[*Information processing: proceedings of the International Conference on Information Processing*, Unesco, Paris 1959] 163

Research on automatic translation at the Harvard Computation Laboratory[1])

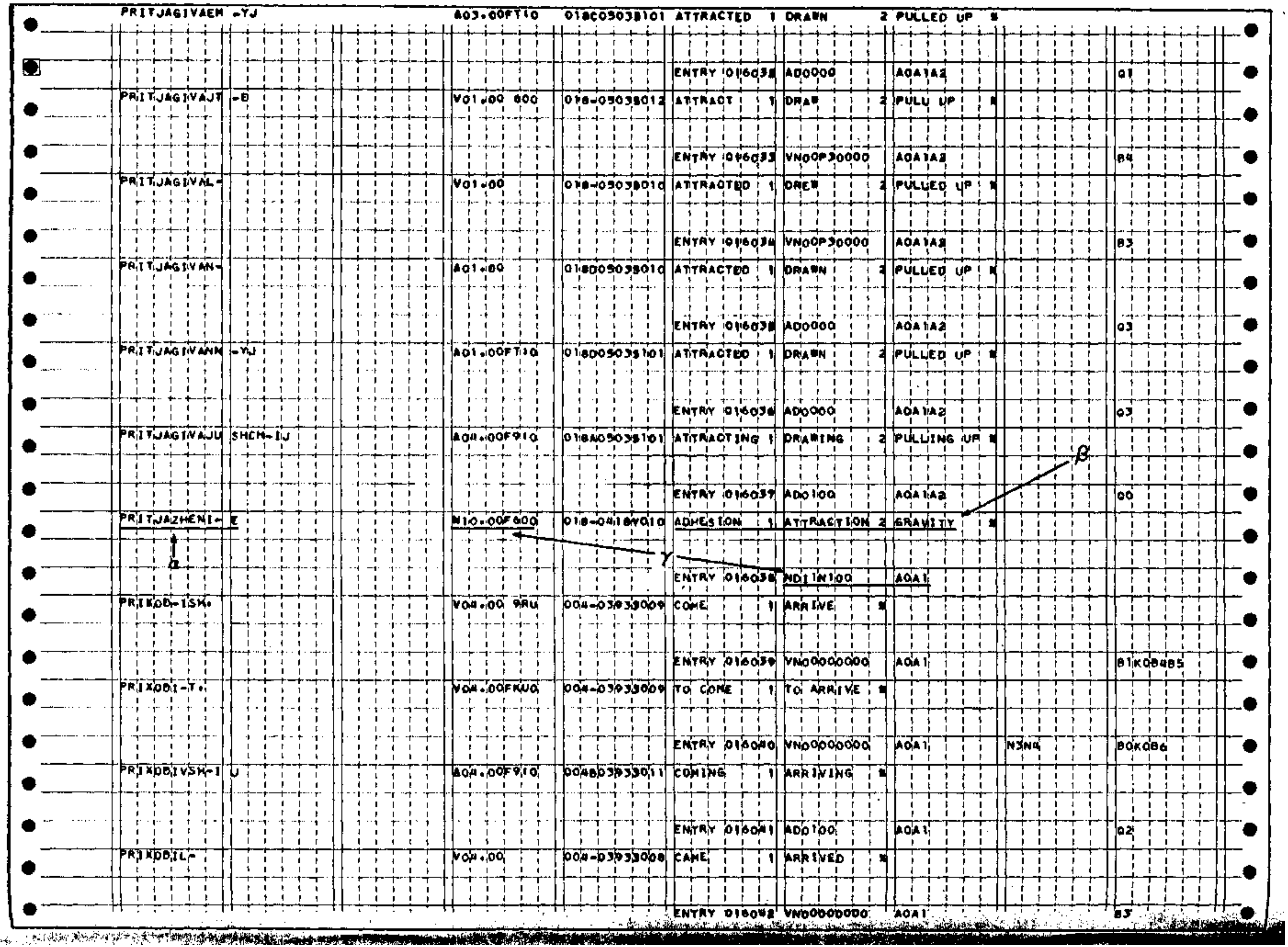
By V. E. Giuliano and A. G. Oettinger,
The Computation Laboratory of Harvard University, Cambridge, Massachusetts (USA)

An automatic Russian-English dictionary of electronics and mathematics, comprising over 10,000 distinct Russian words represented by 22,000 stem entries recorded on magnetic tape, is now being used for the automatic processing of Russian scientific and technical texts. The mode of operation of the dictionary is described, and samples of the dictionary output products are shown.

Immediate practical applications of the dictionary are suggested, and evaluated in the light of preliminary experimental results. The dictionary output products are potentially useful to students, to professional translators, and to technical editors, as aids in their work.

The automatic dictionary is primarily a tool for research on the syntactic algorithms necessary for effecting accurate and smooth automatic translation. Coded grammatical information entered in the dictionary provides, in explicit form, some of the lexical data required for the automatic execution of algorithms. The analysis of Russian syntax is aided by the output products of the dictionary, and by semi-automatic procedures for deriving, applying and evaluating syntactic algorithms.

[1]) This work has been supported in part by the National Science Foundation, the United States Air Force, and the Harvard Foundation for Advanced Study and Research.



```
PRITJAGIVAEM -YJ            A03.00FT10    018C05038101 ATTRACTED   1 DRAWN      2 PULLED UP  *
                                                       ENTRY 016032 AD0000        AOA1A2               Q1
PRITJAGIVAJT -D             V01.00 600    018-05038012 ATTRACT     1 DRAW       2 PULU UP    *
                                                       ENTRY 016033 VN00P30000    AOA1A2               B4
PRITJAGIVAL-                V01.00        018-05038010 ATTRACTED   1 DREW       2 PULLED UP  *
                                                       ENTRY 016034 VN00P30000    AOA1A2               B3
PRITJAGIVAN-                A01.00        018D05038010 ATTRACTED   1 DRAWN      2 PULLED UP  *
                                                       ENTRY 016035 AD0000        AOA1A2               Q3
PRITJAGIVANN -YJ            A01.00FT10    018D05038101 ATTRACTED   1 DRAWN      2 PULLED UP  *
                                                       ENTRY 016036 AD0000        AOA1A2               Q3
PRITJAGIVAJU SHCM-IJ        A04.00F910    018A05038101 ATTRACTING  1 DRAWING    2 PULLING UP *
                                                       ENTRY 016037 AD0100        AOA1A2               Q0
PRITJAZHENI- E              N10.00F600    018-0418Y010 ADHESION    1 ATTRACTION 2 GRAVITY    *
                                                       ENTRY 016038 ND1IN100      AOA1
PRIKOD-ISK.                 V04.00 9RU    004-03933009 COME        1 ARRIVE     *
                                                       ENTRY 016039 VN00000000    AOA1                 B1K08485
PRIXODI-T.                  V04.00FKUO    004-03933009 TO COME     1 TO ARRIVE  *
                                                       ENTRY 016040 VN00000000    AOA1       N3N4      B0K086
PRIXODIVSH-I U              A04.00F910    004B03933011 COMING      1 ARRIVING   *
                                                       ENTRY 016041 AD0100        AOA1                 Q2
PRIXODIL-                   V04.00        004-03933008 CAME        1 ARRIVED    *
                                                       ENTRY 016042 VN00000000    AOA1                 B3
```

Fig. 1. Entries in the Harvard Automatic Dictionary

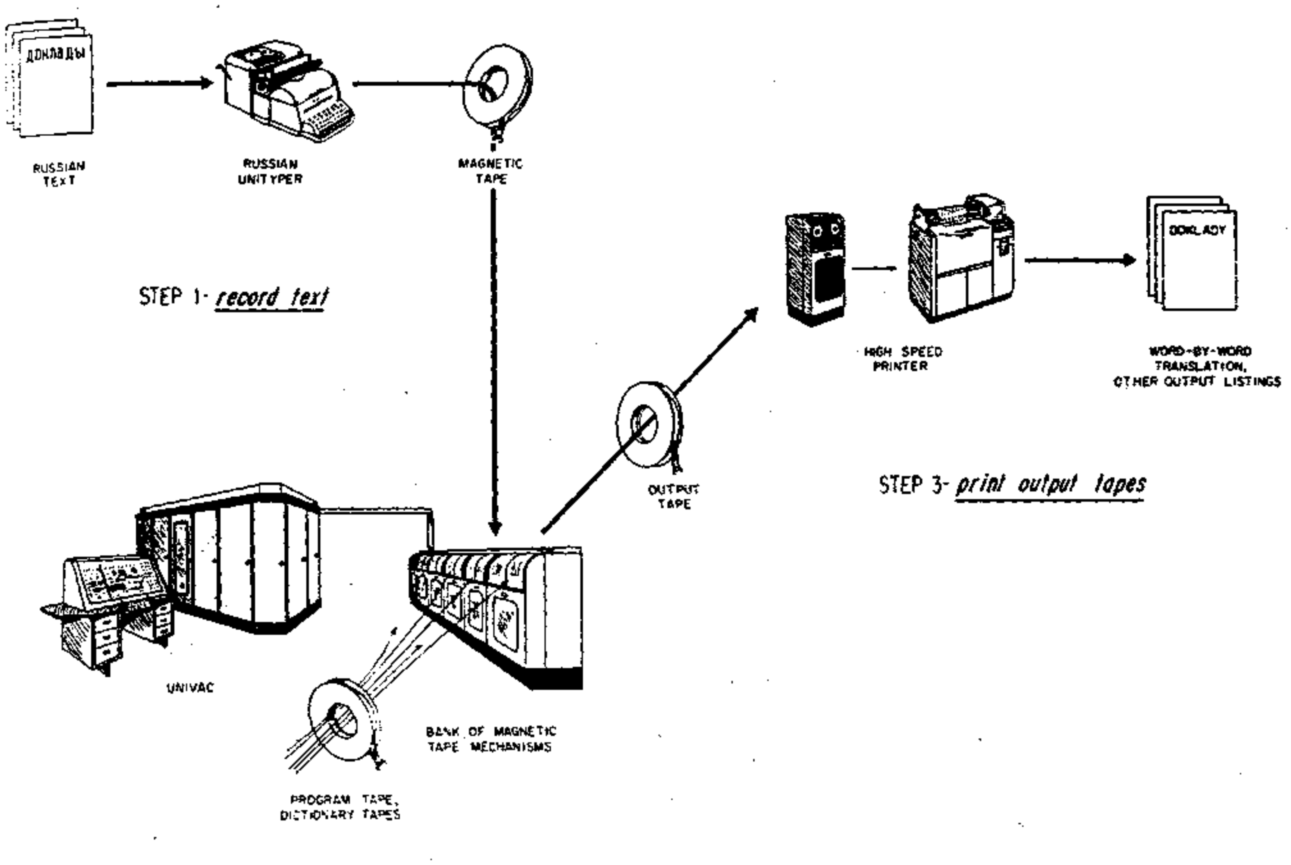

ДОКЛАДЫ
RUSSIAN TEXT
RUSSIAN UNITYPER
MAGNETIC TAPE
STEP 1- record text
HIGH SPEED PRINTER
DOKLADY
WORD-BY-WORD TRANSLATION, OTHER OUTPUT LISTINGS
OUTPUT TAPE
STEP 3- print output tapes
UNIVAC
BANK OF MAGNETIC TAPE MECHANISMS
PROGRAM TAPE, DICTIONARY TAPES
STEP 2- automatic computer run (COR)

Fig. 2. The processing of Russian texts with the Automatic Dictionary

Research on automatic translation at the Harvard Computation Laboratory initially centered around the compilation and experimental operation of a fullscale Russian-English automatic dictionary. An automatic dictionary is a machine system capable of producing word-by-word translations of texts, i.e., translations in which one or more words in an output language are simply substituted for each word in the source language; it is a necessary part of any automatic translating system. The dictionary will soon be a component of a more complex machine system serving as a tool for research on Russian and English syntax. At present, it produces rough English translations of Russian technical texts and a number of by-products valuable for research. This paper is concerned with the operation of the Harvard Automatic Dictionary, its output products and their immediate applications, and with the instruments being developed for more advanced research in automatic translation.

The semi-automatic compilation of the Russian-English file of the Harvard Automatic Dictionary has been described in an earlier paper [1]. A printed section of the dictionary file, illustrating the arrangement of information into Russian-English entries, is shown in fig. 1. Each entry contains a Russian stem (transliterated and marked α), a set of English meanings (β), and coded grammatical data (γ). There are twenty-two thousand stem entries in the present dictionary file, representing more than ten thousand distinct Russian words or about a hundred thousand inflected forms of these words. The file is stored on six reels of Univac magnetic tape. The vocabulary is primarily scientific; it contains a large number of technical terms drawn from the areas of mathematics and electronics. The scientific vocabulary is supplemented by a basic vocabulary of high-currency words selected f rom a general dictionary.

1. Processing Russian texts

Three steps are required to process a text with the automatic dictionary (fig. 2). A Russian text, up to approximately three thousand words in length, is recorded on magnetic tape (step 1) by means of a special Unityper capable of recording Russian as well as English [2]. In the upper-case, the keyboard resembles that of a standard English typewriter and in the lower-case that of a standard Russian typewriter. The typist must be familiar with the Cyrillic alphabet, but need not have a deeper knowledge of Russian; the Russian is simply copied in a normal running format. A sample section of the proof copy produced on the Russian Unityper is shown in fig. 3. The typist has written ad-lib English comments to describe equations and illustrations, and has bracketed these with dollar signs. The wedge symbol "<" represents a space. Punctuation marks are bracketed with asterisks for machine identification.

In laying out the Unityper keyboard and in designing the computer program to which it furnishes input, special efforts were made to allow the typist as much freedom as possible. Special rules have therefore been kept to a minimum; the typist may use any number of spaces to separate words, English comments may be of any length, etc. This freedom is important, since it allows typists to proceed at nearly normal typing speeds without special training, and since it minimizes the human effort at present required for input. The Unityper will be replaced by an automatic print-reading machine, if and when a sufficiently flexible device of this type becomes available. Meanwhile, the Russian Unityper enables the encoding of texts with both reasonable cost and accuracy.

