*Mechanized Information Storage, Retrieval and Dissemination — North-Holland (1968)*

# THE USE OF THE "SLC" SYSTEM IN AUTOMATIC INDEXING

S. PERSCHKE
*EURATOM C.C.R., ISPRA CETIS*

Abstract: The basic concept of the system consists in the simulation of a hypothetical special-purpose computer for the processing of natural languages. (SLC = Simulated Linguistic Computer). The "hardware" of this computer includes the general language data management operations (input-output and dictionary look-up) and a set of instructions which have been defined for the specific needs of this kind of problems.

The SLC system has been conceived by A.F.R. Brown (Georgetown University) for automatic language translation and has been made available to Euratom by a research contract. The Russian-English machine translation system operational on the 7090 uses the SLC.

The Scientific Information Processing Center (CETIS) of EURATOM, actually, is developing a new version of the system for the IBM 360. It maintains the basic idea of the SLC, but introduces important changes in the structure of the programming language, to permit a greater flexibility by exploiting the facilities of the new computer, and by introducing new functions which take better into account the specific indexing operations. Further, the strategy of the dictionary look-up procedure has been redesigned for a more efficient processing of very large dictionaries, using special list processing techniques.

An example of automatic indexing is to illustrate the principal functions of the over-all system and the flexibility of the SLC programming language.

## 1. INTRODUCTION

In many fields of applications, the compiler programming languages such as FORTRAN, COBOL, ALGOL, etc., do not meet completely the needs of the users, while the use of the basic computer assembler languages often requires too much time for coding and debugging. Therefore, it appears advantageous to create

special-purpose programming systems, if the problem is well defined and sufficiently large to justify the programming effort.

Problems connected with the processing of natural languages, especially those which require some kind of dictionary or glossary look-up, as, for instance, machine translation or automatic indexing, or also basic research in syntax, resolution of structural and lexical ambiguities with statistical methods, and many more of them, have all a common procedural basis which can be defined as follows:

(1) The machine-readable source text is read by the computer, and matched against the dictionary. As long as the fast direct access storage even of the large data processing systems is too small to contain a large dictionary, this part of the procedure implies rather involved indexing and sort processes to permit a rapid access to information recorded serially on magnetic tapes or discs.

(2) The words (items) of some logical unit of text (a sentence, a chapter, an abstract, etc.) must be made accessible and processed according to the rules previously defined.

(3) The final results, and in experimental research also intermediate results must be output for further automatic processing or for other uses.

Usually, only the second stage of this process is of real interest to the researcher, while the other stages have a purely instrumental function of providing and controlling the necessary information. But to define the overall process these parts require the major programming effort and are essential for the processing efficiency.

## 2. METHODS

The basic idea of the SLC (Simulated Linguistic Computer) was to define in a possibly optimal way the overall procedure, in a way that the investigator was relieved of the data processing and housekeeping problems and could concentrate on his specific problem. The system was conceived and developed in connection with the Georgetown University Machine Translation Project by A.F.R. Brown, and is now the basis of the Russian-English MT system used at EURATOM in Ispra.

### 2.1. *Components of the SLC system*

The system has the following components:

(1) The Symbolic Assembly Program which converts the symbolic coding into an absolute form;

(2) The Dictionary Updating Program which merges the new dictionary entries with the old dictionary or creates a new one;

(3) A kind of Linkage Editor Program which reads the assembler output of the problem programs and writes it as a file after the dictionary on the system tape;

(4) The Translation Program properly, which carries out the dictionary look-up phase, loads the items of the text to be processed by the problem program, and outputs the results.

From the point of view of the linguist this translation program is to be considered as the "linguistic data processing system". His program is written in the SLC language, and processes as operands the dictionary information corresponding to each item of one sentence or any other logical text unit. The instructions of the language are conceived in a way to resemble operations carried out by a linguist on a sentence as scanning of the text for certain characteristics, examining single items, inserting and deleting items, changing the word order, establishing relations between items, etc. Just to meet better the needs of any particular application, the terminology of classification and instructions can be freely chosen and defined by the linguist himself.

