[Fron: *Handling of Nuclear Information*…Symposium, IAEA, Vienna 1970]

IAEA-SM-128/58

MACHINE TRANSLATION: A NECESSARY COMPONENT OF AN INTERNATIONAL INFORMATION SYSTEM

S. PERSCHKE
CCR, CETIS,
EURATOM,
Ispra. Italy

Abstract

MACHINE TRANSLATION: A NECESSARY COMPONENT OF AN INTERNATIONAL INFORMATION SYSTEM.

Because of the shortage of qualified translators, the growing need for translations for both scientific research and international co-operation presents a serious problem. An essential contribution to its solution could be the mechanization of translation by electronic data processing systems. The experience of several institutions proves that machine translation even at the present relatively low level of quality can be used already in certain fields of applications. In the EURATOM Joint Research Centre at Ispra a regular Russian-English machine translation service has been operational since 1963. On the basis of experience with this service and of the design of a new system the possibilities of further progress in the machine translation field and its future applications are discussed. Emphasis is given to the role which machine translation can play within a large-scale scientific information system.

1. INTRODUCTION

In the Instructions for Submitting Abstracts to the International Nuclear Information System (INIS), the IAEA requests from the reporting centres that at least one version of the abstracts is in one of the four languages of the Agency and recommends a version in the original language. It is up to the customers of the system to make the material accessible through translation, if one does not prefer to teach all the scientific community all the four working languages [1].

The volume of material and the distribution of the languages in INIS is not available, but they should not be essentially different from the general distribution in scientific publications [2, 3]:

E :	43.6 %
R :	24 %
F :	5 %
S :	1.5%

Even if we assume an almost general knowledge of English in the R&D community, the rest of the material, some 56%, must be translated either into the language of each single member state, or into English. In any case those contributors whose language is not one of the IAEA working languages must translate the material which they submit to the Agency. The investment necessary for translation is rather important, and in addition there is the time lag between the distribution of the abstracts and their availability in the language chosen, and the difficulty of forming a sufficiently large and qualified staff of scientific translators.

There is a general trend to mechanize those processes that absorb a large amount of human labour, so that it can be devoted to creative work. In our specific case, the problem is that translation is an intellectual activity, only in part subject to mechanization, and therefore the Fully Automatic High Quality Machine Translation (FAHQMT) of Bar-Hillel [4] is not a realistic objective. Consequently, the question is asked, how useful is a less-than-high-quality translation for information and documentation purposes, and what are the real prospects of progress in the near future.

2. OPERATIONAL MT SYSTEMS

At the Ispra Establishment of the EURATOM Joint Research Centre, Russian-to-English machine translation has been used since 1963 for current awareness and information purposes [4-7].

The translation system used is that developed by Georgetown University, Washington, D.C. It was one of the first to become operational and does not claim high quality. The experience in Ispra, which is similar to that of the two other places where MT is being used - the Oak Ridge National Laboratory (ORNL) and the US Airforce Foreign Technology Division (FTD) [8 -10] - seems to prove that MT in certain applications can replace the human translator without a significant loss of information, and with an important gain in time and cost. A typical example of MT can be seen in Annex 1.

2.1. An experimental MT service

As this experience was made in a relatively small community whose members had little contact with the MT group for consultation, it was felt that the conclusions might be somewhat subjective and not suitable to be extrapolated to general validity. Therefore we established in co-operation with the ETC (European Translations Centre) in Delft an experimental MT Service with CETIS producing, free of charge, the translations requested, and the customers reporting on the usefulness of the service. This experiment is still in progress, but the responses already obtained are mostly (90%) positive. The results of this experiment will eventually be published.

The major complaints in the responses are of lack of specialized terminologies, the clumsiness of the English style and the longer time necessary for reading through the translation. But, as far as the last point is concerned, it is pointed out that this time could be considerably reduced through familiarity with the style of MT and also through the inserting of formulae, symbols, etc., which cannot be reproduced by the computer character set.

