems are based upon spoken languages, so that they tend to have close correspondence to phonemic but not to sememic systems. On the other hand, written symbolic systems that have been developed independently of spoken language, such as symbolic logic, mathematics, and programming languages, do not have this property. Here we find a very close correspondence between writing and meaning, and little stratification.

Hjelmslev (1943), influenced by Saussure (1916), recognized the stratification of language, but he did not go far enough. His system has EXPRESSION SUBSTANCE (which corresponds to the phonetic stratum), EXPRESSION FORM (corresponding roughly to phonology, in the sense of that term used in this paper), CONTENT SUBSTANCE (which corresponds to the semantic stratum), and CONTENT FORM. But when one studies the linguistic data more closely, one finds that Hjelmslev's content form ranges over what are really three separate systems. These may be called morphology, lexology, and semology (see below).

On the other hand, American linguists have often tried to get along without any explicit stratification at all, except in separating phonetic and phonemic strata; or else, in some cases they have, like Hjelmslev, recognized separate structural strata, but too few of them.

The stratum is a type of level, but it is a different type of level from others which linguists frequently talk about. It is the type which is concerned with realization,[1] and it must be kept distinct from other kinds of levels, such as combinatory and classificatory levels, with which it has often been confused.

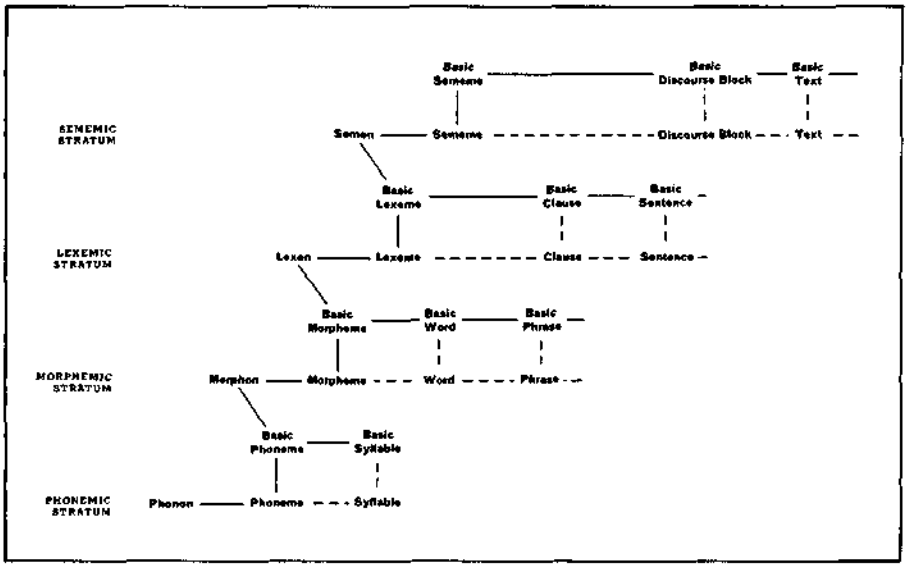
The combinatory or size level is the type of level which is concerned with combinations of linguistic units, the type one is speaking of when one says that the phrase is at a higher level than the word or that the sentence is at a higher level than the clause. Such levels exist within strata. They can be kept distinct from other kinds of levels by being called RANKS, as is done by Halliday (1961). Combinations of linguistic units exist on the same stratum as those units. Thus each stratum has a series of ranks. For example, the syllable is at a higher rank than the phoneme, but on the same stratum. Similarly, lexemes and combinations of lexemes such as clauses and sentences are all on the lexemic stratum.

The stratification hierarchy is also to be distinguished from taxonomic or classificatory hierarchies. The relationship between one stratum and an adjacent one is NOT that units of the one are classes whose members are units of the other. That is, the stratificational view is not the same as that which holds that a morpheme is a class of allomorphs and that a phoneme is a class of allophones. That view is too simple to fit the empirical data. Classes of linguistic units, like combinations, exist on the same stratum as those units. For example, the class of vowel phonemes (which really is a class) is on the phonemic stratum, and the class of nominal lexemes is on the lexemic.

As the concept of stratification is not yet well established in linguistics,

the terminology that is appropriate to it is in a state of disarray. Neither I nor anyone else can even make a pretense of presenting the terminology that is standard in linguistics. Instead the terminology used here is presented as my own, with the hopeful assertion that much of it will be found to agree more or less with those of other linguists. The names I use for some principal units and their strata are shown in Figure 5, in which the vertical dimension represents stratification, with higher strata shown higher in the diagram, and the horizontal dimension is used for different ranks, with potentially smaller units on the left, potentially larger ones on the right.