One of the facilities of the system is a rather elaborate automatic monitoring system which can dump the dictionary entries at any point of the problem program and trace the execution of each SLC instruction, which is most useful for the debugging of a program. Further, there is a set of instructions which permits the printing of messages and intermediate results.

### 2.2. *Control of execution*

Control of the program execution is on 4 levels:

(1) Ordinary branches inside one program;

(2) Subroutine entries inside one program and return relative to the calling instruction;

(3) Subroutine calls into other independent programs;

(4) Request and execution of independent programs with priority scheduling, i.e. when the execution of some program terminates with the return to the processor, the request chain is scanned for the program with the highest priority, which then is executed.

2.3. *Execution*

The execution of the SLC instructions is interpretative, i.e., the processor derives from the absolute instruction code, the function and the operands, executes a subroutine which accomplishes the action requested, and branches to the next SLC instruction addressed in the program.

The dictionary entries are also coded in symbolic form. One of the facilities of the system consists in the possibility of coding procedures together with the dictionary entry (so called local operations) and to define the moment of their execution by the priority number. This permits to code, for instance, decision routines for the resolution of ambiguities and to execute them only if the word occurs really in the sentence. For further detail about the SLC system and programming language see refs. [1,2].

The SLC system is actually operational with the IBM 7090 data processing system. Its only application, by now, has been the Russian-English machine translation, which yields acceptable translations of Russian scientific literature and is largely used by the researchers of the Joint Nuclear Research Center in Ispra. The 7090 system works, with a dictionary of some 35000 entries, at a speed of some 60 000 Russian input words translated into English in one hour. This rate is strictly related to the length of the dictionary, since in the actual system, the dictionary look-up is carried out independently for every 2 000 current words of input text. Actually, the translation cost, including keypunching of the input text and computer time, is some $ 7 per 1 000 Russian words.

The recent installation of a more powerful computer (IBM 360/65) has posed the problem of re-programming the system. It appeared unsatisfactory to replace just the 7090 coding by the 360 coding, because it would not exploit optimally the resources of the new computer on the one hand; on the other hand, the actual dictionary look-up strategy is not satisfactory, if one increases the dictionary significantly, and CETIS has at its disposal a new dictionary of some 180000 entries, which is to be implemented in the system in order to improve the translation quality. Further, the different structure of storage organization requires modifications of the absolute form of the SLC instructions and dictionary entries, and the use of SLC language over several years has shown that a new, hopefully improved version would be desirable.

## 3. EXPERIMENTS

It has been felt that the new version of the SLC language should not only base on the experience in machine translation, but should also take into account other applications. The automatic indexing experiment which is presented here was primarily carried out to show the feasibility of the system for applications other than machine translation and to obtain some experience about the specific needs of such an application which should be implemented in the new version of the language. I would emphasize here that the present experiment in no way claims for the solution of any of the specific problems involved with automatic indexing.

At CETIS actually, investigations are carried out by Fangmeyer, which are to determine to what extent major or minor severity of the conditions for the identification and the assignment of keywords, and different assignment strategies can influence the indexing quality. The results of this research will certainly be useful to decide a definitive automatic keyword assignment procedure.

It is evident that a short-range project for the creation of an operational system for the assignment of English keywords to Russian documents, for instance on the basis of the EURATOM THESAURUS, certainly could not take into account all the semantic and associational relations used for manual indexing. But it is sure that eventual lacks of indexing quality would be compensated by the high timeliness of an automatic system, since the time lag between the publication of Russian periodicals and their reference by NSA or cover-to-cover translation, which normally are used as input for manual indexing, is at least 6 months.

The following source material was used for the experimental system: We have chosen some 70 abstracts from two Russian reviews (Jurnal Tekhnicheskoj Fiziki, and Izvestiya Akademii Nauk - Seriya Fizicheskaya), which are quoted in the Nuclear Science Abstracts and analyzed manually by the EURATOM Documentation Centre CID. As basic glossary source we used the EURATOM THESAURUS [3].

