MACHINE TRANSLATION—
THE EVALUATION OF AN EXPERIMENT

by

A. J. SZANSER, M.Sc., F.I.L.

Senior Scientific Officer, National Physical Laboratory

1. Introduction

THE machine translation (MT) project was started at the National Physical Laboratory, Teddington, in 1960 and was carried on until its conclusion in the past year. This work has been described in some detail in a previous article.[1] The team engaged on the project numbered, at the most, seven people including two electrical engineers (one of them being the leader of the group), one mathematician, two linguists, one programmer and one assistant-operator. The equipment used for the purpose has already been briefly characterized in the article mentioned above, the only addition being a Flexowriter, which allowed for printing the output in both upper and lower case.

Early in 1966 the project reached a stage when the first operational test became possible. About that time the decision had been taken to perform the test and to terminate the basic research, except for such secondary improvements as might suggest themselves during the evaluation. This supplementary work will be described in a later section.

The present article is intended to give the evaluation of the NPL experiment in MT as objectively as possible, and at the same time to bring into focus a wider problem, that of the usefulness of MT in general. This problem has been made topical by the coincidence that within a couple of months from the conclusion of the NPL project, a report was published in the U.S.A. on these matters. It would, therefore, seem pertinent to summarize its findings.

2. American views on MT

The greatest effort towards the realization of MT, expressed both in the number of people engaged in it and in the sums spent on its support, has probably been made in the U.S.A.* It was not, therefore, unexpected that after many years of lavish expenditure from official sources, and the quality of the results being incommensurate with expectations, a reaction followed, aimed at a revaluation of the financial support given to various MT projects.

In 1964, at the initiative of the Joint Automatic Languages Processing Group, responsible for the co-ordination of the officially supported projects, a committee was formed by the National Academy of Sciences — National Research Council to examine the problem in its entirety. The committee, under the chairmanship of J. R. Pierce of the Bell Telephone Laboratories, carried out an extensive study and published in 1966 a report, entitled "Language and machines ".[2]

The main conclusion of the report was that, although limited research in computational linguistics should be continued for the sake of improving the knowledge of language, the financial support for MT projects on the present scale was unjustified. This conclusion was based on the low quality of unedited (i.e. fully automatic) MT on the one hand, and the lack of a real need for mass translation on the other. The latter argument was the result of an extensive investigation into the state of the translating industry in the U.S.A., and was backed by many figures and statistics, such as the fact that at the U.S. Government Employment Service in Washington there were about 500 translators on a waiting list and not a single vacancy. These and similar figures led the Committee to the conclusion that there was a definite excess of supply over demand in the translation market.

As regards the quality of MT, the fully automatic (or, as the report calls it, "raw") translation has been found on the whole unsatisfactory, and the post-edited one was classified as of fair quality, while being at the same time slower (if waiting time, post-editing and producing are included) and more costly than human translation. Some comments on the report will be given in further sections.

3. Recent improvements in the NPL MT system

With reference to the first article[1] it has to be realized that at the time it was written (February 1966) only the preliminary syntactic procedures, viz. the nominal group and the predicate translation routines (l.c., sect 3, p. 103) had been implemented for use with the automatic dictionary. The remaining syntactic procedures (l.c., sect. 4-5. pp. 104, 106) had been worked out, programmed and tested on the simulated dictionary output only. To re-program them for full use would have required much more work, which was impossible with the limited staff and facilities in the time available.

As, however, the team was anxious to improve the output (which then corresponded to the sample included in the paper[1], p. 105) by using a more refined syntactic

* The comparison with the USSR, where research in all branches of computational linguistics is highly developed, is difficult to make, mainly because of a different structure of financing the research.

analysis, it had been decided to select the parts of those procedures which corresponded to the resolution of the most frequent and, at the same time, most disturbing ambiguities found in the texts. In such a way, for example, parts of the co-ordinate blocking or the third person pronoun routines were introduced into the main process in a simplified form.

On the other hand the actual work brought to the surface a number of deficiencies which had been suspected before but whose extent was only now revealed. In the first place there was the matter of the automatic dictionary itself. The original version of this dictionary, compiled at Harvard University, was unsuitable for the NPL project. A first cycle of revision, aimed at reducing the number of equivalents where possible, had been completed. A second, more thorough cycle of revision was then undertaken, and about two thirds of the dictionary had been covered by the end of the project, the remaining third being largely nouns. Moreover, about 1,350 entries were added to the dictionary, filling many of the more serious deficiencies.