Figure 5



The elementary unit of a stratum is named with the suffix -ON. Thus the elementary units of the phonemic stratum, i.e. the components of phonemes, may be called phonons. The English phoneme ^{P}/m/ may be analyzed as composed of the phonons ^{P}/Cl/ 'Closed', ^{P}/Lb/ 'Labial', and ^{P}/Ns/ 'Nasal'. The syllable ^{P}/men/ may be written phonically as

Cl	Vo	Cl
Lb	Fr	Ap
Ns		Ns

Phonemes occur not in haphazard combinations but in arrangements having definite patterns (which of course vary from language to language). The patterns of arrangement on any stratum may be accounted for by tactic rules. The basic or elementary unit for purposes of such rules may be named the basic Xeme, where X is phon-, morph-, lex-, or sem-, as the case may be. (A basic phoneme may alternatively be called a morphophoneme, and basic morphemes and basic lexemes may similarly be called lexomorphemes and semolexemes, respectively). These basic emes are set up by the gram-

marian in such a way that the tactics will be as simple as possible (without being incorrect or incomplete). Basic emes (e.g. basic phonemes) may differ from actual emes (e.g. phonemes) in that a given basic eme may be realized as different actual emes in different environments (diversification) and in that a given actual eme may be the realization of different basic emes in different occurrences (neutralization). For example, the English basic phoneme ^{BP}/a/ in combination with another which may be written ^{BP}/:/ is realized as ^{P}/ey/, as in *sane*, otherwise when stressed as ^{P}/a/, as in *sanity*, and when unstressed as ^{P}/ə/ (cf. automation, automatic, automaton; catastrophe, catastrophic).

When the tactics for a given stratum of a given language is made as simple as possible it reveals one or more important types of tactic units of higher rank than the basic eme, i.e. neatly structured combinations of basic emes. Such a unit in the phonology of at least most natural spoken languages is the basic syllable. The realization of a basic syllable, which is defined by the realization rules pertaining to its components, is a syllable. Although Figure 9 does not show any names of larger units than syllable at the phonemic stratum, such larger units exist, since any of the larger units of the upper strata has a phonemic realization.

Basic phonemes can differ from morphons, which are components of morphemes, in arrangement and in certain other ways too technical to be discussed here. An example of difference in arrangement is furnished by certain realizations of the past tense lexon of English, such as the one occurring with *take*. Morphonically, this realization follows *take*, but its realization in terms of basic phonemes (and in terms of actual phonemes) occurs within that of the verb. Such difference in arrangement, which is commonly found between neighboring strata, may be called anataxis.

«For a more recent treatment of the phenomenon referred to in this article as ANATAXIS, See Lockwood 1972c (' ''Replacives'' without Process') in this volume.»

Whereas phonotactic rules describe the composition of all possible basic syllables for a language in general terms, the composition of the morphemes must be described individually for each one. Morphological realization rules may be used for this purpose as well as for stating the relationships between the actual morphemes and the basic morphemes. For example, the realization rule for the basic morpheme *good* would state that when followed by the comparative suffix *-er* it is realized as ^{M}/behd/, when followed by the superlative *-st* it is realized as ^{M}/be/, while elsewhere it is realized as ^{M}/gud/. Note that this unit ^{M}/gud/ as a whole is the morpheme and that its components, ^{M}/g/, ^{M}/u/, and ^{M}/d/, are the morphons.

Morphotactic rules describe how basic morphemes are combined into basic words and basic phrases, thus indirectly describing how morphemes are combined into words and phrases. (Of course, higher-ranking units, such as morphological realizations of sentences and texts, also exist). Whereas such tactic categories as 'obstruent', 'vowel', etc. are relevant to the tactics of the phonology, in the morphology the tactic categories

of basic morphemes of a typical Indo-European language would have such labels as 'verbal prefixes', 'noun bases', 'deverbative nominalizing suffixes', 'case suffixes', etc.

Basic morphemes occur in different arrangements from those of lexons, since the latter are components of lexemes, whose patterns of arrangement are governed by lexotactic rules. As an example, the perfect tense lexeme L/have > en/ of English is continuous, like all lexemes, but part of its morphological realization, namely *have*, precedes the morphological realization of the verb while the other part, the past participle suffix, follows.

A lexeme is composed of one or more lexons and is the realization of a basic lexeme. As an example, the past tense lexeme of English and the perfect tense lexeme may be analyzed as alternate realizations of a single basic lexeme since they occur in mutually exclusive environments, when environments are characterized in terms of basic lexemes.[2] The rule for this basic lexeme would state that it has the realization L/have > en/ when occurring in combination with other tenses or in non-finite verb expressions (e.g. *he would like to have gone*), and the realization L/ > ed/ elsewhere. Lexotactic rules characterize the set of basic clauses and that of basic sentences for a given linguistic structure in terms of basic lexemes as the ultimate constituents. The determination as to whether a given combination of lexons composes one lexeme or more than one or less than one is provided by the lexotactics, just as in morphology the determination as to whether a given combination of morphons composes one morpheme or more or less than one is provided by the morphotactics. For example, L/wide th/ *width* is one lexeme, while L/tall ness/ is two since, unlike *width*, it is formed according to a lexotactic construction. The English lexeme L/ness/ occurs freely with adjectives, including polylexemic ones (e.g. 'the *many-sidedness* of the Khrushchev personality', said by a television news commentator), so that the only way to accurately characterize its occurrence is to specify that it occurs with members of a particular distribution class; but the lexon L/th/ occurs with only about a dozen English adjectives, which do not constitute a distribution class on any other grounds and hence must be listed individually (in realization rules) to specify their occurrence with it. Therefore *tallness* is two lexemes, while *width* is one, which is composed of two lexons.

