[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University, 1957]

Some Ideas On Inter-structural Syntax

Jane A. Pyne
Institute of Languages and Linguistics

Many theories and applications of syntax are discussed in the literature of machine translation. Various linguists' utilization of syntactic analysis in MT research represents diverse points of view, indicating far-reaching interest in syntax since the latter is perhaps the most essential tool for reducing the translation process to mechanical procedures. In his paper on the Georgetown-IBM Experiment included in the Locke and Booth collection,[1] L. E. Dostert suggested the possibility of developing a coded general or core syntax, common to several languages, into which the syntax of these natural languages would be programmed for purposes of multi-lingual translation by machine. This concept has been further revised and developed by him during the current research project at Georgetown University.

It should be noted in this connection that the question of multilingual core syntax has particular relevance to a future stage in the research, that is, when we hopefully will have come to a consideration of machine translation of more than two languages. At the present time our focus of study in the Georgetown project is one-directional and bilingual only. Consequently this paper deals with certain problems of structural transfer from Russian to English, and is based on the research in syntactic analysis being carried out by M. Zarechnak and myself.

We are attempting to develop a mechanical procedure for effecting structural transfer from Russian chemical discourse to its English translation. The sensing of functional units is considered essential for the machine to be able to transfer meaning adequately, in that translation is defined as the transfer of meaning from the linguistic pattern of one language to that of the other. Functional units, in turn, are defined as structural or syntactic units of language, as opposed to the linear commutativeness of individual words or morphemes. I will refer to these functional units later in connection with a definition of translation units at various levels, from the lexical to phrase level. In an article entitled "Transfer Grammar",

[1] Locke and Booth, *Machine Translation of Languages,* The Technology Press, M.I.T., John Wiley and Sons, N.Y., Chapman and Hall, Ltd., London, 1955.

Z. S. Harris states, and I quote: "translating the morphemes ("'word-by-word") is in any case not enough for translation, since the grammatical interrelation of the morphemes in each language is a matter of the subdivision of the sentence into constituents (in successive inclusion) which will often differ in the two languages; and the order of the morphemes within each constituent will often differ. The analysis of a sentence into successively included constituents, and the composition and order of smaller constituents (down to morpheme classes) is therefore necessary for any method of translation that is to be reducible to mechanical procedures."[2]

Although any method of translation, whether human or mechanical, requires the substitution of the morphemes of one language for those of the other, the nature of linguistic structure precludes *linear* substitution. The morphemes or words of English cannot be linearly substituted for the morphemes or words of Russian because the grammatical interrelationships are not identical. Furthermore,, a given Russian morpheme or word may have more than one possible equivalent in English, and thus be translationally ambiguous either lexically or grammatically or both. The problem of linguistic analysis for MT consists in separating from the start those translations which are *free variants* of transfer from those which are *positional variants*. Positional variants are those where the choice is determinable from some lexical or grammatical item in the determining environment. These can be called positional variants of transfer, and constitute the data of MT, as opposed to the unimportant choice between two translations which are free variants for the transfer. Free variants refer to those whose selection in any environment is a matter of style or individual preference.

Because not all choices among given positional variants are cued by the presence of some item within a predictable and definable distance from the ambiguous item, and because a linear method of translation does not solve problems of rearrangement—such as when the Russian verb precedes the noun, or may even be zeroed, and English structure requires different order—for these reasons, it is necessary to view the transfer operation in terms of a machine-programmable *analysis and transfer* of successively included constituents within the sentence. The goal of our research is to prove

[2] Harris, Z. S., "Transfer Grammar", International Journal of American Linguistics, Vol. XX, No. 4, October, 1954, pp. 259-270.

that the sentence can be handled by the machine in terms of its constituents in successive inclusion, so that the composition and order of smaller constituents (down to morpheme classes) can be adequately translated.

The Russian sentence can be defined, as previously discussed in this morning's session, as a bicomponential function, where H is the independent variable in the nominative case and P is the dependent variable, a verbal form or its substitutes. Once the H and P as nuclei are handled, it is relatively straightforward to translate the elements surrounding the nuclei since the majority of these elements are in direct dependence relation to the nuclei. All Russian sentence types are expressible in terms of H, P and three features of grammatical relation: agreement, government and apposition.

Since we are attempting to achieve mechanical translation from Russian to English, not the reverse, any approach to English syntax is in terms of its minimum difference and maximum similarity to Russian. This difference, determined through comparison of Russian structural types and their English translations, can be defined as the number and content of grammatical instructions needed to generate the English sentences out of the Russian. To refer again to Harris' conceptual framework of transfer grammar, we use the criterion of translation as equivalence, and postulate a transfer relation between each sentence of Russian and its English translation, and then construct transfers between paired items within the sentence. Detailed examples of such pairs are given in a separate Georgetown Work Paper on MT. (MT-38)

The English synthesis part of the research is based on the syntactic theory of the construction, transformation and kernel, in that all sentence structures are combinations and/or transformations of just a few simple sentence structures, the kernels of the grammar. We are particularly concerned with non-linear transfer, that is, cases in which a given Russian unit cannot be translated directly or component-by-component into English. The problem is to establish and describe the regularities in the transforming operations needed to obtain the proper transfer for all Russian structural units.

The sheet which you were given in connection with Mr. Zarechnak's paper shows the basic comparison of Russian and English kernels. Languages are in general much more similar to each other in their kernel sentences than in their final resultant sentences

(that is, after transformations).[3] The factorization of Russian chemical discourse and its English translation into kernels and transformations has enabled us to establish the regularities of insertion and rearrangement operations necessary for English in contrast to the Russian original.