3. MT DEVELOPMENT AT CETIS

Past experience makes it clear that the existing MT systems already produce useful translation, but it is highly desirable to improve them to make MT more reliable and easier to use and also to enlarge the scope of its applications. With these aims, a new project has been designed at CETIS, which at present is being implemented to increase considerably both the quality and savings in costs of MT.

3.1. Basic software for natural-language text processing (SLC-II)

The first version of the SLC (Simulated Linguistic Computer) is the software basis of the operational release of the Georgetown MT system conceived and implemented by A.F.R. Brown [11-13]. The basic concept of the system consists in the simulation of a hypothetical computer whose hardware is able to perform all basic functions of natural-language processing. In practice, the scope of the system has been restricted to a class of applications which requires natural text input and dictionary search. Those operations which present no major linguistic problems, but are rather intricate from the point of view of data processing, are performed by a set of optimized computer programs, while a special-purpose programming language — the SLC — is provided for coding the linguistic data and operations which are the actual subject of the work. The SLC language, at its present stage of development, is on the level of a macro-assembler language with an interpretative executor.

In the design of the new system, the following points were emphasized: considerable improvement of the performance; elasticity and language independence during dictionary search; computer independence of the programming language; flexibility of applications, especially in view of documentation.

At present, the system is being implemented for IBM System/360, OS.

3.2. Linguistic improvement

In all discussions concerning the future of MT, it is the so-called "semantic barrier" which is principally used as an argument against the possibility of further improvement.

A rather awkward situation has been created in this kind of discussion: as one term of comparison, one takes the existent operational MT systems - the Georgetown one, or the IBM Mark-II translator - whose linguistic basis is admittedly unsatisfactory, and, as the other term, one puts the postulate of high-quality translation which demands a deep understanding of the text being translated, along with a perfect linguistic knowledge. Of course, one finds a complete incompatibility between those two extremes, and deduces that the MT quality cannot become competitive with the human.

There is a bias in this reasoning. Of course, we cannot claim to build a complete model of the human mind necessary for high quality translation, with the cultural and technological background necessary for evaluating the correctness of the translation, which at least presupposes the understanding of the text translated. But the linguistic basis of the existing system, which is a dictionary approach with a few syntactic and transfer rules, is not the final aim of a non-cybernetic approach to translation, but only a rather poor point of departure.

3.2.1. Short-term objectives of linguistic improvement

Computer linguistics, which might be called a child of MT research in the 1950s, has made noticeable progress in the last few years, but its results were never applied in practice to MT, nor was its actual effect on MT verified.

On the side of input text analysis, two essential points can be considered to be acquired which do not exist in the present MT systems:

(a) The complete parsing of the source text, the so-called surface syntax, which exploits the word classes, the word forms and word orders. Although at present the general trend is towards the Chomsky model and, to a lesser extent, towards the set-theoretical model of Kulagina, it was decided to apply a syntactic model developed by the Italian Operational School [14].

(b) The first level of the so-called deep structure of syntax, which assigns a certain semantic valency to the single parts of speech and describes the syntactic relations as complements.

The full exploitation of these two points, which may be enforced by a morphological analysis which copes not only with word inflection, but also with word derivation rules, should make the transfer function in MT much less hazardous than it is now. The rneta-linguistic basis, which is the part of discourse accessible to formal description, is considerably enlarged, and permits a more reliable transfer.

Progress is much more evident in the target language: the result of the transfer functions is a completely formalized description of the target text, and the so-called generative grammars are perfectly able to produce a target text according to the grammatical rules of the target language.

In this process the only weak point is the transfer function which has the task of making good the lack of source-text analysis, i.e. to resolve at least part of the ambiguities which remain, by means of ad-hoc rules, word lists, statistical criteria, etc.

