[From: Paul L.Garvin (ed.) *Natural Language and the Computer* (New York: McGraw Hill, 1963)]

DAVID G. HAYS

Research procedures in machine translation

The symbolic nature of language is probably responsible for the widely held but erroneous view that linguistics is a branch of mathematics: a string of symbols "looks like" a mathematical formula. And, of course, a high school language textbook, with its rules, looks rather like a mathematical handbook. Mathematicians are forced to adhere to the rules of mathematical systems by the high cost of mistakes (i.e., of variations from the rules). Most speakers of most natural languages never learn the textbook rules, and those who do learn them discover soon enough that the cost of breaking many of the stated rules is negligible. In fact, whereas mathematical systems are defined by their axioms, their explicit and standard rules, natural languages are defined by the habits of their speakers, and the so-called rules are at best reports of those habits and at worst pedantry.[1] There is good reason for moderate pedantry in language teaching, as G. B. Shaw—lately with the collaboration of Lerner and Loewe—preached. But processing natural language on a computer calls for precise, accurate, voluminous knowledge of the linguistic behavior of the speakers or authors whose utterances or writings are to be processed. Here we shall consider acquisition of that knowledge.

TYPES AND SOURCES OF INFORMATION

Any language-data processing system has a purpose. A system for machine translation (MT) is expected to accept text in some natural language, perhaps Russian, and to produce text in another language, perhaps English. The output text should convey the same information as the input text; if it describes a chemical experiment, a chemist should be able to read the translation and reproduce the experiment with no more difficulty than if he had read the original report. Moreover, he should be able to

[1] See the discussion beginning with the words, "I am concerned with regularities; I am not concerned with rules," in Paul Ziff, *Semantic Analysis*, Cornell University Press, Ithaca, N.Y., 1960, pp. 34-38.

read the translation as easily as if it had been written by a person fluent in the output language—for example, Russian documents should be translated into versions that might have been written by Americans.[2] Other systems—for indexing, abstracting, automatic programming, sociological or historical research, legal documentation, and so forth—have other purposes, but here we shall concentrate on a detailed treatment of machine translation.

Knowing what a system must accomplish tells the designer—clearly or not—what information must be supplied it. An MT system[3] must include a list of source-language words with their target-language equivalents; when it becomes apparent that many words have alternative equivalents, and that choosing among them causes trouble for the reader (confusing him or at least slowing him down), the designer realizes that he must supply the system with information about equivalent choice—under what circumstances each equivalent is chosen.

Even if it were possible to translate every word accurately without reference to context, readers would be dissatisfied with the results. Individual words have meanings, but it is only by putting words together in sentences and paragraphs that authors can communicate useful ideas. In a source-language text, the relationships among the words in each sentence are indicated by natural devices belonging to the syntax of the input language. Translating word by word does not carry over the indicators of relationships, since natural languages share syntactic devices only to about the same extent that they share words; there are cognate words that can be recognized in French or German text by an American who knows no French or German, and there are cognate syntactic devices that make word-by-word "translations" partially understandable, but to rely on them would make reading the MT output like solving a word puzzle. Thus the designer must furnish his MT system with information about the syntactic structure of the input language and the output language and about the correspondence between them.

For sources of information the system designer will naturally turn first to published grammars and dictionaries. A grammar[4] lists categories (of words) and rules for combining categories; it purports to describe the syntax of its language. A dictionary lists words and specifies for each (the

[2] For a broad discussion of the problems involved in evaluating MT output, see George A. Miller and J. G. Beebe-Center, "Some Psychological Methods for Evaluating the Quality of Translations," *Mechanical Translation,* vol. 3, no. 3. pp. 73-80, December, 1956.

[3] An early sketch of an MT system was presented by Victor H. Yngve, "A Framework for Syntactic Translation," *Mechanical Translation,* vol. *4,* no. 3. pp. *59- 65,* December, 1957.

4 For example, H. Poutsma, *A Grammar of Late Modern English.* 2d ed., P. Noordhoff, Groningen, 1928 (2 parts in 5 vols.) .

categories to which it belongs; each entry also contains a discussion of the meaning of the word or a list of its equivalents in a second language. Taking grammars and dictionaries together, it should be possible to read and write grammatically correct sentences, translating each word accurately. Unfortunately, published grammars and dictionaries of the best sort are inadequate, even though they are vast compilations based on the prior original work of many linguists.

The largest dictionaries are intended to meet the needs of laymen, not of professional linguists; consequently, they omit reference to many categories that the layman can either recognize intuitively or disregard when he sees an unfamiliar word in text. The most detailed grammars are written for linguists who, recognizing that new words can be added to existing categories, make no attempt to list every word in every category. In general, until computational linguistics was conceived, no one needed a fully detailed account of any language for any purpose. Now that the need has arisen, new data must be collected and analyzed.

There are qualifications, of course. Fully detailed accounts of language have scientific value for linguistics, since they permit more exact tests of theory than gross statements about general tendencies could support. Furthermore, the major grammatical treatises dealing with Western languages—English, Russian, and others—contain many lists of words with special properties; these lists can be used to elaborate dictionaries by noting, in the dictionary entry for each word on the grammarian's list *X*, that the word has property *x*. But even a combination of information from multiple existing sources does not lead to a final, complete dictionary and there is still information to be gained from research.

The linguist has two sources of information beyond published studies. He can consult persons who speak the language, called *informants*. He can also study text, either written in the language or transcribed from conversations spoken in it. Of course, the published studies go back to exactly the same sources in the end. The two kinds of data sources can be used in tandem, with the informants serving as *editors* who comment on the text. Moreover, it is possible to obtain or create parallel texts in two languages, perhaps one known and one unknown, or one the input and the other the output of a proposed translation system.

The traditional methods of linguistics are based on the use of informants, or the alternate use of text and informants.[5] For non-Western languages, at least, it is fair to say that the success thus far achieved in scientific linguistics is the result of rich technical development and careful application of the informant method. Western languages have been studied by text methods and also with informants; often the linguist

[5] Zellig S. Harris, *Methods in Structural Linguistics,* University of Chicago Press, Chicago, 1951.

serves as his own informant when he is studying his own native language. The largest, most detailed grammars now in existence are the text-based grammars of Western languages, and it seems inevitable that text must supersede the informant when the details are to be filled in, simply because no one knows every particular of his language. Certainly no one knows any modern language, well developed as a medium for scientific and scholarly communication, in all its specialized ramifications. The informant learns his language by formal training and, more importantly, by constant exposure to its use. He cannot repeat to the linguist what he has never seen or heard. A sufficiently diverse set of informants would serve for any language, but the practical difficulties are obvious.

Moreover, data collected by textual research have a certain validity that data obtained from informants can never possess. An MT system, or any other automatic language-data processing (ALDP) system, will be called on to process segments of text from a definable stream. Predictions about the nature of that stream can be made, by the ordinary logic of statistical inference, from samples of it. Predictions can also be made from the responses of informants, but then the logic of inference must take into account the informant as a device that gathers information, summarizes, forgets, distorts, and reports.[6] The linguist should wonder whether he could not design a procedure that would process the same material as the informant more accurately and with less distortion.

The question of procedures for linguistic research always founders in discussion of the informant's intuition. The informant is more effective than a computing machine as a device for linguistic data reduction, according to this argument, because he understands the text to which he is exposed. The argument seems to come down to two points. First, the informant has a rough-and-ready grammar for his own language, which he uses as a framework on which to hang whatever new grammatical details come to him in reading or listening to new material. Second, he uses semantic analysis of text in deciding what its grammatical structure must be. As we shall see, the first point does not differentiate between computers and informants, since the linguist establishes some sort of grammatical framework at the very beginning of his research and commits it to machine memory; the framework may come from specific knowledge of the language to be studied or from a theory of linguistic universals, but it is essential. The second point is more significant: Can the gram-

[6] The methodological and technical problems raised by the use of informants are enormous. In psychological and sociological research, a sizable literature has grown up. See, for example, Robert L. Kahn and Charles F. Cannell, *The Dynamics of Interviewing,* Wiley, New York, 1957; Herbert Hyman, *Survey Design and Analysis,* The Free Press, Glencoe, Ill., 1955; Warren S. Torgerson, *Theory and Methods of Scaling,* Wiley, New York, 1958.

matical structure of a language be determined without reference to its semantic structure? If this question receives a positive reply, as it does from some but not all linguists,[7] then *should* grammar and semantics be kept apart? We cannot even begin to answer this question until we have looked into the nature of grammar, in following sections. In any case, research procedures based on text can be formulated with whatever admixture of informant intuition is considered appropriate.