The abstracts were first scanned for Russian terms which could have some pertinence to thesaurus terms. These words were classified as "potential keywords". For the identification of such a "potential keyword" the following conditions were posed:

(1) all inflectional forms are accepted;

(2) word deviations and prefixed words were accepted as long as they did not create a new thesaurus term (e.g. IONS and IONIZATION);

(3) if a single Russian word referred to some compound thesaurus entry the first component was classified as such and was given the reference to the "lexical number" of the subsequent components. The other components were just classified as "potential keywords".

As output equivalent of a Russian dictionary entry, i.e. the term which was to be assigned to the document in which it occurred, the following criteria were applied:

(1) no difference was made between "keywords" and "accepted non-keyword terms";

(2) terms referred to by USE were also coded as term to be assigned;

(3) SEE references to other terms and more complicate relation systems like the "graphic displays", correlational and association factors, etc., were just disregarded because their implementation in the present experimental system would be far beyond its scope.

The collection used for the experiment contained some 500 Russian words which had to be re-coded for indexing purposes. The coding was somewhat complicated by the fact that it should not affect the current Russian-English machine translation system.

## 4. INDEXING

The indexing procedure was organized as follows:
The source-document, normally the abstract of a Russian article, was first translated. The translation of the document is optional and can be eliminated by a switch setting, but it has been felt that in a bilingual indexing system, used for practical purposes as input for an information retrieval system or selective dissemination of information, it would be useful to have also the translation of the document. Anyway, in the present system, with a large dictionary, there is no significant time and cost saving if the document is only indexed.

When the translation arrives at the end of a document which is identified by a conventional record, those words which previously were classified as "potential keywords" are loaded as if they were a sentence, and instead of the linguistic operations necessary for translation, the specific indexing operations are executed. In the present experimental system, the indexing operations are rather elementary, and certainly should be refined, if one would develop an operational system for practical purposes. The following routines were programmed:

(1) The "normal" translation of the words is eliminated, and words which occur more than once are deleted.

(2) The thesaurus terms coded together with a single dictionary entry are assigned to the words as "output equivalent".

(3) For the identification of compound terms the simplest possible procedure was applied: the mere presence of the components in the document, in any order, was considered a sufficient criterium.

(4) Multiple assignments are deleted, and the assignments are displayed one per line for printing.

This procedure gives the results which one can observe in the sample machine output (fig. 1).

Though there was no intention of using the results of the experimental indexing system for a comparison with the manual keyword assignments of CID, this idea could appear quite suggestive (fig. 2). But is has proved that a direct comparison is not feasible, since most of the abstracts of Russian documents published in NSA are not just the translation of the Russian abstracts of the documents, but original abstracts which usually differ significantly from the Russian ones and are much more detailed. They are also longer, usually two or three times, than the original abstracts. Thus, no direct comparison and evaluation is possible. The NSA abstracts corresponding to the Russian sample and the keyword assignments made by CID are to illustrate these differences (figs. 3 and 4).

The experiment of automatic assignment of English keywords to Russian documents has shown that the SLC system is highly efficient also for such applications. It is evident that the very indexing strategy used in the experiment would be inadequate for the creation of an operational system on the basis of, for instance, the EURATOM THESAURUS.