Another effort consisted in adding "quick-return" programs which would provide for some inadequacies of the dictionary.* Here belonged two programs: anglicizing and prefix-recognizing. The anglicizing program rests on the assumption that many new words in scientific Russian have international or western roots. The program searches for certain standard suffixes (in addition to already recognized grammatical inflexions) and adds the appropriate English equivalents to the transliterated root. The result is sometimes far removed from the English word, but often its meaning can be guessed without difficulty, for example: in the Russian word "динамических" (supposing it is not found in the dictionary), after splitting off the case inflexion "-их" the infix "-ическ" is recognized as corresponding to the English adjectival ending "-ic", the word, therefore, is output as "dinamic". To improve the transformation, non-standard transliteration rules are used, for instance "К" becomes "C", or "КС" (KS) becomes "X". Thus, the Russian word "оксидирование" is anglicized as "oxidation", which in this case is identical with the actual English equivalent. The anglicizing routine including these transliteration changes is not applied to proper names (recognized as such automatically, under special rules).

The second program to improve the treatment of "not-in-dictionary" words is the prefix-recognizing routine. This searches for some common prefixes†, both international, such as "радио-" ("radio-") or "электро- " ("electro-"), and Russian, as "много-" ("multi-") or "полу-" ("semi-"), using a special list including some 130 prefixes. The prefixes are identified and separated from the remainder of the word at a preliminary stage. During the dictionary look-up both the original, full form and the remainder are checked and, if the former is not found, the remainder is either translated (if found) or submitted to the anglicizing routine, and added, with a hyphen, to the translated prefix.

In both the anglicizing and prefix-recognizing routines, if the root (or one of the roots in a compound word) is Slavonic and not recognizable in transliteration, the remaining part usually contains useful information, even if it gives only the syntactic role of the word, which helps the reader to understand the sentence as a whole.

4. Evaluation: organization and method

During the NPL "Open Days" in May 1966, a number of interested visitors from the universities, Government research stations and industry were invited to take part in an experiment by sending in selected articles from Russian scientific journals for translation. The subject of these articles should be, if possible, electronics, towards which the automatic dictionary was oriented, or an allied subject.

The response was quite satisfactory, 44 texts, ranging in size from 160 to 2,300 words, were received. A number of these papers were not accepted as being too far removed from the field requested* or as having been received too late. In all, 26 papers were accepted and translated. In order to obtain more comments, many translated papers were sent to institutions other than those which provided them, in all to 45 places. The response was again good, in the form of 39 comments, of which five had to be discounted as too vague, and 34 formed the basis for evaluation.

The evaluation of any complex experiment is difficult, and when there is any controversy about its basic principle, even more so. The U.S. committee, having studied two previously proposed methods[3,4], rejected them both as too laborious and too unreliable. Their own method consisted in preparing a number of translations by various means, both human and mechanical, and submitting them to a selected group of undergraduates to be classified according to two specially prepared scales. The translations were compared either with the original Russian text, or with a "model" translation, and graded accordingly. One of the scales assessed the intelligibility of the translation, the other the "informativeness" of the original as compared with the translation.

In the NPL work none of the above methods would be applicable, first of all because the primary assumption had been from the beginning that MT is to provide for readers not having access to either the original Russian version†, or to any "model" translation. Instead of two scales, only one was adopted. It was meant to express

* Even if the dictionary were practically complete, it would never be absolutely so, since scientific papers, more especially in advanced research (the basic kind of text for MT), contain numerous neologisms, or words applied in a new sense.

† " Prefix" is meant in a wide sense, that is including the initial parts of compound words.

* A few papers in very remote subjects were translated for experimental purposes, but not included in the evaluation.

† Or, having access, not being able to read Russian.

the degree of usefulness, as found by the ultimate (expert) reader himself. As the scale had to be standardized, so that it may be applied to the comments and the reactions sent in by all the users, the following gradings were adopted:

8 Fully adequate. Meaning immediately clear even though not always conventionally expressed.

6 Mostly very good. A few sentences obscure, so that something essential may be lost, normally clear enough.*

4 Fair. Takes a great deal of time to extract meaning, and even then there is no great confidence in it, which may result in a partial understanding.