The invention of techniques using text alone, with no help of any kind from informants, is one of the most exciting problems in linguistics today, and stimulation of work along this line may prove to be the most important contribution of the computer to the science of language.[8] The problem is to give an adequate characterization of the object of grammatical research without reference either to the intuitions of the informant or investigator or to the infinite *corpus* (body of text) that would resolve all questions if it could be written and studied.[9] (Grammatical statements often have the form *Item X can—or cannot—be used in context Y.* Such a statement would have an obvious empirical interpretation with reference to an infinitely long text in which everything occurred that could occur.)

Edited text can be used with less inventiveness; it is therefore a more practical material for the investigator who wants immediate results in the form of at least approximate knowledge about the speech habits of authors using a certain natural language. Given a text, editor informants can be asked to translate it, to paraphrase it, to describe the grammatical relations within each of its sentences, and so on.[10] The editor certainly uses his ideas about grammar, his semantic understanding of the text, and all his "intuition," in this process. The linguist's task is to generalize and formalize the informant's intuitive analyses of single sentences into a description

[7] Noam Chomsky, "Semantic Considerations in Grammar," *Georgetown University Monograph Series in Languages and Linguistics,* no. 8, pp. 141-154, Washington, D.C., 1955.

[8] Two papers on this subject have recently been published: Paul L. Garvin, "Automatic Linguistic Analysis: A. Heuristic Problem," and Sydney M. Lamb, "On the Mechanization of Syntactic Analysis," in 1961 *International Conference on Machine Translation of Languages and Applied Language Analysis,* vol. 2 pp. 655-686, H. M. Stationery Office, London, 1962. See also O. S. Kulagina, "A Method of Defining Grammatical Concepts on the Basis of Set Theory," *Problemy Kibernetiki,* no. 1, pp. 203-214,1958.

[9] A point raised by I. I. Revzin, "On the Notion of a 'Set of Marked Sentences' in the Set-theoretic Concept of O. S. Kulagina," in N. D. Andreyev (ed.), *Abstracts of the Conference on Mathematical Linguistics, Leningrad, 1959.* Translation 893-D, U. S. Joint Publications Research Service, Washington, D.C., 1959.

[10] The latest edition of the guide used in this work at RAND is Kenneth E. Harper et al., *Studies in Machine Translation—8: Manual for Postediting Russian Text,* RM-2068, The RAND Corporation, Santa Monica, Calif., 1960.

of the language as a whole, testing along the way for consistency, completeness, and simplicity.[11]

This discussion, therefore, is largely devoted to research methods based on text. Informant-centered methods are well described in the current literature, and text-based methods have definite advantages.

Text-based methods also have disadvantages that must not be forgotten. A large amount of text has to be processed before the investigator collects an adequate number of occurrences of any but the few commonest words or constructions. The cyclic method, to be described below, avoids this difficulty so far as possible by using a computer for much of the processing work. Another problem is the influence of the general environment on the content of any text. Caesar never wrote about television, yet no linguist would believe that the rules of Latin grammar prevented him. If there are no "octagonal whales" in our text, is it because of grammatical rules or not? The answer can only be that the distinction between grammatical rules and rules of other kinds is somewhat arbitrary, and will often be decided in terms of formal criteria without help from intuition. Only a dogmatist invariably knows a grammatical regularity when he sees one.

GRAMMAR

Grammar is a branch of linguistics. In a coherent treatment of the science or of a language, the study of grammar follows discussion of phonetics and phonemics—dealing with the sound system by which language is communicated orally—and of graphetics and graphemics—dealing with the writing system. Grammar itself has two main branches, morphology and syntax. Beyond syntax lies semantics, which will be considered later.

Morphology has to do with the analysis of words and *forms* of words. In some but not all languages the word forms that occur in text can be subdivided into repetitive fragments; that is, relatively few fragments combine and recombine in many ways to yield a large vocabulary of forms. In an MT system it is economical to avoid storing repetitive data if they can be reconstructed by a simple program from a smaller base; hence storage of fragments instead of full forms is usually advocated by system designers.[12]

More than economy is involved, however, since morphological analysis lays the foundation for syntax. Typically, the forms of a language can be segmented into prefixes, stems, and suffixes. For example, *inoperative* = *in* + *operate* + *ive*. A single form can consist of no prefixes or one or more prefixes, one or more stems, and no suffixes or one or more suffixes.

[11] Louis Hjelmslev, *Prolegomena to a Theory of Language* (Francis J. Whitfield, tr.), Univ. of Wisconsin Press, Madison, Wis., 1961, pp. 16-18.

[12] L. R. Micklesen, "Russian-English MT," in Erwin Reifler (ed.) , *Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian into English,* Univ. of Washington Press, Seattle, Wash., 1958, p. 5.

It seems to be a universal feature of natural languages that if forms can be segmented, some of the segments are involved in syntactic rules. Thus *operate* is a verb, but the *-ive* suffix converts it into an adjective; *boy* is a singular noun, *boy + s = boys*, a plural noun. In Latin, Russian, and other languages, noun forms can be segmented into stems and case-number endings; the case endings are involved in syntactic agreement with verbs, prepositions, etc.

The morphological classes in a language are classes of prefixes, stems, and suffixes. The classification is established by noting that some stems occur with certain prefixes and suffixes attached, but not with others. A noun stem, morphologically, is a stem that occurs with suffixes belonging to a definite set—the noun suffixes of the language. A verb stem is one that takes verb suffixes, an adjective stem one that takes adjective suffixes, and so forth. Prefixes are sometimes peculiar to nouns, verbs, adjectives, etc., and sometimes are attached to stems in categories that cut across morphological parts of speech.

A form, consisting of certain definite segments, can be assigned to a morphological *form class* according to the class memberships of its components. This classification of the forms in a language is the eventual contribution of morphology to syntax; any procedure for syntactic research can begin with form classes rather than with individual forms.

Syntax has to do with the analysis of sentences and the relations that obtain among the forms that occur in them. The structure of a sentence can be described in several ways; the theory of *dependency*, as used here, is familiar to anyone who has studied grammar in school. Tesnière elaborated the concept,[13] Lecerf contributed to the theory,[14] and the present author and his colleagues are using it in studies of Russian.[15] According to dependency theory, a partial ordering can be established over the occurrences in a sentence. One occurrence is independent; all the others depend on it, directly or indirectly. Except for the independent occurrence, every occurrence has exactly one *governor*, on which it depends directly. The diagram of relations among occurrences in a sentence is a tree, an example of which is given in Figure 1.

[13] Lucien Tesnière, *Eléments de Syntaxe Structurale*, Klincksieck, Paris, 1959.

[14] Y. Lecerf, "Programme des Conflits, Modèle des Conflits," *La Traduction Automatique*, vol. 1, no. 4, pp. 11-20, October, 1960, and vol. 1, no. 5, pp. 17-36, December, 1960.

[15] Kenneth E. Harper and David G. Hays, "The Use of Machines in the Construction of a Grammar and Computer Program for Structural Analysis," *Information Processing*, UNESCO, Paris, 1960, pp. 188-194. David G. Hays, "Grouping and Dependency Theories," in H. P. Edmundson (ed.), *Proceedings of the National Symposium on Machine Translation*, Prentice-Hall, Englewood Cliffs, N.J., 1961, pp. 258-266. Haim Gaifman, *Dependency Systems and Phrase Structure Systems*, P-2315, The RAND Corporation, Santa Monica, Calif., 1961.

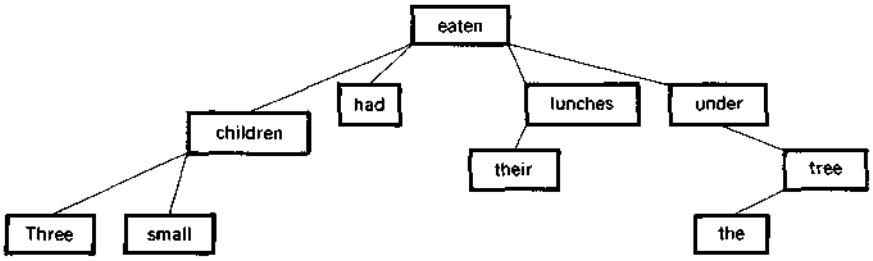

FIGURE 1 *Dependency structure.*

The syntactic structure of a sentence also includes a typification of each dependency link. Each dependent serves some definite syntactic *function* for its governor; one governor can have several dependents, all serving distinct functions, but it can have only one dependent serving any single function. (Of course, a given function can be served by several conjoined occurrences or by two or more occurrences in apposition.)

A sentence printed on a page is a linear array of letters, marks of punctuation, and spaces; morphological analysis converts it into another linear array, this one consisting of occurrences of segments grouped into forms and punctuated. If a sentence has a syntactic structure, it must be deducible from this array. The *indicators* that are available in natural language, the grammatical devices mentioned earlier as requiring translation along with the "words" in a text, include inflection, function words, occurrence order, and punctuation (in written language) or intonation (in spoken language). The use of these indicators is controlled by syntactic rules.

