EUR 4256 e

EUROPEAN ATOMIC ENERGY COMMUNITY - EURATOM

1968 MEETING OF EUROPEAN LIBRARIANS WORKING IN THE NUCLEAR FIELD

A selection of papers read at the 5th annual meeting of scientific librarians organized by the Centre for Information and Documentation (CID) and the Scientific Information Processing Center (CETIS) at Stresa (Italy), April 24-25, 1968

1969

The Use of Machine Translation in Documentation

S. PERSCHKE

Abstract

The applications of the Russian-English MT system at CETIS as an instrument for information and documentation are presented. Four principal points are discussed:

- the Russian-English MT service at the request of the customers;
- current awareness with the automatic translation of the tables of contents of Russian periodicals;
- SDI with automatically translated abstracts from Russian periodicals;
- automatic indexing of Russian abstracts to be used as input for an IR system or as a key for SDI with user profiles.

At the end, the new Russian-English MT system which is at present being implemented at CETIS is presented.

1 . Introduction

Although MT, from the linguistic point of view is to be considered just at the beginning of the development, the translation quality presently obtained is sufficiently high for practical use.

CETIS is the only institution in Europe and one of three all over the world to provide a Russian-English machine translation service. The MT system in operation at CETIS, originally had been developed by the Institute for Languages and Linguistics of the Georgetown University, Washington (1). Its primary version became available to Euratom through a research contract in 1963, and it has since been steadily enlarged and improved. The translation system, at present, works at a speed of app. 60,000 Russian words translated per hour, and the over-all cost of some seven dollars per 1,000 words is highly competitive with that of human translation.

2. Applications

The practical applications of the MT system at CETIS are realized on different levels and range from a Russian-English translation service to advanced documentary applications which should operate on an automatic information system. In the following, the principal applications are discussed.

2.1. The MT service

To provide for translations of documents written in a language unknown to the user is a basic function in an information system. This function is fulfilled at Ispra, as far as Russian is concerned, by means of the MT system (2).

