[From: *The computer in literary and linguistic studies*, ed.A.Jones and R.F.Churchhouse (Cardiff: University of Wales Press, 1976)]

PROJECT BABEL: MACHINE TRANSLATION WITH ENGLISH AS THE TARGET LANGUAGE

T. D. Crawford

The aim of this machine translation project is the production of a computer program which will translate scientific and technical material into English from a number of other important languages, for each of which a separate data-set will be supplied. The present stage of development involves translation from Russian to English, but the system is being designed to accept any source language which is written in the Roman alphabet or which can be easily transliterated into that alphabet by a person who has no knowledge of the language in question.

With Russian as the source language, the input text must be transliterated at the same time as it is prepared on punch-cards. Otherwise the text is transferred to cards in precisely the form in which it appeared in the published source, except that the difference between upper and lower case letters is ignored, and a space is left before as well as after punctuation marks. No pre-editing is permitted, except that the user may indicate with a special sign the beginning and end of each paragraph so that the lay-out of the original can be more accurately preserved in translation.

The program itself is written in FORTRAN 4, not because this is a particularly suitable language (it is not), but because it is a language available on a large number of computers. Therefore it should be possible to operate the BABEL system on other machines without encountering too many compatibility problems.

The input text is processed by the program sentence by sentence. Words found in the text are stored in the core of the computer until the end of the sentence is reached, and they are then located in a dictionary stored on disc in alphabetical order. The look-up procedure consists of two binary searches, the first to find the block in which the word is registered, the second to locate the word itself within that block. Each block on the disc contains 25 Russian words, with each of which is stored an index number, a grammatical category number, an indication as to whether the word is a homograph, details of any grammatical sub-categorization which may be necessary, and a provisional English translation. This information about the words in the source sentence

is retained in the core for processing by the grammar. Items in the sentence which are not found in the dictionary are assumed to be proper names, arithmetical expressions, chemical formulae, and so on, and eventually pass into the output text untranslated.

The dictionary does not consist of roots and affixes separately, as was the norm in machine translation projects of the 1950s, since the increased storage available on modern devices makes it practicable to treat each morphological variant as a separate word, thereby saving the time which would otherwise be required for the correct decomposition of the words of the input text into their constituent morphemes. It may, however, be necessary to introduce some method of word breakdown for source languages such as German which allow a considerable degree of freedom in the combination of existing vocabulary elements to form new compounds. In Russian the only case in which decomposition is really worthwhile is where the negative NYE is prefixed to adjectives and adverbs.

The dictionary search produces a string of grammatical categories corresponding to the words of the input sentence. This string is then subjected to analysis by a grammar of the phrase structure type, which may be thought of as consisting of rules of the sort which one finds in the phrase structure component of a classical transformational grammar, but working in reverse order, e.g. NP + VP → S.

Of course, these rules have to be far more numerous and complex than the phrase structure rules of a transformational grammar, since they must serve for the correct analysis of derived as well as of kernel sentences. The rules are divided into a number of sets according to the order in which they are applied; e.g. T + N → NP would belong to an earlier set than V + NP → VP.

The grammar is operated by comparing the left hand side of the rules in the first set with the beginning of the string representing the input sentence, and applying any rule where the two coincide. The machine then passes on to the first remaining element in the input string, and again tries to apply the rules. If no rule is applicable, the first remaining element in the input string is passed over, and the procedure is then repeated. When the end of the input string is reached, the machine moves to the beginning of the new string which has been produced by application of the preceding set of rules, and analyses it in a similar manner by means of the next set. This procedure continues until the final set of rules has been applied. The input sentence should now have been resolved into the equivalent of the conventional S symbol, i.e. there should be only one element in the final string. If there are more, the analysis has been at least

partially unsuccessful. In this case a check is made as to whether there are any homographs present in the sentence, and if one is found the alternative interpretation of it is inserted and the grammar re-run. In a sentence which contains several homographs (which, in practice, is a fairly rare occurrence), this may involve ringing the changes on a large number of possible combinations, but the user could limit this process arbitrarily if he were prepared to save computing time at the expense of accuracy.