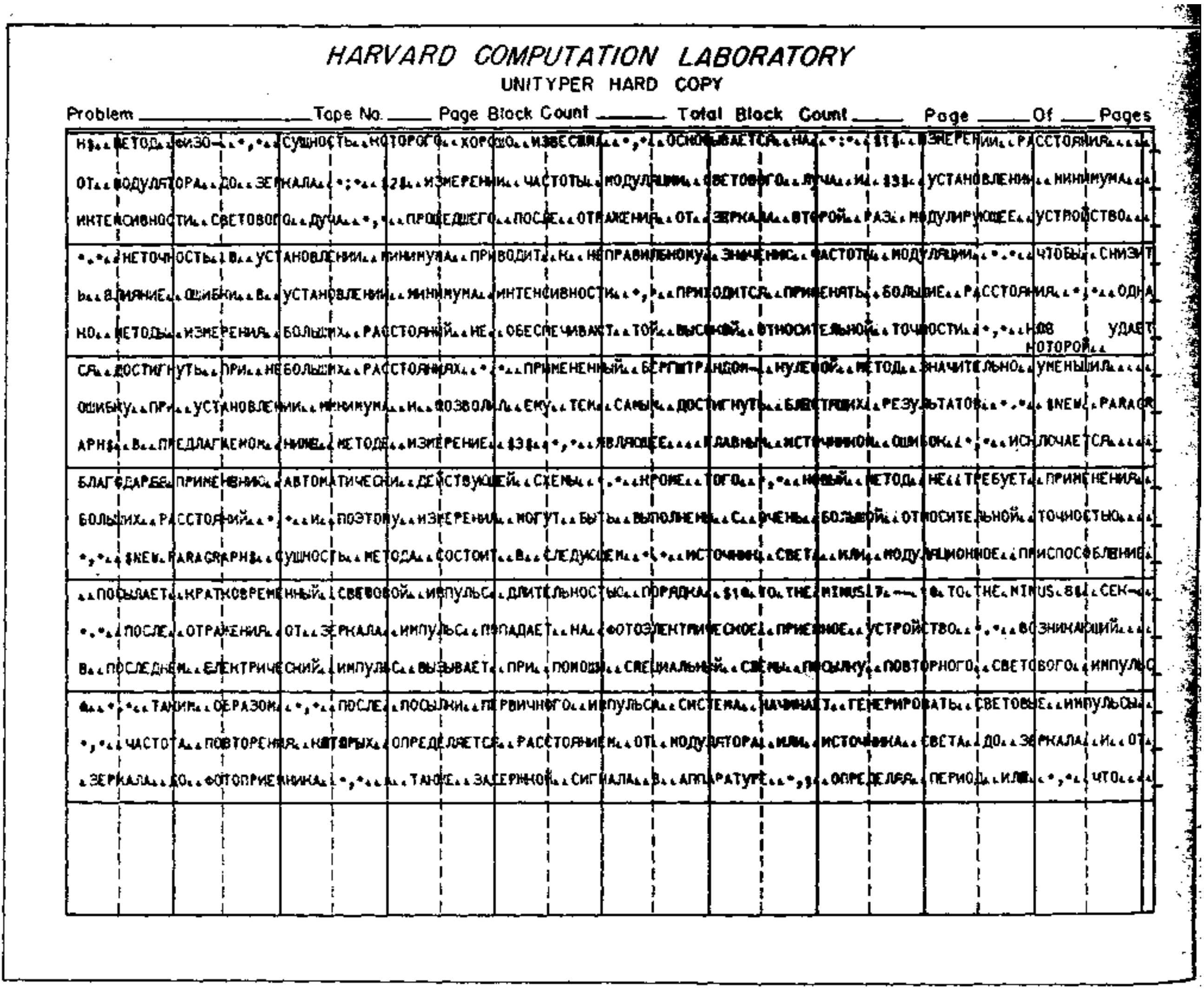
The reel of tape containing the typed text and reels containing the Russian-English dictionary are mounted on tape units of a Univac computer. The operating program of the automatic dictionary is read into the computer from another reel of tape, and the machine is started (step 2 in fig. 2). The computer then operates continuously and automatically until the text is processed. The structure of the operating program is outlined in Section 3. Long texts can be processed in parts; several short texts can be processed in a single run. The present experimental program produces the word-by-word translation of a four-page text (roughly one thousand words) in an hour-long computer run. An equivalent production program designed to operate on a faster computer with a larger internal memory could produce translations at ten to a hundred times this rate. The output tapes are finally removed and printed, using a Univac high-speed printer (step 3).

2. Outputs of the Harvard Automatic Dictionary

The outputs of the Harvard Automatic Dictionary are listed in fig. 4. The main output for research purposes is made essentially by substituting an appropriate dictionary entry for each Russian word in the original text. This output may be called a "text-ordered subdictionary" or, more briefly, an "augmented text". The augmented text combines the available lexical information with the original textual information, and hence contains the data necessary for any further automatic analysis of the text. In experimental work all texts are initially processed with the automatic dictionary and all programs performing syntactic or semantic transformations use augmented texts as inputs. Basic experimental applications of augmented texts are discussed in Section 5.

Because of the large amount of coded data contained in dictionary entries, augmented texts are not normally printed. They are retained on magnetic tape for research purposes, but the information in them is automatically edited into several different formats suitable for printing The outputs numbered 2 through 7 on fig. 4 are the various edited prints produced by the automatic dictionary run. A segment of a word-by-word translation produced by the automatic dictionary is shown in fig. 5. The text is read from left to right; alternative English correspondents for the same Russian word are printed in a column. When the same stem occurs in more than on dictionary entry, the correspondents of all entries are given and marked with the symbol "+ + + + (HOMOGRAPH OF PREV)". Words in the text but not in the dictionary are simply transliterated into Roman characters and marked with the symbol "# # #". English comments made by the typist are reproduced, and marked with the symbol "* * *".

The word-by-word translation is also produced in a format where the original Russian is transliterated and interlineated with the English correspondents. A sample section is shown in fig. 6.



HARVARD COMPUTATION LABORATORY

UNITYPER HARD COPY

Problem ____________ Tape No. ____ Page Block Count ____ Total Block Count ____ Page ____ Of ____ Pages

Fig. 3. Proof copy produced on the Russian-English Unityper

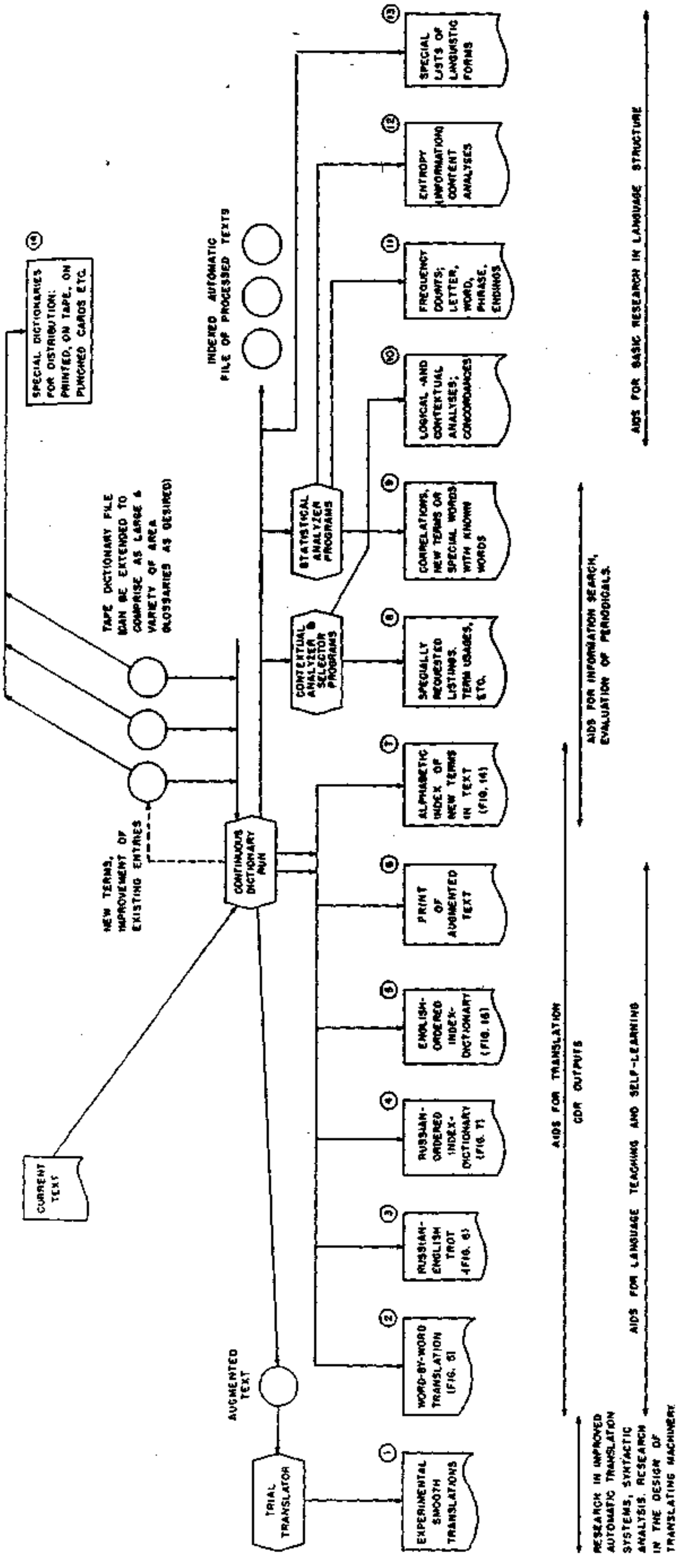

CURRENT TEXT
TRIAL TRANSLATOR
AUGMENTED TEXT
CONTINUOUS DICTIONARY RUN
NEW TERMS, IMPROVEMENT OF EXISTING ENTRIES
TAPE DICTIONARY FILE CAN BE EXTENDED TO COMPRISE AS LARGE & VARIETY OF AREA GLOSSARIES AS DESIRED)
14
SPECIAL DICTIONARIES FOR DISTRIBUTION: PRINTED, ON TAPE, ON PUNCHED CARDS ETC.
INDEXED AUTOMATIC FILE OF PROCESSED TEXTS
CONTEXTUAL ANALYZER & SELECTOR PROGRAMS
STATISTICAL ANALYZER PROGRAMS
1
EXPERIMENTAL SMOOTH TRANSLATIONS
2
WORD-BY-WORD TRANSLATION (FIG. 5)
3
RUSSIAN-ENGLISH TROT (FIG. 6)
4
RUSSIAN-ORDERED INDEX-DICTIONARY (FIG. 7)
5
ENGLISH-ORDERED INDEX-DICTIONARY (FIG. 16)
6
PRINT OF AUGMENTED TEXT
7
ALPHABETIC INDEX OF NEW TERMS IN TEXT (FIG. 14)
8
SPECIALLY REQUESTED LISTINGS: TERM USAGES, ETC.
9
CORRELATIONS, NEW TERMS OR SPECIAL WORDS WITH KNOWN WORDS
10
LOGICAL AND CONTEXTUAL ANALYSES; CONCORDANCES
11
FREQUENCY COUNTS; LETTER, WORD, PHRASE, ENDINGS
12
ENTROPY (INFORMATION) CONTENT ANALYSES
13
SPECIAL LISTS OF LINGUISTIC FORMS
AIDS FOR TRANSLATION
CDR OUTPUTS
AIDS FOR LANGUAGE TEACHING AND SELF-LEARNING
AIDS FOR INFORMATION SEARCH, EVALUATION OF PERIODICALS.
AIDS FOR BASIC RESEARCH IN LANGUAGE STRUCTURE
RESEARCH IN IMPROVED AUTOMATIC TRANSLATION SYSTEMS, SYNTACTIC ANALYSIS, RESEARCH IN THE DESIGN OF TRANSLATING MACHINERY

Fig. 4. Outputs of the Harvard Automatic Dictionary

The edited prints numbered 4, 5, and 7 on fig. 4 are alphabetic "index-dictionaries", containing only words from the particular text being processed. A section of an index-dictionary is shown in fig. 7. Each entry contains:

a) A Russian stem (A).

b) The set of English correspondents assigned to the stem in the dictionary (B).

c) The number of occurrences of the stem in the text (C).

d) Serial numbers that give all the locations in the text where the stem occurs (D).

e) The endings with which the stem occurs at the various text locations (E).

f) Coded grammatical information present in the dictionary (F).

g) Space in which to write new English correspondents, or to modify existing correspondents (G).

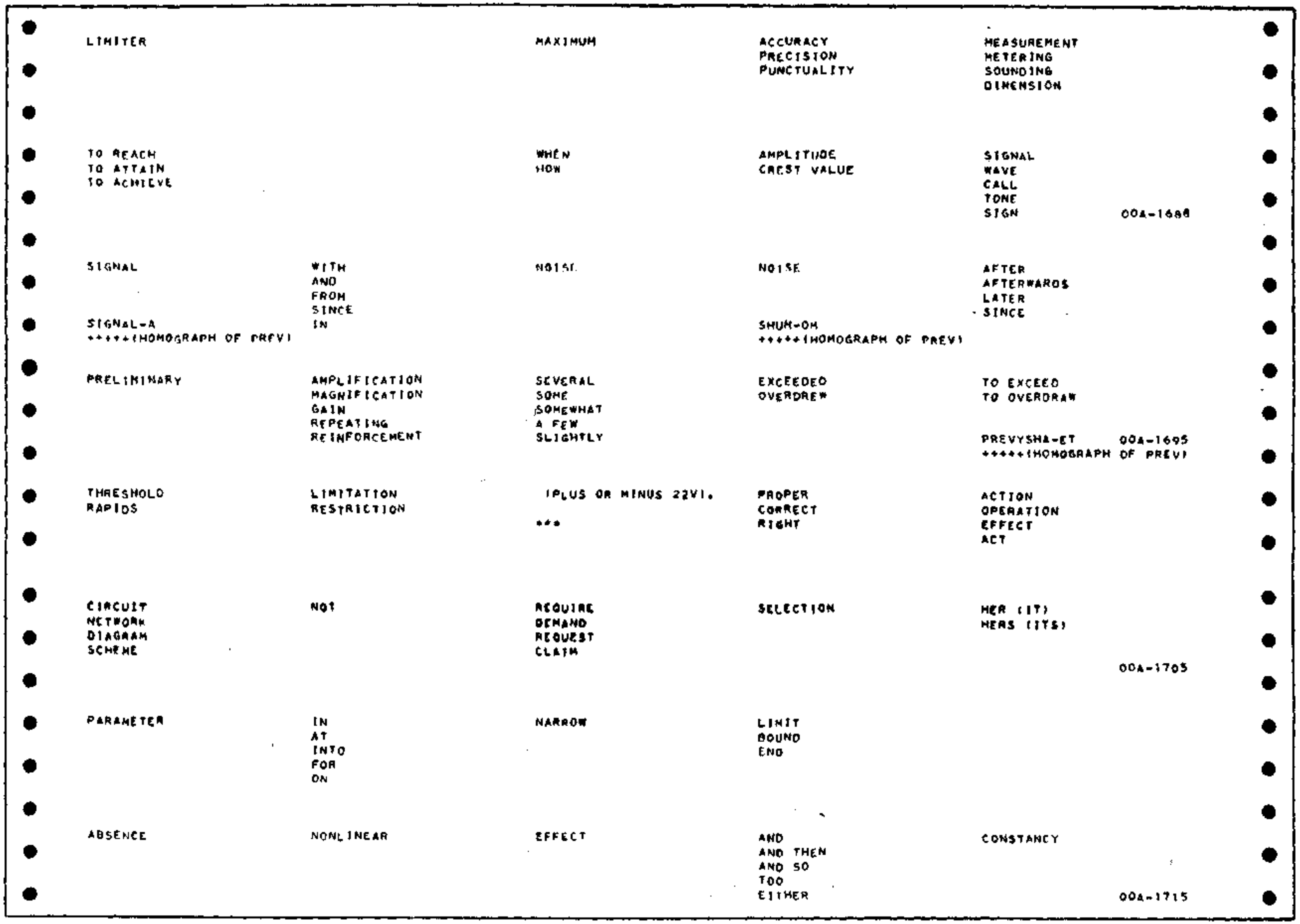
The index-dictionary can be prepared in the alphabetic order of the Russian stems, as shown in fig. 7, or in the alphabetic order of the first English correspondent listed in the entry. An alphabetic index showing the Russian words present in the text but missing from the dictionary is also prepared. This index can be made in Cyrillic alphabetic order, for use by individuals who know Russian, or in the Roman alphabetic order of the transliterated Russian, for use by individuals who do not know Russian. Finally, output 6 of fig. 4 is an edited print of the augmented text, showing dictionary entries in their textual order of occurrence.