Inflection is used to show that a word that can serve several alternative functions in the language is in fact serving one in particular in this occurrence. For example, in Russian, a noun is inflected to show case: nominative when it functions as subject of a verb, accusative when serving as object, etc. Inflection is also used to show concord; a Russian adjective agrees with the noun it modifies in number, gender, and case, although one would not say that it has different functions corresponding to the different noun genders.

Function words are used in many languages; they have little or no *meaning,* in the ordinary sense, but serve only as indicators of syntactic structure. Prepositions, for example, differentiate functions more precisely than the case system can do; Russian has half a dozen cases and about fifty prepositions.

If each sentence contained no more than one word capable of governing any given function word or inflectional category, occurrence order would be almost irrelevant; an accusative noun in Russian, for example, might be recognized in any position as the direct object of the verb in the same sentence but for the fact that accusative nouns can serve other functions and other potential governors can occur along with a verb governing the accusative. (There is also the problem that some noun forms are ambiguous; they may be accusative or some other case.) Some prepositions, for example, govern the accusative, and a noun, a preposition, a verb, a comparative adjective or adverb, etc., can govern a genitive noun. Occurrence order therefore has to indicate which of several potential governors is actually served by a given occurrence. Occurrence order even differentiates functions; in English, the subject of a verb ordinarily comes ahead of it, whereas the object ordinarily follows, and the two are not morphologically distinguished except when one agrees with the verb in number and the other does not.

Punctuation serves sometimes to enforce a connection (as in hyphenated combinations), sometimes as a barrier to connection, sometimes to set off a semiparenthetic portion of the sentence. Intonation, historically the ancestor of punctuation, serves somewhat the same indicative role in spoken language.

One further kind of indication is given by word-class membership. The inflected forms of a word often share properties that help to indicate sentence structure. For example, words that govern objects (a syntactic function) can be taken as a class, and words that govern, as objects, accusative nouns are a subclass.

The *syntactic type* of a complete form is given by listing the functions that it can serve for all possible governors, the functions that possible dependents can serve for it, and the properties involved in agreement with potential governors or dependents. This information takes into account word-class membership and inflectional category; each function word in a language is likely to have a syntactic type peculiar to itself. Represented in a glossary by a grammar-code symbol, the syntactic type of a form is its whole contribution to the indication of the structure of any sentence in which it occurs.[16]

We can now see what the grammatical part of a machine-translation system must do: Using the indicators of a natural language—syntactic types, occurrence order, and punctuation, in conjunction with syntactic rules—the system must determine the structure of each input sentence,

[16] See, for example, K. E. Harper, D. G. Hays, and D. V. Mohr, *Studies in Machine Translation-6: Manual for Coding Russian Grammar*, RM-2066-1, The RAND Corporation, Santa Monica, Calif., 1958 (rev. I960). A. S. Kozak et al., *Studies in Machine Translation-12: A Glossary of Russian Physics*, RM-2655, The RAND Corporation, Santa Monica, Calif., 1961.

i.e., the dependency links and their functional types. Then, given the structure of a sentence, the system must find devices in the output language with which to indicate that structure. On the input side, there may be ambiguities; sentence-structure determination can end with more than one possible interpretation of a given sentence. Semantic analysis, as we shall see, can reduce this ambiguity in many or most cases. On the output side the system should be designed to avoid introducing new ambiguities, although it seems likely that that goal can never be fully accomplished.[17]

SEMANTICS

Sounds or letter sequences indicate what forms occur in a text. Grammatical devices indicate what syntactic relationships obtain among the form occurrences. And the words and syntactic relationships in a text indicate its meaning. The concept of syntactic structure can be formalized, perhaps as outlined in the preceding section, and the grammatical devices of a language inventoried. When we turn from syntactic theory to semantic, we face a blank wall; no adequate formulation of semantic structure is available today. Nevertheless, we are already able to survey at least some of the problems with which a semantic theory must cope and to offer at least some specific characteristics that a semantic theory must possess.

The segmentation of forms into prefixes, stems, and suffixes does not imply that those segments are the units to be translated. As we have already seen, some segments are used in the input text to indicate syntactic relationships, and it is those relationships that have to be translated, by means of appropriate indicators in the output language, not the segments themselves. Other individual segments do in fact have to be translated, but it is sometimes most convenient to translate combinations of segments within one form and occasionally combinations that include segments of several forms. The choice of units is connected with the determination of meanings.

Much evidence goes to show that the words of natural languages are ambiguous—i.e., have multiple meanings.

In translation, as from Russian into English, French, German, etc., a given Russian word may have many different equivalents in each output language, and its English equivalents may not translate unambigu-

[17] This is only a brief statement of grammatical theory; for a more complete treatment see Charles F. Hockett, *A Course in Modern Linguistics*, Macmillan, New York, 1958.

ously into French even if the correlation with a Russian word is known.[18] Monolingual dictionaries give multiple definitions for individual words, and, as Kaplan has shown, native speakers given context can "resolve the ambiguities" by assigning dictionary definitions to form occurrences.[19] Here it is only the fact of interinformant reliability that is convincing; no one informant could convince us that a real difference exists between two dictionary definitions of the same word, but if several informants, consulted independently, agree that occurrences *A, B, C,* . . . , take the first definition, whereas occurrences *X, Y, Z,* . . . , take the second, the difference clearly exists for speakers of the language. In conducting a test of this type it is necessary, of course, to remember that informants can be ignorant of distinctions that other users of their language make with regularity and precision. On the other hand, dictionaries are not infallible either, and they undoubtedly contain distinctions that are not known to speakers of the language, at the same time missing distinctions that are widely known.

A third line of evidence suggested by Harris[20] is that words with the same meaning should occur in the same range of contexts (have the same distribution, in the linguistic sense). It follows that a word with two meanings should occur in two distinct, separable ranges, i.e., its distribution should have distinguishable parts corresponding to the two meanings. All known suggestions for the resolution of ambiguity in ALDP systems, as well as all suggestions conceivable in computing systems limited to textual input, are based on this notion. Our point for the moment is simply that if a word occurs in two distinct ranges of contexts, and grammatical theory does not explain its distributional peculiarity, then semantic theory must be adduced.

The evidence that establishes multiple meaning as a linguistic phenomenon does not provide for determining exactly how many meanings each word has and how the boundaries are to be drawn. Informants may agree that a certain word has two meanings, yet not agree on its meaning in certain contexts, or a large group of informants may agree that it has two, while a subgroup divides one meaning into two, making three altogether. Translation into one language may require two equivalents for a certain word, into another three, and it may be argued that

[18] I. A. Mel'chuk, "Machine Translation and Linguistics," in O. S. Akhamanova et al., *Exact Methods in Linguistic Research,* Moscow University Press, Moscow, 1961. Translation: Univ. of Calif. Press, Berkeley, 1963.

[19] Abraham Kaplan, "An Experimental Study of Ambiguity," *Mechanical Translation,* vol. 2, no. 2, pp. 39-46, November, 1955.

[20] Zellig S. Harris, "Distributional Structure," *Word,* vol. 10, pp. 146-162, 1954.

some of the equivalents differ only stylistically or syntactically. Distributional evidence likewise ranges from strikingly clear to suggestively vague. In point of fact, a search for precision by any of these methods is likely to be thwarted, since all of them are indirect.

The three lines of evidence so far mentioned are all linguistic, whereas semantics must deal with the relations between language and reality, or, if reality is elusive, cognitive and cultural elements. Reality, as far as we now know, is infinitely complex, and languages, like science and all of culture, are finite. On a smaller scale, it would be nonsense to claim that the English word *hat* has as many meanings as there are, have been, or will be hats (headgear) in the world. All those hats are simply different referents for a single meaning of the word. No more does *bird* have as many meanings as there are species or varieties of Aves; one meaning covers them all. If a badminton bird is something else, it is because the culture has an organization independent of the language, and egg-laying birds are culturally differentiated from feathered hemispheres at a very deep level. It is *not* primarily a linguistic fact that the properties characteristic of birds (robins, canaries, etc.) and the properties characteristic of (badminton) birds are practically nonoverlapping. This fact pertains to the culture, to the cognitive systems of persons bearing the culture. Reality influences culture, and culture influences language; better said, the nonlinguistic part of culture influences the linguistic. Hence linguistic evidence, though indirect, can be used in the study of meaning.[21]

Each meaning of a word, then, is a cultural unit corresponding to a segment of reality that the culture regards as relatively homogeneous. A formal theory of meaning will have to go further, relating meanings to one another and giving an exact theoretical account of "relative homogeneity." One possible method is to list properties that the culture employs in forming concepts of reality; then a segment is relatively homogeneous if it can be distinguished from other segments by many properties but only subdivided by a few. Or it may be necessary to recognize that some properties are more significant to a culture than others and to decide homogeneity on the basis of the significance of the properties that isolate a segment as against the significance of those that cut it into subsegments. As yet we can say no more than this about the formal analysis of ambiguity.