2 Poor. Could only be useful to someone prepared to struggle hard and even then he would often be disappointed.

The extrapolation of the scale at both ends, viz. to "10" (absolutely perfect) and "0" (absolute nonsense) was clearly unnecessary. The odd numbers of the scale were to provide intermediate gradings.

The above scale was not sent to the readers, in order not to force upon them any standard expressions. They were, therefore, able to express their reactions as they saw fit and in their own words. The comments were instead scrutinized independently by four NPL workers, who were prepared to "read between the lines" if, for example, courteous wording obscured the issue, and who actually applied the gradings of the scale to the comments and, subsequently, compared their assessments in order to obtain averages. It is satisfactory to observe that the differences between the assessments were insignificant: normally no more than one point, and two points only in a few exceptional cases.

The samples show typical passages taken from the translated papers. Sample I is from a paper on a subject related to the dictionary field and does not show any transliterated words, whereas Sample 2 exemplifies an "unfamiliar" text. The latter passage comes from a paper, which for this very reason was not accepted for evaluation. It is shown here to provide examples of the action of the anglicizing and prefix-recognizing routines. All transliterated words are marked with an asterisk. The name inserted in handwriting was originally in Latin script and, therefore, was not punched in by the operator.

* In Fig. 1 gradings "8" and "6″ have been marked "v.good" and "good" respectively, for brevity.

5. Evaluation: the results

The papers translated contained, in all, 34,480 text words, of which 1,252* were not found in the dictionary. The latter figure included: 506 proper names, 610 "alien" terms (belonging to other fields) and rare words, and 136 which, in our view, should have been included in the dictionary. The last figure represents 0.39% of the text words.

The dependence of the relative number of missing words (proper names being excluded) on the subject of the paper can be seen in the following table:

Field	*% of missing words* (a)
Electronics (and allied subjects)	0.76
Radio engineering	1.26
Mathematics	1.68
Applied physics	1.91
Engineering (other)	2.10
Nuclear physics	2.12
Cybernetics	2.82
Hygiene (b)	3.78
Physical chemistry (b)	4.24
Palaeontology (b)	11.04

(a) All occurrences of "not-in-dictionary" words.
(b) These papers were not included in the evaluation.

* Occurrences, not different words.

SAMPLE 1

Metal melted into furnace(s) is possible to present in the form of continuous
in

block / assembly / define / determine , but then to cut out from it elementary cube of any dimension and to engrave size also its resistance.

Having replaced elementary cube of melted metal by unit of electrical circuit of
node / knot
model, is possible to reveal distribution of current in it and, having
also

modelled thus all bath of furnace, is possible to recognize character of
learn
distribution of current in melted metal.

Constructional grid model represents geometrically similar volume of bath in constructive

significantly decreased scale.

SAMPLE 2 Significant interest presents immediate bactericide effect of aero-ions.
direct *

E. A. Chernyavskii observed decrease of bacterialic illnesses onto
* supervised * for
on

chlopchatnic, subjected to influence of artificial aero-ionization.
*

L. M. Dobrolet installed favourable influence of negative
* established unfavourable

aero-nionization onto micro-flor of cozhic ran.
* for * * *
on

Cruger (Krueger) and co-authors in experiments with stafilococc,
* also *

suspenzated in small droplets of water, found, that high concentrations of
*

positive and negative ions speed up breakdown of microbes, immediately
unfavourable directly

acting onto cages and raising velocity of evaporation of droplets.
for also rate

The frequency distribution of the degrees of usefulness, as assessed from the comments received, is shown in Fig. 1. The distribution provides the mean usefulness 5.6 (slightly less than "good") and the median 5.5. There was only one comment corresponding to a grading lower than "fair", hence unsatisfactory, and there were seven comments higher than "good".