Outputs 2 through 7 of fig. 4 are produced by the main operating program of the automatic dictionary and are made for each text processed. Outputs 8 through 14 may be produced optionally by auxiliary programs. Immediate and long-range applications of some of the outputs are discussed in Section 4. More complete discussions can be found in references [3], [4], and [5].

3. The operating program of the Harvard Automatic Dictionary

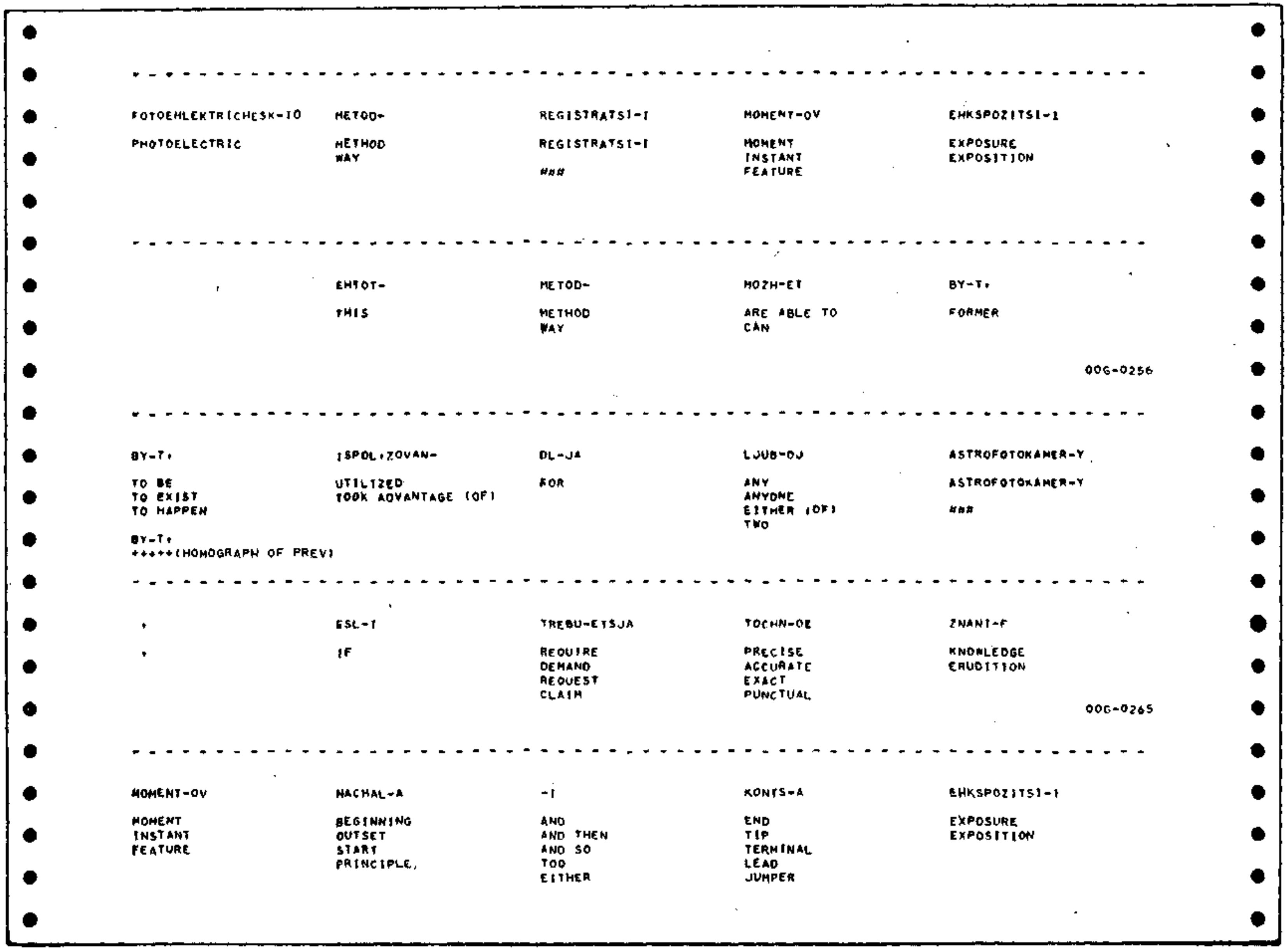
The searching of automatic dictionaries may eventually be accomplished by relatively simple special-purpose machines[2]). At present, however, fast-access information storage devices with the necessary high storage capacity are still being developed. The operating program of the Harvard Automatic Dictionary is designed for the Univac I computer at Harvard, This machine, like most other large-scale computers now operating, combines a relatively small high-speed memory with ample magnetic tape storage capacity. If a computer of this type is to be used for the operation of an automatic dictionary, certain serious problems of word-retrieval arise. As these problems have been discussed elsewhere [7], only the salient points will be mentioned here:

[2]) A promising special-purpose machine for dictionary storage and search is described in [6].



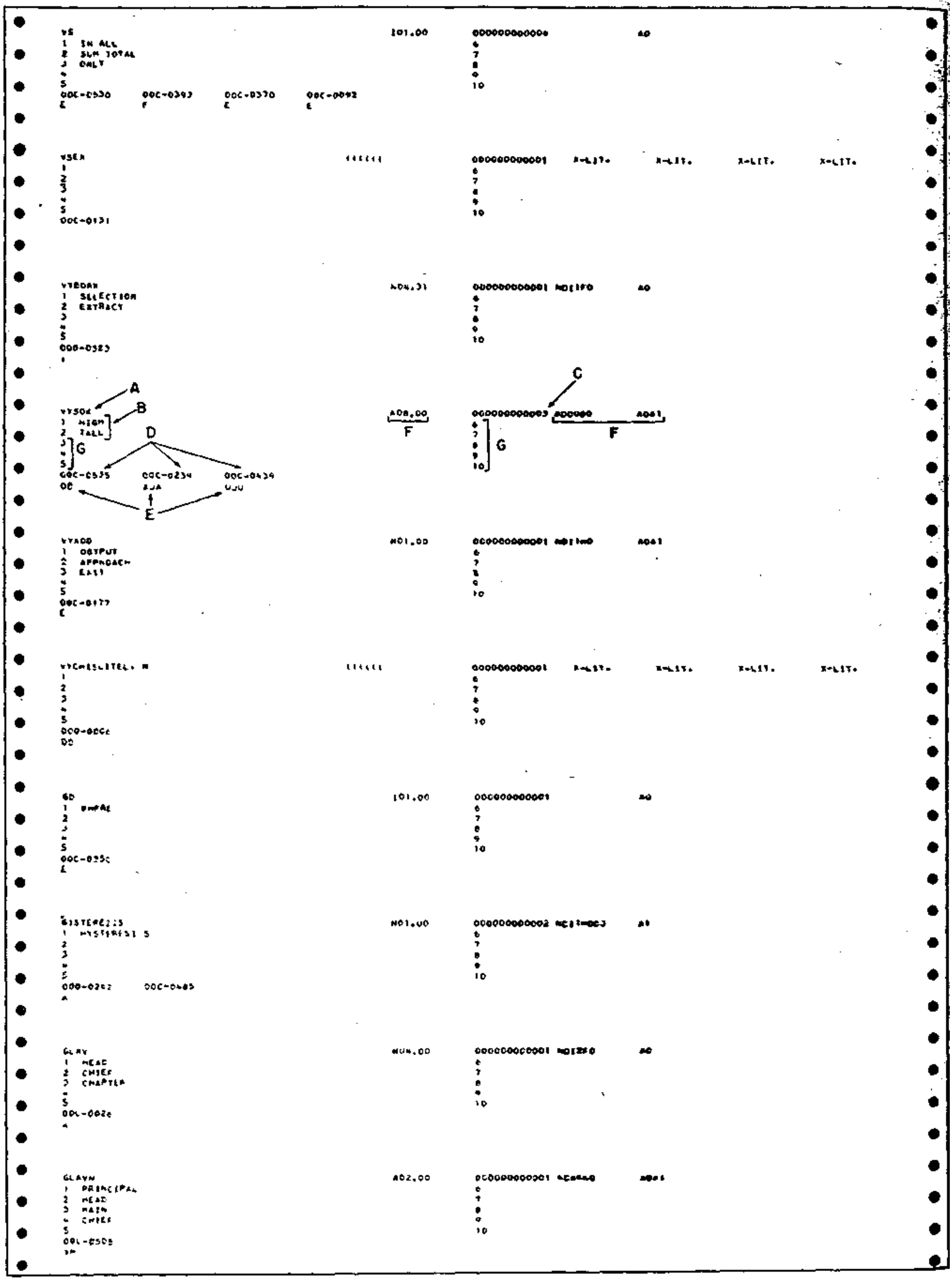
```
LIMITER                                               MAXIMUM                    ACCURACY                   MEASUREMENT
                                                                                 PRECISION                  METERING
                                                                                 PUNCTUALITY                SOUNDING
                                                                                                            DIMENSION

TO REACH                                              WHEN                       AMPLITUDE                  SIGNAL
TO ATTAIN                                             HOW                        CREST VALUE                WAVE
TO ACHIEVE                                                                                                  CALL
                                                                                                            TONE
                                                                                                            SIGN            00A-1688

SIGNAL                     WITH                       NOISE                      NOISE                      AFTER
                           AND                                                                              AFTERWARDS
                           FROM                                                                             LATER
                           SINCE                                                                            SINCE
SIGNAL-A                   IN                                                    SHUM-OM
+++++(HOMOGRAPH OF PREV)                                                         +++++(HOMOGRAPH OF PREV)

PRELIMINARY                AMPLIFICATION              SEVERAL                    EXCEEDED                   TO EXCEED
                           MAGNIFICATION              SOME                       OVERDREW                   TO OVERDRAW
                           GAIN                       SOMEWHAT
                           REPEATING                  A FEW
                           REINFORCEMENT              SLIGHTLY                                              PREVYSHA-ET     00A-1695
                                                                                                            +++++(HOMOGRAPH OF PREV)

THRESHOLD                  LIMITATION                  (PLUS OR MINUS 22V).      PROPER                     ACTION
RAPIDS                     RESTRICTION                                           CORRECT                    OPERATION
                                                      ***                        RIGHT                      EFFECT
                                                                                                            ACT

CIRCUIT                    NOT                        REQUIRE                    SELECTION                  HER (IT)
NETWORK                                               DEMAND                                                HERS (ITS)
DIAGRAM                                               REQUEST
SCHEME                                                CLAIM
                                                                                                                            00A-1705

PARAMETER                  IN                         NARROW                     LIMIT
                           AT                                                    BOUND
                           INTO                                                  END
                           FOR
                           ON

ABSENCE                    NONLINEAR                  EFFECT                     AND                        CONSTANCY
                                                                                 AND THEN
                                                                                 AND SO
                                                                                 TOO
                                                                                 EITHER                                     00A-1715
```

Fig. 5. Word-by-word translation



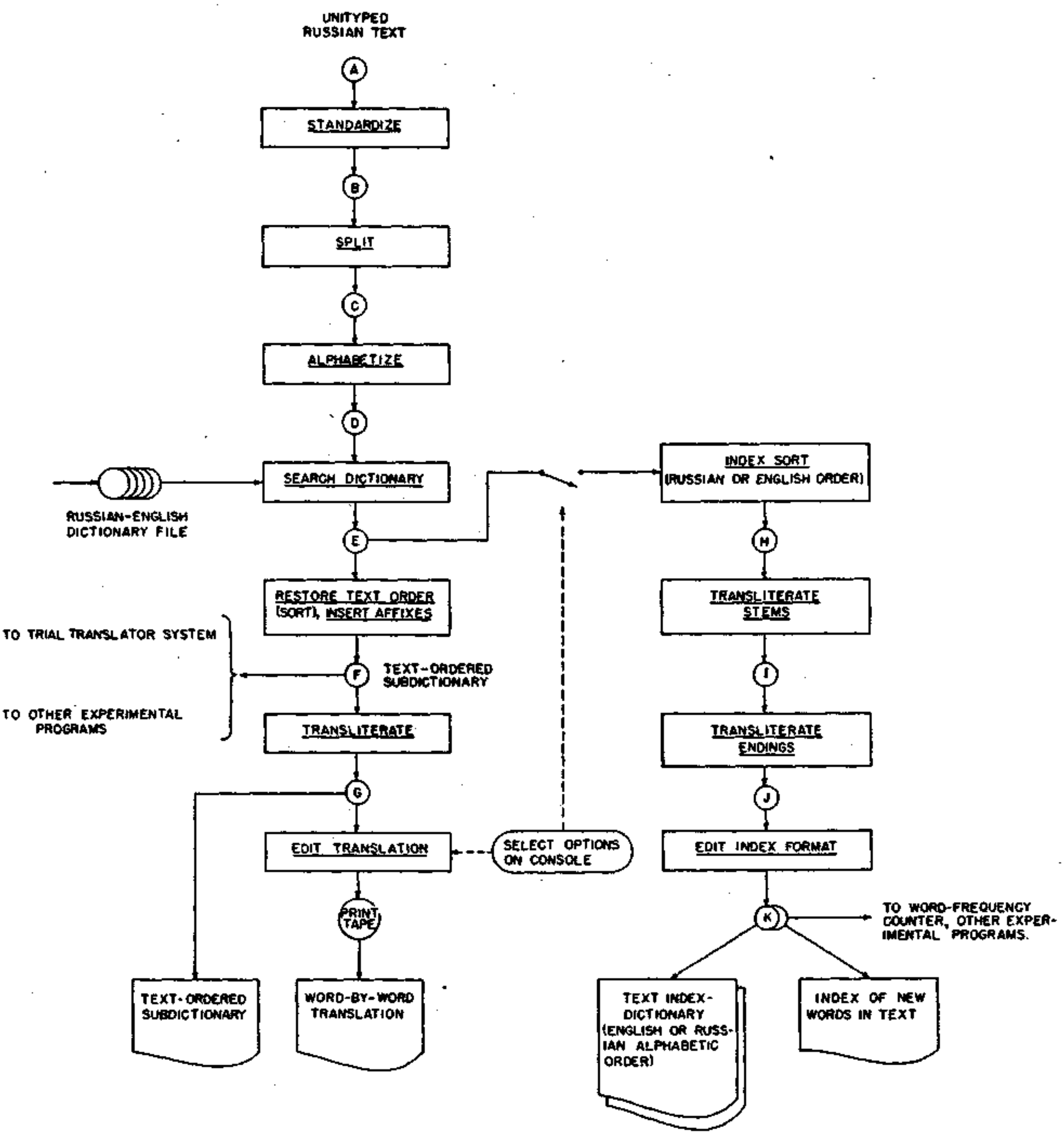
```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

FOTOEHLEKTRICHESK-IO     METOD-                   REGISTRATSI-I            MOMENT-OV                EHKSPOZITSI-I

PHOTOELECTRIC            METHOD                   REGISTRATSI-I            MOMENT                   EXPOSURE
                         WAY                                               INSTANT                  EXPOSITION
                                                  ###                      FEATURE



- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

                         EHTOT-                   METOD-                   MOZH-ET                  BY-T+

                         THIS                     METHOD                   ARE ABLE TO              FORMER
                                                  WAY                      CAN

                                                                                                         00G-0256

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

BY-T+                    ISPOL+ZOVAN-             DL-JA                    LJUB-OJ                  ASTROFOTOKAMER-Y

TO BE                    UTILIZED                 FOR                      ANY                      ASTROFOTOKAMER-Y
TO EXIST                 TOOK ADVANTAGE (OF)                               ANYONE
TO HAPPEN                                                                  EITHER (OF)              ###
                                                                           TWO
BY-T+
+++++(HOMOGRAPH OF PREV)

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 +                       ESL-I                    TREBU-ETSJA              TOCHN-OE                 ZNANI-F

 +                       IF                       REQUIRE                  PRECISE                  KNOWLEDGE
                                                  DEMAND                   ACCURATE                 ERUDITION
                                                  REQUEST                  EXACT
                                                  CLAIM                    PUNCTUAL
                                                                                                         00G-0265

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

MOMENT-OV                NACHAL-A                 -I                       KONTS-A                  EHKSPOZITSI-I

MOMENT                   BEGINNING                AND                      END                      EXPOSURE
INSTANT                  OUTSET                   AND THEN                 TIP                      EXPOSITION
FEATURE                  START                    AND SO                   TERMINAL
                         PRINCIPLE,               TOO                      LEAD
                                                  EITHER                   JUMPER
```