## 5. CONCLUSIONS

For the new version of the SLC system and programming language the following conclusions can be drawn from the experiment:

(1) Actually, only those words which previously were classified as relevant for indexing are loaded and processed. If one wants to apply more severe conditions for the definition of compound terms, certain relations, etc., it is necessary that all words of a document be available to processing. For this reason control over the loading of the items for linguistic processing should be given to the SLC programmer and not to the processor program. This, may be,

THE INVESTIGATION OF DEIONIZATION OF PLASMA UPON
ATMOSPHERIC PRESSURE FROM ULTRAHIGH FREQUENCIES

S. I. ANDREEV AND B. M. SOKOLOV

THERE IS DESCRIBED THE METHOD , WHICH PERMITS TO DETERMINE
VARIATION IN THE TIME OF ELECTRONIC CONCENTRATION AND
TEMPERATURES , AND ALSO TO OBTAIN CERTAIN DATA CONCERNING
THE CHARACTER OF DISTRIBUTION OF THESE MAGNITUDES
ACCORDING TO THE LENGTH OF THE COLUMN OF PLASMA .
DIRECTLY IN VOLUMETRIC RESONATOR THERE IS MEASURED
VARIATION IN THE TIME OF CAPACITY AND LOSSES IN FLAT
CONDENSER , WHICH WAS FILLED BY DECAYED PLASMA UPON
THE HIGH PRESSURE OF GAS .
THERE ARE LED THE EXPERIMENTAL DATA FOR THE CASE OF SPARK
DISCHARGE IN AIR .
DISCOVERED , THAT THE PROCESS OF DEIONIZATION IN THIS CASE
OCCURS SHARPLY UNEVENLY ACCORDING TO THE LENGTH
OF THE DISCHARGE INTERVAL .

KEYWORDS ASSIGNED TO THE ABOVE DOCUMENT

IONIZATION
PLASMA
ATMOSPHERE
PRESSURE
MICROWAVES
FREQUENCY
VARIATIONS
TIME
ELECTRONS
CONCENTRATION
TEMPERATURE
ELECTRON DENSITY
VOLUME
RESONANCE
MEASUREMENT
ENERGY
ELECTRON GAS
CAPACITORS
LOSSES
CONDENSERS
DECAY
ELECTRIC DISCHARGES
ULTRA-HIGH FREQUENCY
GASES
SPARKS
ELECTRON TEMPERATURE
AIR

**ИССЛЕДОВАНИЕ ДЕИОНИЗАЦИИ ПЛАЗМЫ ПРИ АТМОСФЕРНОМ ДАВЛЕНИИ С ПОМОЩЬЮ СВЕРХВЫСОКИХ ЧАСТОТ**

*С. И. Андреев и Б. М. Соколов*

Описывается метод, позволяющий определить изменение во времени электронной концентрации и температуры, а также получить некоторые данные о характере распределения этих величин по длине столба плазмы. Непосредственно в объемном резонаторе измеряется изменение во времени емкости и потерь в плоском конденсаторе, заполненном распадающейся плазмой при высоком давлении газа. Приводятся экспериментальные данные для случая искрового разряда в воздухе. Обнаружено, что процесс деионизации в этом случае происходит резко неравномерно по длине разрядного промежутка.

Fig. 1. Sample output of automatic indexing. The NSA abstract (21572) corresponding to this document and the manual keyword assignments made by CID are reproduced on fig. 2.

could also permit to abandon the concept of a predefined glossary and to develop some indexing procedure with a free vocabulary based on linguistic and statistical approaches. However, these problems are outside the scope of the present experiment.

21572 ULTRAHIGH FREQUENCY INVESTIGATION OF PLASMA DEIONIZATION AT ATMOSPHERIC PRESSURE. S. I. Andreev and B. M. Sokolov. Zh. Tekhn. Fiz., 35: 101-7(Jan. 1965). (In Russian)

An ultrahigh frequency method is described by which the time variation of the electron density and temperature of a plasma can be determined and some information can be obtained concerning the distribution of these quantities along the plasma column. This method was used to investigate the deionization following a spark discharge in air at atmospheric pressure, and the results are presented and discussed. A 76-ohm coaxial resonator was used, loaded with an adjustable internal capacitance so that its resonant frequency could be varied slightly from the 750 Mc/sec exciting frequency. The spark discharge took place within the resonator, and the characteristics of the resulting plasma were determined from the shift in resonant frequency and the change in the Q of the cavity. The theory of these effects is discussed, and it is shown that an average value of the electron concentration and temperature can be determined and some information can be obtained concerning the deviation from uniform electron density distribution. It was found that the electron density following a spark discharge in air is very unevenly distributed over the length of the gap. The volume recombination coefficient at electron concentrations between $10^6$ and $10^{10}$ $cm^{-3}$ was found to vary from $2 \times 10^{-6}$ to $1.5 \times 10^{-4}$ $cm^3/sec$, depending on the length of the gap and the energy of the discharge. (ATD)

