[From: *Linguistics* 8 (1964), 62-71.]

LINGUISTIC THEORIES UNDERLYING THE WORK OF VARIOUS MT GROUPS

MARY LU JOYNES AND W. P. LEHMANN

1.. Linguistic analysis is being carried on in various centers devoted to MT or computational linguistics. Since many of the scholars carrying on this work are not by profession linguists, and others have entered linguistics from such fields as mathematics or social sciences, the terminology used in these centers often differs from that met in standard linguistic publications. With the energy and resources available, these centers should produce findings of general interest to linguists. These findings are, however, often framed in special terminology which may make their contributions obscure. Also, literature in this new field is seldom readily obtainable in linguistic journals or in scholarly publications widely available to linguists. There is usually a great time-lag between research and publication, frequently amounting to years. Such publication, when it does appear, is often in the form of privately distributed work-papers, reports to sponsoring organizations, or various types of intragroup ephemeral communications. However, most of the general theories on which these groups base their work have changed relatively little from their original designs, and in most cases can be brought up to date by direct communication with the research group in question.[1] This survey is designed to outline some of the theories held and to relate them to the extent possible to general linguistic theory.

[1] In addition to the specific articles cited, see especially *Bibliografija zarubežnyx robot po mašinnomu perevodu* (1960-1961 gg.), edited by I. A. Mel'čuk and G. S. Cvejg, or its translation in *Foreign Developments in Machine Translation and Information Processing,* No. 122 (available from the U.S. Department of Commerce Office of Technical Services, Joint Publications Research Service, Ohio Drive and Independence Avenue, S.W., Washington 25, D.C.). The Mel'čuk-Cvejg study is a listing of abstracts of work done in various parts of the world to that date. The *Foreign Developments in Machine Translation and Information Processing* series is a continuing set of translations of significant work being done by non-English speaking groups in the field.

The journal *Machine Translation,* edited by Victor Yngve of the MIT group in Cambridge, Massachusetts, appears from time to time with abstracts and original articles. A book edited by Paul L. Garvin, *Natural Language and the Computer,* is of great interest to the linguist. For a periodical handbook of the work and publications

Current work in machine translation has essentially two different points of origin: the disciplines of logic, especially the type of symbolic logic developed in this century by mathematicians and philosophers, and linguistic theory, which has also been developing in this century as the result of the experience of workers in many different areas of human behavior. In many respects the two disciplines coincide, but in others they diverge widely. Perhaps the greatest difference, and the one which has least often been discussed openly in the literature of machine translation, is the specification of what actually constitutes the language to be analyzed. If the language in question is an artificial one, such as a mathematical or chemical formula or a symbolically-stated logical proposition, the problem of analysis into components and the conventions of arrangement of components can be solved with very little difficulty. On the other hand, if the text under consideration is "natural language", especially if it is carefully transcribed actual speech, not edited or normalized at all, the problem will be a very different one. In actual practice, both of these extremes have been avoided so far by the expedient of centering work on bodies of scientific text, which are generally written so as to eliminate as far as possible some of the ambiguities and gaps in spoken language. This editing, of course, removes many of the signalling devices from the spoken language, but it also allows a reader to re-scan a line or a page as a substitute for the question he might be able to ask a live speaking-partner. Thus, to an extent, a piece of technical prose is an artificial language, and to the extent that it has been regularized in this manner it becomes more amenable to the logician's formulations and presumably needs less of the linguist's efforts in analysis.

2. Of the groups now working in machine translation, those headed by Oettinger[2] of Harvard and Rhodes[3] of the National Bureau of Standards are most obviously of the logic-oriented type. Although they never really

of various projects, see the National Science Foundation's *Current Research and Development in Scientific Documentation* (Office of Technical Services, U.S. Department of Commerce, Washington 25, D.C.). As a supplement, the National Science Foundation publishes a bi-monthly newsletter, *Scientific Information Notes,* which keeps *Current Research* up-to-date, publishes notices of new projects, and provides general current information about developments of interest in the field.

[2] Oettinger, Anthony G., "Automatic Syntactic Analysis and the Pushdown Store", *Proceedings of the Twelfth Symposium in Applied Mathematics* (Providence, R.I., American Mathematical Society, 1961), pp. 104-129; Oettinger, Anthony G., et al., *Mathematical Linguistics and Automatic Translation* (= *NSF-7)* (Cambridge, Mass., The Computational Laboratory of Harvard University, 1961).