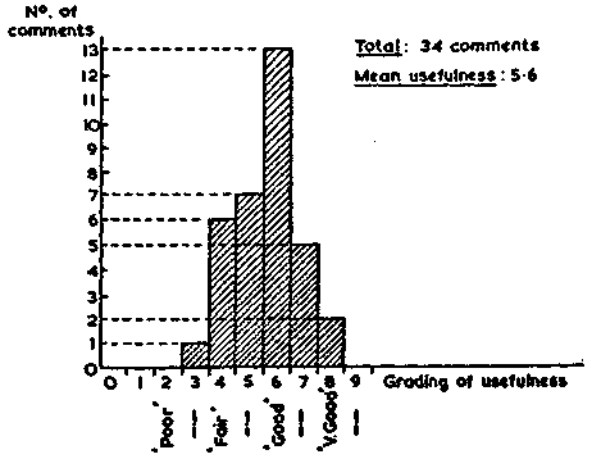

FIG. I.
ASSESSMENT OF USEFULNESS OF N.P.L. MT OUTPUT.

6. Evaluation: the criticisms

Apart from the opinions as to the general usefulness of translation, the comments contained many particular points of criticism, which are worth recording and, if appropriate, answering.

In general, all these criticisms can be classified into three groups, concerning respectively: (i) the English equivalents offered, (ii) the syntactic resolution, and (iii) the word order.

As regards the equivalents, the most frequent criticism concerned equivalents missing, or inappropriate. It has to be pointed out that in addition to fully justified remarks of this kind (see Sect. 3 above), there were also cases in which the meaning proposed or preferred by the reader was uncommon. Its absence was the result of a preferential choice having been made, a compromise between completeness and simplicity. The other alternative, i.e. including all possible equivalents, would drastically impair readability. The particular solution is often very difficult and can only be achieved to a satisfactory degree after a long practice, for which there was no time.

In other cases there is no obvious preference and the problem is further aggravated by the very high frequency of occurrence of the word. Here belong some special classes, as for example all prepositions, and some very common words as "и", "а", "что", and the like. Prepositions can and should be resolved by considering a preposition together with either the governing word or the governed complement (nominal or otherwise).* For the awkward common words specific syntactic subroutines should be devised.† In practically all cases the solution is unique. Obviously, there was no time to include any of these procedures into the NPL MT system.

Only two readers complained about the necessity of selection among two or three equivalents. This is a matter of preference, but it seems to the writer that for a *bona fide* reader an additional possibility of meaning (if it is not carried too far) is more an asset than a disadvantage, even if it impairs to some extent smooth reading.‡ Until a

* On the lines already used for the recognition of idioms, expanded to include non-adjacent words.
† cf. paper [1], Sect. 6, pp. 106-107.
‡ Much can be said on this point. The readers, no doubt, realize how a velvet smoothness of translation may hide many a grievous fault.

semantic analysis can be achieved, multiple equivalents are bound to stay in MT.

A minor point, but nevertheless worth attention, was to the effect that when multiple equivalents followed each other, the difficulty in understanding increased out of proportion. This was undoubtedly so, and could to some extent be helped by improvements in layout.

The complaint concerning many un-idiomatic translations (e.g. "period of work" instead of "life-time") would obviously be allayed by more work spent on the idiom list, which contained only about 540 items, whereas 1,500 would be a more realistic figure.

Complaints about the inadequate syntactic analysis, leading to ambiguities and wrong resolutions, would have been considerably reduced by an implementation of the syntactic routines, described in paper[1] (Sections 4 and 5). One of the minor but annoying ambiguities, which had been theoretically resolved, was that of adverb/short adjective (l.c., Sect. 5, p. 106). The word and component order, too, can be re-arranged according to the English usage only after the complete analysis had been made.

Among other things criticized was an inadequate treatment of abbreviations and abbreviated units, some of which were covered by dictionary entries, while others were not, and this led to some misunderstandings. Obviously this again is a matter for a more complete dictionary.*

Lastly, the anglicizing routine was criticized (while appreciating the general idea) for unorthodox transliteration, which made it more difficult to identify the word in a standard dictionary, if necessary.† A partial solution may be to exclude certain word classes, e.g. acronymic abbreviations, which are obviously not suitable objects for the routine (they can be automatically recognized as clusters of capital letters), and so forth.

In the prefix-recognizing routine there is an inherent danger that a "not-in-dictionary" word may have a part of the stem identical with an accepted prefix. This applies in particular to short prefixes, like "не", and there was, in fact, one case where the abbreviated word "нейтр." was rendered as "non-itr". There is no general way of dealing with such words. The best solution, in respect of both routines, seems to be, however, to include in the output both the original (Cyrillic, if possible) and the transliterated versions for all "not-in-dictionary" words.