It is at this point that the "semantic barrier" may be located, and, at present no one is in a position to say how high it is, and how essential it is to overcome it (or what are the priorities) in order to increase the quality of MT. It is the objective of the CETIS MT project, as far as linguistic development is concerned, to reach this barrier, and it will also reveal what it actually is in terms of translation.

3.2.2. Long-term objectives of linguistic research

The outlines of how to overcome this barrier, even if only in a dim way, are already clear:

(1) Introduction of a second level of deep syntactic structure, i.e. the definition of the meaning of the syntactic relations.

(2) Introduction of lexical meaning of the words, i.e. extension of the metalinguistic formalization to the lexical material of the language.

(3) Introduction of a system of associations, i.e. the formal description of general knowledge about the relations of the things to each other**,** which to a certain extent could represent a notional background of the computer.

However, for the time being, these aspects are still in the embryonic stage and are not likely to be applied in short- or medium-term developments.

4. LONG-TERM MT DEVELOPMENT PROJECTS AT CETIS

The CETIS MT project was designed as a pilot project for medium-term development. Its primary objectives have been discussed above. The extensions, for which it should be used as the material and methodologic basis, are :

(1) Development of a reversible multi-language MT system.
(2) Connection of MT and automatic indexing, so as to create an integrated tool for an international information system.

The first objective was suggested by the operation of an international organization such as the European Communities which is obliged to produce all its official documents in the four working languages - Dutch, French, German and Italian - and frequently also in English. For this purpose, a staff of about 500 graduate translators must be maintained. As the character of the texts translated is not so much technical as economic and legal, it was judged that the quality of translation should be considerably raised before MT could be extended for this kind of application.

For scientific and technical information, this improvement is not a sine qua non; however, as was pointed out by the ALPAC report [15] it could become essential if the number of users is considerably increased and the time saved in translation is lost again because of the greater effort and time necessary for reading the translations.

One could state that, at the present degree of linguistic knowledge, the quality of MT can reach a level comparable with that of a translator who uses good specialized dictionaries but has no basic understanding of what he is translating. At this level, MT could even be more advantageous, because it would not attempt to give a, possibly wrong, interpretation to something that it does not understand.

Anyway, at present these expectations should be termed a belief, which may be accepted or not. It is an objective of the CETIS MT project to prove its correctness, and the reason for the decision to attack a project of such a dimension as presented by a reversible multi-language MT system.

5. LIAISON OF MT AND MECHANIZED INFORMATION STORAGE AND RETRIEVAL

One of the objectives of the pilot project, as stated above, is that the SLC-II, as basic software for natural-language processing, should be open for a wide range of applications, especially in the field of documentation. In documentation at present the most urgent problem is probably automatic indexing. Therefore when the system was designed particular attention was devoted to indexing. In 1967, an experimental system was implemented on the basis of the SLC of the Georgetown MT system, which performed automatic translation and indexing of Russian abstracts parallelly.

The results of this experiment were presented at the FID/IFIP Conference on Mechanized Information Storage and Retrieval [16] in Rome June 1967.

In another paper prepared by the division, on automatic methods in documentation and translation of CETIS[1], the results of the automatic indexing project are discussed. At present, the investigations concern the methods of indexing and use a series of ad-hoc programs for testing and evaluating the results [17]. However, it was decided to use the SLC-II for the implementation of an operational automatic indexing system. Thus the connection between translation and indexing becomes evident. This was already emphasized by the findings of the staff of VINITI in a document prepared for the Council of Scientific Unions in connection with the UNESCO World Science Information project, which characterize basic functions involved in the two processes as follows [18]:

	Procedure	Automatic indexing	Automatic translation
1.	Dictionary search	+	+
2.	Morphological analysis	+	+
3.	Local resolution of homonymy	+	+
4.	Syntactic analysis	+	+
5.	Semantic analysis	+	+
6.	Semantic synthesis	-	+
7.	Syntactic synthesis	-	+
8.	Morphological synthesis	-	+

If an abstract is to be translated anyway, automatic indexing becomes a by-product of MT at virtually no extra cost, as the entire phases of data acquisition and handling of both processes are the same.