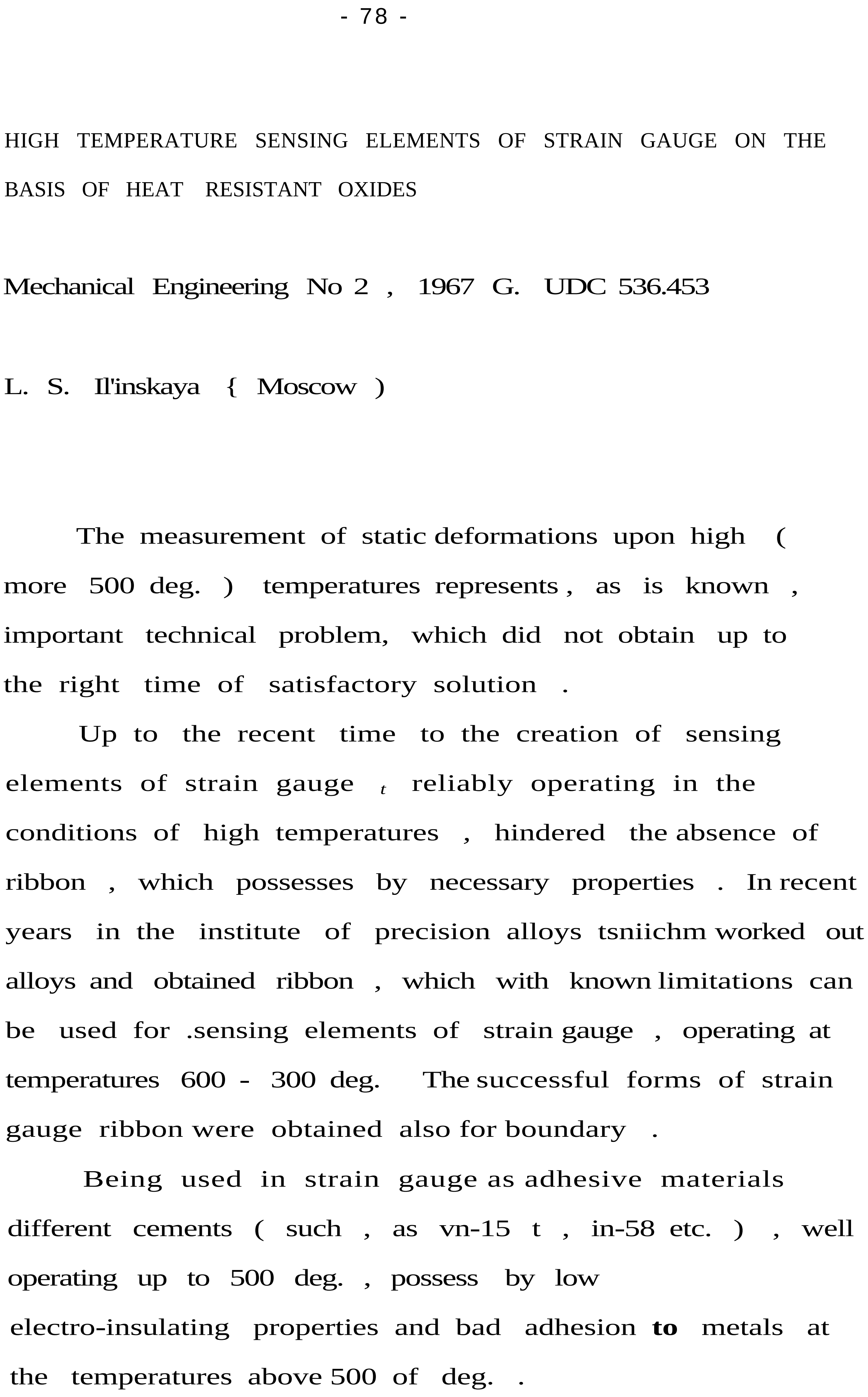
Russian texts are translated automatically at the request of the investigators at Ispra and other Euratom Centres. Although no systematic publicity has been made for this facility, request has been steadily growing since its introduction. Statistics on the number of texts translated are available since 1965 :

1965	app.	30
1966	"	60
1967	"	120
1968 (to April, 1st)	"	30

The 120 documents translated in 1967 correspond to some 700,000 words.

The MT output is delivered to the customers, without any manual editing, together with a copy of the Russian text so as to enable them to identify formulae, equations, graphics etc. which could not be reproduced by the printer of the computer. The MT samples given in Figs. 1-4 may give an idea of the actual MT quality. Although it is all but perfect, experience has proved that, in general, it is sufficient to fulfill its primary function, i.e. to convey the information contained in a document to the customer. Although we did not perform any systematic evaluation of MT quality, a good indication for its acceptability is the fact that the customers normally are satisfied, and never have recurred to the alternate possibility of making retranslate the same text by man, were they dissatisfied with the machine output.

At present, we are examining the possibility of using MT as an aid to translators, so as to increase their efficiency and to produce more, and, possibly, better, translations. It should be possible to produce an equivalent to a scientific translator through post-editing the machine output by English language staff without knowledge of Russian, if there is the possibility of consulting a bilingual translator in the case of doubts.



HIGH TEMPERATURE SENSING ELEMENTS OF STRAIN GAUGE ON THE BASIS OF HEAT RESISTANT OXIDES

Mechanical Engineering No 2 , 1967 G. UDC 536.453

L. S. Il'inskaya { Moscow)

The measurement of static deformations upon high (more 500 deg.) temperatures represents , as is known , important technical problem, which did not obtain up to the right time of satisfactory solution .

Up to the recent time to the creation of sensing elements of strain gauge , reliably operating in the conditions of high temperatures , hindered the absence of ribbon , which possesses by necessary properties . In recent years in the institute of precision alloys tsniichm worked out alloys and obtained ribbon , which with known limitations can be used for .sensing elements of strain gauge , operating at temperatures 600 - 300 deg. The successful forms of strain gauge ribbon were obtained also for boundary .