Although they exist on different strata and are therefore not directly commensurate, one may say that there is a rough correspondence in size between lexons and morphemes in that the number of tokens of each in the realization of a given basic sentence is usually about the same. Similarly, there is a rough correspondence in size between the morphon and the phoneme. There is somewhat less correspondence between morphemes and syllables and between lexemes and words. The syllable is a combination of phonemes and the morpheme is a combination of morphons, but they are combinations of different types. The realization of a morpheme is sometimes larger, sometimes smaller, sometimes the same in size as the syllable. Similarly, the lexeme has no necessary correlation in size with the word. The realization

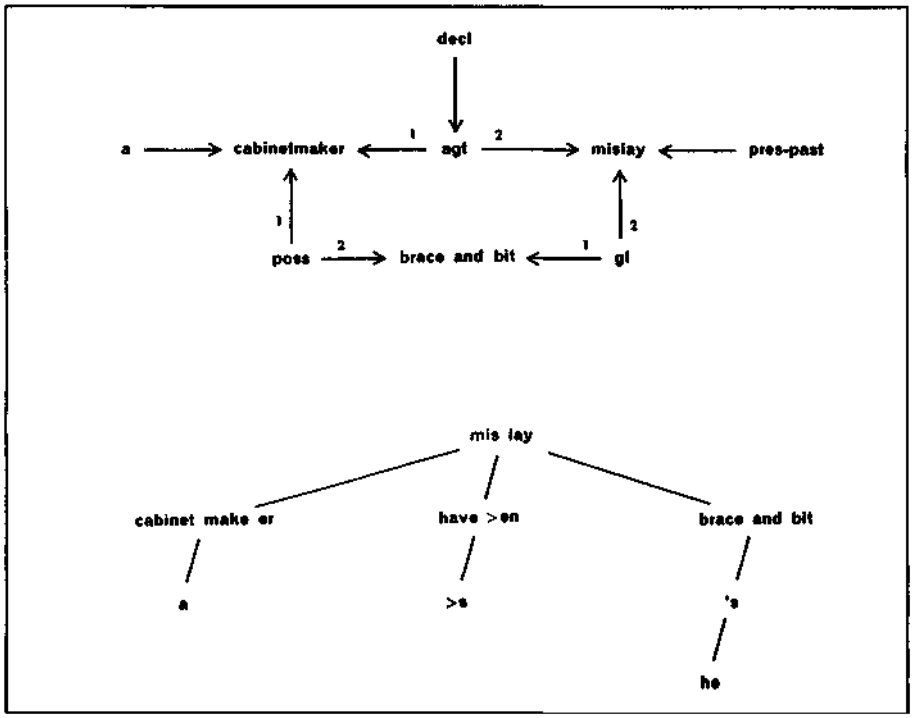
of a lexeme may be a word or it may be smaller than a word or larger (in which case it is called an idiom), or it may even be realized as parts of two different words, as is L/have > en/ in *has taken*.

Basic lexemes can differ from semons in arrangement and in certain other ways. Basic lexemes (of at least some languages) may be analyzed as occurring in trees that are something like dependency trees (cf. Tesnière 1959), Hays 1961, 1963), while sememes (which are composed of semons) occur in networks.

«During the subsequent development of the theory dependency trees were abandoned. The reason for this decision was the fact that dependency trees used elements at their nodes, whereas the nodes of the more modern theory, for which the and-or dichotomy has become basic, do not contain any elements but function purely as connecting devices. Thus dependency trees seemed incompatible with this newer approach.»

Thus in the sememic realizate of *a cabinetmaker has mislaid his brace and bit* (Figure 6), S/cabinetmaker/ is both the agent of S/mislay/ and the possessor of S/brace and bit/, and these two sememes in turn are connected to each other through the S/goal/ realation. But on the lexemic stratum there are two separate realizations of s/cabinetmaker/, namely L/cabinet make er/ (a lexeme composed of three lexons) and L/he/, so that the clause has the form of a tree instead of a network with a closed circuit.