6. APPLICATION IN AN INTERNATIONAL INFORMATION SYSTEM

The application of an integrated automatic system for translation and indexing in an international information system could be profitable whether centralized or decentralized. A regional national centre responsible for the submission of the documentary material to the central agency, and the distribution of the material received among the customers, is confronted with the problem of translation and indexing. The availability of an integrated MT and indexing systems as outlined above could solve a great part of the problems. As is shown in the paper on indexing by Fangmeyer and Lustig referred to above, automatic indexing, at the present stage of development, can reach the level of manual indexing, as has been shown by inter-indexer consistency and the precision-recall ratio of retrieval tests. It has the obvious advantage of absolute internal coherency. Machine translation, although it is unlikely that it will reach the quality of a good science translator, can very well fulfil its primary task, i.e. to give fast and cheap

[1] FANGMEYER, H.. LUSTIG, G.. "Experiments with the CETIS Automatic Indexing System", these Proceedings, IAEA-SM-128/11.

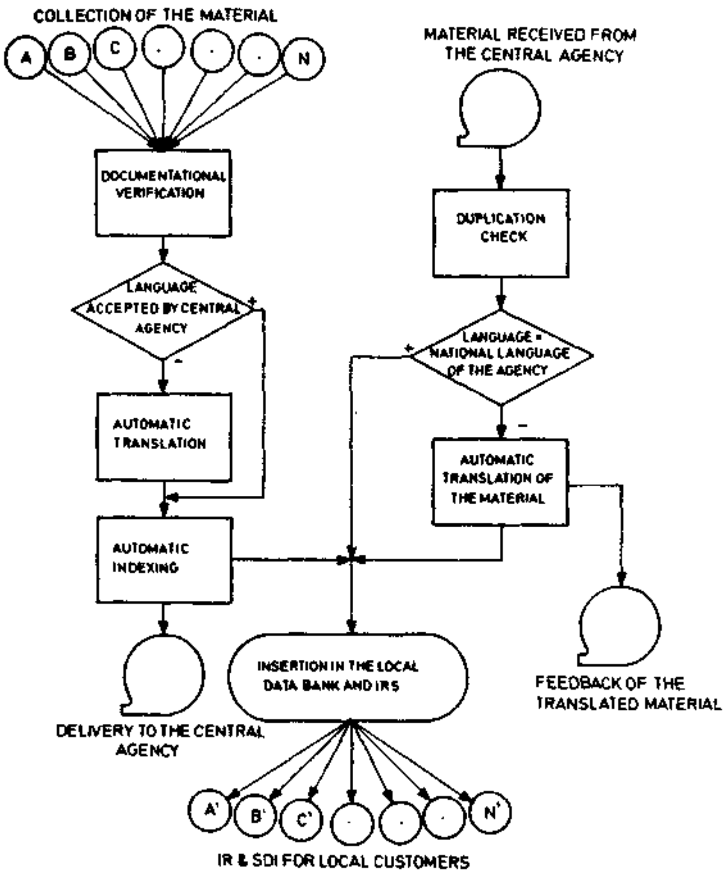
COLLECTION OF THE MATERIAL
A
B
C
.
.
.
N
DOCUMENTATIONAL VERIFICATION
LANGUAGE ACCEPTED BY CENTRAL AGENCY
+
-
AUTOMATIC TRANSLATION
AUTOMATIC INDEXING
DELIVERY TO THE CENTRAL AGENCY
INSERTION IN THE LOCAL DATA BANK AND IRS
A'
B'
C'
.
.
.
N'
IR & SDI FOR LOCAL CUSTOMERS
MATERIAL RECEIVED FROM THE CENTRAL AGENCY
DUPLICATION CHECK
LANGUAGE = NATIONAL LANGUAGE OF THE AGENCY
+
-
AUTOMATIC TRANSLATION OF THE MATERIAL
FEEDBACK OF THE TRANSLATED MATERIAL

FIG.l. Operation of an integrated MT and indexing system in the environment of a decentralized organization.

information. If the source texts are available on a machine-readable support, its cost with the implementation of the SLC-II system will be reduced to about $1.00 per 1000 words translated [19].