Fig. 6. Interlineated Russian-English word-by-word translation



```
VS                                              101.00   000000000004          A0
1  IN ALL                                                6
2  SUM TOTAL                                             7
3  ONLY                                                  8
4                                                        9
5                                                        10
00C-0530      00C-0393      00C-0370      00C-0092
E             F             E             E


VSEH                                        ((((((       000000000001   X-LIT.     X-LIT.     X-LIT.     X-LIT.
1                                                        6
2                                                        7
3                                                        8
4                                                        9
5                                                        10
00C-0131


VYBOR                                           A04.31   000000000001 NOEIFO   A0
1  SELECTION                                             6
2  EXTRACT                                               7
3                                                        8
4                                                        9
5                                                        10
000-0323
I


VYSOK                                           A08.00   000000000003 A00000   A061
1  HIGH                                                  6
2  TALL                                                  7
3                                                        8
4                                                        9
5                                                        10
00C-0575      00C-0234      00C-0434
OE            AJA           OJO


VYHOD                                           H01.00   000000000001 H0EIHO   A061
1  OUTPUT                                                6
2  APPROACH                                              7
3  EXIT                                                  8
4                                                        9
5                                                        10
00C-0177
E


VYCHISLITEL' N                              ((((((       000000000001   X-LIT.     X-LIT.     X-LIT.     X-LIT.
1                                                        6
2                                                        7
3                                                        8
4                                                        9
5                                                        10
000-006E
DO


GD                                              101.00   000000000001          A0
1  WHERE                                                 6
2                                                        7
3                                                        8
4                                                        9
5                                                        10
00C-0250
E


GISTEREZIS                                      H01.00   000000000002 NE13=0C3 A1
1  HYSTERESI S                                           6
2                                                        7
3                                                        8
4                                                        9
5                                                        10
000-0262      00C-0485
A


GLAV                                            H04.00   000000000001 H0EZF0   A0
1  HEAD                                                  6
2  CHIEF                                                 7
3  CHAPTER                                               8
4                                                        9
5                                                        10
00C-0026
A


GLAVN                                           A02.00   000000000001 AEA4H0   A061
1  PRINCIPAL                                             6
2  HEAD                                                  7
3  MAIN                                                  8
4  CHIEF                                                 9
5                                                        10
00C-0505
JP
```

Fig. 7. Alphabetic index-dictionary of words in a particular text


UNITYPED RUSSIAN TEXT
A
STANDARDIZE
B
SPLIT
C
ALPHABETIZE
D
SEARCH DICTIONARY
RUSSIAN-ENGLISH DICTIONARY FILE
E
RESTORE TEXT ORDER (SORT), INSERT AFFIXES
TO TRIAL TRANSLATOR SYSTEM
TO OTHER EXPERIMENTAL PROGRAMS
F
TEXT-ORDERED SUBDICTIONARY
TRANSLITERATE
G
EDIT TRANSLATION
SELECT OPTIONS ON CONSOLE
PRINT TAPE
TEXT-ORDERED SUBDICTIONARY
WORD-BY-WORD TRANSLATION
INDEX SORT (RUSSIAN OR ENGLISH ORDER)
H
TRANSLITERATE STEMS
I
TRANSLITERATE ENDINGS
J
EDIT INDEX FORMAT
K
TO WORD-FREQUENCY COUNTER, OTHER EXPERIMENTAL PROGRAMS.
TEXT INDEX-DICTIONARY (ENGLISH OR RUSSIAN ALPHABETIC ORDER)
INDEX OF NEW WORDS IN TEXT

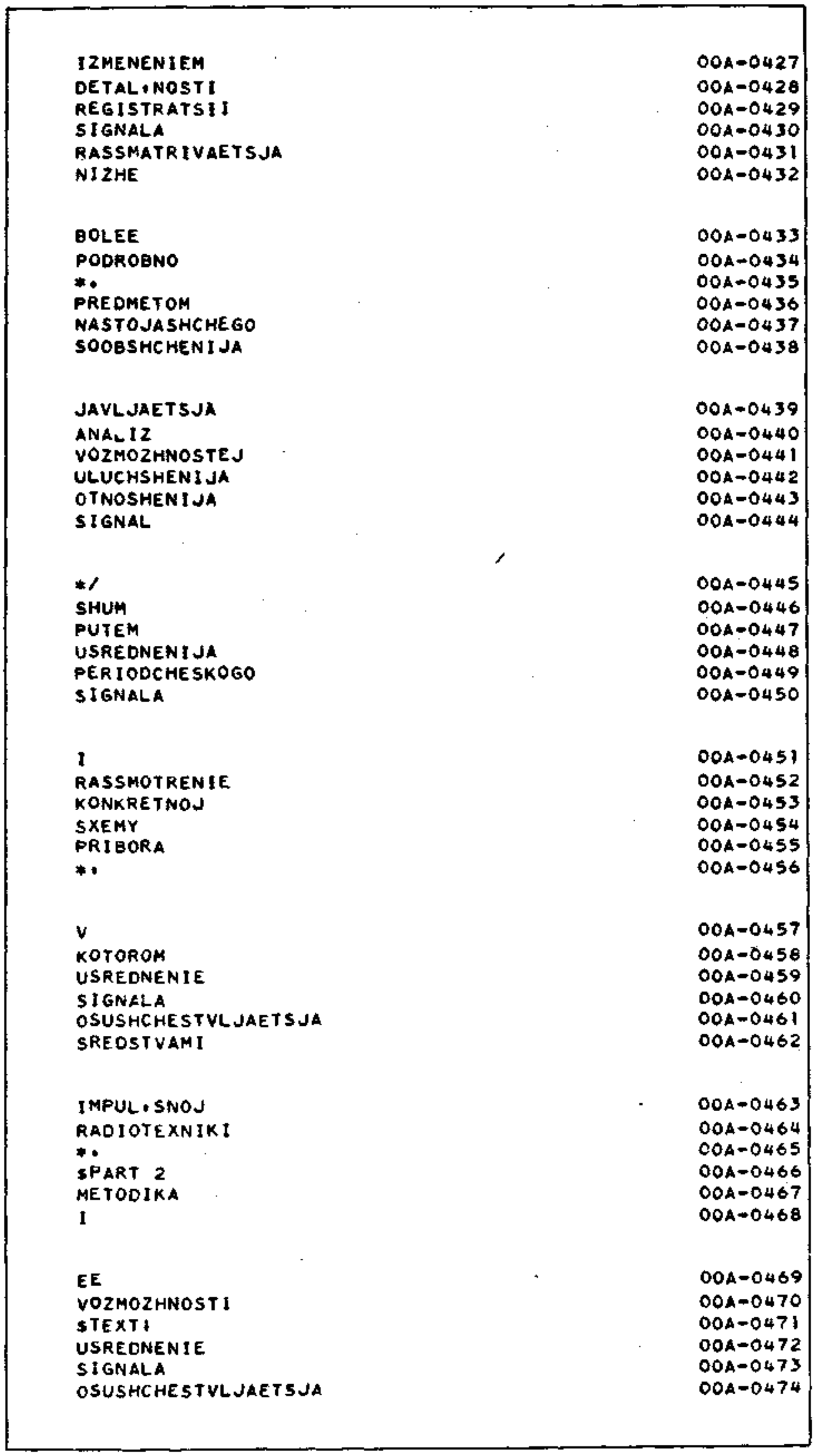
Kg. 8. The operating program of the Harvard Automatic Dictionary

a) An information-store of well over ten million bits is required to hold a sizable Russian-English dictionary. The thousand-word operating memory of the Univac is far too small to hold a significant portion of such a dictionary; the necessary storage can be provided only by several reels of magnetic tape. Dictionary search must therefore involve the mechanical motion of reels of tape.

b) The word-sequence in a Russian text defines a poor ordering for purposes of dictionary search. It is not practical to search for individual Russian words in a magnetic tape dictionary, in the sequence in which they occur in a running text. This procedure would require repositioning a dictionary tape for each word, with a delay of perhaps several minutes per word.

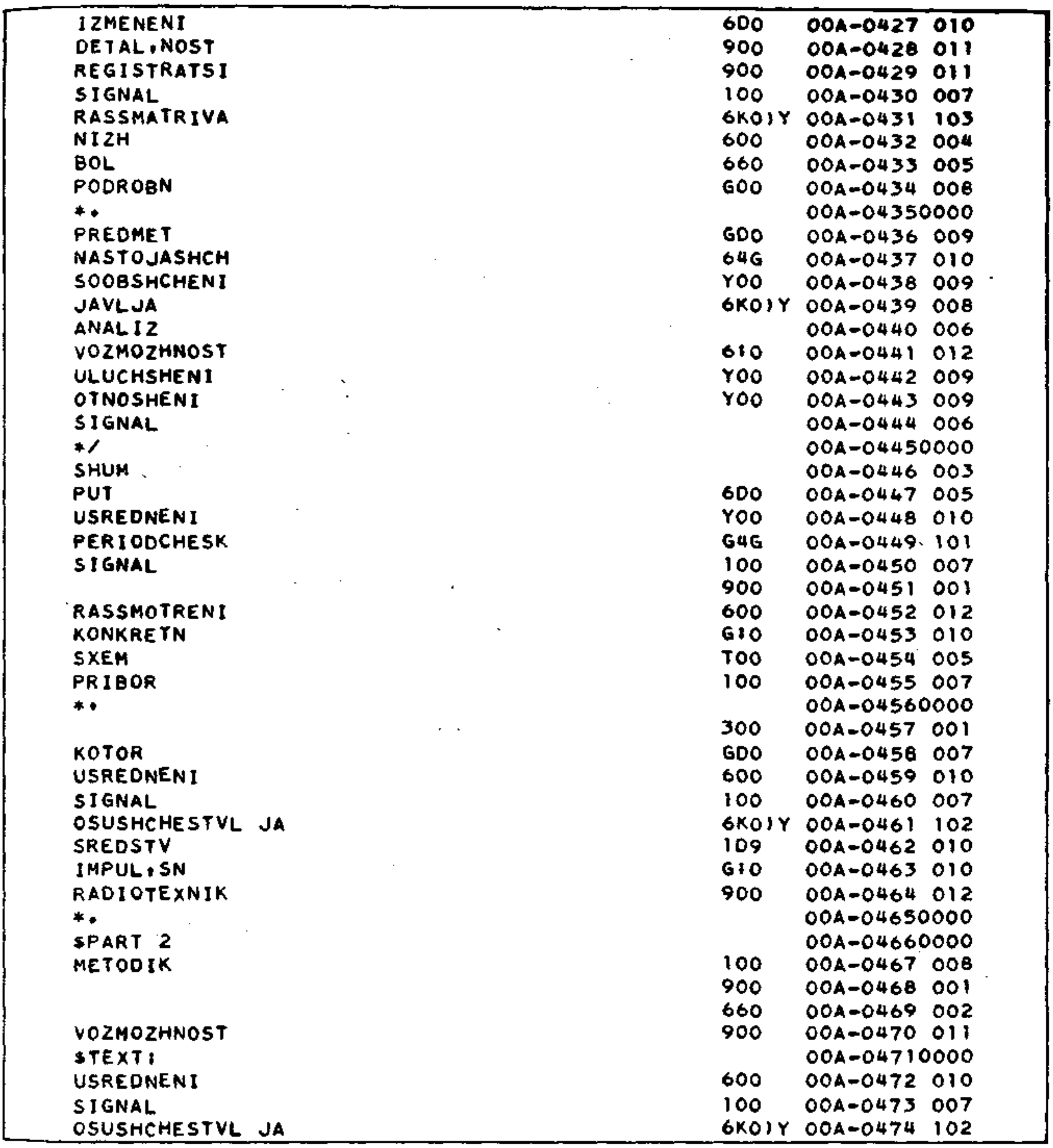
c) If a magnetic tape dictionary is to be used, it is necessary to sort the text prior to dictionary search. The entries in the dictionary file are arranged in normal Russian alphabetic order. When a text is processed, the words are first numbered in textual order, and then sorted into dictionary order. Dictionary search consists of a comparison of the alphabetic text list with the alphabetic dictionary file. Since the lists are arranged in the same order, the search requires only a single pass through the dictionary. The entries containing the English correspondents must finally be re-sorted, to restore the original textual order defined by the word numbers.