7. Machine translation vs. machine-aided translation

The American report contrasts MT, which it considers inadequate, with systems based on machine-aided translation. The latter consist essentially of relieving the (human) translator from the tedious task of dictionary searching. Readers will be familiar with the idea from articles on this subject, that appeared in *The Incorporated Linguist*[6,7]. "Language and machines" reports on three methods currently in use.

The first one is in operation at the Federal Armed Forces Translation Agency, Mannheim (Germany) and concerns English to German translation only. The words required in the English text are underlined by the translator, reduced to their standard form (without inflexions), and keypunched by an operator. The punched cards are fed to a computer, which returns a printed-out list of these words, together with their German equivalents, in the original order (hence the name: "Text-related glossaries"). If a part of a compound word is also not understood, it may be underlined by a second line, in which case the computer retrieves this in addition. Words not found in the dictionary are returned with a suitable remark. The system is said to save considerable time in the translation process and to work, in general, very satisfactorily.*

Another system of machine-aided translation has been worked out by J. A. Bachrach of the European Coal and Steel Community, Luxembourg, in co-operation with Mme. L. Hirschberg of the Free University, Brussels. It is much more sophisticated than the German one and supplies translations of whole sentences containing the required terms (or those nearest in meaning), from one of the four languages used by the Community (viz. French, Dutch, German and Italian) into the remaining three languages. The writer is glad to have learnt recently that this system will be described in an article to be published in *The Incorporated Linguist.*† There is, therefore, no need to go into further detail, apart from the mention that the approach seems to be very promising. "Languages and machines" gives it also a very high rating—"excellent", as compared with "fair to good" for ordinary human translation in general.‡

As the third instance of machine-aided translation, the U.S. report quotes the U.S.A.F. Foreign Technology Department system, although it describes that work in an earlier section as a post-edited MT. This is, in fact, an example of the elasticity of terms, since the dividing line in such a case is, indeed, difficult to draw.

Should machine-aided translation replace MT? No, because it is and will remain real human translation. It should, by all means, be cultivated and developed, but as the writer has already expressed (in the previous paper) the two kinds of translation should be complementary and not vie with each other.

8. Vistas in MT

Apart from the improvements, or more exactly elimination of faults and weaknesses of the system which were discussed in Section 6 above, it is worthwhile to devote some attention to more basic problems which at

* With a few exceptions, however. Thus 'B' may be very troublesome, as regards the choice between the preposition and the abbreviation unit ("volt"), without a special syntactic sub-routine.

† This criticism clearly implied some knowledge of Russian.

* Report 2, pp. 25-26 and App.12.

† Private communication from Mr. Bachrach, dated 28.2.1967.

‡ Report 2, pp. 27-28 and App. 13.

the present time remain unsolved. These belong in the main to the semantic field.*

In order to make the following discussion more precise, it is proposed to distinguish between two different aspects of semantics. The *lexical aspect* refers to the meaning (or meanings) of a text item which may be a word or an idiomatic group, as defined within the (given) language as a whole, or at least within a particular register, or a field, of it. The *pragmatic* aspect, on the other hand, aims to recover the meaning within a particular text in which the item occurs, which may be anything from a single phrase to a book. To illustrate the above a noun, adjective or a verbal idiom are primarily the objects of lexical study, while the meaning of a pronoun is only fully recovered by a pragmatic analysis. And again, a noun may acquire a different meaning if the pragmatic aspect is considered. In what follows the writer will refer to the lexical aspect only.†

Certain possible approaches to semantic analysis have been mentioned in the first article. Other new methods, concerning either the semantic resolution of syntactic ambiguities[8], or the resolution of semantic ones[9] have been recently described. As regards more fundamental studies, known to the writer, those of Halliday[11] and Sinclair[12] are of most interest. Of earlier ones, a deep insight into the problem is shown by a study of Lyons' "Structural semantics".[10]

Dr. Yates, in a chapter on lexis in his thesis quoted above, gives a clear summary of the subject in its present state, as well as some very interesting original ideas on its possible application in MT. He draws a distinction between *general lexis,* the study of the collocation of items, and *relational lexis,* the study of their co-occurrence in a specific grammatical relationship. An organized body of knowledge of what collocations of these types are in fact found, could be useful in the resolution of both kinds of ambiguity mentioned above. The questions arise: would the amount of data involved be too much for a computer, and how could the computer distinguish collocations which are acceptable, but happen not to have occurred before, from those unacceptable? Yates shows how a lexical "score" measuring probable acceptability, might be calculated for any pair of items in either language‡ and how these scores might be used to help resolve the two kinds of ambiguity.