Figure 6

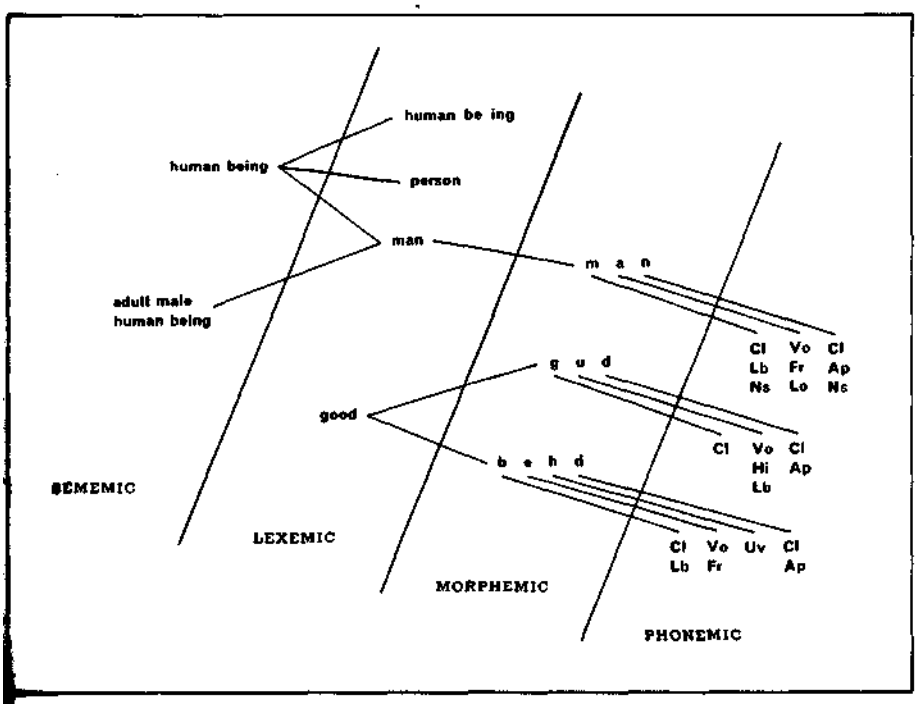


Just as some lexemes are simple (e.g. ^{L}/ > ed/ and ^{L}/find/ in *Joan found her hat*) while others are complex (e.g. ^{L}/be > ing/, ^{L}/look for/, and ^{L}/pocket book/ in *Joan is looking for her pocketbook*), so sememes can be either simple or complex. ^{S}/possessive/, as in *her hat*, is a simple sememe, and an example of a complex sememe is the sememic realizate of *may I ask* in *May I ask who's calling?* (said by secretaries on the telephone). If the realizate were taken as polysememic instead of a single sememe, then the appropriate answer would be *yes* or *no*, to which (if it were *yes*) the secretary would respond *Who's calling?* (Cf. the lexeme ^{L}/under stand/ which, if it were polylexemic instead of single lexeme, would mean to stand underneath). Sememes are realizations of basic sememes. For example, *may I ask* is one realization of a basic sememe of politeness occurring with the interrogative basic sememe, of which an alternate realization is *may I tell him.*

Finally, the combinations of sememes which are well-formed according to a given linguistic structure, namely the texts and structured portions of texts, which may be called discourse blocks, may be accounted for (or generated) by means of semotactic rules, which apply to the basic sememes as simple tactic units and which characterize basic discourse blocks and basic texts as complex tactic units.

An illustration which covers all four structural strata is given as Figure 7. In it the lexemes, morphemes, and phonemes are shown in terms of

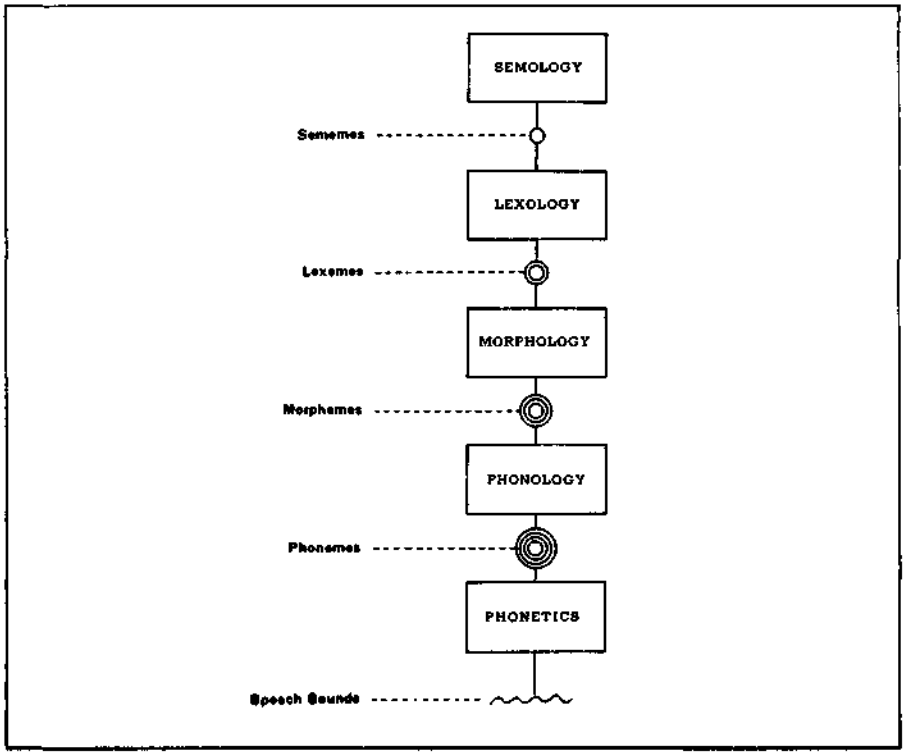
Figure 7



their components, i.e. their lexons, morphons, and phonons, respectively. The sememe ^{S}/human being/ is shown with three lexemic realizations; ^{L}/human be ing/, ^{L}/person/, and ^{L}/man/, of which two consist of single lexons while one is composed of three. The lexon ^{L}/man/, an indivisible unit, is realized by the morpheme ^{M}/man/, which is composed of three morphons, while for ^{L}/good/ two of the three morphemic realizations are shown. Phonemic realizations are shown (in terms of phonons) for ^{M}/man/, ^{M}/gud/, and ^{M}/behd/.

A full description of a language, according to stratificational theory, has semological, lexological (in traditional terms, lexical and syntactic), morphological, and phonological components, as well as a phonetics, which relates the structure to actual speech sounds (cf. Figure 8). Each of these components may be divided into two major sections, covering (1) tactics and (2) realizations. (Even the phonetics has a primitive tactics, which deals with the composition of segments). Realization rules are a means of describing the relationships between basic emes and actual emes as well as the componency of the actual emes.[3] Thus phonological realization rules relate basic phonemes, phonemes, and phonons to one another, while morphological realization rules relate basic morphemes, morphemes, and morphons, etc.