When a successful analysis has either been made or proven impossible, the next stage is to transform the source language sentence structure into something which is both semantically equivalent and grammatically acceptable in English. Of course, there may be some structural common ground between the two languages (between Russian and English there is a good deal), but the system does not assume this. Therefore a set of rules is given which will effect the necessary transformations. This may be compared to the transformational component of a classical transformational grammar, except that all the rules are obligatory where applicable, and that whereas the left hand side of any rule represents a source language structure, the right hand side represents the equivalent target language structure. The difficulties inherent in the artificial convention for the assignment of derived phrase markers after a permutation are not of practical importance here, since no subsequent transformation is carried out on the resulting string at the same level. If the previous grammatical analysis has failed to construe the input sentence accurately, it may not be possible to apply some of the transformational rules which should be applicable, and the resulting English output will be ungrammatical. Only the failure of the analysis at a very trivial level is likely to result in wrongful application of a transformational rule, and the system has already reached a stage of development at which this normally occurs only if the input text is corrupt.

When all the applicable transformations have been made, the English text is ready for output on the line-printer. In order that the result shall not be a series of disconnected sentences, only full lines are output, except at the end of paragraphs; otherwise the remaining words are stored pending translation of the next sentence.

The chief limitation on this machine translation system (and probably on all others at the present time) is our lack of an adequate method for formalizing semantic information. This means in effect that many problems involving homographs are at present insoluble. The analytical grammar should, when completed, be able to cope in most cases with homographs whose aspects belong to different grammatical categories,

but where a homograph has two or more aspects falling in the same grammatical category (e.g. in Russian, KOMANDA means either a TEAM or an ORDER), the grammar, which is purely formal and would regard the Russian equivalent of COLOURLESS GREEN IDEAS SLEEP FURIOUSLY as a perfectly acceptable sentence, is unable to differentiate, and the translation must therefore include alternative readings. In scientific and technical texts such insoluble homographs are none too frequent, but any extension of the use of the system to other more literary topics is necessarily out of the question at the present time. A semantic component could readily be written into the system if further basic research resulted in an adequate method of formalization, but the inclusion of inadequately formulated semantic criteria might easily lead to the type of comic mistranslation which rather frequently enlivens papers on this topic!

In its limited role as a scientific and technical translation system, BABEL has very recently been made available to staff at University College, Cardiff, for translations from Russian to English. The dictionary contains at the moment something like 17,500 words, and although these cover much of the basic Russian vocabulary, each new text presented naturally contains words which have not yet been entered in the dictionary. Therefore it is necessary first to scan the text by means of an auxiliary program and to list all words which the dictionary does not contain, so that the latter can be enlarged as necessary. Unfortunately this places me in the position of the man with the red flag who walked along in front of cars in the late Victorian era, and it reduces the speed of the translation process to very much the same extent! However, just as the man with the flag was dispensed with after a few years, so I hope that when the dictionary has been expanded to cover a much wider vocabulary, the user will be able to transliterate his own text (or—who knows?—we may even be able to input Cyrillic by then!) and have direct access to the system.