The flow diagram in Fig. 1 characterizes schematically the function of such a system in the framework of a national or regional agency. Of course, all its functions could also be centralized, but it seems that this would give no essential advantages.

7. CONCLUSION

The role of MT at present is limited in the languages available, applications, quality, quantity and diffusion. However, it is evident that translation, in the framework of a large information system, is essential and must be mechanized so as to avoid a dangerous bottle-neck in the information flow. The Russian-to-English MT project of CETIS is not only an end in itself, but has also an important task as a pilot project for the preparation of a reversible multi-language MT system and the integration of translation in the operation of an information system.

There is little evidence in favour of the possibility and feasibility of such a system, and in fact there are quite strong opinions against it. Therefore the CETIS project is intended also as a feasibility test, and was designed as an open-end system so that improvements and extensions could be gradually incorporated.

ANNEX 1

SAMPLE OF RUSSIAN-TO-ENGLISH MACHINE TRANSLATION

INFORMATION SEARCH

T. V. Otradinski , M. D. Kravchenko

Concerning One Method Of Indexing Of Complete Texts

UDC 002.54 // 681.3

In contemporary conditions great value is given to automation and to the mechanization of processes of information activity ; there is automatized and also the process of indexing of the documents .

Quality and the labor-consumption of indexing scientific - the technical documents from informationly - prospecting languages (IPYA) descriptor type it is possible to count depending on semantic power IPYA , if one understands under semantic power IPYA : 1) quantity of dictionary composition IPYA , 2) the degree of expression of the basic relations between units IPYA and 3) the presence of grammatical means in IPYA . However , if the quality of indexing is , as rule , in straight relation to enumerated factors , then between the labor-consumption of indexing and by semantic power IPYA there is no univocal relation .

In the right time known several methods of automatic indexing , in the number of which enters and also automatic KLASSIFITSIROVANIE (/ 1 /) . Most known from number realized on practice is the method KWIC (/ 2 /) . In the given work is studied one of the possible methods of

indexing and the prospects of its automation . The determination of used method of indexing closest to following determination , which was given in work (/ 3 . page 251 /) : " Indexing there usually is called the procedure of expression of the main meaning contents of the document in the tens IPYA . In addition to that at first in evident or latent form there is constituted abstract (summary) the document , in which there is formulated its main meaning contents . Then the text of this abstract (summaries) is transferred on corresponding IPYA "? .

Up to the exposure of results of experiment expedient to consider : 1) the main characteristics being used IPYA ; 2) the peculiarities of process of translation on given IPYA and 3) which part of texts of the documents is necessary to transfer on informationly - prospecting language .

1 . IPYA Objectly - ASPEKTNYKH Descriptors

As IPYA , on which there is transferred usual or latent abstract upon the formation of prospecting ways of the documents (UNDER) , there is used language objectly - ASPEKTNYKH descriptors , which briefly it is possible to characterize by following way : the language contains " the object "? and " ASPEKTNYE "? descriptors . Object descriptors there are counted words , forming the simple heading in alphabetically - the object indicator of the Multipurpose Decadic Classification (see footnote) .

****

The multipurpose Decadic Classification . Alphabetic object indicator . M. , 1966

****

. These descriptors work for the reflection of special terminology of a certain object field . As ASPEKTNYMI descriptors work words , the forming headings , which were used to many object fields and indicate the aspect of consideration of subject . By basis for the creation of ASPEKTNYKH descriptors served list the standard technical subtitle of the object catalog of the library im. V. I. LENINA , in which in alphabetic order were collected the most important terms , suitable for the reflection of points of view on subject . Object subtitle in much coincide with the sub-headings of the cluster-sowing headings in alphabetically - object indicator ///UDC . More detailed informations concerning the gradation of the terms on object and also ASPEKTNYE descriptors were given in work (/ 4 /) .

