[*Automatic Translation of Languages*: NATO Summer School…July 1962]

SYNTACTIC INTEGRATION CARRIED OUT MECHANICALLY*

IDA RHODES

National Bureau of Standards, Washington 25, D.C., U.S.A.

1. THE UNATTAINABILITY OF A PERFECT TRANSLATION

IT is easy to prove that a perfect translation from one language into another cannot be achieved. Faced with a set of occurrences in the source language, the translator must first use them as clues, to ascertain the *purport* which they were intended to express. From the very start of his task, the translator is beset by a huge number of obstacles, some of which we list below.

A. Morphological Ambiguities

For example, the English word 'book' has no less than seventeen grammatical interpretations in the source—namely,

1. Noun Singular.
2-3. Adjective, Singular or Plural.
4. Verb Infinitive.
5-6. Verb Imperative, Singular or Plural.
7-11. Verb Present Indicative, 1st or 2nd Person Singular or All Persons Plural.
12-17. Verb Subjunctive, All Persons and Numbers.

B. Syntactic Ambiguities

For each grammatical interpretation, there exist, in general, a considerable number of possible functions which it can have in a sentence. For instance, an English noun can be, among other things, the

1. Subject of a Clause.
2-3. Direct or Indirect Object of a Verb.
4. Complement of a Preposition or Adjective.
5. Appositive of an Expression.

C. Semantic Ambiguities

These are so formidable, as to be responsible for much of the misunderstanding, distrust, and suspicion which exist even among people speaking the same language. A partial list follows.

1. Polysemia.
2. Misconception.
3. Obfuscation.
4. Lapses in Grammar or Spelling.
5. Localisms, such as patois, argots, slang.

Even if the translator were able to carry out this part of his assignment with a high degree of success, he is bound to flounder in the next phase, while attempting to convey the original intention by means of the target language.

* Presented at the NATO Advanced Study Institute on Automatic Translation of Languages, Venice, 15-31 July, 1962.

When *two* languages are involved, not only is the number of difficulties, encountered within the source language, multiplied by similar vexations, found in the target language, but a great many new obstacles arise to obstruct the path toward a correct translation.

The lack of correspondence between two languages is shown in:

1. *The Difference Between Morphological Features.* For instance, some languages do not incorporate cases in their grammar, others have no articles, still others lack certain tenses in their verbs, or have no verbs at all.

2. *Inflection.* A target language, which is more highly inflected than is the source, causes an increase in morphological ambiguities. This is witnessed by the fact that the English word 'book', which has seventeen ambiguities in the source, possesses no less than fifty-four morphological interpretations when rendered into Russian. Moreover, highly inflected languages are quite compact, i.e. economical of words. When these are rendered into a less inflected language, many additional target words must be inserted into the translation.

3. *Order of Occurrence.* This is so vital a feature of each individual language, that a translation unaccompanied by suitable rearrangements of the targets is barely comprehensible.

4. Idioms, puns, distortions for the sake of humor or satire are too well known to need more than mere mention, but we must bear in mind that the translator is also expected to be familiar with the native lores of the ethnic groups using the languages under consideration.

2. PREDICTIVE ANALYSIS

The above facts are, we believe, sufficient to indicate that a faithful translation cannot be achieved even when carried on by the most knowledgeable and competent human being. The situation becomes even more desperate when we try to enlist the aid of the present-day electronic processors for this task. These machines are magnificent tools when called upon to simulate a formal, systematic mental process, of whose nature the programmer is completely aware. In this category belong computational schemes connected with the technical sciences, as well as actuarial functions encountered in management tasks.

Language, on the other hand, far from being based upon any formal, universally agreed-upon axioms, is actually a notoriously lawless, arbitrary, capricious, wayward child of the human mind. One has to delve very patiently and deeply into one's own brain to discover even a single mechanizable process underlying human speech. Fortunately, we have been able to find one such process, which enabled us to program a routine for the mechanical *syntactic* integration (i.e. parsing) of a sentence. Correct parsing involves the resolution of all morphological and syntactic ambiguities, and only rarely that of semantic ambiguities. Our scheme has proved to be fairly successful as regards the first two aspects, but in the region of semantics our achievements are very slight.