Another semantic problem that we must consider is the calculation of the meaning of a sentence from the meanings of its constituent words or word segments. Syntax is needed in language to reveal semantic connections among the parts of sentences. In most sentences, for example, interchanging the subject and object of a verb alters the meaning of the

[21] The author is indebted to Dunne G. Metzger and A.Kimball Romney for the point of view adopted here.

whole in striking fashion; when the propaganda organization of a dictatorship announces that "Nation A has committed acts of aggression against nation B," interchange of A and B in such an announcement would be treasonable. Semantic relations are not identical with syntactic relations, however, and the same problems of identifying distinct meanings and resolving ambiguities arise with relations that we have already considered for words. We can begin with syntactic functions and attempt to determine how many different semantic relations can be indicated by each function. As before, we can use textual methods in research, but we must remember that these methods are indirect; the meaning of a syntactic function is a kind of relation that is identified by the culture and isolated from other kinds of relations.

With a theory of semantics in view, we can return to the problem of isolating translatable units in language. For some—but not necessarily all—of the segments that he isolates in a language by morphological methods, the linguist can determine one or more independent meanings. He certainly excludes those segments that serve only to indicate syntactic relations, since he must deal with them separately. He next considers word forms made up of combinations of segments, always excluding segments of purely grammatical (syntactic) significance. If the meaning of a word form can be calculated from the meanings of the component segments by a standard rule—i.e., a rule that holds for many forms in the language—then the segments are translatable units. If not, the form itself must be taken as a unit for translation.[22] Thus there are meaningful morphological relationships in language as well as meaningful syntactic relationships; each permits determination of the meaning of a combination from the meanings of the parts. Again, the linguist must examine combinations of forms in the language, testing whether the meaning of the combination can be calculated from the meanings of the forms and the syntactic functions that tie them together. When a combination appears with meaning that cannot be calculated in this fashion by a general rule, the combination must be treated as an *idiom,* or translation unit larger than a single form. The general rules correspond to semantic relations one to one; a single rule may not suffice for all occurrences of a single syntactic function and therefore would show multiple meaning: the syntactic function can- indicate more than one semantic relation, each associated with a rule.

Consider now the requirements of the semantic part of a machine-translation system. Taking sentences with known syntactic structures as input, the system identifies the translatable units and determines both the meaning of each unit and the semantic relations that obtain among the units. Then, given a representation of the meaning of each input

[22] The author is indebted to Martin J. Kay for discussions of this point.

sentence, the system must find words and semantic relations in the output language that express the same meaning. The output-syntax system operates on the results to produce sentences in which the meaning is indicated as clearly as possible. There may be ambiguities, of course, and two or more possible meanings may be discovered for a single sentence. Until semantic theory and research have progressed and additional systems are elaborated to go beyond semantics, the MT program can only offer alternative output sentences or a single sentence with the same ambiguity as that of the input.

SENTENCE-STRUCTURE DETERMINATION

In the input section of an MT system (or, apparently, any other ALDP system), a necessary step is determination of the syntactic structure of each input sentence. Computer programs for this purpose, called *parsing* or sentence-structure-determination (SSD) programs, can be written in many ways.[23] Different theories of syntax call for different programs, but a theory of syntax does not automatically suggest a definite method for SSD. The designer and programmer must attempt to satisfy several criteria. A good SSD program should always discover the correct structure of a sentence; at the same time, it should not propose more than the unavoidable minimum of incorrect structures. The minimum is fixed by the grammar of the language, but in practice good programming is needed, in addition to good linguistics, to approach the minimum. Second, the SSD program should operate economically; all other things being equal, a faster program is better than a slow one that wastes computing time. Finally, the SSD program should be economical to teach, modify, or adapt to a new language, even if these requirements can only be satisfied by a program that is not as fast as possible; fortunately, simplicity and speed go together to some extent. The need to teach programs is obvious. Modification is needed as empirical knowledge and formal theory advance; no formalism can be called final and complete. Adaptation to new languages is certainly to be expected, and this requirement means that a good SSD program is designed around "universal" features of language.

[23] For examples, see Ida Rhodes, *A New Approach to the Mechanical Translation of Russian,* Report No. 6295, National Bureau of Standards, Washington, D.C., 1959; papers by A. G. Oettinger, M. E. Sherry, M. Zarechnak, and P. Garvin in Edmunson (ed.), *op. cit.,* fn. 15; papers by I. Sakai, A. F. Parker-Rhodes, M. Corbe and R. Tabory, and F. L. Alt and I. Rhodes in 1961 *International Conference on Machine Translation of Languages and Applied Language Analysis, op. cit.,* fn. 8; D. G. Hays and T. W. Zeihe, *Studies in Machine Translation-10: Russian Sentence-structure Determination,* RM-2538, The RAND Corporation, 1960; and Lecerf, *op. cit.,* fn. 14.

The syntactic indicators that we have noted (agreement rules involving inflection, stem classes, and function words; occurrence order; and punctuation or intonation) can be taken, roughly, as linguistic universals. Not all languages use all these devices, and certainly not to the same degree or in the same fashion, but the list is exhaustive, and each device is used in many languages. These devices have different logical characteristics that entail different treatment in a program.

Every language seems to have some degree of grammatical classification, and it has been suggested that every language must have at least two classes. But the classification schemes of natural languages vary enormously, and the agreement rules vary correspondingly. Hence tabulation of grammatical classes is a new task for each new language, and whatever program is used for SSD must be able to accept new agreement systems. One scheme that will most probably work for any language is to use a table of pairs (or perhaps triples, etc.) of syntactic classes in testing agreement. A grammar-code symbol is assigned to each class in the language. The table of *dependency types* is entered with a pair of grammar-code symbols; if their agreement indicates syntactic connection, they are listed with an indication of which member of the pair is governor and what function the dependent serves.

More sophisticated, simpler, and faster schemes have been proposed.[24] For example, each form can be assigned to a part-of-speech class according to the syntactic functions that it can govern and those that it can serve as dependent. Part-of-speech pairs such that one member of each pair can serve a certain function for the other are listed in a small table. Then a set of additional tables is used to determine whether two forms agree in those respects that are significant for the possible function. This technique is advantageous for forms that have many grammatical properties most of which do not relate to any particular function; this condition is satisfied for many languages, perhaps for all.

Occurrence order can be factored into *direction* and *separation.* Direction is similar to grammatical classification; its indicative function is specific to the grammatical classes involved and therefore varies from one natural language to another. This aspect of occurrence order can be effectively handled by tabulating it with grammatical classes in the table of dependency types.

Separation is the subject of a putative universal. In terms of de-

[24] A. F. Parker-Rhodes, "A New Model of Syntactic Description," *Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis,* op. cit., fn. 8, vol. 1, pp. 26-60. See also RamoWooldridge reports, *Machine Translation Studies of Semantic Techniques,* 1960 and 1961; the grammar-coding scheme reported there is due to Paul Garvin.

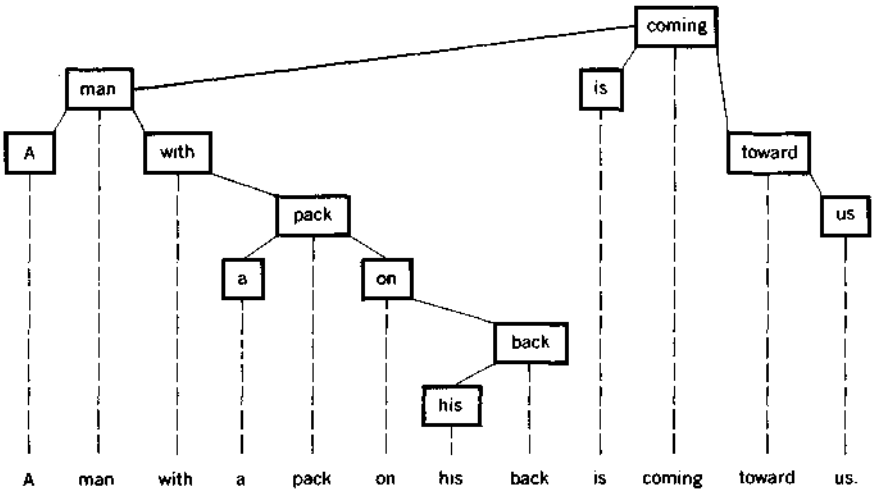

FIGURE 2 *A sentence with projective dependency structure.*

pendency theory, Lecerf calls this feature of language *projectivity.*[25] According to this rule two connected occurrences are separated only by occurrences that depend on them, directly or indirectly. In Figure *2,* a sentence is displayed with its dependency structure; governors are placed higher than their dependents, and occurrence order is followed from left to right. Dropping a projection line downward from each node in the tree to the occurrence below it, we see that no two dependency connections intersect and that no connection line intersects a projection line; hence the name projectivity.