The limited size of the Univac memory also requires the operating program to be subdivided into routines that fit in the high-speed memory. The operating program is block-



```
IZMENENIEM                         00A-0427
DETAL+NOSTI                        00A-0428
REGISTRATSII                       00A-0429
SIGNALA                            00A-0430
RASSMATRIVAETSJA                   00A-0431
NIZHE                              00A-0432

BOLEE                              00A-0433
PODROBNO                           00A-0434
*.                                 00A-0435
PREDMETOM                          00A-0436
NASTOJASHCHEGO                     00A-0437
SOOBSHCHENIJA                      00A-0438

JAVLJAETSJA                        00A-0439
ANALIZ                             00A-0440
VOZMOZHNOSTEJ                      00A-0441
ULUCHSHENIJA                       00A-0442
OTNOSHENIJA                        00A-0443
SIGNAL                             00A-0444

*/                                 00A-0445
SHUM                               00A-0446
PUTEM                              00A-0447
USREDNENIJA                        00A-0448
PERIODCHESKOGO                     00A-0449
SIGNALA                            00A-0450

I                                  00A-0451
RASSMOTRENIE                       00A-0452
KONKRETNOJ                         00A-0453
SXEMY                              00A-0454
PRIBORA                            00A-0455
*.                                 00A-0456

V                                  00A-0457
KOTOROM                            00A-0458
USREDNENIE                         00A-0459
SIGNALA                            00A-0460
OSUSHCHESTVLJAETSJA                00A-0461
SREDSTVAMI                         00A-0462

IMPUL+SNOJ                         00A-0463
RADIOTEXNIKI                       00A-0464
*.                                 00A-0465
$PART 2                            00A-0466
METODIKA                           00A-0467
I                                  00A-0468

EE                                 00A-0469
VOZMOZHNOSTI                       00A-0470
$TEXT$                             00A-0471
USREDNENIE                         00A-0472
SIGNALA                            00A-0473
OSUSHCHESTVLJAETSJA                00A-0474
```

Fig. 9. Standardized and numbered text words (Tape B)



```
IZMENENI              6D0    00A-0427 010
DETAL,NOST            900    00A-0428 011
REGISTRATSI           900    00A-0429 011
SIGNAL                100    00A-0430 007
RASSMATRIVA           6KO)Y  00A-0431 103
NIZH                  600    00A-0432 004
BOL                   660    00A-0433 005
PODROBN               G00    00A-0434 008
*.                           00A-04350000
PREDMET               GD0    00A-0436 009
NASTOJASHCH           64G    00A-0437 010
SOOBSHCHENI           Y00    00A-0438 009
JAVLJA                6KO)Y  00A-0439 008
ANALIZ                       00A-0440 006
VOZMOZHNOST           610    00A-0441 012
ULUCHSHENI            Y00    00A-0442 009
OTNOSHENI             Y00    00A-0443 009
SIGNAL                       00A-0444 006
*/                           00A-04450000
SHUM                         00A-0446 003
PUT                   6D0    00A-0447 005
USREDNENI             Y00    00A-0448 010
PERIODCHESK           G4G    00A-0449 101
SIGNAL                100    00A-0450 007
                      900    00A-0451 001
RASSMOTRENI           600    00A-0452 012
KONKRETN              G10    00A-0453 010
SXEM                  T00    00A-0454 005
PRIBOR                100    00A-0455 007
*.                           00A-04560000
                      300    00A-0457 001
KOTOR                 GD0    00A-0458 007
USREDNENI             600    00A-0459 010
SIGNAL                100    00A-0460 007
OSUSHCHESTVL JA       6KO)Y  00A-0461 102
SREDSTV               1D9    00A-0462 010
IMPUL,SN              G10    00A-0463 010
RADIOTEXNIK           900    00A-0464 012
*.                           00A-04650000
$PART 2                      00A-04660000
METODIK               100    00A-0467 008
                      900    00A-0468 001
                      660    00A-0469 002
VOZMOZHNOST           900    00A-0470 011
$TEXT:                       00A-04710000
USREDNENI             600    00A-0472 010
SIGNAL                100    00A-0473 007
OSUSHCHESTVL JA       6KO)Y  00A-0474 102
```

Fig. 10. Stems derived from the words of fig. 9 (Tape C)

diagrammed in fig. 8. The various routines, denoted by boxes, are applied consecutively in the order indicated bv the arrows. Each routine requires a separate tape pass; the circles represent magnetic tapes containing input or output data. Since each routine contains terminal instructions sufficient to initiate the operation of the next, the whole program runs automatically. The present correspondence between routines and memory loads is not essential. Given a computer with a larger internal memory, several of the routines could be combined in a single memory load and the number of tape passes correspondingly reduced. Because of the freedom allowed in unityping, the format of the input tape is not well suited for automatic processing. The machine cannot readily locate Russian words on this tape, since the boundaries of Russian words and machine words do not necessarily coincide. The first routine therefore standardizes the format of the text once and for all. This "Standardize Routine" places each Russian word, and each punctuation mark or English comment, in a separate item five machine-words long. At the same time, each item is assigned a unique serial number defining its position in the text. A print of a section of the output tape, labelled B, is shown in fig. 9. The Russian is transliterated in this and subsequent figures to make the prints legible.

The Standardize Routine rewinds tape B immediately after it is produced, and initiates the operation of the next routine, called "Split". The Split Routine applies inverse inflection algorithms [8] to the Russian items contained in tape B and produces a tape C, containing Russian forms divided into stems and affixes. The Split Routine is a modified version of the split routine used to compile dictionary entries [1]. The same inverse inflection algorithms are used, but the split operation is inhibited for comment and punctuation items, identified by dollar signs or asterisks in their initial character positions. A section of tape C, showing split stems for the sample of fig. 9, is shown in fig. 10. Since the same algorithms are used to prepare both dictionary stems and text stems, a stem derived from a text word can be used to locate the stem entry for that word in the dictionary.

The next routine called into play is "Alphabetize", a sorting routine that arranges the items on tape C in Russian alphabetic order. Keys for this sort are the same as those used in compiling the dictionary; the Russian stem is the primary key, the affix the secondary key. The output is a tape D containing alphabetized five word items.

The next routine actually searches the dictionary. The routine simultaneously advances tape D and the dictionary tapes, comparing Russian items. Since both tapes are in the same order, they need move only in a forward direction. Each time a stem on tape D is successfully matched with a dictionary stem, an output item is prepared. The output item is the complete dictionary entry, modified to include also the text serial number and the affix occurring in the text. A transliterated section of an output tape E is shown in fig. 11.

The "Search Routine" prepares an output item for every item in tape D. If an input item does not correspond to an entry in the dictionary, a dummy entry is manufactured. For example, dummy entries for the Russian stems "март-" and "мансимальн-", which are in the text but not in the dictionary, appear in fig. 11. Special markers are used in the dummy entries, "((((((((((((" for machine identification and "X-LIT" for later visual identification. Words so treated eventually appear in the index of words missing from the dictionary. Punctuation marks and English comments are likewise incorporated in dummy output items.

The Search Routine has provisions for reading through any number of consecutive reels of dictionary tape at the approximate rate of one reel in four minutes. This is a little more time than is required for the mechanical motion of the tape alone. The read time does not depend on the length of the text nor on the number of words found in the dictionary. Since the dictionary reels are mounted on two tape mechanisms that are read alternately, there is no loss of computer time due to reel changing.

Tape E is the input to the "Restore Text Order Routine". This routine sorts the text, now consisting of complete dictionary entries, back into the original text order. The sort key is the serial number assigned by the Standardize Routine. Since the sorting instructions treat regular entries and dummy entries alike, punctuation marks and comments are restored to their original positions. The Sort Routine performs an auxiliary function during the first tape pass. The affixes split from Russian words prior to the alphabetization of the text are reinserted next to the stems, with hyphens interposed to show where the words were split. The output of the sort assembly, tape F, contains the augmented text (text-ordered subdictionary). The augmented text tape is retained for later use in the automatic production of improved translations.

The next routine in the dictionary operating program transliterates the hyphenated Russian words in tape F into Roman characters, to make the words readable on the output prints. The final routine in the main portion of the operating program edits the information in the augmented text into the output formats of figs. 5 and 6. The two edited versions of the word-by-word translation, with and without the transliterated Russian, are generated simultaneously on separate tapes.

When the augmented text and the word-by-word translations have been produced, the computer operator may either terminate the dictionary run or allow it to continue and to produce index-dictionaries like that shown in fig.7. The routines in the index-producing part of the program have been described elsewhere [4, 5].

4. Post-editing the dictionary outputs

The most immediate application of the automatic dictionary is the production of the output prints of fig. 5, 6, and 7 for use as aids in translating Russian scientific and technical texts. A word-by-word translation can be converted into a smooth and idiomatic translation by a post-editor, who works directly on the machine-produced prints. A sample page of post-edited material is shown in fig. 12 The post-editor has indicated with arrows his choice of English correspondents and of word-sequence. Besides drawing arrows, a post-editor occasionally inserts words, modifies them, or supplies correspondents either missing from the dictionary or else better than those printed out A typist can readily transcribe a post-edited text into conventional format; she simply copies the words at the heads of the arrows in the sequence indicated. A section of typist's copy including the material shown in fig. 12 is shown in fig. 13.

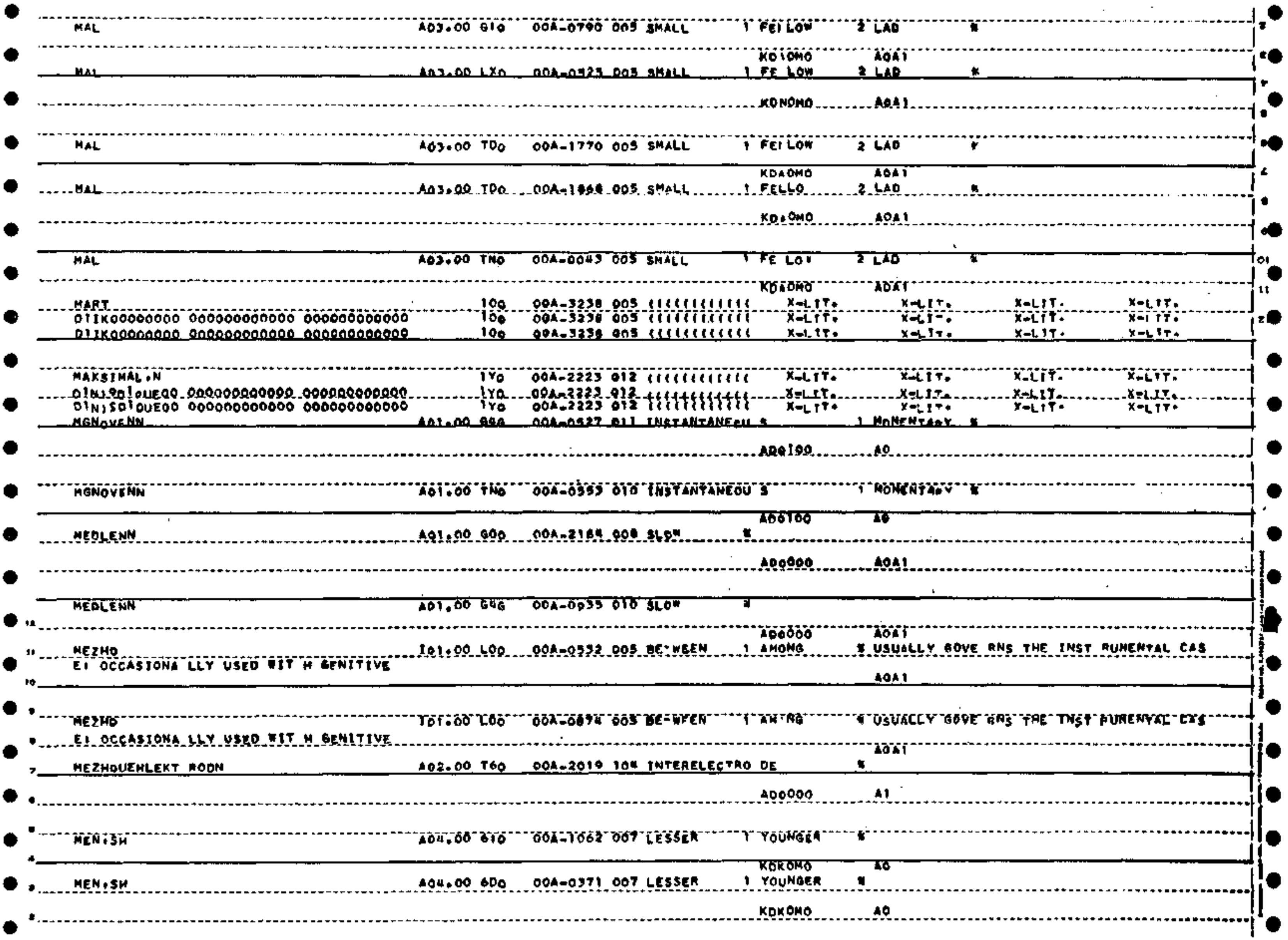
Experiments are being conducted to find out the value of the dictionary outputs to translators, scientists, and students of technical Russian. At the time of writing, experimental word-by-word translations have been made for more than two-dozen short texts, each one-thousand to three-thousand words in length. The texts were selected for machine processing by volunteer post-editors, mostly graduate students in the sciences, and treat technical sure jects of interest to the participating individuals. Each volunteer was given an interlineated word-by-word translation, like that shown in fig. 6, a set of post-editing instructions, and a set of index-dictionaries. The volunteers were asked to post-edit the texts, and to furnish suggestions for improving the dictionary outputs.

The experimental post-edited translations have not yet been systematically evaluated. Preliminary observation indicates that the outputs are much more useful to some post-editors than to others, depending largely on their backgrounds and abilities. There is some evidence to support the following preliminary conclusions:

a) The outputs of the automatic dictionary can be very useful to an individual having a scientific or technical background, a knowledge of the rudiments of Russian, and, above all, a desire to read Russian technical material in his own field. Indications are that such a person can understand and translate texts much more rapidly when dictionary searching is done for him automatically.