Figure 8



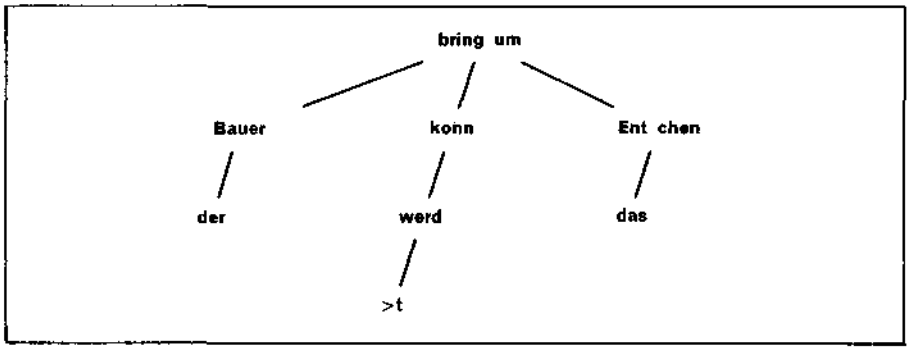
The semology generates or characterizes the (infinite) set of well-formed sememic networks for a language. The lexological component, if left to itself, i.e. if uncontrolled by the semology, generates the (infinite) set of grammatical sentences of a language, including both nonsensical and 'sensical' sentences (i.e. the larger of the two circles under the lexology in Figure 8). The smaller (but still infinite) set of sentences which are both grammatical and sensical is generated by the lexological rules if the choices allowed by them are governed by sememic networks instead of being made at random. A typical sememic network gives rise not to just one sentence but to a sequence of two or three or more (cf. Gleason, 1964:93–95), and the morphological component, which generates words and phrases, commonly provides more than one phrase for each sentence which comes to it from the lexology. In the normal use of language, the morphology operates under the control of the lexology, which in turn is controlled by the semology, thus generating the morphological realizations of the grammatical and sensical lexemic trees (represented by the smallest of the circles under morphology in Figure 8); if controlled by an uncontrolled lexology, the morphology generates a still larger set (the outer circle) which includes ungrammatical sequences (in the outer circle but not the middle one) in addition to the grammatical ones specified by the lexology. Lexological control is exercized by the lexons, which constitute specifications of what choices are to be made at points where there are alternatives. When the need arises, the morphology can provide new lexemes from its generatable stock or morphologically well-formed words and phrases. Similarly, the phonology generates sequences of syllables, normally under control of the controlled morphology, but when not under such control it generates nonsense syllables (as well as sensical ones), and with a certain type of relaxed control the result is jabberwocky, while with control by various esthetic factors in addition to more or less relaxed morphological and lexological control, the result is poetry (or attempted poetry). In addition, the phonology can be called upon, as it were, to provide new morphemes, just as the morphology makes possible the creation of new lexemes.

As there is a tactics associated with each stratum, the patterns of arrangement found on one stratum do not directly correspond to those of neighboring strata. An illustration of the type of difference in arrangement between the sememic stratum and the lexemic is given above in Figure 6, which also provides an indication of the type of difference to be found between the lexemic and morphemic strata, since the morphemic arrangement can be analyzed as a linear chain in which the order of the morphemes can be seen in the written realization *a cabinetmaker has mislaid his brace and bit*. A more striking difference between the lexemic and morphemic arrangements is exhibited by the simple interrogative version of the same clause. For it, the sememic network can be set up with the sememe ^{S}/int/ 'interrogative' in place of the ^{S}/decl/ 'declarative' of Figure 6. This sememe in this environment is realized by a feature which may be symbolized '>' (the same as that which appears in ^{L}/have > en/) attached to the head

of the subject phrase. This element specifies to the morphology that whatever it accompanies (including anything under it in the tree) is to be delayed until the following word (cf. *has mislaid* [BM]/have z + mislay en/). Thus the subject phrase *a cabinetmaker* is delayed one word and the interrogative order is *has a cabinetmaker mislaid his brace and bit?*

More striking still is the situation involving word order in German clauses, which has plagued students through the years because of its seeming complexity. But actually what appears to be a very complicated pattern of arrangement when viewed solely at the morphemic stratum (or lower) turns out to be very simple when it is related to the lexology. The basic clause structure is very similar to that of English on the lexemic stratum; but German has a special rule used by its morphology in realizing lexemic trees, to the effect that only the first word of the verb phrase is realized in the expected order, and any remaining words are delayed in a temporary push-down store until the end of the clause; and if it is a subordinate clause, then every word of the verb phrase including the first is delayed in the push-down store. A push-down store has the property that the last item to go in is the first one out while the first in is the last out, etc. Thus the tree of Figure 9 is realized as *der Bauer wird das Entchen umbringen können* 'the farmer will be able to kill the duckling' (with infinitive suffixes supplied

Figure 9



by the morphology as empty morphemes and *wird* as the realization of [BM]/werd t/). Without the pushdown store delay the order would have been *der Bauer wird* [*können bringen um*] *das Entchen* (with brackets enclosing the portion that is delayed). If [L]/werd/ is not present, then (without any other difference in the lexemic tree) the order is *der Bauer kann das Entchen umbringen*, and if neither [L]/werd/ nor [L]/könn/ is present then the order is *der Bauer bringt das Entchen um*. And if the clause as shown in Figure 9 is introduced by *dass* 'that', then all words of the verb phrase are delayed, and the order is *dass der Bauer das Entchen umbringen können wird*. Thus the gross differences in linear order seen on the surface are accounted for by a simple delay rule which is merely a matter of morphological realization of a simple and constant pattern of lexemic arrangement.