The practical significance of this observation in SSD, whether attempted by a computer program or by a human listener, is that it reduces the labor involved in the process. A program can be written on the basis of the projectivity rule alone, without reference to grammatical classes and agreements, that identifies pairs of occurrences as connectable or not; if one occurrence is separated from another only by occurrences that depend on one or the other, the first is said to precede the second. Agreement need be tested only for precedence pairs, and not for all pairs in the sentence. Without regard to projectivity or rules of agreement,

[25] Lecerf, *op. cit.,* fn. 14, is based on a Euratom report published at a conference in February, 1960; the SSD program given by Harper and Hays, *op. cit.,* fn. 15. relies on projectivity, but mentions the property without naming or analyzing it. Lecerf and the present author met and learned of each other's work only late in 1960.

5×10^{24} distinct tree diagrams can be drawn for a sentence of twenty form occurrences; the rule of projectivity reduces this number to 4×10^{13}.[20] The program that tests for precedence can be written just once and applied to any language characterized by projectivity.

It would be possible, but uneconomical, to list all the projective trees with a certain number of nodes and then, using agreement rules, test each of them against a particular sentence. Since the number of trees is in fact too large, an iterative plan is required instead. In its simplest version, this plan lists the precedence pairs in a sentence of which nothing is known; each of these pairs consists of two occurrences lying side by side. Then agreements are tested and some connections made which lead to establishment of precedence between new pairs of occurrences; new agreements can be tested, new connections made, and so forth.

The general organization of an SSD program determines whether it yields all grammatically possible structures for a given sentence or only one. If it were possible to guarantee that the one would always be correct, it would be economical to choose a plan of the latter type, but there are at least temporary advantages in finding all possible structures. One plan is to form all possible subtrees of two nodes, then all those of three nodes, etc., until finally all possible trees have been formed containing nodes for all the occurrences in the sentence.[27] Economy demands that no subtree be formed by two or more paths; this demand can be satisfied by accepting two restrictions. First, no new dependent is added to any node in a subtree other than the unique independent node. Second, the order in which dependents are attached to any governor is fixed; for example, after any dependent is added that precedes the governor, no further following dependents are attached.[28] (Since projectivity precedence requires nearer dependents to be attached earlier than those more distant, this rule makes the order of attachment unique.) Disregarding agreement rules, a sentence of three occurrences can have seven distinct projective tree structures (see Figure 3). Following the above sequencing rules those seven structures can be obtained with connections made in the order shown in the figure; without the sequencing rules, three structures would be obtained in two ways each (those labeled with an asterisk). As the number of occurrences in a sentence increases, the saving increases disproportionately.

Without projectivity, a search for all possible structures would be time consuming indeed. It is well known that nonprojective sentences

[26] Y. Lecerf, "Analyse Automatique," in *Enseignement Préparatoire aux Techniques de la Documentation Automatique*, Euratom, Brussels, 1960, pp. 179-245.
[27] This point was communicated to the author by John Cocke of IBM Research.
[28] This suggestion comes from Lecerf.

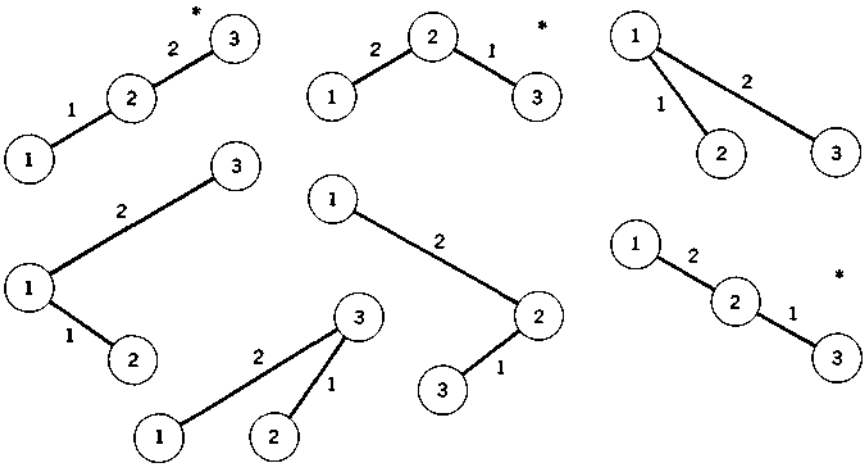

FIGURE 3 *The seven possible projective structures for a three-occurrence sentence. *See text for alternative.*

occur frequently in some languages, and occasionally even in languages that are largely projective, such as English and Russian. A survey of nonprojective structures appears necessary as a possible means of discovering regularities that can be used in SSD.[29] The present computational cost of assuming that every sentence is nonprojective would be too high, and the approximation—in many languages—is too good, to permit outright rejection of the concept.

SEMANTIC RECOGNITION

When the syntactic structure of a sentence has been determined, and the minimal units with independent meaning have been identified, the meaning of each occurrence in the sentence and the nature of the semantic connections among them have to be determined. Work in this field does not yet enable us to describe procedures of proven effectiveness, but some suggested methods can be reported.

It may be possible to assign meanings to semantic agreement classes.[30] In that case, a table could be used much as a table of dependency types is used in SSD. An entry would consist of a pair of semantic-class symbols and

[29] Lydia Hirschberg of the Free University of Brussels is engaged in such a study.

[30] See, for example, Kenneth E. Harper, "Procedures for the Determination of Distributional Classes" in 1961 *International Conference on Machine Translation of Languages and Applied Language Analysis, op. cit.,* fn 8, vol. 2, pp. 687-700.

an indication of a syntactic function. The question would then be, *Can an item of this class serve that function for an item of the other class? If so, what is the semantic relation between the two?* For example, *Can the name of a person serve as the (syntactic) subject of a verb of communication?* The answer would be, *Yes; the person is actor.* The classes in this example, are, of course, not necessarily those that would appear from empirical research.

Following this plan leads to success if one and only one meaning of each occurrence in a sentence agrees with the meanings of neighboring occurrences and if each syntactic connection is resolved to a unique semantic relation. If the sentence has more than one possible syntactic structure, semantic disagreements may rule out some or all of them. The semantic classes and agreement rules therefore have to be designed to determine a unique meaning for each word occurrence and each syntactic relation in every sentence, to eliminate all but one possible structure for each sentence, and to assign to every (intuitively acceptable) sentence a semantic description that can be translated or otherwise manipulated to the satisfaction of whatever external criteria are applied. As yet there is no evidence that any semantic agreement system can approach this design standard.

Another proposal is to organize the vocabulary of *meanings* hierarchically; formally, the organization would be a lattice.[31] Choosing any set of meanings arbitrarily, we find that there is some set of meanings in the hierarchy that includes all of them; in fact, there may be many sets with that property, one having a smaller count of meanings than any of the others. The occurrences in a sentence have meanings that can be found in this hierarchical system, but each occurrence can have one or more meanings. In a three-occurrence sentence, for example, let the first occurrence have two meanings: 1*a* and 1*b*. Let the second have two meanings also—2*a* and 2*b*—but let the third occurrence be unambiguous—call its meaning "3." We must choose one meaning for each occurrence; we can choose: 1*a*, 2*a*, 3; 1*a*, 2*b*, 3; 1*b*, 2*a*, 3; or 1*b*, 2*b*, 3. Trying each set of meanings in turn, we learn the size of the smallest class in the hierarchy that includes all meanings in the set; e.g., how large is the smallest set that includes 1*a*, 2*a*, and 3? We thus obtain four quantities, one associated with each set of equivalent choices for the sentence, and we take the set of equivalents associated with the smallest of those quantities, since semantic *homogeneity* is to be expected in an ordinary text. Ties are possible, however, which lead to semantic ambiguities.

[31] Margaret Masterman, "Potentialities of a Mechanical Thesaurus," *Mechanical Translation*, vol. 8, no. 2, p. 36, November, 1956. See also her paper in *Proceedings of the International Conference on Machine Translation and Applied Linguistic Analysis, op. cit.*, fn. 8, vol. 2, pp.437-474.

The difficulty with this model is that syntax cannot be combined with it in any obvious way. In fact, the proper solution to semantic problems could be a combination of the two methods that we have described—the first takes advantage of local context, the second uses broad context. The ambiguities not eliminated by one might then be resolved by the other.