Going beyond these brief notes would lead us outside the scope of this article. The writer would like to return to this fascinating subject some time in the future. For the time being, however, the conclusion remains that, far from being completely explored, the vistas in MT stand out wider and more inviting as the linguistic research is advanced.

* cf. paper 1, Sect. 8, pp. 108-109.

† The above terminology is not generally accepted. Yates (5) makes a rather similar distinction, but uses the terms "lexical" and "semantic" respectively, while not reserving any special name for the common notion.

‡ In a bilingual MT.

9. Conclusion

In the above summary of the several years* effort in MT research at the NPL, the readers may have observed, it is hoped, a continued progress. Although the project has been closed, the research has not reached a dead point or untimely death. What, in fact, has happened can, in more justice, be likened to opening or breaking the ground for further research, wherever this may be undertaken.

The report "Languages and machines" does not query the essential validity of this research. The stress in that report is laid on the lack of sufficient justification for the financial support given, on a generous scale, to various such projects in the U.S.A. This conclusion is based on two main arguments: that the MT systems operating in the U.S.A. are both more costly* and of lower quality than human translation, and that the supply of (human) translators there exceeds the demand.

From the writer's point of view neither of these arguments can be applied in the case presented in this article. First of all, neither cost nor the quality of the production material is an overriding factor in a pilot stage research. Secondly, the excess of the supply of translators in today's U.S.A. seems to result from the massive immigration from Europe, especially as an aftermath of the last World War. Whether this phenomenon is permanent even there, is open to doubt; certainly it is not universal.

Another and very important aspect of the problem is the attraction which a new idea exerts on human minds. Once even a *theoretical* possibility of attaining the target is admitted†, it will remain a challenge for man's intellect and skill, and for this reason alone is worth pursuing. We have witnessed many examples of this, whether in the aviation of yesterday or in the "space race" of today. Let us hope, therefore, that MT research will also be resumed somewhere in this country, at some time in the future.

The work described above (Sections 1, 3-6) has been carried out at the National Physical Laboratory.

* This applies to a post-edited MT.

† It would, perhaps, be worth repeating that the target is not perfect or literary MT, but a practical and limited version (cf. paper[1], Sect. 1, p.102).

References

1. Szanser, A. J.—Machine translation research at the National Physical Laboratory, Teddington. *The Incorporated Linguist,* **5** (4), Oct., 1966.
2. *"Language and machines—Computers in translation and linguistics".* A report by the Automatic Language Processing Advisory Committee. Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Publication 14l6.Washington, D.C., 1966.
3. Miller, G. A. & Beebe-Center, J. G.—Some psychological methods for evaluating the quality of translation. M*achine Translation,* **3**(3),Dec.,1958.
4. Pfafflin, S. M.—" Evaluation of machine translations, etc.". *Machine Translation,* **8** (2), Feb., 1965.
5. Yates, D. M.—*"A linguistic model for Russian-English machine translation ".* Ph.D. thesis. University of London, Nov., 1966.
6. Shibayev, V.—"Translation: man and machine". *The Incorporated Linguist,* **1** (1), Jan., 1962.
7. Arthern,P.J.—" An electronic dictionary". **6** (1), Jan.,1967.
8. Leont'eva, N. N. and Nikitina.—"Kontekstual'nye znacheniya", *Nauchno-tekhnicheskaya Informatsiya,* (12), 1966.
9. Zholkovskii, A. K. and Mel'chuk, I. A.—" O sisteme semanticheskogo sinteza", *Nauchno-tekhnicheskaya Informatsiya,* (II), 1966.
10. Lyons, J.—" *Structural semantics".* 1963. Oxford, Blackwell
11. Halliday, M. A. K.—" Lexis as a linguistic level", in the symposium "*In memory of J. R. Firth ",* eds. Bazell, et al. Longmans. 1966.
12. Sinclair, J. McH.—"Beginning the study of lexis" (See 11)