Since most or all current machine translation research is concerned with written languages, we should now turn to consideration of how their structure relates to that of the spoken languages on which they are based. In the first place the upper strata are generally substantially the same. Written languages differ from their spoken models in that at some point, usually relatively low in the generative hierarchy, the encoding is to rules which will lead to written marks instead of speech sounds. In alphabetic languages, the written characters—i.e. letters—take the place, roughly, of either basic phonemes or phonemes; or more commonly the orthography is of mixed character, reflecting partly phonemes and partly basic phonemes. In syllabaries the characters stand for basic or actual syllables. In Chinese, the characters are the written alternatives to morphemes of the spoken language for representing lexons. Clearly, therefore, the first stages of machine translation systems having Russian and Chinese as source languages will differ from each other. In Chinese there is in fact less to do since the input is already in the form of morphemes.

In the case of Russian, the graphemes, which are to be considered the elementary units in terms of which the input is given to the computer, correspond in several respects more closely to the basic phonemes than to the phonemes of spoken Russian in that they fail to exhibit some of the phonologically conditioned alternation found among the phonemes, e.g. alternation of vowels under varying stress conditions. It is therefore efficient to treat the graphemes of written Russian like basic phonemes in the more general structural model described above, so that with Russian, as with Chinese, we can avoid the stage of phonological decoding. (But the head start is not as great as for Chinese since the input units for the latter are morphemes).

B. Stratificational translation

A machine translation system must have the linguistic information and computer programs necessary for a decoding capability in one language (the source language) and an encoding capability in the other. In other words, most of the machine translation process involves programs and information whose usefulness is not limited to translation. Only that which is in the middle is specifically concerned with translation, while automatic linguistic decoding and encoding have numerous uses. To speak of the process of automatic translation, then, is mainly to speak of automatic decoding and encoding. Decoding is the process of going through the linguistic structure from bottom to top, while encoding is the reverse process. Each may be separated into sections on the basis of stratification (and in fact neither can be efficient if such separation is not made), so that we may speak of phonological decoding, morphological decoding, etc. That is, corresponding to each of the components of a linguistic structure described above (cf. Figure 8) there is a decoding process and an encoding process. Inasmuch as these 'ologies' all have the same type of internal organization (cf. Figure 5 above), each having tactics and realizations, with rules of similar forms,

it might be expected that the processes of encoding and decoding are basically the same from one of them to another. And this is indeed the case. That is, the same basic decoding process applies for phonological, morphological, lexological, or semological decoding, and the same is true for encoding. Moreover, the processes of encoding and decoding for any of the 'ologies' are quite similar to each other, and may be regarded as variations on the same basic process; and the linguistic information needed for their execution is the same for either process and can be organized in roughly the same way. This organization of the linguistic information is also basically the same as that which can be used for an economical description of the structure. That is, it is not the case that the linguistic information must be organized in one way for decoding and in another for encoding, and that neither of these is the same as that which would be used for a linguistic description. It is not the case that the organization to be used for efficient description is unsuitable as a basis for production or decoding (as is the case for, e.g., the transformational approach to linguistics).

Encoding with respect to the Xology, where X is phon-, morph-, lex-, or sem-, consists of the formation of a combination of basic Xemes in accordance with the Xotactic rules, and the realization of the basic Xemes as actual Xemes, consisting of Xons, in accordance with the Xemic realization rules. Of course since the Xotactics (as information) generates (in the abstract sense) infinitely many combinations of basic Xemes, the production process must be under some kind of control to enable it to produce just a particular one at a given time. In experiments with computers involving a single stratum, control can be supplied by random numbers, as in the work of Yngve (1962);[4] and under various special circumstances or for special purposes, human language-users may provide various kinds of special control; but for normal speech and writing the control comes from the stratum above, or in the case of semological production it comes from the communicative intentions of the speaker (or writer), the circumstances in which he is speaking, and features of what he is speaking about.

The kind of control which an upper stratum exercises on the tactics of the next lower one is most easily seen in morphemic control of the phonotactics, since phonology is the simplest of the 'ologies'. Wherever a choice is available to the phonology, a morphon specifies which of the alternatives is to be selected. In other words, a morphon is not to be thought of as an object to be replaced or otherwise operated upon but rather as a specifier or selector, whose occurrence consists of activating a particular connection in the phonology. The phonology generates (again in the abstract sense) all well-formed syllables and combinations thereof, and a morpheme consists of the specifications leading to the production of a specific syllable or portion of a syllable or combination of syllables, as opposed to all the others which could be produced. For example, the morpheme which may be transcribed ^{M}/mor/ consists of the specifications leading to the specific syllable ^{P}/mor/ (orthographically, *more*). Although for notational convenience one can use the same letters 'mor' for ^{M}/mor/, ^{BP}/mor/, and ^{P}/mor/, the units symbolized

by these letters are quite different. In ^{P}/mor/, each letter stands for a bundle of phonons, while in ^{M}/mor/ each letter stands for a choice specifier, and in ^{BP}/mor/ each letter stands for the designation of a specific phonological realization rule selected jointly by the tactics and the corresponding morphon. The ^{M}/m/ specifies not a particular basic phoneme selected from the whole set of basic phonemes, but rather a particular possibility for the onset position of the syllable. Following it, i.e. after a nasal in the onset position, the phonology requires that a vowel come next, and the next morphon specifies a particular vowel as opposed to the others. In other words, what is symbolized as ^{M}/o/ in the notation ^{M}/mor/ selects one possibility from the set of vowels, while ^{BP}/o/ designates one member of the whole set of phonemic realization rules, i.e. a set with a larger number of members.