It is interesting to note that the kernel analysis of a given Russian sentence is remarkably similar to that of its English translation. Where the languages differ is largely in respect to the transformations employed, just as this factor causes vast stylistic variation within one language itself.

The approach in question here, however, is not the mechanical factorization and translation of kernels (although this has wide interest for information retrieval procedures) but rather the transfer of Russian structure in its original transformed (or put-together) state into the corresponding English translation.

Machine operations depend on the sensing or delimitation of translation units in the source language and their transfer into the target. Translation units are describable at three levels: lexical, syntagmatic, and syntactic. On the lexical level a translation unit may be a single morpheme or several words. On the syntagmatic (or subsentence) level a translation unit may be a suffix function or a prepositional phrase and the like. On the syntactic level there are only two basic translation units, the noun-phrase and the verb-phrase.

Any translation unit may be subject to selection and/or arrangement, which in the transfer procedure means the following three operations may be involved in translating any translation unit: 1) choice between positional variants 2) insertion and 3) rearrangement. All translation units must be delimited by the machine, which, in view of the fact that they overlap—because they are successively included—presents considerable but not insurmountable problems. The lexical and syntagmatic translation units may consist of one unit sensed (a word between spaces) or part of one or more than one. The search area for delimiting these units rarely constitutes the complete stretch of input, the sentence. Conversely, the syntactic translation unit is delimited by examining a search area including the entire input, the complete sentence including punctuation.

[3] See Z. S. Harris, "Transformations Manual", mimeographed booklet, not yet published. (University of Pennsylvania)

We have concentrated to date mostly on the problem of the sensing delimitation and transfer of syntactic translation units, and have arrived at a procedure for machine transfer (at least on paper) of the basic structure or kernel of any Russian sentence into English.

The procedure followed in the research was to reduce to symbolic formulas a very large sample of Russian chemical discourse and its English translation. Compression techniques were applied to each sentence, that is, a method was devised to express any sentence type in terms of its particular transfer features. Thus a given HP relation in Russian is transferred into English structure according to a specific operation. Although the scope of this paper does not permit a detailed explanation of the techniques of Russian analysis and English synthesis, a brief summary of the transfer syntax procedure would seem to be pertinent at this point. The procedure involves a cut of each minimal Russian sentence into two parts, verbal and nominal. These are in turn handled first in terms of their head words, the H and P respectively, and then the rest of the components fall into a string-type operation. Altogether there are three basic transfer instruction operations: H, P and S. H operation extracts the head of the noun-phrase, and P operation the head of the verb-phrase. Their relative morphological composition and order determines the particular transfer instruction for English structure. The S operation (meaning string, or chain) completes the transfer and is directly dependent on the first two. Locating the H and P is necessary for the delimitation of the unit boundaries.

We have formulated search sequences for extracting the H and P which are the structural nuclei, and the next crucial step is to discover the quantitative relation of H and P, namely, to discover how many H's and P's are in the stretch. The result of this tally determines the particular type of operation to be employed for cutting the stretch into its components and for performing the transfer instructions. The standard, or nominal sentence in Russian has one noun-phrase and one verb-phrase, each having a head word whose relative position and morphology determine the transfer instruction for a given structural type. This we label a 1-1 type. There may be no H and/or no P, and these cases take a particular transfer instruction procedure. Where there is more than one H and P, (called a 2-2 type) the search types do *not* (and note that this is a departure from the usual classification) follow the sentence groupings of simple, complex, dependent, etc. Instead, as a result of the gathering of a large corpus and of extensive testing, we found that the crucial

feature separating one search type from another was simply the relative number of H's and P's. Only after a 2-2 type is discovered, for example, is it necessary to separate conjoined from complex structure.

The complete series of questions for the component distribution search (the title we give to the process of finding out how many H's and P's are found in the stretch) and all the series of operations for the various search types and transfer instructions, will be presented in a forthcoming Seminar Work Paper.

In conclusion, however, I will touch briefly on the types of questions which are "asked" in sequence to discover the number of head words, namely nouns in the nominative case and verbal forms or their substitutes.

Punctuation helps to mark structural breaks. Separators, such as ; : , function as positional signals and enable one to make cuts within which a certain search should be initiated. For example, the search stops immediately at a semi-colon. Delimiters mark off inserted structures. I refer to parenthesis, brackets, quotations and the like. These have no positional function in terms of extracting the H and P components, and whatever is found to be an inserted structure is functionally another and separate search area.

Nouns, pronouns, and numerals in the nominative case function as H. They can be pulled out simultaneously, whereas verbal forms or their substitutes must be looked for in sequential order, namely: verbs with person markers first, then short participles or short adjectives, then full adjectives or adverbs which have predicative function.

Next comes the count of H's and P's, and then the corresponding operation, either 0-1, 0-0, 1-1, 1-2, 2-2 and so forth. These differ markedly in the search sequences applied to make the cut between the noun-phrase and verb-phrase, as well as in the instructions for transfer. The first problem is to discover the structural type, as in 0-1,where H is lacking, which must be separated for transfer according to whether it translates as NV or simply as V. 1-0, on the other hand, where P is lacking is not subject to any modification beyond the syntagmatic level. 1-2 may be either NV plus NV (conjoined sentences) or NV plus V. 1-1 is perhaps the most basic transfer operation, representing the minimal sentence in Russian. It is basic because many more-than-one types are reducible to 1-1, whether

they turn out to be conjoined or complex structures. After the structural type is determined, the instructions for rearrangement and insertion go into operation.

Again let me apologize for the rather summary character of these remarks by referring to our forthcoming work paper on the transfer syntax procedure, which will clarify and amplify these statements.