Our method has become known as ‘predictive analysis’ and is based upon the universal habit on the part of the listener to *anticipate* the type of word which a speaker is about to utter. The thinking speed of the former is so much greater than the speaker’s rate of enunciation, that an incredible number of processes are carried on within his mind while he is listening. These activities bear at least three aspects; foresight, hindsight, and association of ideas. We shall illustrate each by an example in the English language.

Let our first sentence start with the word ‘He’. The listener expects at least one other word to follow, because a clause is not complete without a predicate. But he may not be consciously

aware of the fact, that there are only a few types of occurrences which may follow the word ‘He’. These consist of either (1) a verb in the third person singular of any tense in the indicative mood, or of (2) a member of the set of ‘unpredictable occurrences’, namely, the adverbs, prepositions, conjunctions, and certain punctuation marks. The last-named set is of little interest in the resolution of ambiguities, since its members are not inflected in any of the languages with which we are familiar, and thus present no difficulties. Thus, if the word ‘He’ is followed by ‘can’, the mind will automatically reject the possibility of the latter being considered as either a noun or an adjective (as in ‘can-opener’) and accept it only as a verb in the present indicative, agreeing with the subject.

The pair of words ‘He can’ continues to keep the listener in a state of anticipation, because ‘can’ falls into the category of *modular* terms that call for a verb in the infinitive. Should we now add the word ‘book’ to the previous two occurrences, it becomes clear that this word would now have only one grammatical interpretation. In other words, if ‘book’ were rendered into Russian in the present context, its fifty-four original morphological possibilities would be reduced to just one.

Nor is the listener satisfied that the three words starting our sentence constitute a complete utterance. His mind is conditioned to expect a direct object. This game of ‘teasing’ the listener can be continued indefinitely, for, theoretically, it is *always* possible to add to a sentence a new occurrence, which will demand that at least one more follow it.

During the above partial utterance, the listener had occasion to exercise ‘foresight’ only. But suppose we had started our sentence with ‘This book’. At first, the listener is conditioned to accept the expression as the subject of the clause, and he anticipates a predicate in agreement with it. If, however, the next word of the sentence is ‘I’, the listener’s mind revises his previous prediction, realizing that the last word must be the subject instead. He now expects the first two words to be the object of some transitive verb still to be uttered, and a different form for the predicate. All of the complicated backward and forward oscillations take place in his brain before the speaker utters the next word, ‘intend’. Since the last is a modular word, the listener is not surprised when it is followed by the infinitive ‘to read’. Moreover, the last word fulfils his earlier expectation of a transitive verb.

3. THE MACHINE-GLOSSARY

We simulated mechanically the processes of foresight and hindsight by incorporating 'predictions' into our machine glossary. Before explaining these in detail, we should like to record some general observations concerning the construction of a machine glossary.

Conventional dictionaries were prepared to be utilized by the sublime human mind and not for the benefit of a brainless, senseless, lifeless, man-made contraption. The human translator carries a huge amount of information within his brain before he ventures to consult a bilingual dictionary. In order to achieve *mechanically* the quality of human translation, it would therefore be necessary to feed into the machine, in addition to a dictionary, all that extra store of knowledge which is contained in the capacious human mind. This miracle we cannot accomplish for a number of reasons. The human brain contains some ten billion neurons which receive and store impressions. Furthermore, it possesses the superb feature of being able to associate these impressions in every possible way, to form concepts. A conservative estimate for the total number of combinations attainable with so vast a number of impressions yields a quantity which exceeds the number of all elementary particles in the entire universe!

Of course, we do not know what portion of this fantastic number of concepts is involved in the act of translating, but it is easy to conjecture that no machine in the foreseeable future would be adequate for such a task. We shall therefore have to lower our sights and resign ourselves to the necessity of accepting mechanical results of much lower quality than even the far-from-perfect one achieved by the human translator.

There is still another, and even more potent, reason why we cannot achieve high quality machine translation, at least at the present time. We have not succeeded in achieving a detailed analysis of the principles which enable the human mind to resolve semantic ambiguities, that is, the manner in which it associates ideas.