Being used in strain gauge as adhesive materials different cements (such , as vn-15 t , in-58 etc.) , well operating up to 500 deg. , possess by low electro-insulating properties and bad adhesion **to** metals at the temperatures above 500 of deg. .

Figure 1
Machine translation of a Russian article

2.2. Current awareness of Russian literature

One of the difficulties of the access to Russian publications (as well as for other little known languages) is the fact that also the titles are uncomprehensible to the customer. Only part of the Soviet periodicals publishes tables of contents in English. For the rest, the investigator is bound to wait for references in secondary literature as NSA, or citations in other reviews.

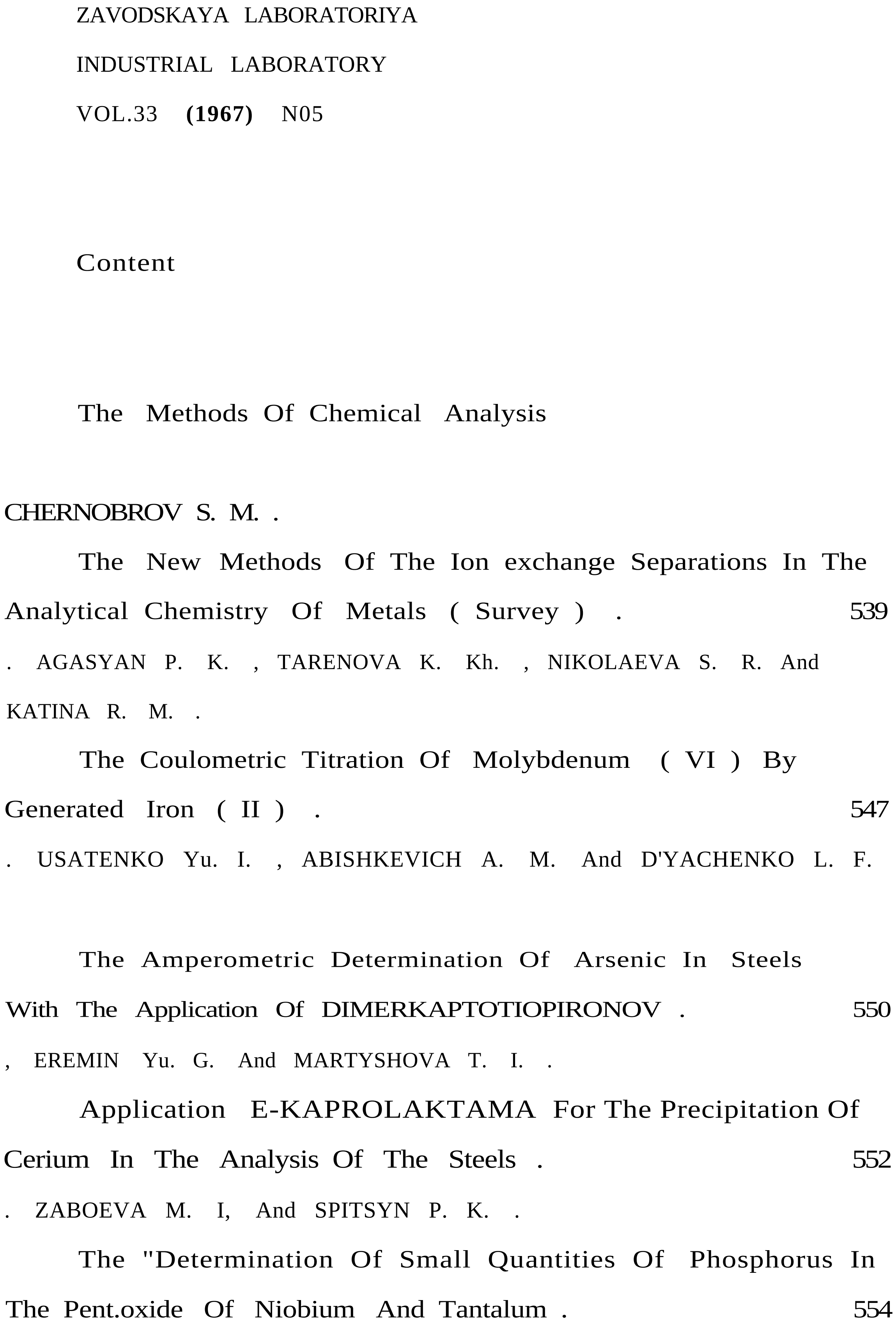
To facilitate access to Soviet publications, therefore, in the environment of the Ispra Centre, the tables of contents are translated automatically and diffused in the Centre with the internal publication "NEW TITLES" which appears more or less weekly (Fig. 2 is to illustrate the presentation of the titles translated). This service was introduced at the end of 1966, and, as far as the MT service is concerned with, it had a double effect : on one hand, the demand for translations was doubled within 1967 ; on the other hand, the time lag between publication and access (i.e. request of translation) was considerably reduced. This can be seen from the following table :