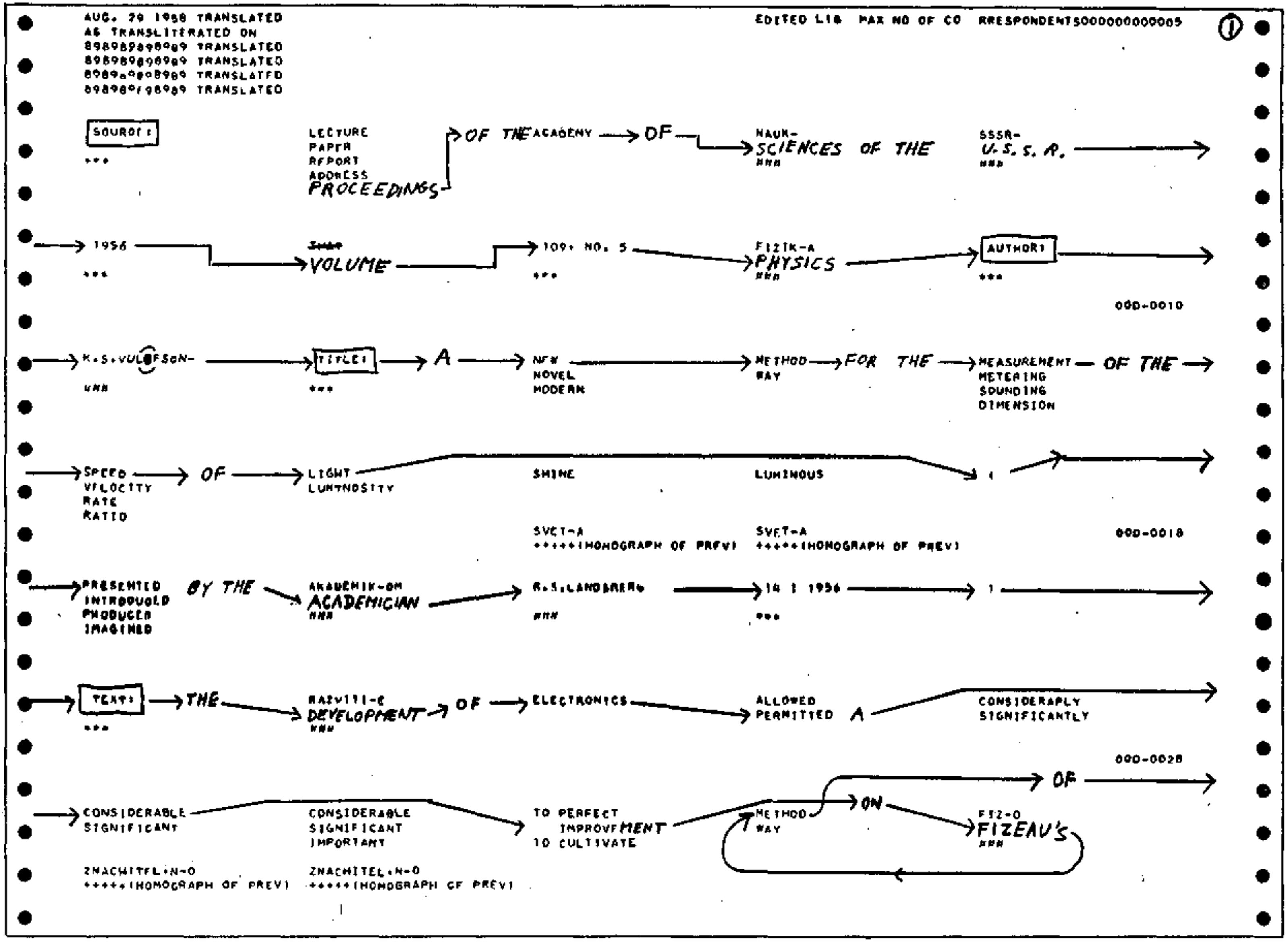
b) Capable and technically qualified individuals who have never studied Russian can produce passable translations of texts selected from their own fields if they are sufficiently interested in doing so. The post-editing process is typically more time-consuming for these individuals than for those who know a little Russian, and a few sentences are usually left only partially translated. All volunteers were, however, able to complete their translation in a matter of hours, whereas days might have been required otherwise. Significantly, the complaint registered most often was that five or ten percent of the words in the text were not yet in the automatic dictionary, and had to be located in a standard desk dictionary.

c) The special technical vocabulary of the automatic dictionary makes it marginally useful to individuals who



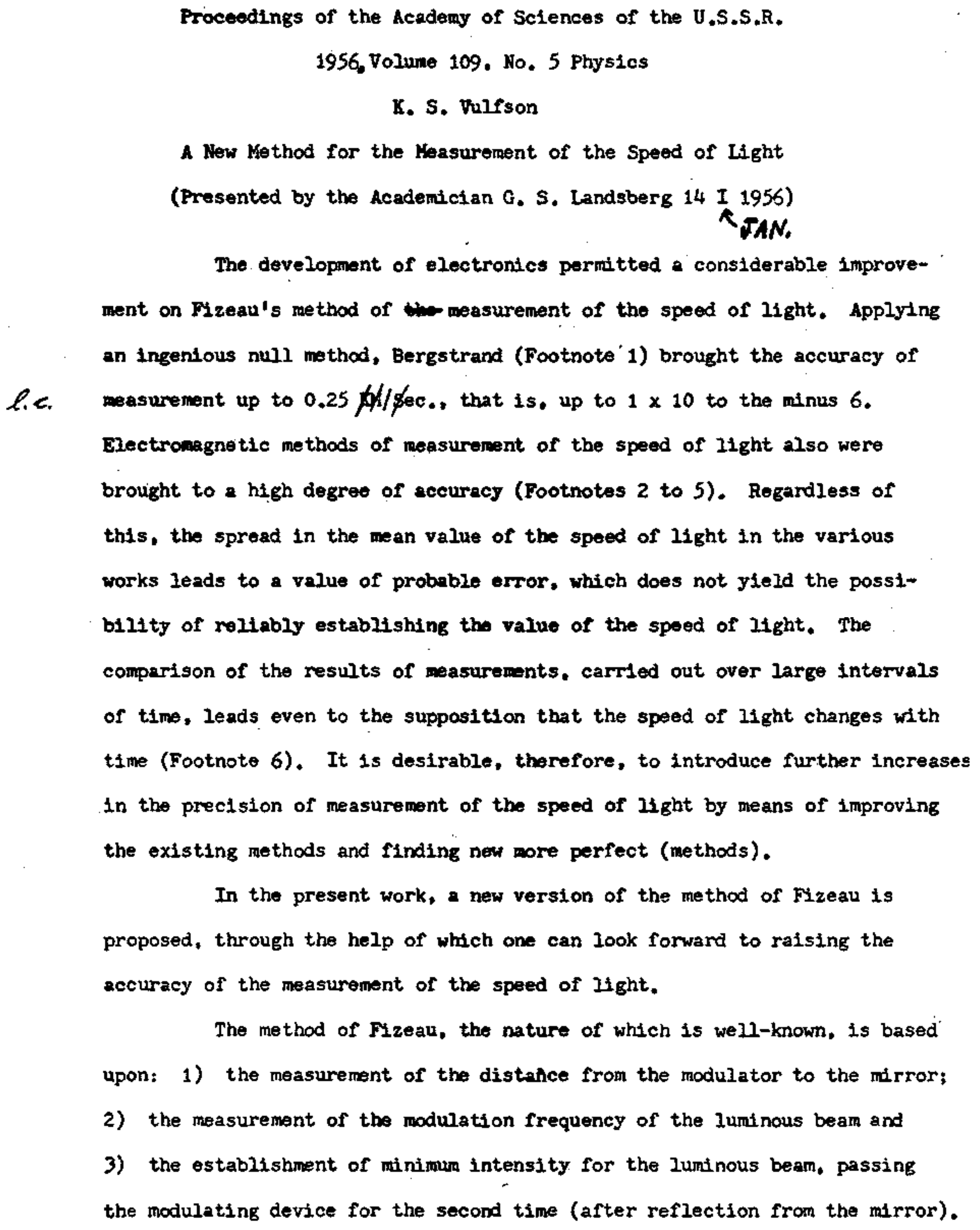
```
MAL                                        A03.00 G10   00A-0790 005 SMALL       1 FEI LOW     2 LAD        *
                                                                                 KO1OMO        AOA1
MAL                                        A03.00 LX0   00A-0925 005 SMALL       1 FE LOW      2 LAD        *
                                                                                 KONOMO        AOA1

MAL                                        A03.00 TD0   00A-1770 005 SMALL       1 FEI LOW     2 LAD        Y
                                                                                 KDAOMO        AOA1
MAL                                        A03.00 TD0   00A-1868 005 SMALL       1 FELLO       2 LAD        *
                                                                                 KD.OMO        AOA1

MAL                                        A03.00 TN0   00A-0043 005 SMALL       1 FE LO*      2 LAD        *
                                                                                 KDAOMO        ADA1
MART                                            100     00A-3238 005 ((((((((((((   X-LIT.        X-LIT.        X-LIT.        X-LIT.
 O1IK00000000 000000000000 000000000000         100     00A-3238 005 ((((((((((((   X-LIT.        X-LI-.        X-LIT.        X-LIT.
 O1IK00000000 000000000000 000000000000         100     00A-3238 005 ((((((((((((   X-LIT.        X-LIT.        X-LIT.        X-LIT.

MAKSIMAL.N                                      1Y0     00A-2223 012 ((((((((((((   X-LIT.        X-LIT.        X-LIT.        X-LIT.
 O1N)90!OUE00 000000000000 000000000000         1Y0     00A-2223 012 ((((((((((((   X-LIT.        X-LIT.        X-LIT.        X-LIT.
 O1N)SO!OUE00 000000000000 000000000000         1Y0     00A-2223 012 ((((((((((((   X-LIT.        X-LIT.        X-LIT.        X-LIT.
MGNOVENN                                   A01.00 880   00A-0527 011 INSTANTANEOU S             1 MOMENTARY   *
                                                                                 AD0190        A0

MGNOVENN                                   A01.00 TN0   00A-0353 010 INSTANTANEOU S             1 MOMENTARY   *
                                                                                 AD0100        A0
MEDLENN                                    A01.00 G00   00A-2164 008 SLOW      *
                                                                                 AD0000        AOA1

MEDLENN                                    A01.00 G4G   00A-0935 010 SLOW      *
                                                                                 AD0000        AOA1
MEZHD                                      I01.00 L00   00A-0532 005 BE-WEEN     1 AMONG       * USUALLY GOVE RNS THE INST RUMENTAL CAS
E1 OCCASIONA LLY USED WIT H GENITIVE
                                                                                               AOA1

MEZHD                                      I01.00 L00   00A-0874 005 BE-WEEN     1 AMONG       * USUALLY GOVE RNS THE INST RUMENTAL CAS
E1 OCCASIONA LLY USED WIT H GENITIVE
                                                                                               AOA1
MEZHDUEHLEKT RODN                          A02.00 T60   00A-2019 10* INTERELECTRO DE           *
                                                                                 AD0000        A1

MEN.SH                                     A04.00 610   00A-1062 007 LESSER      1 YOUNGER     *
                                                                                 KOKOMO        A0
MEN.SH                                     A04.00 6D0   00A-0371 007 LESSER      1 YOUNGER     *
                                                                                 KOKOMO        A0
```

Fig. 11. Alphabetic output of the Dictionary Search Run (Tape E)



```
AUG. 29 1958 TRANSLATED                                              EDITED LIB  MAX NO OF CO  RRESPONDENTS000000000005  1
AS TRANSLITERATED ON
898989898989 TRANSLATED
898989898989 TRANSLATED
8989a9@08989 TRANSLATFD
898989f98989 TRANSLATED

SOURCE:              LECTURE         OF THE ACADEMY  OF   NAUK-                   SSSR-
***                  PAPER                                SCIENCES OF THE         U.S.S.R.
                     REPORT                               ###                     ###
                     ADDRESS
                     PROCEEDINGS

1956                 ~~SHAH~~        109, NO. 5           FIZIK-A                 AUTHOR:
***                  VOLUME          ***                  PHYSICS                 ***
                                                          ###
                                                                                              000-0010

K.S.VULOFSON-        TITLE:  A       NEW                  METHOD  FOR THE         MEASUREMENT  OF THE
###                  ***             NOVEL                WAY                     METERING
                                     MODERN                                       SOUNDING
                                                                                  DIMENSION

SPEED     OF         LIGHT           SHINE                LUMINOUS                ,
VELOCITY             LUMINOSITY
RATE
RATIO
                                     SVET-A               SVET-A                              000-0018
                                     +++++(HOMOGRAPH OF PREV)  +++++(HOMOGRAPH OF PREV)

PRESENTED  BY THE    AKADEMIK-OM     R.S.LANDSBERG        14 1 1956               1
INTRODUCED           ACADEMICIAN     ###                  ***
PRODUCED             ###
IMAGINED

TEXT:  THE           RAZVITI-E   OF  ELECTRONICS          ALLOWED   A             CONSIDERAPLY
***                  DEVELOPMENT                          PERMITTED               SIGNIFICANTLY
                     ###
                                                                                              000-0028
                                                                                  OF
CONSIDERABLE         CONSIDERABLE    TO PERFECT           METHOD  ON              FIZ-O
SIGNIFICANT          SIGNIFICANT       IMPROVEMENT        WAY                     FIZEAU'S
                     IMPORTANT       TO CULTIVATE                                 ###

ZNACHITEL.N-O        ZNACHITEL.N-O
+++++(HOMOGRAPH OF PREV)  +++++(HOMOGRAPH OF PREV)
```

Fig. 12. Post-edited word-by-word translation



Proceedings of the Academy of Sciences of the U.S.S.R.

1956, Volume 109, No. 5 Physics

K. S. Vulfson

A New Method for the Measurement of the Speed of Light

(Presented by the Academician G. S. Landsberg 14 I 1956)

JAN.

The development of electronics permitted a considerable improvement on Fizeau's method of ~~the~~ measurement of the speed of light. Applying an ingenious null method, Bergstrand (Footnote 1) brought the accuracy of measurement up to 0.25 M/sec., that is, up to 1 x 10 to the minus 6.

l.c.

Electromagnetic methods of measurement of the speed of light also were brought to a high degree of accuracy (Footnotes 2 to 5). Regardless of this, the spread in the mean value of the speed of light in the various works leads to a value of probable error, which does not yield the possibility of reliably establishing the value of the speed of light. The comparison of the results of measurements, carried out over large intervals of time, leads even to the supposition that the speed of light changes with time (Footnote 6). It is desirable, therefore, to introduce further increases in the precision of measurement of the speed of light by means of improving the existing methods and finding new more perfect (methods).

In the present work, a new version of the method of Fizeau is proposed, through the help of which one can look forward to raising the accuracy of the measurement of the speed of light.

The method of Fizeau, the nature of which is well-known, is based upon: 1) the measurement of the distance from the modulator to the mirror; 2) the measurement of the modulation frequency of the luminous beam and 3) the establishment of minimum intensity for the luminous beam, passing the modulating device for the second time (after reflection from the mirror).

Fig. 13. Transcribed edited word-by-word translation

know literary Russian and English well, but who have little or no technical background in the subject of the material translated. A person with a technical background must still scan the post-edited translations to pick out errors due to technical ignorance. However, the technical editor need not know Russian since he can consult the machine-produced prints whenever doubts arise about the edited translation.

d) There is a small number of expert technical translators who have no need for the automatic dictionary outputs. These individuals are senior research scientists having excellent command of scientific Russian and English, and extensive experience in technical writing. Usually, a translator of this caliber can rapidly write or dictate a final translation of a technical paper in his field while reading the Russian original. He is more likely to be hampered than assisted by the automatic dictionary outputs in their present form. Unfortunately, the number of individuals in this category is very small and few of those properly qualified can take time from their scientific work to do a significant amount of translating.

e) Individuals without technical qualifications, or who have difficulty in expressing themselves in English, cannot produce good translations, with or without the aid of the automatic dictionary. A successful post-editor must have a good command of English and some background in the technical field of the material translated. Preferably, he will also have some knowledge of basic Russian.

f) The present format of the word-by-word translations is an important source of difficulties. Specifically, more text must be put on each page to avoid the need for time consuming and irritating turning of pages.

The difficulties in evaluating translation were found to be formidable, corroborating the views of Miller and Beebe-Center [9] on this subject.

The dictionary outputs will be used as experimental teaching aids in a Russian course at Harvard in the spring of 1959. In response to a questionnaire circulated during fall registration, forty of the students taking the course expressed an interest in using the automatic dictionary. The interested individuals are all science majors or graduate students in the sciences, and many of them are taking the course to learn how to read scientific Russian. Each student will be asked to select a text for processing and, finally, to post-edit the machine-produced translation.