When a speaker says: "Niels BOHR BORE a grudge against the unmitigated BORE, who again told that story about the wild BOAR which BORE down upon him, as he was cleaning the BORE of his gun", the English-speaking listener has no difficulty in distinguishing the various meanings of the six identically sounding words. Unable as we are to grasp the process by which the brain accomplishes this feat, we cannot, of course, mechanize it. Attempts to apply Roget's *Thesaurus* won't work for the machine. It is a heartbreaking enough task to deal, mechanically, with single occurrences. Where semantic *groups* are involved, only well known idioms can be handled with some degree of success, because their members appear consecutively and thus may be recognized without much difficulty. But the vast majority of polysemantic terms will, in our opinion, not be resolved mechanically for many decades to come, if ever.

However, it is incumbent upon us to strive to attain the utmost that the current state of development in electronic equipment would permit. We must prepare an entirely new, revolutionary type of machine glossary which will supply to the machine as many of the *lacunae*, exhibited by the existing dictionaries, as our knowledge, ingenuity, and financial resources would allow.

Our Russian to English glossary is made up of two sections. In the smaller section, the source entries are *complete* words, the vast majority of which are uninflected words. In the second section, the entries consist of stems only. Alongside of each entry, we list all the morphological, syntactic, and semantic information that we have been able to cull from available sources, including the cranial contents of the members of our staff. This information includes *predictions*. Each of the latter contains at least two portions:

1. A number indicating the degree of expectation for the fulfilment of the predicted occurrence; and

2. The grammatical features of the predicted occurrence.

Since we are attempting to translate the highly inflected Russian into the slightly inflected English, we often have to add another portion:

3. A code indicating the English word (or words) which must be inserted before the listed target (or targets).

We allow room for as many as seventy-four predictions per entry, but the average number will probably be in the neighborhood of three.

4. THE TABLE OF ENDINGS

As an adjunct to the stem glossary, our routine contains a Table, whose arguments are the eighty-three Russian endings. Listed alongside of each, are all of its possible morphological interpretations, as well as the markers which allow the machine to distinguish which subset is to be linked to the stem under consideration.

In this manner, each word of the original source text, which had been decomposed into its stem and ending, is assigned a set of morphological possibilities (only rarely does it possess just a single one) which we call Temporary Choices. These constitute the morphological ambiguities which our routine is constructed to resolve.

5. THE PROFILE

A sentence is, in general, made up of several clauses, and each of these must, of course, contain a subject and predicate (explicit or implied). To mechanize the syntactic integration of the entire sentence, we must first be able to locate the boundaries of each clause, so as to predict the oncoming of a new subject-predicate pair. Unfortunately, clauses are frequently *nested* one with another, and it becomes a Herculean task to ascertain, to precisely which clause each occurrence belongs.

Our routine incorporates a rather sophisticated scheme for disentangling the various clauses and phrases of each sentence. In this part of our work, we guide ourselves by certain signals which, too, are stored in our glossary alongside our entries, indicating what role (if any) the entry plays in causing a clause either to start or to end. Punctuation marks are particularly helpful in this connection, since the Russian language is more 'ruly' in the use of these marks, than is the English. On the other hand, the Russian habit of omitting the present indicative of the verb 'to be' creates many hardships, as do ellipses and long lists of appositive phrases.

The translation problem is so vast and complicated that no detailed explanation of our method for handling it can be given in writing. The serious student is referred to our reports and is urged to try out the method, using any pair of languages with which he is thoroughly familiar. We shall be glad to offer aid, if he runs into difficulty. Moreover, we shall be immensely grateful for any suggestions that would lead to the improvement of our method.

REFERENCES

[1] IDA RHODES: "A New Approach to the Mechanical Syntactic Analysis of Russian", *Mechanical Translation,* 1961, 6, 33-50.

[2] FRANZ L. ALT and IDA RHODES: Recognition of clauses and phrases in machine translation of languages. *Proceedings of the First International Conference on Machine Translation of Languages and Applied Language Analysis.* London: Her Majesty's Stationery Office. (In Press.)

[3] FRANZ L. ALT: The outlook for machine translation. *Proceedings of the Western Joint Computer Conference,* Vol. 17. San Francisco, Calif., (May 1960).

[4] FRANZ L. ALT and IDA RHODES : The hindsight technique in machine translation of natural languages. *J. Res. Nat. Bur. Stand.* April-June 1962. Section B, 3.