[3] Rhodes, Ida, "A New Approach to the Mechanical Syntactic Analysis of Russian", *Mechanical Translation* (1961), 6:33-50.

specify their frames of reference or basic units as most linguists would do, a survey of their working procedures indicates that their primary units are glossary entries, either roots, prefixes, suffixes, whole "words", or occasionally longer units. The glossary listings of these units, in turn, carry morphological and syntactic information which sets up a range of predictions as to what other types of units may be expected to follow in left-to-right sequence through the string. This "foresight pool" or "prediction pool" usually supplies more than one prediction; from these alternatives one is chosen as a working hypothesis and the others stored in a "hindsight pool", from which corrections may be made as needed on the basis of subsequent information gained from items following in the string. After the source or "input" language-string has been "analyzed" in this procedure, it may be matched to corresponding structures in the target or "output" language. Actually, since both the NBS and Harvard systems are on a language-to-language correspondence basis, much of this matching is done during the analysis procedure itself.

Rhodes has listed some of the defects and problems in this approach.[4] Reduced to a few small categories in linguistic terms, they are chiefly problems that arise when the text becomes less formalized (or less an "artificial language") than the skeleton grammar and the word-by-word glossary can cover. Two of these difficulties, polysemy and ambiguity, can be expected to increase as the artificiality and formalization of the text decrease. At this time, the NBS expects little of the translation scheme other than a transfer of information in unambiguous form from the source language to the target language, accepting such compromises as the regularization of certain morphological features *(foots* for *feet),* providing alternative lexical choices for a given item, and using a stylized syntax.

Operating within roughly the same frame of reference for determining relationships and segmentation as the Rhodes-Oettinger systems are the "dependency" systems of Hays of RAND[5] and to a lesser extent of

[4] *Ibid.,* p. 33.

[5] In reply to a question about the relationship of his work to that of NBS, Hays is quoted as saying: "It appears to me that each of her predictions is equivalent to a potential dependency connection, but that a prediction may be a prediction of a governor or a prediction of a dependent. ... I think that Mrs. Rhodes' syntactic theory is very close to ours and thus different to some degree from the more frequent phrase-structure conception of grammar, and that her algorithm is considerably more like the one that we have in operation than like any other in the field". This quotation, along with many longer papers and discussions, appears in *Proceedings of the National Symposium on Machine Translation,* H. P. Edmundson, ed. (Prentice-Hall, Inc.: Engelwood Cliffs, N.J., 1961). The specific quotation is from page 46.

Pacak of Georgetown. The Pacak[6] usage of dependency relation statements is directed chiefly at the problems of segmentation and analysis working from a language (Russian) which signals by means of inflectional endings, unambiguously in most instances, the dependency relationships involved. Pacak has developed probabilistic rules derived from word-order for resolving the situations of ambiguous signalling. In the term "dependency" both Pacak and Hays include concepts familiar to linguists for many years. Dependency includes the relationships which were traditionally called agreement and government. These dependency relationships, stated in the form of "trees", represent a description which must follow a previous analysis to determine the dependent-governor or lower-node-higher-node relationship. In Russian, as Pacak has demonstrated, a typical relationship is subject-verb-object, with the dependency relationship usually indicated by inflectional endings and to a lesser extent by inflectional markers. Hays seems more concerned than Pacak with the theoretical problems and implications of two types of analysis (or more accurately, two types of analytic description):

> Immediate-constituent analysis and dependency analysis (two theories of syntactic description) are based, respectively, on the topologies of grouping and of trees. A correspondence between structures of the two types is defined, and the two topologies are compared, mainly in terms of their empirical applications.[7]

Hays states very clearly one of the chief differences between his dependency-tree description and the form of immediate constituent analysis which leads to phrase-structure descriptions:

> Let us conclude with some examples from natural language. In a sentence such as *Children love candy,* whose form is N-V-N, immediate-constituent analysis groups verb and following noun into a verb phrase, rendering the sentence as N-VP. Dependency analysis makes the two nouns dependents of the verb. A passive transform, *Candy is loved by children,* with the form N-is Ven-by N, would be grouped into N-VP. Note that the groupings reflect grammatic properties clearly enough, but disregard meaning; *candy* goes into VP one time but not the other. Constancy of meaning behind the grammatic transformation is reflected more clearly, as we believe, by two trees, in which *children* and *candy* are dependents of *love* in both active and passive forms of the sentence.
>
> Again, consider the naming of phrases. An adjective plus a noun form a noun phrase, and an adverb plus an adjective form an adjective phrase. The

[6] Pacak, Milos, "Loci of Agreement and Government Structures in Slavic", *Estratto Revista Methodos* (1962), 14:75-86.