Without a semantic component, the BABEL program will not be able to achieve the ultimate aim of machine translation research, which is fully automatic high quality translation. If we could be certain that the problems of formalizing semantic information would be solved in the reasonably near future, there would be every justification for postponing further research in automatic translation until that day. But given the current state of linguistic theory, we would be waiting for Godot with no firm assurance about when or even if he would arrive. As an interim expedient, a system which falls short of perfection but which has marked advantages over the simple dictionary look-up systems may yet prove worthwhile.

```
## AKADYEMIYA NAUK SSSR . #
## INSTITUT RUSSKOGO YAZIJKA . #
## OBZOR RABOT PO SOVRYEMYENNOMU RUSSKOMU LITYERATURNOMU YAZIJKU ZA 1966 – 1969
   GG. . #
## RUSSKIY YAZIJK V ISSLYEDOVANIYAKH PO AVTOMATICHYESKOMU PYERYEVODU . #
## POD RYEDAKTSIYEY CHLYENA-KORRYESPONDYENTA AN SSSR F.P. FILINA . #
## ( MATYERIALIJ DLYA OBSUZHDYENIYA ) . #
## MOSKVA 1973 . #
## RYEDAKTOR VIJPUSKA S.K. SHAUMYAN . #
## AVTORIJ : YU. D. APRYESYAN ( CHAST' II. ) , I.A. MYEL'CHUK ( CHAST' I. ) . #
## OGLAVLYENIYE . #
## PRYEDISLOVIYE 5 . #
## SPISOK LITYERATURIJ 8 . #
## CHAST' I. 15 . #
## I. MORFOLOGIYA 21 . #
## II. SINTAKSIS 23 . #
## 1. PRYEDSTAVLYENIYE SINTAKSICHYESKOY STRUKTURIJ 24 . #
## 2. OBNARUZHYENIYE SINTAKSICHYESKOY STRUKTURIJ 27 . #
## III. SYEMANTIKA 39 . #
## IV. RUSSKIY YAZIJK V DYEYSTVUYUSHCHIKH SISTYEMAKH AP 50 . #
## CHAST' II. 56 . #
## 1. PRAVIL'NAYA SINTAKSICHYESKAYA STRUKTURA 56 . #
## 2. SINTAKSICHYESKAYA OMONIMIYA 64 . #
## 3. ZAKLYUCHYENIYE 69 . #
## PRYEDISLOVIYE . #
# V POSLYEDNIYE GODIJ VSYE UVYELICHIVAYETSYA POTOK RABOT , POSVYASHCHYENNIJKH
SOVRYEMYENNOMU RUSSKOMU LITYERATURNOMU YAZIJKU , RUSIST UZHYE NYE MOZHYET
VNIMATYEL'NO SLYEDIT' ZA VSYEY ETOY LITYERATURNOY , FONYETIST CHASTO NYE ZNAYET
NOVIJKH DOSTIZHYENIY SLOVOOBRAZOVATYEL'NOY TYEORII , SPYETSIALIST PO
SLOVOOBRAZOVANIYU NYERYEDKO NYE VLADYEYET NOVIJMI IDYEYAMI V OBLASTI SINTAKSISA
I T.D. . .
```

FIGURE 1.

Transliterated Russian Text

FIGURE 2.
BABEL Translation

BIBLIOGRAPHY

Crawford, T. *The Cardiff Machine Translation Project;* paper read at the Conference of the Association of Teachers of Italian, Penarth, March 25th 1972. Copies available from the author.

Crawford, T. *Machine Translation from Italian to English;* Ph.D. thesis. University College, Cardiff, 1973 (unpublished).

Crawford, T. *The Computer as a Translating Machine;* paper read at the UCC/UWIST Joint Research Seminar in Phonetics and Linguistics, Cardiff, May 16th 1973. Copies available from the author.

Crawford, T. *A Review of the Cardiff Machine Translation Project;* Proceedings of the 1973 International Conference on Computational Linguistics, Pisa, Italy (in preparation).

Crawford, T. *Machine Translation;* in *Cardiff Joint Computing Centre Annual Report for the period 1 August 1972-31 July, 1973,* pp. 17-19.

Delconte, G. *Problemi e Metodo;* in *Delta,* No. 9, Sept. 1968.

Gamberini, S. *Analisi Grammaticale Automatica;* in *Delta,* No. 9, Sept. 1968.

Gamberini, S., Delconte, G. and Patrone, E. *Studi per un Vocabolario delle Frequenze dell' Italiano Scritto Contemporaneo;* in *Delta,* No. 5, January 1967.