The words , which are contained in indexed texts , but are not contained in the lists of object or ASPEKTNYKH descriptors , are counted the indefinite terms .

From the number of the indefinite terms are excluded so-called not signifying words , i.e. professional parts of speech and words , which do not bear the meaning load . By the example of the list of not signifying words there can work list words , not used as key by the composition of PERMUTATSIONNOGO indicator KWIC in journal " Chemical "? . Given the the ready list of such words not signifying words are removed from indexed texts still up to the comparison of texts with the lists of descriptors . Composition the

printing error of not signifying words is changed upon transition from one object field to other . In the result of experiment (/ 5 /) showed , that upon the formulation of the list of not signifying words in this category is necessary to register such indefinite terms (i.e. words , which were not found in the lists ASPEKTNYKH and object descriptors) , which are contained in small general dictionaries , for example in brief bilingual dictionaries

2 . The Peculiarities of Process Of Translation On IPYA

The process of translation usual , and and also " latent "? abstract on IPYA differs significantly from the process of translation from one natural language on other according to its/their results . Any natural language has much greater semantic power , than IPYA . The process of translation from natural language on IPSA can be univocal . For example , if one has in form the process of POSLOVNOGO translation on IPYA objectly - ASPEKTNYKH descriptors , then slightly INDEKSATOROV , as rule , perform such a translation equally . If there is necessary back translation , i.e. translation UNDER into natural language , then upon the utilization of variant IPYA , described in work (/ 4 /) , such a translation cannot be univocal .

This is explained by the fact that upon translation on IPYA significant part of meaning contents of the document is lost an d remains only this , which facilitates the finding of the corresponding publication in the mass of the

other documents upon any formulation of inquiry . Is , that along with increase of semantic power IPYA the texts of inquiries and their prospecting ways (and also UNDER) become more adequate , and , consequently , the translation of prospecting way of inquiry into natural language and back translation more univocal .

3 . The Peculiarities Of Indexing Of Complete Texts

From above-given determination follows , that the process of indexing is in a certain sense derivative from the process of abstracting , and in that case , when ready abstract there is no , is constituted latent abstract (which nowhere . besides memory INDEKSATORA , is not fixed and at once is transferred in UNDER) . Upon the treatment of the abstracts there usually is used not all of the abstract , but the indicative fragments of abstracts , for example first 2-3 suggestion and the title (/ 6 /) or the first suggestion and the phrases , the containing words " the main goal investigations "? , " the goal investigations "? , " are studied "? , " are described "? , " there are evaluated "? and other words from the fixed list (/ 7 /) .

Being used in the absence of ready abstract " latent abstract "? - this usually quasiabstract (i.e. several suggestions , extracted from the text of the document in unchanged form) , abstract " telegraph style "? (NAZYVNYE suggestions , containing key words) or the collection of the indicative phrases . Conventional abstracts are

REFERENCES

[1] INTERNATIONAL ATOMIC ENERGY AGENCY. INIS: Instructions for Submitting Abstracts. IAEA Rep. IAEA-INIS-4 (Rev.0) (1969) 12.

[2] BREE, R., Erfassung und Verbreitung kerntechnischer Inforrnationen. Atomwirtsch. 12 7(1967) 365.

[3] PLANNING RESEARCH CORPORATION, Survey of the Need for Language Translation. Survey Rep. RC-634(1962).

[4] BAR-HILLEL, Y., Can translation be mechanized?, Methodos 7 25-26 (1955) 45.

[5] PERSCHKE, S., Automatic language translation - its possibilities and limitations, EURATOM Bulletin 2 (1967).