The basic principles of decoding with respect to any of the 'ologies' may be illustrated with morphological decoding, which is the first part of a machine translation system in which Russian is the source language. For the general model discussed above, morphological decoding covers the decoding operations leading from basic phonemes to basic morphemes. As discussed above, the Russian orthography is in large part oriented to the basic phonemes of Russian rather than the actual phonemes, so that it is efficient to treat the input chain of graphemes as composed of basic phonemes, handling what graphological alternation there is as if it were morphological alternation.

Decoding up to actual morphemes may be accomplished by means of a dictionary look-up process. For the most efficient dictionary look-up procedure that has been designed so far, the distinction between morphons and basic phonemes (or graphemes of a language like Russian) is utilized by coding the morphemes (i.e. the realizations in the morphemic realization rules, the strings which appear in them at the right of '/') collectively in a tree structure, which for computational efficiency is stored in two parts, namely (1) the 'letter tables', which can be directly addressed using the first few graphemes of a morpheme successively as addresses, and (2) numerous short 'truncate lists', in which the search for a matching morpheme can be completed. The first letter of a word is used to refer directly to its entry in the first-letter table, the contents of which are the address of the second-letter table for all morphemes beginning with that first letter; and the second letter determines the addressing of an entry in that second-letter table, etc. The procedure was described in an early version by Lamb and Jacobsen (1961) and later by Veillon (1963), who worked out some valuable improvements in certain features of the original technique. A version with further improvements is being programmed by the Machine Translation Project of the University of California, Berkeley, but it has not yet been described in print. It is hundreds of times as fast in its operation as what was said by some a few years ago to be the 'absolute maximum' speed that could be achieved for dictionary look-up by computer.

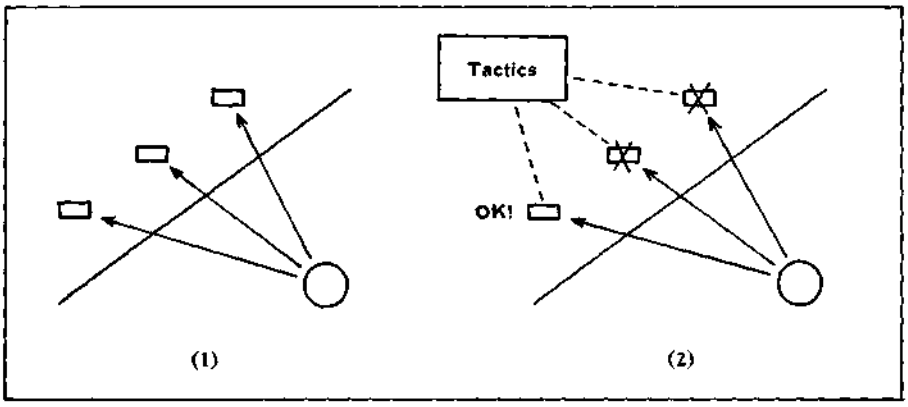
Of course ambiguities can be encountered in this look-up phase, since some chains of basic phonemes (in at least most languages) have more than one possible segmentation. For example, English *unionized* can be

segmented *union-ize-d* or *un-ion-ize-d*. Wherever such alternatives are encountered, all possibilities are to be carried forward. Any phase of the decoding process may encounter ambiguities, in which case multiple possibilities are passed on the next phase, where they can hopefully be resolved. An unresolved ambiguity remaining after the final phase of decoding reflects a true ambiguity in the text, provided that the linguistic information in the decoding system is complete and correct.

Upon finding each morpheme, the program is in effect given a direct reference to every basic morpheme of which that morpheme can be a realization. There may be two or more for which it is a neutralized realization. How then, when there are such multiple possibilities, does the program determine which is the correct one? When a basic morpheme has multiple realizations (i.e. when diversification is present), the realization rules specify which one is correct for any occurrence by giving a conditioning environment for each realization. But such conditioning environments are not directly suitable for 'upward' conversion to basic morphemes from neutralized actual morphemes, because they are expressed in terms of basic morphemes (not actual ones) and because in any case—i.e. even if restated in terms of actual morphemes—they would not in general suffice to resolve all cases of this type of ambiguity, since different realizates do not have to be in complementary distribution. In other words, realization rules are oriented towards encoding but not towards decoding. So there is a two-fold problem here: First, how then is such ambiguity to be resolved? Second, how is the information in realization rules concerning conditioning environments to be utilized in decoding? (Surely it must be necessary in some way, for if not then decoding would require less information than encoding, which hardly seems likely). The answer is that at this stage of the decoding process, the information concerning conditioning environments is simply to be ignored; but it will be used a little later. Bypassing the conditioning environments, then, the program is given directly all possible realizates for each possible morpheme, according to each possible segmentation (since all possible segmentations are carried forward to the next stage, as mentioned above). And what happens next is that each of the possibilities is taken through the morphotactics, which has the effect of rejecting all those that are not morphotactically well-formed (i.e. usually most of the multiple possibilities). The process is illustrated in Figure 10. In part one, it goes directly to all possible realizates of the ambiguous chain that are given by the realization rules. In part two all such possible realizates are tested by the tactics, which (in the usual case) rejects most of them (in the diagram, all but one). The Berkeley MT project has a tactic decoding procedure which allows multiple possibilities to be tested against the tactic rules in parallel, i.e. simultaneously in a single left-to-right pass through the chain, rather than serially (one after another), which would be far more time-consuming.