time lag between publication and translation

years	less than 6 months	6 - 12 months	more than 12 months
1966	10 %	20 %	70 %
1967	30 %	30 %	40 %

2.3. SDI of Russian abstracts

The titles, as it is well known,contain insufficient information about the usefulness of an article. Therefore, we are at present examining the modalities of enlarging the current awareness service with some kind of SDI with automatically translated abstracts of Russian periodicals (Fig.3 is a sample)



ZAVODSKAYA LABORATORIYA

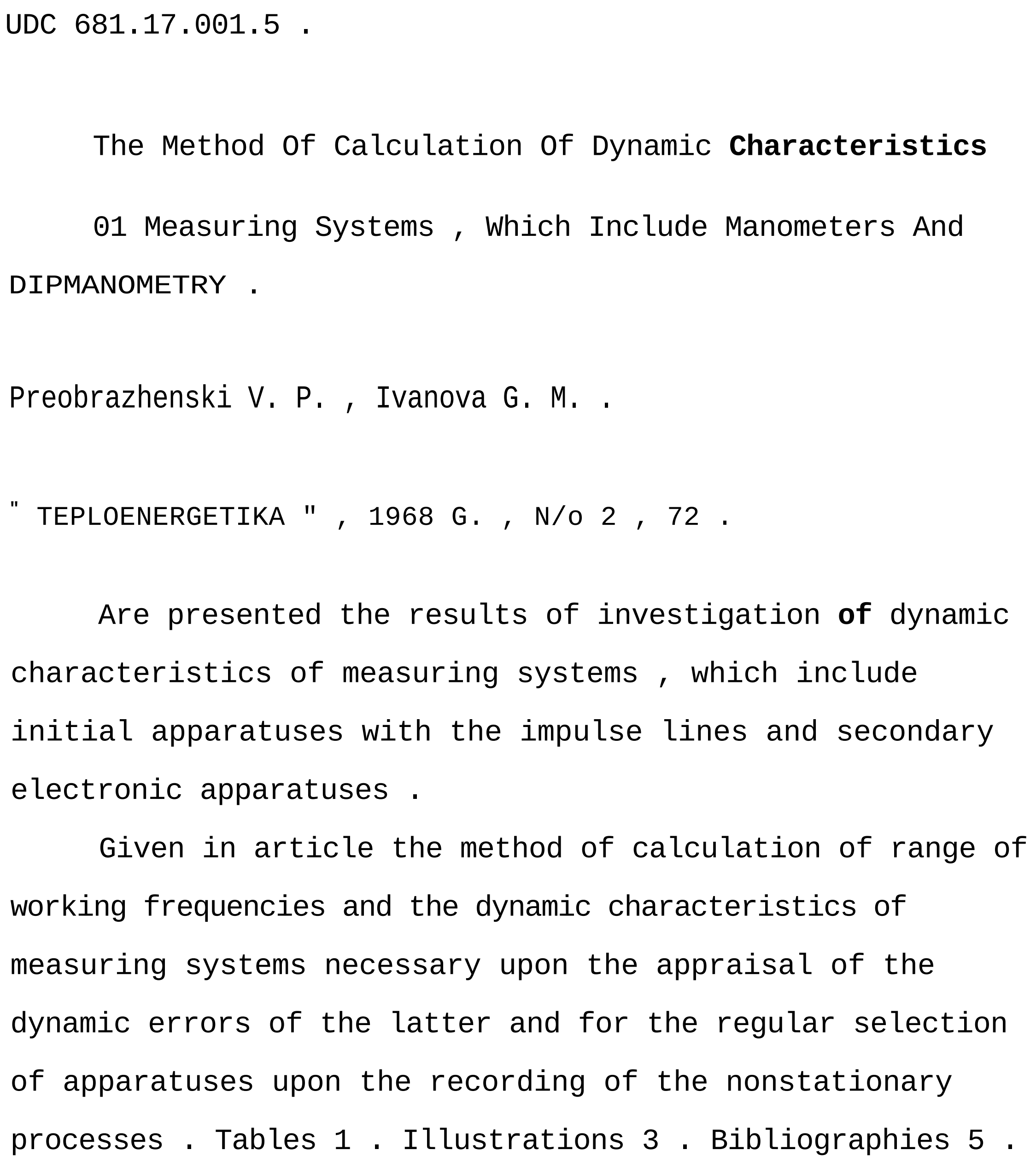
INDUSTRIAL LABORATORY

VOL.33 **(1967)** N05

Content

The Methods Of Chemical Analysis

Figure *2*
Machine translation of *a* table of contents.



UDC 681.17.001.5 .

The Method Of Calculation Of Dynamic **Characteristics**

01 Measuring Systems , Which Include Manometers And DIPMANOMETRY .

Preobrazhenski V. P. , Ivanova G. M. .

" TEPLOENERGETIKA " , 1968 G. , N/o 2 , 72 .

Are presented the results of investigation **of** dynamic characteristics of measuring systems , which include initial apparatuses with the impulse lines and secondary electronic apparatuses .

Given in article the method of calculation of range of working frequencies and the dynamic characteristics of measuring systems necessary upon the appraisal of the dynamic errors of the latter and for the regular selection of apparatuses upon the recording of the nonstationary processes . Tables 1 . Illustrations 3 . Bibliographies 5 .

Figure 3

Machine translation of a Russian abstract.

It is known that abstract journals as NSA, for soviet publications, have a time lag of at least 6 months. Therefore, a properly organized MT service for abstracts could very well close a gap in the existent information systems. We did not make any decision about the modalities and distribution of such a service, but we feel that it might be useful even for a larger community than the Euratom Joint Research Center.

2.4. Automatic Indexing of Russian Abstracts

Last year, we carried out an experiment of assigning automatically English keywords of the EURATOM Thesaurus to original Russian abstracts (3). The experiment was rather limited - it comprised a collection of some 70 abstracts from the field of plasma physics and astrophysics which were also referenced by the NSA and indexed manually within the framework of the EURATOM Nuclear Documentation System (4). The analysis of the documents produced some 500 indexing terms which were integrated into the dictionary. The experiment proved that, from the technical point of view, bilingual indexing does not present particular difficulties. The indexing procedure itself which had been adopted - mainly the formal match of text words with indexing terms and the application of glossary relations (5) - was rather brute-force, however, we believe, it should be improvable, especially in connection with the automatic indexing project of CETIS which is basically language-independent (6).

The principal advantage of such an application is, again, the timeliness and also the economy. If the abstracts are also translated - which is highly desirable - indexing is a by-product at practically no extra cost.