THE CYCLIC METHOD

Linguistic research, if it is to be based on text at all, must be based on a very large corpus, since many rare words and constructions must be discovered and described. The advent of the automatic digital computer makes a very large amount of text-based linguistic research economically feasible for the first time. However, if informant editors are to be used in the system, care must be taken to keep their task under control. The cyclic method makes as much use of computers as possible: Each cycle is based on a fresh batch of text and consists of automatic application of what is known about the language followed by editorial review and reduction of the new data to a form that can be used in the next cycle.[32]

Glossary expansion furnishes a simple illustration of the cyclic approach. A glossary is, for the moment, an alphabetic list of all the forms that have occurred in text. A form is a distinct sequence of alphabetic characters that occurs preceded and followed by spaces or marks of punctuation. When a fresh batch of text has been prepared, it is segmented into form occurrences, and each form occurrence is matched, by spelling, against the known forms of the language—that is, against the old glossary based on previous batches of text. The unmatched items constitute new information about the language; they are alphabetized and merged with the old glossary. Thus a cycle ends, and a new one can be initiated with the preparation of a fresh batch of text.

If the glossary contains segments—prefixes, stems, and suffixes—rather than forms, each form occurrence must be identified as a sequence of known segments that are described in the glossary as capable of cooccurring in a form or as partially, perhaps entirely, unknown. A new form might consist, for example, of known prefix, unknown stem, and known suffix. If the prefix and suffix can cooccur around a stem, the morphological type of the stem may be deducible. The glossary-development program can list, for an informant editor, the new stems and their possible morphological types. The editor, who also needs a list of the full forms in which the new stems occur, must decide whether the segmentation is valid; if it is not, he corrects the segmentation and code assignments. His

[32] This method was adopted at RAND in 1957; see H. P. Edmundson and D. G. Hays, "Research Methodology for Machine Translation," *Mechanical Translation*, vol. 5, no. 1, pp. 8-15, July, 1958.

notes are returned to the computer and the new material is added to the glossary in preparation for the start of another cycle.

To find new idioms, the editor must read the text in full. Each batch of text is machine translated; each sentence is translated by whatever approximation to the MT system is available at the time. Some occurrences are translated by sophisticated techniques, some are translated in idiomatic combinations, and others are simply given whatever equivalents are in the glossary. The text and its "translation" are printed out in parallel columns, matched item by item. The editor reads the printout, looking for all kinds of errors. In particular, he looks for new idiomatic combinations—sequences of form occurrences that are mistranslated individually but can be given an accurate translation as a group. He writes in the translation that he wants, marking the extent of the idiomatic sequence. His notes are keypunched and entered into the computer along with the original text and its machine translation.

An analytic program selects all occurrence sequences marked as new idioms and sorts them, alphabetically or otherwise. If the same form sequence has occurred several times, and the same idiomatic equivalent used each time, the occurrences can be summarized. A summary list of new idioms is prepared, inspected by linguists, and returned to the computer for another operation.

The final step automatically merges the new idioms with the old, creating a new idiom list, and adds to the glossary of forms or form segments whatever indicia are required for idiom recognition.

The addition of new idioms is a complete, albeit simple, illustration of the cyclic method as practiced at RAND and elsewhere. In each cycle, text is prepared for computer input, submitted to machine translation, and postedited. Analytic routines reduce the postediting data in accordance with linguistic requirements, and linguists inspect the result. The new information is finally merged automatically with old information of the same type. This sequence is characteristic of the method; it gives the computer a major share of the work but also permits both informants and linguists to apply their intuition. Cutting out a single task, such as the search for new idioms, is also characteristic. Economy dictates that the editor read each corpus as few times as possible, but analytic operations have to be isolated for theoretical reasons. Moreover, the whole MT system—comprising a sequence of processes applied to text one after another, each using the results of those already completed—suggests separation of the research cycle into subcycles. Glossary expansion can be completed before idiom lookup is performed; sentence-structure determination can be postedited before its output is submitted to semantic recognition procedures. Stepwise postediting of this variety saves editorial labor by preventing carrying over mistakes from step to step. On the other hand, the technique would leave, at each step, ambiguities that

could be eliminated at the next, and it requires the informant editors to read each text more than once. The stepwise editing plan has not been tried out as yet, but the cyclic method with posteditors has been in use for some time.

TRANSLATION OF WORDS: SEMANTICS

Standardization of Equivalents Let us next turn to a more general view of the problem of pairing meaningful units in the input and output languages. Suppose that we must begin with a new pair of languages, for which the only available information is that contained in published dictionaries. We are to proceed by the cyclic method, processing text in successive batches. At first the relatively few most common words in the language will dominate our lists of new forms; many forms found in the first batch of text will occur frequently both in that batch and in succeeding batches. Other words in the first batch, and most new words in later batches, will be rare. The basic plan for assigning equivalents can therefore reasonably change as the number of batches processed increases.

The first batch is prepared, and an alphabetic list of the forms that it contains is made. It is convenient to collect forms into groups, when the input language is highly inflected, since the translational equivalents of different inflected forms of the same word will usually be identical (except for what we will regard as grammatic variations). Now each form or word in the text-based list can be looked up in a published bilingual dictionary and the equivalents listed for it copied into machine storage. The list of forms, each accompanied by one or more equivalents, is an initial glossary.

The next step is to list the first batch of text, with its machine translation in parallel. But of course the translation is merely a statement—for each form occurrence—of the equivalents shown for that form in the initial glossary. Now the editor informant, who should be well acquainted with the subject matter of the text, selects an equivalent for each occurrence in the first batch. He can mark one of the listed equivalents, write in a new one, or identify an idiom. His marks are keypunched and correlated with the machine-stored text and translation.

The number of times that the editor selected each equivalent for each form in the first batch is easily determined by an automatic process. Then the equivalents of each form, including those inserted in advance and those added during editing, can be ordered automatically by frequency of choice. Although separate records must be kept for each *form,* the ordering should be done for each *word;* i.e., the equivalents of all forms of a word should be kept in the same order. Now the typical glossary entry consists of a form and an *ordered* set of equivalents, together with a code symbol if the form was used idiomatically.

Treatment of the second, third, etc., batches of text proceeds in the same fashion, with two modifications. Beginning with the second batch, the editor is instructed to use the first equivalent listed with each occurrence whenever it is substantively accurate. When the second and subsequent equivalents of a word are never used by editors working under this instruction, the linguist can be sure that the alternatives differ only stylistically in the stream of text that he is processing. On the other hand, if one (or more) of the alternatives is still used, its meaning is substantively distinct from the meaning of the most frequent equivalent, and the linguist can look for contextual indicators of the difference.

The second change in procedure is omission of the first step—the insertion of tentative equivalents found in a published dictionary. This change is justified when the average number of occurrences of each new form is small enough; generally, the reasonable level is an average of two occurrences in a batch of text. There are several reasons for this modification. For one thing, the new words are hard to find; some are in the dictionaries, some are not. For another, the proportions of cognates and proper names increase. And for another, the equivalents obtained grow less and less reliable. Altogether then, it seems best to add equivalents only during editing, once a good glossary has been developed.

Input-language Inflection and Choice of Equivalents Most words, or form groups, have uniform translations, but not all. Some Russian verbs have one English equivalent in their nonreflexive occurrences, another (not passive of the first, which would be considered the same equivalent, modified grammatically) in reflexive occurrences. Some nouns have one equivalent in the singular, the same equivalent or another in the plural. These exceptions to the general rule must be discovered and taken into account. The procedure is simple and straightforward. A file of equivalent-choice data, tallied by form and grouped by word, is required. With each form, the file must include a grammatic description. The procedure is applied to each word that satisfies three tests: (1) at least two forms have occurred; (2) at least two nonidiomatic equivalents have been chosen; and (3) enough occurrences of the word have been processed for reliable conclusions to be drawn.

The procedure is to sort the forms of a word into grammatic categories and for each equivalent test whether it occurs equally often in each category—that is, in proportion to the total number of occurrences of the word in each category. A statistical test for nonproportionality should be applied; although the satisfaction of its underlying assumptions is by no means clear, the chi-square test is perhaps appropriate.

The exceptional words can be listed and the findings installed in the glossary so that when new text is processed only applicable equivalents will be printed with each form occurrence.

Equivalent Selection by Contextual Criteria Perhaps the majority of substantive equivalent-selection problems can be resolved by reference to grammatically related occurrences in context. A verb, for example, may take one equivalent or another depending on its subject, its object, or a modifier. An adjective is most likely to be influenced by properties of the noun it modifies, since adjectives usually occur without dependents. In any event, establishment of rules for the determination of equivalents must include analysis of context.

It appears that analysis of related occurrences should be organized by kind of relation—i.e., by grammatic function. The procedure would be applied to each word with two or more equivalents, both applicable to at least some forms, when sufficient occurrences had been processed to permit anticipation of reliable results. All occurrences of the multiple-equivalent form are collected; the required file of information includes *what* words were related to the given word, and with what function, as well as the choice of equivalent that was made.