New technical terms and meanings, suggested by the scientifically trained individuals who read the dictionary outputs, are being fed back into the dictionary. For example, a post-edited list of words in a text but not in the dictionary is shown in fig. 14. Each English correspondent was simply written into the first blank space provided in the index; no attempt was made to modify its case or tense. Information fed back from post-editors is also being used to modify and improve the existing entries in the dictionary. Sections taken from a post-edited index dictionary for a sample text are shown in fig. 15. The post-editor has suggested several revisions in the existing entries, mostly to make them more suitable for technical texts. The initial sets of meanings in the automatic dictionary were drawn from existing technical dictionaries [10, 11]. The suggested revisions speak for themselves. At present, new Russian words are entered into the automatic dictionary through the semi-automatic compilation procedure described in [1]. Existing entries are corrected and updated by standard tape-editing routines drawn from the Univac library.

Since the transformations made by post-editors are precisely the transformations that must eventually be automatized, the experiments in post-editing are important also from a long-range viewpoint. An automatic system that makes use of the valuable information available in the post-edited outputs is discussed in Section 5.

5. Advanced research in automatic translation

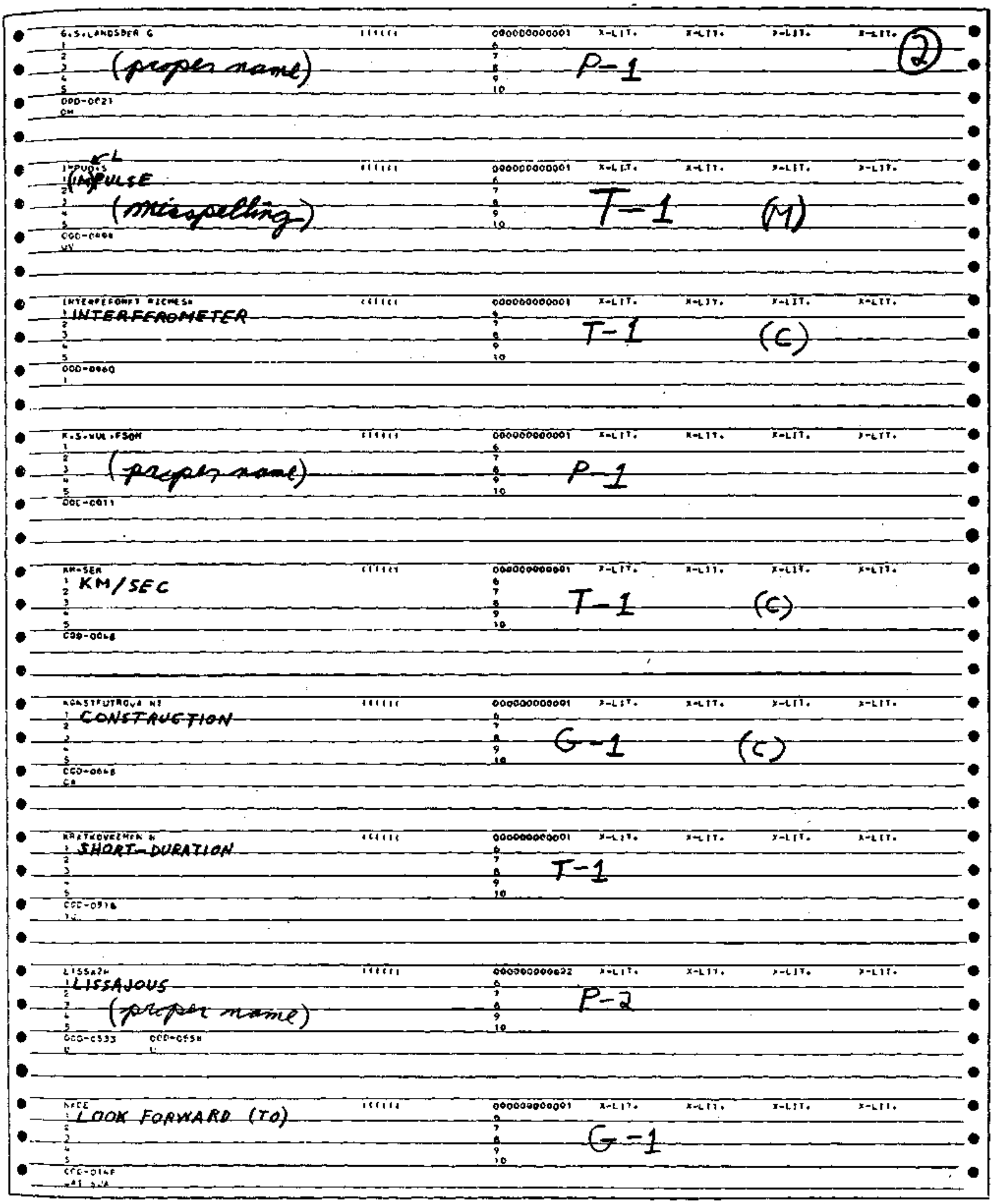
The basic experimental system to be used for advanced research on automatic translation is block-diagrammed in fig. 16. The automatic dictionary program is a fundamental part of this system; it is extended by two additional programs, the "Affix Interpreter" and the "English Inflector" Both of these programs operate on individual entries to make explicit all information about a text that can be obtained by considering individual words in isolation from their neighbors. The resolution of the ambiguities remaining in the augmented texts produced by these programs requires the examination of contexts. The program labelled "Trial Translator" is a research instrument that will enable linguists to test syntactic and semantic translation algorithms on large bodies of Russian text. The program called "Formula Finder" will make use of a primitive machine learning process for the automatic derivation of certain classes of translation algorithms.

The coded information in the dictionary entries of augmented texts is of great importance for all phases of research involving syntax. The information in each stem entry characterizes both the morphology of the Russian word containing the stem and certain properties of the stem not explicitly reflected in its morphology. As an example, the information marked "γ" in the sample entry in fig. 1 will be considered in detail. The morphological class marker, N 10.00, indicates the fact that the word притяжение belongs to a category of nouns usually accepting the inflectional affixes "е, я, ю, ем, и, й, ям, ями"and "ях". The code characters NDI1N100 convey the following information: the word functions as a noun (N), it is declinable (D), it belongs to a certain subclass of inanimate nouns (I1), it is neuter (N), it generally occurs in the singular only (1), and it has no special forms (00). The markers A0 and Al indicate that the word can be found in two standard Russian-English dictionaries [10, 11].

5.1 *Affix interpretation* [12,13]

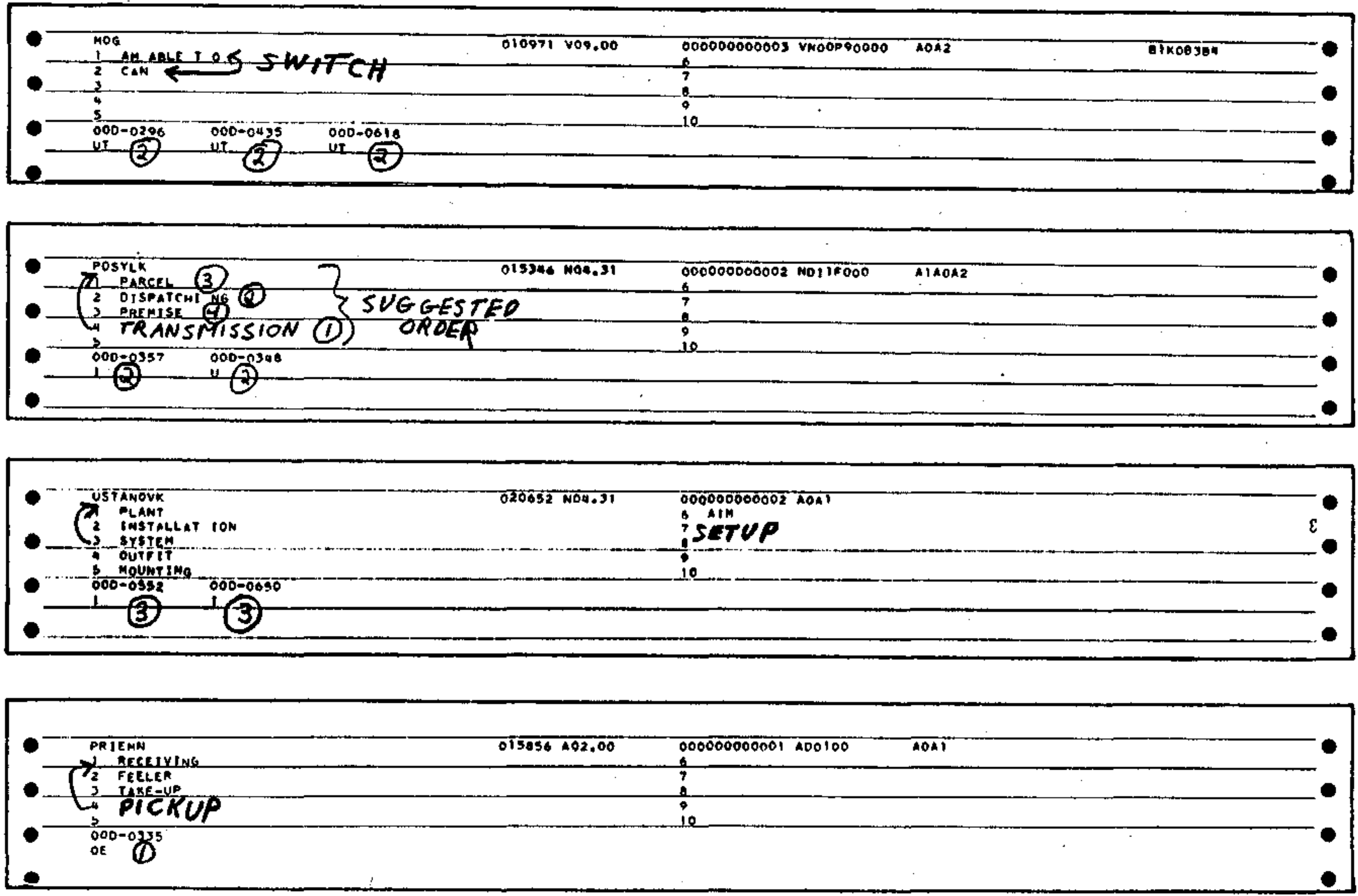
The present dictionary is a stem dictionary, and the information in an entry characterizes the stem or its paradigm rather than the distinct inflected forms that can occur in texts. To characterize a text form, it is necessary to interpret the affix associated with the stem. For example, suppose that the Russian form притяжением occurs in a text. The stem entry of fig. 1 will therefore appear in the augmented text. The coding in the entry indicates that the word is a noun, neuter, declinable, etc., but does not explicitly convey the information that the ending ем denotes the instrumental singular. This information is essential for research purposes, and a machine program is therefore being provided for the automatic interpretation of affixes in augmented texts.

The input to the affix-interpreting program will be an augmented text containing stem entries and affixes split from text words. The output of the program will be an interpreted copy of the augmented text, containing statements of the case, number, tense, etc., determined by each stem and affix. Interpretations will be based on the coded data in the entries. In the sample entry of fig. 1, the class marker "N10.00" and the subclass marker "I1" are sufficient to indicate to the program that the ending е denotes either the nominative or the accusative singular, the ending ю denotes the dative singular, the ending ем denotes the. instrumental singular, etc. The portion of the program for handling nouns is already operating; the routines for handling verbs and adjectives are currently being programmed. The diagram of fig. 16 shows that affixes in augmented texts will normally be interpreted before the texts are used as inputs to the more advanced translating programs. Words that are normally distinct in all their inflected forms sometimes lead to the same dictionary stem. For



```
G.S.LANDSBER G          IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.   2
1                                  6
2                                  7
3     (proper name)                8        P-1
4                                  9
5                                  10
000-0021
OM

       L
I*PUD.S                 IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 (IM)PULSE                        6
2                                  7
3     (misspelling)                8        T-1              (M)
4                                  9
5                                  10
000-0404
UY

INTERFEROMET RICHESK    IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 INTERFEROMETER                   6
2                                  7
3                                  8        T-1              (C)
4                                  9
5                                  10
000-0060
I

K.S.VUL.FSON            IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1                                  6
2                                  7
3     (proper name)                8        P-1
4                                  9
5                                  10
000-0011

KM-SEK                  IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 KM/SEC                           6
2                                  7
3                                  8        T-1              (C)
4                                  9
5                                  10
000-0048

KONSTRUIROVA NI         IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 CONSTRUCTION                     6
2                                  7
3                                  8        G-1              (C)
4                                  9
5                                  10
000-0668
CA

KRATKOVREMEN N          IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 SHORT-DURATION                   6
2                                  7
3                                  8        T-1
4                                  9
5                                  10
000-0916
YU

LISSAZH                 IIIIII     000000000022   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 LISSAJOUS                        6
2                                  7
3     (proper name)                8        P-2
4                                  9
5                                  10
000-0533       000-0558
U              U

NADE                    IIIIII     000000000001   X-LIT.   X-LIT.   X-LIT.   X-LIT.
1 LOOK FORWARD (TO)                6
2                                  7
3                                  8        G-1
4                                  9
5                                  10
000-0148
JAT SJA
```

Fig. 14. Post-edited index of new words in a text but not in the dictionary



```
MOG                                       010971 V09.00   000000000003 VN00P90000   A0A2            B1K0B3B4
1  AM ABLE T O  SWITCH                                    6
2  CAN                                                    7
3                                                         8
4                                                         9
5                                                         10
000-0296     000-0435     000-0618
UT  2        UT  2        UT  2
```

```
POSYLK                                    015346 N04.31   000000000002 ND11F000     A1A0A2
1  PARCEL  3                                              6
2  DISPATCHI NG 2        SUGGESTED                        7
3  PREMISE  4             ORDER                           8
4  TRANSMISSION 1                                         9
5                                                         10
000-0357     000-0348
1  2         U  2
```

```
USTANOVK                                  020652 N04.31   000000000002 A0A1
1  PLANT                                                  6  AIM
2  INSTALLAT ION                                          7  SETUP
3  SYSTEM                                                 8
4  OUTFIT                                                 9
5  MOUNTING                                               10
000-0552     000-0650
1  3         1  3
```

```
PRIEMN                                    015856 A02.00   000000000001 AD0100       A0A1
1  RECEIVING                                              6
2  FEELER                                                 7
3  TAKE-UP                                                8
4  PICKUP                                                 9
5                                                         10
000-0335
OE  1
```