| | |
|---|---|
| AIR | FREQUENCY |
| CAPACITORS | INTERACTIONS |
| CIRCUITS | IONIZATION |
| CYLINDERS | MEASUREMENT |
| DENSITY | PLASMA |
| DISTRIBUTION | PRESSURE |
| ELECTRIC ARCS | RECOMBINATION |
| ELECTRIC DISCHARGES | RESONANCE |
| ELECTRONS | TEMPERATURE |
| ENERGY | TIME |
| EXCITATION | VARIATIONS |

Fig. 2. Reproduction of the NSA abstract 21572 corresponding to the document indexed automatically on fig. 1 and the corresponding manual assignments of keywords made by CID.

(2) The use of the information looked up in the dictionary and added during the processing of the text as arguments for operations on other items of the sentence (such as comparisons) should be emphasized.

(3) In the present system, arithmetic functions are practically missing. In purely linguistic analysis as it is used for machine translation they are not required. But if one introduces some kind of correlations for the decision whether certain terms are to be assigned or not, these factors must be evaluated numerically.

(4) The actual procedure for the segmentation of compound words which primarily is defined by the sequential organization of the dictionary must be modified in a way that the segmentation is

CONCERNING THE OSCILLATIONS OF CURRENT IN THE THERMIONIC TRANSFORMER OF ENERGY

I. P. STAXANOV AND A. S. STEPANOV

THERE IS PROPOSED MECHANISM FOR THE EXPLANATION OF OSCILLATIONS OF CURRENT IN THE THERMIONIC TRANSFORMER OF ENERGY , BASED ON THE EMERGENCE OF BEAM INSTABILITY AT CATHODE .
THERE ARE REGIMES , IN WHICH EXIST OSCILLATIONS . OSCILLATIONS , WHICH ARISE IN SURROUNDING THE CATHODE RANGE , EXCITE KVAZINEITRAL6NYE THE LONGITUDINAL OSCILLATIONS OF PLASMA IN INTERELECTRODE SPACE , THE PHASE VELOCITY OF WHICH OF ORDER OF IONIC THERMAL VELOCITY , AND DAMPING IS DETERMINED BY COLLISION WITH NEUTRAL ATOMS .
FROM THE CONDITION OF RESONANCE THERE IS THE FREQUENCY OF THESE OSCILLATIONS , WHICH TURNS OUT TO BE BACK PROPORTIONAL TO THE DISTANCE BETWEEN ELECTRODES UPON LOW PRESSURES AND TO THE SQUARE OF DISTANCE UPON HIGH PRESSURES .
THE CALCULATED VALUES OF THE FREQUENCY ARE COMPARED WITH THE EXPERIMENTAL DATA .

KEYWORDS ASSIGNED TO THE ABOVE DOCUMENT

OSCILLATIONS
CURRENTS
ENERGY
COLLISION FREQUENCY
ATOMIC ENERGY
ATOMIC BEAMS
BEAMS
STABILITY
CATHODES
PLASMA
ELECTRODES
SPACE
VELOCITY
IONS
DAMPING
COLLISIONS
ATOMS
RESONANCE
FREQUENCY
DISTANCE
PRESSURE
MEASUREMENT