Fig. 4 is to illustrate the output of an abstract translated and indexed automatically. Fig. 5 reproduces the same document abstracted in NSA and indexed by CID. As one can see, a direct comparison between the two samples is not possible, because the document in NSA is not a mere translation of the Russian abstract written by the author, but much more detailed. The

5 V290

CONCERNING ONE POSSIBILITY OF INVESTIGATION OF COMPOSITION OF PRIMARY COSMIC RADIATION OF ULTRAHIGH ENERGY

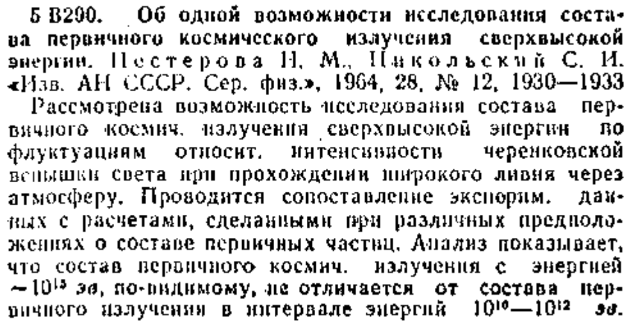
NESTEROV N. M. , NIKOL6SKI S. I.

((BULL. ACAD.SCI. USSR .
SER. PHYS.)) , 1964 , 28 , N/O 12 , 1930 - 1933

WAS EXAMINED THE POSSIBILITY OF INVESTIGATION OF COMPOSITION PRIMARY COSMIC THE RADIATIONS OF ULTRAHIGH ENERGY ACCORDING TO FLUCTUATIONS REL. THE INTENSITIES OF CHERENKOV FLARE OF LIGHT UPON THE PASSAGE OF WIDE SHOWER THROUGH ATMOSPHERE .
THERE IS CONDUCTED THE COMPARISON EXPERIMENT THE DATA WITH COMPUTATIONS , WHICH WERE MADE UPON DIFFERENT PREMISES CONCERNING THE COMPOSITION OF PRIMARY PARTICLES .
ANALYSIS SHOWS , THAT COMPOSITION PRIMARY COSMIC RADIATIONS WITH ENERGY EV , ACCORDING TO-VISIBLE , DOES NOT DIFFER FROM THE COMPOSITION OF PRIMARY RADIATION IN THE RANGE OF ENERGIES EV .

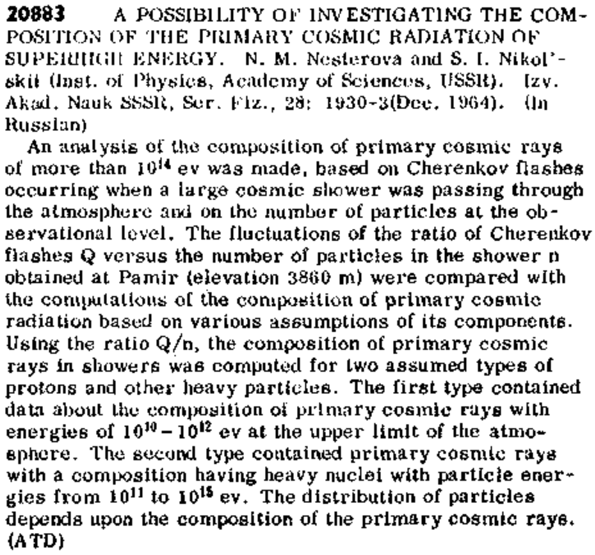
KEYWORDS ASSIGNED TO THE ABOVE DOCUMENT

PRIMARY COSMIC RADIATION
COSMIC RADIATION
RADIATIONS
ENERGY
EXTENSIVE AIR SHOWERS
COSMIC SHOWERS
ENERGY RANGE
SHOWERS
ATMOSPHERE
MEASUREMENT
NUMERICALS
PARTICLES
ANALYSIS
EV RANGE



5 В290. Об одной возможности исследования состава первичного космического излучения сверхвысокой энергии. Нестерова Н. М., Никольский С. И. «Изв. АН СССР. Сер. физ.», 1964, 28, № 12, 1930—1933

Рассмотрена возможность исследования состава первичного космич. излучения сверхвысокой энергии по флуктуациям относит. интенсивности черенковской вспышки света при прохождении широкого ливня через атмосферу. Проводится сопоставление эксперим. данных с расчетами, сделанными при различных предположениях о составе первичных частиц. Анализ показывает, что состав первичного космич. излучения с энергией $\sim 10^{15}$ эв, по-видимому, не отличается от состава первичного излучения в интервале энергий 10^{10}—10^{12} эв.