Fig. 15. Section of a post-edited index-dictionary

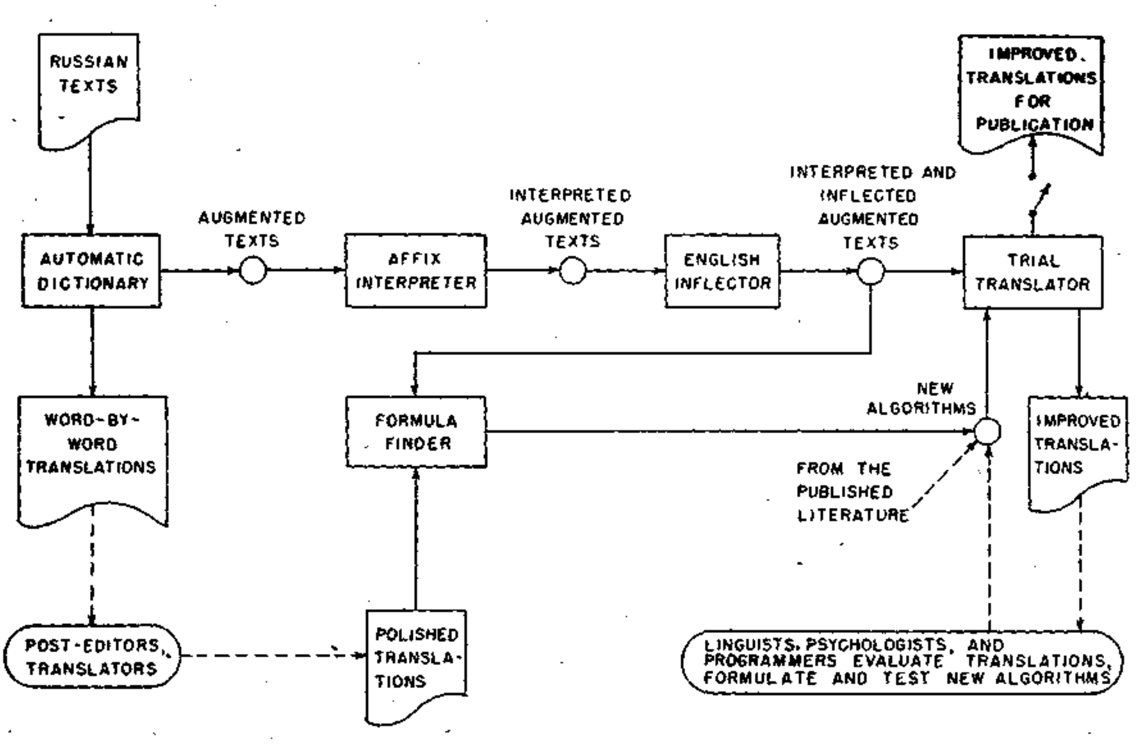

RUSSIAN TEXTS
AUTOMATIC DICTIONARY
AUGMENTED TEXTS
AFFIX INTERPRETER
INTERPRETED AUGMENTED TEXTS
ENGLISH INFLECTOR
INTERPRETED AND INFLECTED AUGMENTED TEXTS
TRIAL TRANSLATOR
IMPROVED TRANSLATIONS FOR PUBLICATION
WORD-BY-WORD TRANSLATIONS
FORMULA FINDER
NEW ALGORITHMS
FROM THE PUBLISHED LITERATURE
IMPROVED TRANSLA-TIONS
POST-EDITORS, TRANSLATORS
POLISHED TRANSLA-TIONS
LINGUISTS, PSYCHOLOGISTS, AND PROGRAMMERS EVALUATE TRANSLATIONS, FORMULATE AND TEST NEW ALGORITHMS

Fig. 16. Basic experimental system for advanced research in automatic translation

example, the Russian nouns гран and грань meaning respectively "grain" and "border," both lead to the dictionary stem гран. At present, when either word is encountered in a text, two stem entries are extracted from the dictionary, one for "grain," the other for "border." The two entries appear consecutively in the augmented text, and in the word-by-word translation. Duplicate entries of this type can be seen in the word-by-word translations of figs. 5 and 6, marked with the indicator "+ + + + + (HOMOGRAPH OF PREV)."

Once the affix-interpreting program is operating, ambiguities due to stem homographs will no longer appear in the output of the dictionary. The program will reject spurious entries containing unwanted stem homographs as a by-product of its normal operation. For example, the program will recognize the affix ом in граном as compatible only with the entry for "grain," and the entry for "border" will not appear in the output augmented text.

In a certain number of instances, of course, a definitive interpretation of affixes cannot be made on a strictly word-by-word basis. For example, the affix я in a noun of class N10.00 can denote either the genitive singular or the nominative or accusative plural, and a definitive interpretation of the affix requires an analysis of contexts. By doing the interpretation first on a word-by-word basis, however, the ambiguity is narrowed to that inherent in the actual inflected form.

5.2. *English inflection*

The English correspondents in the dictionary and augmented texts are initially expressed in certain standard inflected forms corresponding to the Russian stems. As a result, distinctions of case, number, and tense present in Russian endings are not always reflected correctly in the English correspondents of an augmented text. For example, the stem entry to which the third person singular of "делать" is referred contains the standard correspondent "to do" instead of "does." It is only natural, then, to make provisions for the automatic inflection of English correspondents in augmented texts.

Algorithms have already been developed for the automatic inflection of Russian, and have been used in the compilation of the Harvard Automatic Dictionary [1]. Since English, unlike Russian, is not a highly inflected language, the development of analogous inflectional algorithms for English should pose no fundamental problems. Research is underway in this area. An English-Russian dictionary has been automatically compiled from the Russian-English dictionary and is being used as a research tool for the investigation of the English inflectional system.

Algorithms for the automatic inflection of English will eventually be incorporated into an English-inflecting machine program, to operate on interpreted augmented texts (fig. 16). Each input entry will contain an indicator, inserted by the affix-interpreting program, specifying the case, number, tense, etc., of the inflected Russian form. The English inflector program will sense these indicators; all English correspondents in an entry will be inflected into the indicated case, number, tense, etc. Remaining ambiguities will be resolved at a later stage.

So long as it is necessary or desirable to use stem dictionaries, the affix interpreter and English inflector programs must be applied to every text that is to be translated. Should it prove feasible and worthwhile to use dictionaries in which every distinct inflected form is represented by a separate entry, these programs may be applied once and for all during dictionary compilation. The semiautomatic-methods for compiling stem dictionaries described in [1] may therefore be extended to encompass the compilation of full-paradigm dictionaries. Given the classical canonical form of a Russian word (e.g. nominative singular for nouns, infinitive for verbs) to be included in a full-paradigm dictionary, English correspondents and grammatical coding may be assigned manually, and a complete set of inflected entries may then be generated automatically.

5.3 *Trial translating*

While several syntactic and semantic rules for producing smooth Russian-English automatic translations have been proposed in the literature [14, 15, 16], published experimental results have been conspicuously absent[3]): Until very recently, the reported use of automatic machines has been confined to the applications of ad-hoc computer programs, tailored to the processing of particular sentences or of carefully selected texts. Since few if any of the algorithms proposed for Russian-English translation have been tested on large bodies of Russian text, it is exceedingly difficult, if not impossible, to evaluate them objectively.

Trial translating is the process of applying experimental translation algorithms to representative Russian texts, of examining the results, and of evaluating the algorithms. For such a process to be practical and meaningful, the algorithms must be applied by a machine. One of the writers has proposed an automatic programming system designed to put the computer readily at the disposal of linguists, Slavic scholars, and other individuals not usually trained in computer programming. This system, called the "Trial Translator", has been described elsewhere [17]; only its essential features will be mentioned here.

The inputs to the trial translator are a set of experimental syntactic and semantic algorithms expressed in a notation similar to that of the propositional calculus. Each of the algorithms is expressed in the form: "if the logical condition P is satisfied, then apply the transformation Q". P is an expression compounded of the logical variables that determine when the transformation Q is to be applied to a text. If k is a serial number defining the position of an entry in the augmented text, typical variables might be "N(k)" standing for "text word k is a noun", "GP(k)", standing for "text word k is in the genitive plural", "PREP(k-l)" standing for "the text word preceding word k is a preposition", etc. A logical expression P is constructed by connecting variables with the connective functors of the propositional calculus: "•" standing for "and", "v" standing for "or" and "~" standing for "not". Typical transformations might be "PERM (k, k + l)" standing for "permute the translations of text words k and k+1", "INS (of the, k)" standing for "insert *of the* before the translation of k", etc. We might then consider the sample algorithm: "If a noun in the genitive plural is not preceded by a preposition, then insert *of the* before its translation". This can be simply abbreviated as "~PREP(k—l)•N(k)•GP(k) →INS(of the, k)". The algorithm is obviously too simple to be valid, and is included here only to illustrate the use of the notation.

The operation of the trial translator is based on the automatic association of algorithms with dictionary entries, the automatic specification of the truth values of logical variables, and the automatic evaluation of logical formulas. The system applies the given algorithms to English-inflected augmented texts (fig. 16). Its outputs are the readable translations resulting from the application of the given rules to the given texts. Linguists, psychologists, and computer specialists examine these translations and suggest modified algorithms that are, in turn, coded in the language of the propositional calculus and applied by the trial translator. The same man-machine cycle will be repeated until a satisfactory set of algorithms is determined, or until it is obvious that some major change must be made in the machine system.

5.4 *Formula finding*

The first.formulas to be tested while experimenting with the trial translator will be either drawn from the existing literature or suggested in the course of experiments with the products of the automatic dictionary. It may also be possible to find algorithms automatically, by means of a primitive machine "learning" system. The writers are currently investigating such a system, called "Formula Finder".

The inputs to the formula finder are an augmented Russian text, and the final post-edited translation of the same text. The English translation must first be transcribed onto magnetic tape by manual or automatic means. When given proper clues by a linguist, the system will deduce algorithms that can be used to transform the augmented text into the edited version.

The clues that must be given to the formula finder are: (a) a list of logical variables that might conceivably determine a certain transformation, and (b) a statement of the transformation being investigated. The variables and transformations are assumed to be stated in the mnemonic notation used as input to the trial translator. The formula finder compares the augmented text and the post-edited text. Whenever a product of the indicated transformation is found in the post-edited text and certain auxiliary conditions are satisfied, the formula finder examines the truth-value configuration of the given variables in the augmented text. After examining all instances of the transformation, the formula finder can ascertain whether the indicated variables can be combined into a logical formula that implies the given transformation. The output of the formula finder is either:

a) A logical formula that always implies the given transformation, thus defining a translation algorithm valid for the given corpus of text. The formula finder will reduce this formula to its simplest logical form and eliminate vacuous variables not actually needed in the algorithm.

b) A statement of the "closest" logical formula in case other variables besides those originally given are required to determine the transformation. The statement will be accompanied by indications of the quality of the approximate formula, and by clues that will help linguists to suggest additional variables for testing.

c) A machine-coded statement of the exact or approximate algorithm, ready to be tested by the trial translator.

Since natural languages are open systems, an algorithm-finding process can never be quite finished. Nevertheless, there is hope that the processes of trial translating and formula finding will eventually lead to an acceptably stable set of algorithms and that, in the interim, the quality of the trial translations will steadily improve. Systems like that of fig. 16 should lend themselves to the economic mass-production of interim translations made by the best set of tested translation algorithms available at the time. It is hoped that these interim translations will serve as valuable aids to professional translators, to students of technical Russian, and to scientists interested in the Russian technical literature.

6. References

[1] Oettinger, A. G., W. Foust, V. Giuliano, K. Magassy and L. Matejka: *Linguistic and Machine Methods for Compiling and Updating the Harvard Automatic Dictionary*. Preprints of Papers for the International Conference on Scientific Information, National Academy of Sciences, National Research Council, Washington, D. C., Part V, 1958, pp 137-160.

[2] Oettinger, A. G.: *An Input Device for the Harvard Automatic Dictionary*. Mechanical Translation, Vol. 5, No.1.,July 1958, pp 2-7.

[3] Jones, P. E.: *A Method of Contextual Analysis Applicable to Linguistic Research.* Papers Presented at the Seminar on Mathematical Linguistics, Vol. IV, 195S. On deposit at Widener Library, Harvard University.

[3]) Discounting several newspaper reports that have never been adequately substantiated in the technical literature.

[4] BARNES, V. L.: *An Index Routine.* Papers Presented at the Seminar on Mathematical Linguistics, Vol. IV, 1958. On deposit at Widener Library, Harvard University.

[5] GIULIANO, V. E.: *An Experimental Study of Automatic Language Translation.* Doctoral Thesis, Harvard University, 1959.

[6] SHINER, G.: *The USAF Automatic Language Translator Mark I.* 1958 IRE National Convention Record, Part 4, pp 296-301.

[7] GIULIANO, V. E.: *Programming an Automatic Dictionary.* Design and Operation of Digital Calculating Machinery, Report No. AF-49, Section I, Harvard Computation Laboratory, 1957; also *Papers Presented at the Seminar on Mathematical Linguistics.* Vol. III, 1957. On deposit at Widener Library, Harvard University.

[8] OETTINGER, A. G.: *A Study for the Design of an Automatic Dictionary.* Doctoral Thesis, Harvard University 1954.

[9] MILLER, G. A., J. G. BEEBE-CENTER : *Some Psychological Methods for Evaluating the Quality of Translations.* Mechanical Translation, Vol. III, No. 3, 1956, pp 73-80.

[10] SMIRNITSKIJ, A. I.: *Russko-Anglijskij Slovar* (Russian-English Dictionary). Gosudarstvennoe Izdatel'stvo Inostrannyx I Natsional'nyx Slovarej, 1949.

[11] *English-Russian, Russian-English Electronics Dictionary.* (TM 30-545), Dept. of the Army, Washington 1956.

[12] MATEJKA, L.: *Grammatical Specifications in the Russian-English Automatic Dictionary.* Design and Operation of Digital Calculating Machinery, Report No. AF-50, Section V, Harvard Computation Laboratory, 1958.

[13] SHERRY, M. E.: *Analysis of Case and Number of Nouns for Automatic Translation.* Papers Presented at the Seminar on Mathematical Linguistics, Vol. IV, 1958. On deposit at Widener Library, Harvard University.

[14] *Papers Presented at the Seminar on Mathematical Linguistics,* Vols. 1-IV, 1955-1958. On deposit at Widener Library, Harvard University.

[15] *Seminar Work Papers* (mimeographed). Georgetown Institute of Languages and Linguistics, Georgetown University, 1957, 1958.

[16] KOUTSOUDAS, A., A. HUMECKY: *Ambiguity of Syntactic Function Resolved by Linear Context.* Word, Vol. 13, No. 3, Dec. 1957, pp 403-414.

[17] GIULIANO, V. E.: *The Trial Translator, An Automatic Programming System for the Experimental Machine Translation of Russian to English.* To appear in the Proceedings of the 1958 Eastern Joint Computer Conference, IRE and ACM.

7. Discussion

D. G. Hays (USA): There are two methods for the building of MT dictionaries. Mr. Oettinger compiles entries from existing dictionaries, then tests their validity by inspecting texts. An alternative is to build the dictionary directly from textual studies.

It is interesting to compare these methods from the standpoint of economy and usefulness.

D. G. Owen (UK): In view of the time and effort required to remove all misprints from the data supplied for translation, is it not worthwhile to program the computer to search for closely similar words when it is unable to locate the word as spelt? It might be assumed that errors of four kinds could occur: a wrong character, a missing character, an extra character, or two consecutive characters transposed.