О КОЛЕБАНИЯХ ТОКА В ТЕР...ЭЛЕКТРОННОМ ПРЕОБРАЗОВАТЕЛЕ ЭНЕРГИИ

*И. П. Стаханов и А. С. Степанов*

Предлагается механизм для объяснения колебаний тока в термоэлектронном преобразователе энергии, основанный на возникновении пучковой неустойчивости у катода. Имеются режимы, в которых существуют колебания. Колебания, возникающие в прикатодной области, возбуждают квазинейтральные продольные колебания плазмы в межэлектродном пространстве, фазовая скорость которых порядка ионной тепловой скорости, а затухание определяется столкновениями с нейтральными атомами. Из условия резонанса находится частота этих колебаний, которая оказывается обратно пропорциональной расстоянию между электродами при низких давлениях и квадрату расстояния при высоких давлениях. Рассчитанные значения частоты сопоставляются с экспериментальными данными.

Fig. 3. Sample output of automatic indexing. The NSA abstract (20913) and the manual keyword assignments made by CID corresponding to this document are reproduced on fig. 4.

performed only with those words which are not completely matched with dictionary entries, and not, as it is done now, before the dictionary look-up a table of possible prefixes, which causes non-sense segmentation in too many cases.

20913 ON THE OSCILLATIONS OF THE CURRENT IN A THERMOELECTRONIC ENERGY CONVERTER. I. P. Stakhanov and A. S. Stepanov. Zh. Tekhn. Fiz., 35: 132-9 (Jan. 1965). (In Russian)

A method proposed for the investigation of oscillations in a thermoelectronic energy converter offers the advantage of explaining the formation of ion clusters and also takes into account the occurrence of collisions. Because a dispersion formula derived from the kinetic and Poisson equations becomes intractable when the distribution functions of electrons and ions in the near-cathode regions are inserted, $f_e$ is approximated by two electron beams, representing the electrons in a quiescent plasma and the perturbation caused by the passage of the current through the diode. The perturbations of the ion distribution are concentrated in one point, and therefore $f_i$ is represented by a single beam. The use of the above approximation leads to a simplified dispersion function for beam instability which shows that, in the region close to the cathode, the instability sets in at sufficiently strong electron currents, or, when the electron current is small, at strong ion current. It is in this near-cathode region that the instability engenders quasi-neutral longitudinal oscillations of the plasma in the interelectrode spacing, the phase velocity of which is of the order of the thermal ionic velocity, while the damping is determined by the collisions with neutral atoms. The frequency of these oscillations is found, and is shown to be inversely proportional to the interelectrode distance if the pressure is low, and inversely proportional to the square of the interelectrode distance if the pressure is high. The results are shown to be in good agreement with experimental data. (ATD)

ATOMS
CATHODES
CONVERSION
CURRENTS
DAMPING
DIFFERENTIAL EQUATIONS
DIODES
DISPERSIONS
DISTANCE
DISTRIBUTION
ELECTRON BEAMS
ELECTRON TUBES
ELECTRONS
FREQUENCY
INTERACTIONS
IONS
LOSSES
OSCILLATIONS
PLASMA
POISSON EQUATIONS
PRESSURE
THERMOELECTRICITY
VELOCITY
WAVE PROPAGATION

Fig. 4. Reproduction of the NSA abstract 20913 corresponding to the document indexed automatically on fig.3 and the corresponding manual assignment of keywords made by CID.

(5) As a linguistic feature at the dictionary look-up phase, it would be most useful not to limit the morphological analysis to the inflectional suffixes, as it is usually done for machine translation, but also to introduce the analysis of regular word derivations, as for instance ION, IONIC, IONIZE, IONIZATION, etc. This could sensibly reduce the size of the dictionary.

## REFERENCES

[1] BROWN, A.: The SLC System and Programming Language for Machine Translation (2 volumes). Euratom Rep. EUR. 2418e, 1965.
[2] PERSCHKE, S.: The Computer Programs of the SLC System for Machine Translation. Euratom Rep. EUR. 2583e, 1965.
[3] EURATOM-THESAURUS, Indexing Terms Used within Euratom's Nuclear Documentation System. Euratom Rep. EUR. 500e, 2nd ed., Part 1, 1966.