[7] Hays, David G., *Grouping and Dependency Theories,* Mathematics Division. The RAND Corporation, 16 February 1960, PR 1910.

naming singles out an element of each phrase, as does the topology of a tree. Grouping - e.g., ((A)(N)) - does not.[8]

But while many of the problems he cites in "translating" one type of description into the other are real, many are superficial and may even conceal the more serious problems of linguistic analysis which must precede the logical statement or presentation of the finished analysis. Perhaps such descriptive presentations for the purposes of machine translation will be an excellent test for the traditional linguist's assumption that each language must be described in terms of its own structure rather than in terms of a pre-designed format.

3. It is in making these analytic judgements that the linguist oriented to natural language will be most cautious. The linguist who habitually deals with natural languages as they are actually used regards these descriptive frame-works and techniques as tools which will vary in useful application from situation to situation. He realizes that any natural language will be describable in a great number of ways, but that for a given purpose, in a given frame of reference, there will probably be one form of description which will cover that body of data for his particular purpose in a more convenient way than another description might.[9] On the other hand, in a different frame of reference the description might not cover all of the data or might even give a false or unworkable result. In this context the question of preference of description, in this instance Hays' contrast of dependency series against immediate constituent analysis, becomes relevant. The real question of applicability, however, is ignored or by-passed in favor of the format or presentation. Presumably the purpose of either type of analysis is to segment the string (composed of as yet unspecified graphemes, morphemes, "words", etc.) into some sort of arrangement which will indicate the relationships among the component elements and the relation of these arrangements of entities to other similar arrangements already observed in the language. The question of the frame of reference becomes extremely important at this point in the analysis. In most logic-based analyses the string to be analyzed is the self-contained unit called the sentence. In the frame of reference usually employed by the linguist it may be the sentence or less,

[8] *Ibid, p.* 11.

[9] This principle was made explicit for phonemics and linguistics generally by Yuen-Ren Chao in "The Non-uniqueness of phonemic solutions of phonetic systems", *Bulletin of the Institute of History and Philology Academia Sinica,* Vol. IV, Part 4 (1934), pp. 363-397. (Reprinted in Martin Joos, ed., *Readings in Linguistics,* Washing ton, 1957, pp. 38-54).

but usually is considerably more, depending upon the type of entities and relationships the linguist is trying to isolate and describe at that particular time. The linguist may reshape his tools as he works; the logician generally has his tools completely operable before he begins work on his string. Put in other terms, the linguist may vary his frame of reference in making his decisions by shifting only one entity or factor in his string at a time, any number of times, while the "machine-logic" of most of the systems now operable or proposed has a limited stock of alternative analyses already built in. If this built-in stock of alternatives does not fit the sentence or sentence part, or worse if it locates only one possible analysis of an ambiguous string, there is no real check upon the procedure involved or any resolution possible from information beyond the string called "sentence". For this reason of convenience of alternate analyses, among others, linguists have found the technique of immediate constituent analysis, and the procedures underlying it, a very useful analytic tool. When grammars of sufficient sophistication and completeness are developed for machine translation to make the entire process workable, such grammars may be statable with great economy in terms of dependency relations. Then the question will merely be one of presenting rather than gathering the data. At this time, however, both logicians and linguists are or should be concerned with gathering the data.

4. In gathering and ordering natural language data it has long been assumed that language is stratificational in nature. There have been many names and descriptions attached to linguistic strata, but the terminology most widely used in the United States today is that of Bloomfield[10] or some modification of it. Unfortunately, many of these modifications of the Bloomfield theory and terminology have not been done literally as more detailed knowledge of the working of natural language has been acquired, but rather new interpretations have continued using the old terminology in new senses. Naturally some confusion and a proliferation of short-lived new terms have arisen in the process along with some clarification of the older terms and theory involved. In the tradition of Bloomfieldian stratificational grammar, Hockett[11] has called attention to one of the defects in Bloomfield's terminology and has given a history of attempts which have been made to solve problems

[10] Bloomfield, Leonard, "A Set of postulates for the science of language", *Language (1926*), 2:153-164. (Reprinted in *Readings in Linguistics, pp.* 26-31.)
[11] Hockett, Charles R, "Linguistic elements and their relations", *Language (1961),* 37:29-53.