But then there must come the final phase of morphological decoding, in which whatever possibilities remain are put through the realization rules, but this time in the opposite direction, i.e. in the encoding direction, in

Figure 10



order to see whether the result of such encoding matches the input chain. It is here that the information concerning conditioning environments is used. (And so the morphotactic rules and the realization rules—i.e. the same linguistic information—are used for both encoding and decoding.) Any provisional decoding for which the test encoding fails to match the input is of course rejected. Whatever provisional morphological decodings remain are now no longer provisional but are to be taken as the morphological decodings. They fit the rules and their encodings match the input. If there is more than one morphological decoding for a given chain, that chain is morphologically ambiguous, and both (or all) decodings are passed on to the lexological decoding. Such a system has, of course, the ability to reject an input chain as being morphologically ill-formed—i.e. as having no valid morphological decoding (e.g. a word with a typographical error).

As a simple example consider the word *liven* of written (not spoken) English. The look-up would result in segmentation *live-n* and in references to multiple basic morphemes for each of these two actual morphemes. For *live* we have as basic morphemes a verb *live* and an adjective *live*. For *en*, there are several basic morphemes, including a verbal prefix, as in *endear, enshrine, enslave, entwine,* the past participle suffix of verbs, as in *taken, proven,* and a verbalizing suffix occurring with adjectives, as in *harden, sweeten*. Considering just these possibilities, we have two possible realizates for the first morpheme and three for the second, giving six (two times three) provisional decodings for the sequence. But these are immediately narrowed down to two by the tactic rules, which allow of the six, only the verb followed by the past participle suffix, and the adjective followed by the verbalizing suffix. Running these two remaining possibilities back through the realization rules in the encoding direction, we see that the provisional decoding of verb followed by past participle suffix is disallowed, since the realization rule for the past participle suffix specifies that with *live* it is realized as *d*. Therefore we have the one remaining decoding for the word, i.e. adjective followed by verbalizing suffix. (Note that if the word had

been *livens*, the tactics alone would have narrowed the possibilities down to one).

Lexological decoding consists of phases corresponding to those described above for morphological decoding. First there is the look-up, but this time in the lexeme dictionary instead of the morpheme dictionary. Just as the morpheme dictionary provides the information making it possible to segment combinations of basic phonemes (or graphemes) into morphemes, giving references to basic morphemes for each such morpheme, so the lexeme dictionary provides the information making it possible to segment combinations of basic morphemes into lexemes, giving references to basic lexemes for each such lexeme. The combinations of provisional lexological decodings are then tested by the lexotactics, which rejects all that are not lexotactically well-formed, and the remaining provisional lexotactic decodings are further tested by being put through the lexological realization rules in the encoding direction for comparison of the test encodings with the input combination of basic morphemes. Whatever provisional decodings now remain are taken as the lexological decodings, and are ready for semological decoding. There may of course be two or more lexological decodings for a sentence, since there are such things as lexologically ambiguous sentences; in fact such sentences are quite common. So, as before, whenever such multiple decodings exist, all are passed on to the next stage, in this case semological decoding.

Semological decoding is similar in its basic design to morphological and lexological decoding (and in fact the same is true for phonological decoding, which is a necessary stage for spoken languages). A preliminary discussion of the use of semotactic information to resolve ambiguities is given in Lamb 1966c.[5] The linguistic analysis needed for this stage is of overwhelming proportions by comparison with any of the first three stages, and several years of research lie ahead in this area, to provide the computer with the linguistic information needed.

Each language has its own semological structure, but it is quite likely that for a translation system for at least some pairs of languages it will be efficient to set up a single compromise semotactics for them, in which case the results of semological decoding will be immediately ready for encoding in the target language, beginning with the sememic realization rules. Such an approach would seem to be desirable in translating from Russian to English, since these two languages have various similarities in semological structure, possibly reflecting the fact that both are Indo-European languages. But it is possible that for some language pairs, such as English and Chinese, it will be desirable to have separate semotactic systems, in which case there will be a stage of conversion from the one to the other before the encoding stages.

In either case, the remainder of the translation process consists of the stages of encoding, from sememic networks to lexemic trees, and so forth, until strings of target language graphemes are arrived at, at which time, of course, we are ready for output to the printer.

NOTES

1. The relationship of realization is described and illustrated in Lamb, 1964a (where it was originally called representation).

2. The evidence for this analysis was brought to my attention by M.A.K. Halliday.

3. Cf. the description of realization rules in Lamb, 1964b. At the stage of the theory represented by that paper, however, phonological realization rules, for example, would be used to describe all of the features of the phonemic realization of MORPHONS (and the phononic composition of the PHONEMES), whereas in the system represented in this paper they are used only for describing the realization of BASIC PHONEMES, and differences between morphons and basic phonemes are accounted for by the tactic rules of the phonology.

4. This work of Yngve illustrates very well the use of the computer and a random number generator for testing grammatical rules, but is not to be taken as an illustration of encoding for any specific stratum, since the types of grammatical rules used are tactic only and account for a mixture of lexotactic and morphotactic phenomena.

5. That paper, however, is probably in error when it states that the phase of test encoding using realization rules is unnecessary in the semology because of absence of diversification.