20883 A POSSIBILITY OF INVESTIGATING THE COMPOSITION OF THE PRIMARY COSMIC RADIATION OF SUPERHIGH ENERGY. N. M. Nesterova and S. I. Nikol'skii (Inst. of Physics, Academy of Sciences, USSR). Izv. Akad. Nauk SSSR, Ser. Fiz., 28: 1930-3(Dec. 1964). (In Russian)

An analysis of the composition of primary cosmic rays of more than 10^{14} ev was made, based on Cherenkov flashes occurring when a large cosmic shower was passing through the atmosphere and on the number of particles at the observational level. The fluctuations of the ratio of Cherenkov flashes Q versus the number of particles in the shower n obtained at Pamir (elevation 3860 m) were compared with the computations of the composition of primary cosmic radiation based on various assumptions of its components. Using the ratio Q/n, the composition of primary cosmic rays in showers was computed for two assumed types of protons and other heavy particles. The first type contained data about the composition of primary cosmic rays with energies of $10^{10} - 10^{12}$ ev at the upper limit of the atmosphere. The second type contained primary cosmic rays with a composition having heavy nuclei with particle energies from 10^{11} to 10^{15} ev. The distribution of particles depends upon the composition of the primary cosmic rays. (ATD)

ABUNDANCE
ANALYSIS
ATMOSPHERE
CHERENKOV RADIATION
COSMIC RADIATION
ENERGY RANGE
EXTENSIVE AIR SHOWERS
LEVELS
MEASUREMENT
NUCLEI
NUMERICALS
PAMIR
PROTONS
SCATTERING
SHOWERS

Figure 5
Reproduction of the abstract 20883 of NSA, Vol. 19, No. 11 and its manual indexing.

automatic assignment of index terms to the Russian abstracts translated should be useful not only in an IR system, but also for the development of a fully automatic SDI system.

3. MT Development at CETIS

In order to increase the efficiency of its machine translation service, since 1963 CETIS has performed the following improvements of the translation system :

- a detailed analysis and description of the computer programs (7), (8), which went along with the re-programming of the system under the control of the IBM 7090 IBJOB monitor system as to reduce considerably operator interventions and to increase the performance of the program ;
- periodical updating of the dictionary and improvement of some linguistic operations which increased somewhat the translation quality ;
- a modification of the input conventions as to permit a more efficient control of processing non-Russian items in the source text, especially, in order to avoid nonsense matches with Russian dictionary entries ;
- an enlargement of the input media as to increase the input capacity. It is now possible to keypunch Russian texts not only on punched cards, but also on paper tape with either Russian or English key-board ;
- the introduction of an output with upper and lower case characters, which highly increases the legibility of the translations and eliminates a certain psychological resistance to the characteristic all-upper case machine output. (Compare the samples given in Figs. 1-3 with the older one in Fig. 4). Since spring 1968, all MT output has been printed with the new facility.

These modifications are very useful for our production purposes, but they do not or only marginally influence the translation quality itself. More important linguistic improvements have not been achieved, the rather poor basis of the actual system making them practically impossible. One should not forget that the Georgetown University system was the first one in the world to be started, and certainly suffered from the unlimited optimism of the pioneer period in which MT was considered basically as a one-to-one term substitution with the addition of a few rules concerning the differences between the source and the target language (9). This concept very soon turned out to be inadequate, but the Georgetown University project never revised it completely. Thus, actually the entire set of linguistic operations in the system is a long series of frequently contradictory ad-hoc solutions, and it is practically impossible to predict the effect of modifications or additions to the analysis performed.