[6] PERSCHKE, S., LUSTIG, G., Automatische Sprachübersetzung - fünf Jahre praktischer Übersetzungsdienst Russisch-English bei EURATOM, Atompraxis 4/5(1968).

[7] PERSCHKE, S., Machine translation - the second phase of development. Endeavour 27 101(1968)97.

[8] PERSCHKE, S., "The use of machine translation in documentation". 1968 Meeting of European Librarians Working in the Nuclear Field, EURATOM Rep. EUR 4256. e (1969) 75.

[9] ARTHUR D. LITTLE, INC., An Evaluation of Machine-Aided Translation Activities at FTD, Contract AF 33(657)-1316 (1965).

[10] KERTESZ, F., Present status and reorganization of the CTC-ORNL machine translation project, Intra-Laboratory Correspondence, Oak Ridge National Laboratory (1969) 8.

[11] BROWN, A., The "SLC" Programming Language and System for Machine Translation, EURATOM Rep. EUR 2418.e, 2 vol. (1965).

[12] BROWN, A.F.R., "Flexibility versus speed", Session 10, Proceedings of the National Symposium on Machine Translation (EDMUNDSON, H.P., Ed.), Prentice-Hall. Inc., Englewood Cliffs, N.J. (1961).

[13] PERSCHKE. S., The Computer Programs of the "SLC" System for Machine Translation, EURATOM Rep. EUR 2583.e (1965).

[14] CENTRO DI CIBERNETICA E DI ATTIVITA LINGUISTICHE. UNIVERSITY OF MILAN, Mechanical Translation: The Correlational Solution. European Office - Office of Aerospace Research USAF., Rep.RADC - TR -(1963) 133.

[15] AUTOMATIC LANGUAGE PROCESSING ADVISORY COMMITTEE, Language and Machines - Computers in Translation and Linguistics, National Academy of Sciences, National Research Council Publication 1416, Washington. D.C. (1966) 124.

[16] PERSCHKE, S., "The use of the "SLC" system in automatic indexing", Mechanized Information Storage, Retrieval and Dissemination, North Holland, Amsterdam (1968) 300.

[17] FANGMEYER. H., LUSTIG. G., "The EURATOM automatic indexing project". Preprint of a paper presented at the IFIP Congress. Edinburgh (1968).

[18] VINITI, Linguistic Problems of Scientific Information. Mimeographed working paper (1969) 31.

[19] PERSCHKE, S., The possibilities of further development of machine translation. Thought and Language in Operations (in press).

DISCUSSION

F. KERTESZ: Similar work is also being carried out at Oak Ridge National Laboratory in co-operation with the Computing Technology Center. We are trying to improve the terminology stored in the machine, using the scientific abilities of the collaborators and restricting our efforts to a few fields, such as nuclear engineering, isotopes, radiation technology, etc. We are doing routine translations of material submitted by the laboratory staff. We found that the scientists will accept rough, unedited translations because they do not have to wait too long. The output is not as elegant as in your case — we do not have upper and lower case.

R.K. WAKERLING: Could Mr. Perschke or Mr. Kertesz give any information on the cost of machine translation compared with human translation?

S. PERSCHKE: The cost of key-punching and machine operation amounts to about $7 per 1000 Russian words translated. No data are available on the cost of system maintenance. There is no uniform cost for human translation. A non-profit organization such as JPRS charges $ 16 per 1000 words of English translation, but this price is located at the lower end of the scale. The computer cost of the Ispra translation system is somewhat lower than at Oak Ridge, since it was integrated in the IBSYS monitor system and requires virtually no set-up time, and the reprogramming of certain portions of the system using an IBM 7090 gave a higher translation speed — 60 000 words/h — than the Georgetown University version.

F. KERTESZ: The cost is comparable with that of human translation - some $15 to $20 per 1000 words. This includes everything — key-punching, placing the tape on the machine, file maintenance and computer time, but not post-editing. The chief advantage at present is the speed, not the cost. The system is there when human translators become "saturated" by the volume of work.