arising from the terminological and theoretical point involved. The question arises from a literal interpretation of Bloomfield's use of the phrase "made up of" in Assumptions 2 and 6. Hockett has demonstrated in a logical format that to say that a morpheme is composed of phonemes would be a contradiction with regard to certain utterances (ex. English *knife/knives)* and has posited a morphophonemic relationship (his *P* relation) as a bridge between the two strata without specifying the one as composing the other.[12] The really significant point of Hockett's argument is mentioned only incidentally to his specific arguments. In regard to one of the devices he had proposed to show the relation between morphology and phonology he remarked:

> Morphophonemes, morphs, phones, and acoustic phones are ARTIFACTS OF ANALYSIS or CONVENIENCES FOR DESCRIPTION, not elements in a language. Likewise, the relation *R,* the resort to which is responsible for such pseudo-elements, is a mistake. It has no status in a language, but is evoked by our desire to make cross-stratum correlations neat[13]

At the end of the paper, and almost as an afterthought to it, Hockett asked one of the most interesting sets of questions openly asked in modern linguistic writing:

> In closing this paper, I must for the sake of honesty mention a suspicion that cannot be followed through in detail here, but that, if verified, is due to undermine the logic of most of our accomplishments in descriptive linguistics since Saussure, Sapir, and Bloomfield, or even an earlier period. The only tenet that might survive the holocaust is the duality hypothesis.
>
> Most of our descriptive linguistic thinking in the past few decades has been based on an unstated assumption: that any utterance in a language, occurring in a specific context involving specific speaker and hearers, has *in that context* a determinate grammatical structure, involving an integral number of grammatical elements in specifiable structural relations with one another. We have quarreled extensively over the exact nature of the elements and the relations, about heuristic criteria, and about their status as objective parts of the universe or as convenient fictions. But the underlying assumption has scarcely been challenged.
>
> There are, in fact, certain types of utterances that should raise serious doubts about the assumption. One of them is *Don't shell so loud!,* something I once said, in angry irritation, to my noisy children. It is clear that an attempt to deal with this sentence without information about the context would yield erroneous conclusions: the item *shell* was not, or was not exclusively, the morpheme that is customarily programmed into that string of phonemes. But if it was not that morpheme, what was it grammatically? We can call this

[12] This form of redefinition was actually begun by Hockett in 1942 in "A System of descriptive phonology", *Language,* 18:3-21. (Reprinted in *Readings in Linguistics,* pp. 97-107.)

[13] *Ibid.,* p. 42.

shell a 'blend' of *shout* and *yell;* but no existing system of grammatical analysis or theory makes provision for the building of a grammatical form by 'blending'. Such utterances are not rare, but extremely common. They occur not only as 'slips of the tongue' (whatever that means), but as planned puns, double entendres, plays on words, and variously in poetry and advertising. We can do three things about them: (1) Ignore them (perhaps as 'ungrammatical'). (2) Regard them as deviations from 'normal' sentences, to be explained with special machinery glued onto our basic theory for 'normal' sentences. (3) Use them as evidence for some new and very different theory of the generation of speech, that will provide at once for such 'deviant' utterances and for all 'regular' utterances. If we are really concerned, in linguistics, with the discovery and description of the place of language in the universe, I believe we most seriously consider the third alternative, no matter how radical may be the revisions that are required in our ways of thinking.[14]

This question of the actual operation of a natural language has been studiously avoided by most linguists, at least in writing. Some speculation has been done privately, of course, but linguists carefully avoid these discussions in print. Such questions as Hockett has asked and further inquiry into machine-language may show that simulation of a natural language is not the same as the actual process by which human beings produce and receive natural language stimuli. Results of these experiments may, in effect, remind linguists that much of the descriptive terminology and apparatus of current linguistic theory is a useful fiction.[15] An approach to machine translation which is closely associated with work on linguistic theory is that of Lamb,[16] who has concerned himself largely with statements of the relations between units in a stratificational grammar. As Bloomfield also specified, Lamb's lower limit of the field of linguistics is sound as sound, and his upper limit is meaning. Everything between is thought of as statable in stratificational terms. His middle layer, and presumably also his starting point in analysis, is, like that of Bloomfield, the morpheme. With the exception of these starting points and boundaries, however, Lamb's system and terminology are much more elaborate and less flexible than those of Bloomfield. The difference, which seems to center around the question of the relation of morpheme to allomorph to morphophoneme, appears to be one of degree of abstraction and relation of degrees of abstraction to the units from

[14] *Ibid.,* pp. 52-53.

[15] W. Freeman Twaddell, in *On Defining the Phoneme (= Language,* Monograph 16, 1935), had a section called "The Phoneme as a fiction", which might easily apply to morphology, syntax, and semology in current linguistic contexts. (Reprinted in part in *Readings in Linguistics,* pp. 55-79.)