Therefore, in order to obtain a sensible improvement of the translation quality, CETIS is developing now a new system (10). Its design bases principally on the experience and the criticism of the Georgetown system. The main objectives of the new system are :

- a new design of the algorithms of linguistic analysis, especially of the syntax, in a way that one exploits first the formal information already contained in the dictionary and in a second phase adds gradually new information, principally of semantic nature. The primary purposes of the design is to make the system open-ended ;
- the integration of a larger dictionary (180,000 entries against 30,000 in the actual system). The strategy of dictionary look-up presently adopted does not permit the use of such a large dictionary. The new strategy bases on special list- processing techniques ;
- a new design of the special-purpose programming language SLC (7), (8), as to make it basically computer-independent and more flexible, in particular for other linguistic applications ;
- an optimal exploitation of the resources of the IBM 360/65 installed at CETIS. This will raise the translation speed to app. 300,000 words/hour.

The synthesis of these objectives permits various improvements of the actual translation procedure as e.g.

- the handling of non-Russian items. While in the Georgetown system, Russian and non-Russian items (as figures, formulae, English words, etc.) are looked-up indifferently in the dictionary (occupying time and space) and can produce accidental nonsense matches (Cu translated as Chew), in the new system they are considered as a mere character string and identified linguistically through a code which is attached to every non-Russian item ;
- the handling of compound words. While with a dictionary recorded on magnetic tape the only economically acceptable solution is the detaching of certain pre-defined prefixes (as pseudo- semi- poly- etc.) during the text input, the use of a disk storage permits very well to analyse compound words during the dictionary search. Thus, free word combinations which in fact are unpredictable and cannot possibly be contained all in the dictionary as "JELEZOXROMOALHMINIEVY1" (iron-chromium-aluminum) can be identified and translated ;
- the handling of homographs - i.e. multiple matches of text words with dictionary entries. While actually, the problem is disregarded, except some accidental cases (e.g. TOM), it is provided to include all possible matches into the trans= lation process, and to reduce the number of alternatives - possibly to one - in the course of syntactic analysis. Homographs which cannot be resolved with the general analysis procedure are treated individually.

Also, the field of documentation will profit from the new system. The improvement of the translation quality and economy will permit a better exploitation of the existing possibilities of application, while the new SLC system, because of its flexibility, will become an appropriate tool for more advanced documentation problems involving linguistic features.

References

(1) Dostert, Leon E. : "The Georgetown-I.B.M. experiment." in : Machine Translation of Languages. The Technology Press of M.I.T., Cambridge (Mass.). 1957.

(2) Perschke, S. ; Lustig, G. : Automatische Sprachübersetzung - Fünf Jahre praktischer Ubersetzungsdienst Russisch-Englisch bei EURATOM. Atompraxis, (1968) Heft 4/5.

(3) Perschke, S. : The use of the "SLC" system in automatic indexing. Proceedings of the F.I.D./I.F.I.P. Conference 1967, North Holland Publishing Company, Amsterdam.

(4) Rolling, L. : A computer-aided information service. Journal of Documentation, 22 (1966) No. 2, p. 93-116

(5) EURATOM-Center for Information and Documentation - CID :
EURATOM-Thesaurus. Keywords used within EURATOM's nuclear documentation system.
Report EUR 500.e
Presses Académiques Européennes, Brussels, 1964/1966

(6) Lustig, G. : The development of an automatic indexing system at EURATOM. These Proceedings

(7) Brown, A.F.R. : The "SLC" programming language and system for machine translation, Report EUR 24l8.e. Presses Académiques Européennes, Brussels. 1965

(8) Perschke, S. : The computer programs of the "SLC" system for machine translation. Report EUR 2583.e. Presses Académiques Européennes, Brussels, 1965

(9) Perschke, S. : Automatic language translation, its possibilities and limitations.
EURATOM Bulletin, (1967) No. 2

(10) Perschke, S. : Machine translation, the second phase of development.
Endeavour, 27 (May 1968) No. 101, p. 97-100