The analysis then takes one function at a time; since every occurrence has a governor, let us start with that. A particular word can serve different functions for its governor in different occurrences: A certain noun can occur in one place as subject of a verb, in another as object of a preposition, etc. If the multiple-equivalent word that we are studying has only one equivalent in each kind of relation that it enters as dependent, the analysis is complete; the problem shifts, as it were, from semantics to grammar. If there is any kind of relation in which the word has two equivalents, the analysis continues by examining each *word* that governs the multiple-equivalent word. If a certain word as governor always—in the processed text—implies a certain translation for its multiple-equivalent dependent, that fact is recorded; if the same can be said of every governor, the evidence suggests that choice of equivalent depends on type of governor. If not, a summary statistic can be computed—that is, the percentage of occurrences for which the correct equivalent can be selected by inspection of the governor.

The summary statistic is computed in the following manner. Let $E_1, E_2, \ldots, E_n$ be the equivalents of a word, and $C_1 C_2, \ldots, C_n$ be criterion classes. Considering only one class of related words (e.g., governors) , assign each related word to class C_i if E_i is chosen more frequently than any E_j, $j \neq i$ when the related word is present. (In case of ties, assign at random.) Then, assuming that E_i is chosen when a word in class C_i is present, the summary statistic is just the number of correci choices divided by the number of occurrences of the multiple-equivalent word. This fraction, which we can call p, is at least as large as the ratio of choices of E_1, the most frequent equivalent, to occurrences of the word under study; and p is no larger than unity, which would indicate com plete accuracy. In fact, since the number of errors with each distinct re-

lated word is limited to $(n - 1)/n$ times the number of occurrences of that word, the expected value of p must increase with the number of distinct words related in the given way. The sampling distribution of p, under the hypothesis that the classification of related words is irrelevant, still has to be calculated, and from it, parameters for normalizing p could be deduced.

Continuing our analysis of a multiple-equivalent word, we would examine words in each possible grammatic relation to it, calculating p, or a normalized variant p^*, for each relation. The relation for which p^* had highest value would deserve the attention of a linguist, since a few errors that might prevent p^* from reaching unity could be due to careless editing. If no value of p^* was high enough to be useful, the automatic analysis would have to continue by combining criteria. Harper, for example, working with a less formal method of analysis, used both governors and objects of prepositions to determine the equivalents required.[33] There is no certainty, of course, that the governors and dependents of an occurrence determine its translation, but it seems plausible that they will often do so.

When the criterion classes can be discovered, their members have to be marked in the glossary. A generalized semantic-recognition program can use these marks to select meanings, and thus equivalents, for occurrences of the words to which the method is applicable.

So far it has been assumed that criterial classes are defined independently with respect to each multiple-meaning word. That plan would eventually call for the storage of a vast amount of information. However, it is desirable to reduce the requirements, or at least to be assured that redundant information is not stored. Furthermore, the criterial classes can reasonably be interpreted as semantic classes only if they are relatively few in number and if word meanings fall into classes that allow use of the same criteria with all members of any given class. The question is therefore whether criterial classes formed in different ways are identical. With finite text no two classes are likely to have exactly the same members, but a degree of overlap exceeding random expectations would be evidence of relatedness. Two classes, criterial for selection of the meanings of two different words, are the same class if every word belonging to one also belongs to the other; no matter how large the corpus, there is always some chance that a sentence will occur in which a member of one class gives an incorrect result when treated as a member of the other class. This possibility must be eliminated before a sound model for stasistical inference can be formulated. If "exceptions" are allowed, an alternative formulation is to coalesce two classes whenever the cost of storing

[33] Kenneth E. Harper, *Machine Translation of Russian Prepositions*, Paper P-1941, The RAND Corporation, Santa Monica, Calif., 1960.

and manipulating one list with known exceptions is less than the cost of storing two lists with no exceptions. To make this alternative attractive, an intuitively acceptable estimation of the relative costs must be made.

SYNTACTIC RESEARCH

Problems of morphology come up in the study of non-Western languages and even in work with Russian or English when it becomes necessary to cover all details with a uniform scheme, but problems of syntax are much more significant in current work on well-known natural languages. In this section we shall assume the existence of a complete and unchangeable morphological description of the subject language; working on that assumption, we consider several plans for syntactic research.

After a sentence-structure-determination program has found all possible structures for a sentence, an editor informant examines them and chooses the correct one if it is listed. Errors in grammatical classification or tabulation of dependency types, as well as failures of the syntactic theory, can cause the SSD program to miss the correct structure; in that case, the editor must add the structure he desires to those listed. His notes, covering both connections and functions, are keypunched and collated with the stored output from SSD in preparation for analysis.

The sentences for which the editor wrote structures not found by the SSD program must be processed first, since they reveal major gaps in the system. The first step is to test for projectivity. The program examines each connection in a sentence and determines whether every occurrence between the members of the connected pair derives from one or the other of them. If not, it marks the connection as nonprojective; such a connection needs further study by a linguist.

In the SSD program considered above, the establishment of connections in a sentence is in a fixed order, which we can call the *recognition order*. The primary sequencing variable is the size of the subtree that results from a connection. Two subtrees are assembled only by connecting the independent element of one with the independent element of the other; dependents of a given node must be attached first on one side and then on the other. For several reasons, it is necessary to alter grammar-code symbols as connections are made; the alterations follow instructions in the table of dependency types. Thus at the time any connection is made, the grammar-code symbols of the occurrences to be connected are the result of all prior connections in which they have participated. When the correct structure of a sentence is not found by the SSD program and if the structure is projective, it must contain a connection that is impossible according to the existing system. To ascertain the cause, the SSD operation has to be repeated.

A controlled SSD program can be used for this purpose. The control is based on knowledge of the correct connections in the sentence; these con-

nections are taken in recognition order and tested in turn against the table of dependency types. Alterations in grammar-code symbols are made as they were originally. When the "impossible" connection is reached, the SSD program has constructed a list of all grammar-code symbols assigned to each of the two connected members as a result of the alterations keyed by prior connections. Considering all possible pairs of these symbols, the program determines whether some alteration is responsible for the failure to find a suitable entry in the table of dependency types. In other words, if the impossible connection could have been made but for the alteration of a grammar-code symbol at the time of a prior connection, the alteration can be blamed for the failure to produce a correct structure for the sentence, and the relevant information must be printed out for a linguist to examine. The linguist can decide whether to change the alteration instructions, change an entry in the dependency table so that the latter connection can be made in spite of the alteration, etc.

If no alteration is responsible for the failure, the difficulty is in the grammar-code symbols, in the dependency table, or in lack of an alteration that should have been made. What is possible at this stage depends somewhat on the organization of the table.

In the simplest case, the table is a list of pairs of full grammar-code symbols. The symbols belonging to any pair of occurrences that have to be connected can be added to the table, but a screening process must eventually be carried out to avoid recognition of an excessive number of false structures for sentences in the future text. The screening program can be exemplified in Russian. In this language there is a morphological category of nouns; noun forms are morphologically subclassified by case. When enough connections between noun governor and noun dependent have been recognized in text, the screening program can detect that the case of the dependent is relevant to the function it serves, whereas the case of the governor is not. To reach this conclusion, the program must consider all morphological categories, testing for morphological diversity within each functional type; finding that every noun dependent serving a given function for a noun governor is in a certain case, the program can conclude that the case of the dependent is relevant. On the other hand, the program finds that a noun governor in any case can take a dependent noun with a given function; hence the case of the governor is irrelevant. The screening program also builds word classes. In a statistical sense, the nouns that can serve a given function when they occur in a given case are lexically diverse—many different nouns are found as dependents with any given function. Governing nouns, by contrast, are lexically restricted; the number of different nouns that govern the instrumental case, for example, is much smaller than the number expected by chance if every noun is capable of governing instrumental nouns. The statistical evidence proves the existence of a syntactic class; membership

in the class is proved only by occurrence in the defining context—in the example, a noun is added to the class when it occurs as governor of an instrumental noun serving a particular function.