[16] Lamb, Sidney, *Outline of Stratificational Grammar* (U. of California, Berkeley 4, California, 1962).

which the abstractions are derived. In order to account for the wide variety of morphophonemic shapes of allomorphs of morphemes, Lamb uses the neutralization and diversification theories which characterized the Prague School's work. He also has used a subsystem with the prefix *mio-* (as in *miosememe, miomorphophoneme,* etc.) which he explains rather obscurely:

> Except where X is *morph,* a component of an Xeme is a *mioXeme.* (Morphemes do not have components). A few additional terms, (some of which, however, lack applicability) are defined by substituting *mioX* for X in the above definitions, except that of *mioX* itself. MioXemes are the minimal units on their stratum. The morpheme is the minimal unit of the morphemic stratum.[17]

Lamb's *Outline* is unfortunately very difficult to evaluate fairly, since it was designed to serve both as a report on his project and as a textbook for a course. For either purpose it would probably be supplemented by materials which are not available to the casual reader. In later writing[18] he has abandoned or modified some of his elaborate terminology and simplified the relationships stated between his strata to a certain extent. Other projects have concentrated on collecting data which may be of great interest to linguistic study, but which have little new contribution for linguistic theory. The group at Wayne State University headed by Josselson has concentrated on analysis of Russian, producing some very interesting statistical data. W. S.-Y. Wang's project at Ohio State University is working on Chinese, chiefly from the point of view of current linguistic theory. One of the important by-products of Wang's investigation is his making available, although somewhat indirectly, linguistic work being done in China on Chinese. Little of this material has been available to readers in the Western world for some time. Moreover, individual studies have been produced which are of great interest for special problems in specific languages, and in this way contribute to linguistic theory.[19] These studies will eventually be reported in standard media, and in this way become accessible to linguists in much the same way as do other results of research. They have accordingly not been listed here. Nor has the research of the Linguistic Research Center at the University of Texas been discussed. Like most of the groups elsewhere. LRC makes available reports on request, and answers questions when con-

[17] *Ibid.,* p. 15.

[18] Lamb, Sidney, *The Sememic Approach to Structural Semantics* (Machine Translation Project, University of California, Berkeley, California, April 1963).

[19] Summary progress reports on these groups and partial listings of their publications are available in *Current Research and Development in Scientific Documentation* cited above.

tacted directly. In general, we may only mention that the theory at the University of Texas is closely related to widely established linguistic theories.

5. The survey of the linguistic theories given above has disregarded all superstructures imposed on analyses taken from areas such as symbolic logic or mathematics. This decision was made on the basis of past linguistic findings: Linguistic analysis is in no way aided by an adoption of procedures from other areas. On the other hand, after a language has been analyzed, after a linguistic analysis procedure has been adopted, for computer processing it may be useful to recast the form of the linguistic descriptions on those developed in logic.[20] But this modification is outside the scope of linguistics, outside the scope of language analysis. It is not unimportant for the presentation and manipulation of the data, in somewhat the same way that the design of a book may assist in displaying more usefully the results of a linguist's analysis. But discussion of book design in a grammar has little benefit; similarly discussion of logical procedures would contribute little to understanding of language, and is therefore completely omitted here.

This survey has also presented linguistic theory with no critique. To the present the only critique on the adequacy of a linguistic procedure has been its success in managing language data. One of the great promises of MT research is the availability of a programming analogue to language, a model, which will either permit the linguist to display his results or will founder. In view of either possibility, any critique of linguistic theory would be premature, and pointless.

Linguistic Research Center ,
The University of Texas
Austin, Texas

[20] A very interesting example of the result of recasting linguistic data into other formats is to be found in the work of Victor H. Yngve, "A Model and an Hypothesis for Language Structure", *Proceedings of the American Philosophical Society* (1960), 104:444-466. While his initial concern was with the temporary storage capacity of the machine, he concluded that the human memory capacity was similarly restricted in the number of depth-units it could handle in specific sequences. In developing his model, Yngve represented in computer form many observations familiar to language teachers from their practical work. Probably the most dramatically illustrated point is the idea that adding material at the end of a string is "easier" for the computer or the brain than adding it earlier in the string. The question which Yngve's presentation will immediately suggest to linguists, is, of course, the nature and size of the units involved. Another interesting facet of Yngve's syntactic analysis is the possibility, as yet unrealized in machine linguistics, of the relation of suprasegmental signals to segmental grammar in languages such as English and German.