Once syntactic word classes are established, the organization of the dependency table can profitably be elaborated. Each grammar-code symbol will be cut into three parts: the morphological part of speech, other morphological properties, and syntactic word-class memberships. When two occurrences are said to be connected and the dependency table cannot connect them, the parts of the grammar-code symbols can be tested in turn. Continuing the previous example, let us suppose that both occurrences are nouns. First, parts of speech are consulted. Second, given the function named by the editor, the relevant morphological properties are sought in the table. If two occurrences of the given morphological types cannot be connected, an entry must be added to the table (but see below). If the morphological requirements are satisfied but a connection still cannot be established, syntactic word-class memberships must be involved in the agreement. If the class memberships of both occurrences are relevant and one belongs to a single relevant class, the grammar-code symbol of the other can be changed in the glossary; the same is true if only one occurrence in the pair must belong to a special category. On the other hand, if both occurrences must belong to particular classes and neither belongs to a relevant class, the glossary entries can be changed only if the classes are unique; otherwise, the pair must be set aside for further analysis. For example, suppose that the connection is possible if the governor belongs to class *A* and the dependent to class *B,* or if the governor belongs to class *C* and the dependent to class *D.* It follows that the governor belongs to one of two classes, *A* or *C,* and the dependent to class *B* or *D,* but the information provided by one occurrence is inadequate to make a definite assignment. Other occurrences can make the choice unique if the linguist assumes that the minimum number of assignments per form is desirable, or he can make the decision for each pair.

The possibility of altering grammar-code symbols during SSD raises further problems that must be recognized in the research procedure. The purpose of alteration, roughly speaking, is to prevent connections that are impossible in a certain context. First, a certain word may be restricted to a given class of governors when it is accompanied by one or more dependents of particular types; for example, the object of a preposition sometimes restricts the range of governors that the prepositional phrase can serve, and a genitive singular noun can serve as the subject of a plural verb only if it is accompanied by a cardinal number. Second, the various dependents of a single governor may impose restrictions on one another; most verbs can take a direct object in the genitive case only if they are modified by a negative particle, and a verb cannot take two direct objects. When an entry is added to the table of dependency types, as described

above, or a grammar-code symbol is changed in the glossary, the possibility of altering a symbol during SSD is not considered. A screening process can be used thereafter.

In deciding whether alteration of a grammar-code symbol is desirable, negative evidence is needed. The evidence is that a connection between two occurrences is allowed by the dependency table but not by posteditors. The false connection can be eliminated in several ways: by semantic procedures, by subclassification of grammatical categories, or by recognition of contextual restrictions. Only the last leads to alteration of grammar-code symbols. If two grammatical categories are connected by the dependency table, sometimes correctly and sometimes not, a test for contextual restriction should be performed on the pair. As indicated above, there are two cases.

The restriction can involve a chain of three connected occurrences. If the data show that an occurrence of type *A* governs one of type *B* only when the latter governs an occurrence of type *C,* then type *B* should be altered to type *B′* when type C is attached, and a dependency-table entry linking types *A* and *B′* should replace the *AB* entry. Type C can be a morphological category, a syntactic word class already established on the basis of other evidence, or a new category established ad hoc, provided that the existence of a class can be shown by the usual statistical evidence of lexical limitation—the number of different words in the class must be less than the number expected by chance.

The restriction can also involve a governor and two of its dependents. If the data show that an occurrence of type *A* governs one of type *B* only when it also governs one of type *C,* then *A* should be altered to *A′* when the first of the two dependents is added. Suppose that the *AB* connection is always earlier than the *AC* connection in recognition order; then *A* becomes *A′* when *B* is attached, and *A′C* replaces *AC* in the dependency table. Type C must satisfy the requirements stated in the preceding paragraph.

The programs described above lead to the establishment of many independent syntactic word classes. Economy demands that the number of distinct classes in the grammar be reduced as much as possible, and it has been suggested that a category is grammatical only if it appears in a number of different rules.[34] The methods and statistical problems of class comparison have been discussed in the section on semantics; the same methods can be applied to syntactic classes, and the statistical problems have to be solved.

One answer to the question of what distinguishes syntactic classes from semantic seems more acceptable than the others. Starting with the notions of morphological classification, function words, occurrence order, and punctuation, the research procedures that have been described here pro-

[34] Edward Klima, personal communication.

duce certain categories of words. All the word classes that can be defined by rules involving them *and* the initial syntactic indicators are taken as syntactic classes; any class that can be defined by rules involving it and a syntactic class is also a syntactic class. The rules are those that differentiate between structures acceptable to editors and structures that editors reject. It is an empirical question whether a program capable of determining a single acceptable structure for almost every sentence in a large corpus—and more structures than one for almost all of the remainder—can be based entirely on syntactic classes, morphological classes, function words, occurrence order, and punctuation. If the answer is affirmative, then semantics (by this analysis) is not required for sentence-structure determination; but if the answer is negative, semantics is required for the elimination of *syntactic* ambiguity.

The possibility of writing a grammar for a language by purely automatic methods, using unedited text as data and an analytic program based on linguistic universals, is currently being raised.[35] Although it is still too early to say what results can be obtained with such methods, an important theoretical difference between methods with and without informant editors should be noted at once and remembered as research progresses.

We have seen three levels, or strata,[36] in language: the level of the writing system, the level of the grammar, and the level of semantics. It is apparently characteristic of editor informants—of all users of language—that they deal with all its levels simultaneously and, for the most part, unconsciously. When an informant is asked whether two sound sequences are "same" or "different," he evidently answers according to the grammatical-level patterns that they indicate; as sounds, the sequences can be quite distinct, yet if they stand for the same string of inflected forms, they are "same" to the informant.[37] Two sentences with different composition at the grammatical level are "same" if they are semantically identical, that is, if they indicate the same semantic content; but consciousness reaches the grammatical level, and it is more difficult to apply the test. The point is that informants use their higher-level understanding of a sentence whenever they are asked to comment on it.

An automatic system for grammatical analysis is usually conceived as working its way upward from level to level. First morphological analysis is carried out in accordance with morphological criteria (and lower-level criteria as well; similarity of sound or spelling is used in deciding whether

[35] See references cited in fn. 8.

[36] Sydney M. Lamb, "The Strata of Linguistic Structure," presented at a meeting of the Linguistic Society of America, Hartford, Conn., December, 1960. (His strata are not identical with these.) Cf. also the three sets of levels in Garvin, "The Definitional Model of Language," earlier in this volume.

[37] See Chomsky, *op. cit.*, fn. 7.

two forms are forms of the same word). Next syntactic analysis is carried out, using morphological and syntactic criteria. Then semantic analysis, using semantic and syntactic criteria, is performed. How far the sequence of levels continues is still an open question, but the proposed automatic analysis programs pass from level to level in one direction only.

If informants and automatic analysis programs operate in exactly opposite directions, are they not certain to yield vastly different results? Perhaps not, for two reasons. A minor point is that informants use criteria at all levels simultaneously; they are not unidirectional. A major point is that language seems universally to have correlated structures on its various levels. The grammatical structure obtained by grammatical criteria corresponds closely with the grammatical structure obtained by semantic criteria. Were this untrue of any language, it would be unspeakably complicated, too complicated for the human organism to learn quickly and use fluently—and if it were learned nevertheless, it would in time be altered for the convenience of its users. Although formal tests of level-to-level structural similarity have never been conducted on a grand scale, the weight of years of linguistic research favors the hypothesis.

Similarity does not imply identity. The syntactically most elegant morphology of a language is not likely to be achieved by following morphological criteria exclusively. The ultimate program for automatic research in linguistics is therefore likely to go forward, then back: A fairly good morphological analysis, based on morphological criteria, paves the way for syntactic analysis; once completed, the syntactic analysis furnishes criteria for adjustment of the morphology. The syntax obtained by using syntactic criteria likewise furnishes the basis for semantic analysis, but the semantic structure, when known, permits refinement of the syntax.[38]

Linguistic methodology is being developed very rapidly; the sound work of recent decades is being tested and enriched by linguists concerned with computers. The criticism sometimes voiced,[39] that computational methods lead to ad hoc schemes unthinkingly propounded and not to understanding of the true structure of language, can be refuted if not silenced by attention to some general principles. First, the temptation to overgeneralize must be denied. The modest samples currently available for computational research permit no general statements about languages, and it may be some years before adequate samples can be obtained and analyzed. Second, the search for linguistic universals must continue. Those that are well supported by evidence and relevant to the research

[38] David G. Hays, "Linguistic Methodology and the Theory of Strata," presented at a meeting of the Philological Association of the Pacific Coast, Santa Barbara, Calif., November, 1961.

39 Dean S. Worth, "Linear Contexts, Linguistics, and Machine Translation," *Word,* vol. 15, no. 1, pp. 183-191, April, 1959. Noted with agreement and expanded by Mel'chuk, *op. cit.,* fn. 18.

now being conducted with computer aid are (1) natural languages can be closely approximated by simple formal models; (2) the appropriate models have recursive features; (3) the appropriate models are multilevel; (4) the appropriate models include simple postulates about occurrence order (at least with respect to separation); (5) the appropriate models include classification of recurrent units (e.g., word classes) ; (6) the classifications are multidimensional; and (7) simplicity and economy are significant criteria in classification as in the structural design of the model. Third, results obtained by various methods of research should not propose to refute results obtained by other methods until a more complete, integrated theory of linguistic research is written.