Linguistic and Machine Methods for Compiling and Updating the Harvard Automatic Dictionary

A. G. OETTINGER, W. FOUST, V. GIULIANO, K. MAGASSY, and L. MATEJKA

Natural languages, unlike some abstract linguistic systems studied by mathematical logicians or some instruction codes devised by designers of automatic programming systems, have no easily described simple structure. Considerable empirical study is therefore necessary to develop for these languages grammars that are sufficiently precise and comprehensive to serve as the basis for any system of automatic translation. The difficulty of such empirical study is attested to by the almost total absence, after nearly a decade of interest in automatic translation, of any but theoretical discussions of the subject. The empirical work that has been reported is generally the result of laborious manual work and, even where machines have been used, results are based on such limited and carefully selected samples that their significance is doubtful.

Significant research on automatic translation presents such massive data handling problems that, unless automatic machines and associated techniques are used as tools to assist in research from the beginning, chaos is a likely result. Research workers with both adequate qualifications in linguistics and experience in the design and operation of automatic information processing machines are relatively scarce. Careful planning is therefore essential in order to enable the performance of large scale routine tasks by a team of clerical and technical personnel assisted by automatic machines. The necessary automatic machines are presently available in the form of general purpose digital calculators. By the application of present techniques of automatic programming, and by the development of new ones, much of the programming of these cal-
Calculators can systematically be reduced to routine processes to be performed by clerical personnel or even by the machine itself.

While enough is already known about automatic translation and allied problems to warrant paying attention to the development of suitable input, output, and storage devices, the processes of translation are hardly well enough defined yet to justify the construction of a complete specially designed translating system. General purpose machines can provide adequate algorithmic power and sufficient storage capacity with a minimum of capital outlay. Different methods can readily be tested merely by writing different programs, suggesting a method of successive approximations, whereby the results of experimental operation of the system up to a given time can be used to improve the mode of operation for future time. Whenever possible, it is desirable to make program modification itself a part of automatic machine functions. In this fashion optimal design parameters for special equipment can eventually be determined. For a variety of field applications, general purpose machines may well continue to be used even for production, although it is likely that the efficient and economical operation of large translation centers will eventually require the use of specially designed equipment.

With these premises in mind, the work described in this paper has been centered on the formulation and practice of efficient techniques for the initial compilation and periodic up-dating of automatic dictionaries. The first Harvard Automatic Dictionary is intended primarily to provide the following three facilities: (a) an immediately useful device for lightening the burden on professional translators, speeding up their work, and improving its accuracy and timeliness; (b) a system of automatic word-by-word translation, serving as a linear first approximation to an automatic translation system; (c) an experimental tool to facilitate the extensive basic research still necessary to develop methods for faithful smooth translation of technical Russian into English.

While automatic translation from Russian to English is the specific object of our research, automatic dictionaries and the techniques for compiling them lend themselves to more general applications, including translation between other pairs of languages, and certain phases of information organization and retrieval. Automatic abstracting, by techniques such as described by Luhn (1), is one example of such an application. An automatic dictionary in which a record is kept of the frequency of use of each entry can provide an accurate and current standard for eliminating the “noise” caused by words common in any text. The inverse inflection algorithm mentioned in Section 2, extended if desirable to account for derivation as well, can be useful in mapping inflectional variants of a stem, or derivatives of a root, into a single class or canonical form. When the source text is in a foreign language, combining automatic abstract-
ing with automatic translation of the abstract is an obvious possibility. In searching texts for the occurrence of key words or word combinations, the use of inverse inflection and derivation algorithms will permit specifying keys in a canonical form, with the guarantee that occurrences of inflected or derived variants will also be detected. As the translation algorithms based on the use of automatic dictionaries grow in sophistication, the ties between automatic language translation and the many areas where syntactic analysis and code conversions are necessary will very likely continue to be strengthened.

1. Word selection

In planning for dictionary compilation, efforts were made to minimize the need for manual intervention, to retain flexibility for experimental purposes, and to provide procedures suitable for periodic up-dating of operating dictionaries, as well as for their initial compilation.

The initial selection of entries for a dictionary can be carried out by two major processes, each having peculiar advantages and disadvantages. The first method may be likened to panning for gold. In this method some number of texts of the type eventually to be translated are scanned, and a glossary of all distinct forms occurring in this sample is compiled. The main advantage of the method is that every form obtained in this way is in current use, and therefore is a suitable candidate for inclusion in a glossary. If scanning is to be performed automatically, the texts must necessarily be transcribed onto some automatically readable medium, hence made available for other purposes, including the eventual testing of translation methods. The major disadvantage of the procedure is that it becomes progressively less and less productive. While nearly all the first few words of the first text scanned are likely to be nuggets, as the work progresses more and more gravel must be handled before another nugget is found. Relatively few word forms account for the vast majority of form occurrences in any text; these forms are found over and over again and must be rejected over and over again. Second, in pure panning, only those particular inflected forms of any word that actually occur in the sample under study will be entered in the glossary. Until all the forms constituting the paradigm of a word have occurred in some text, a complete characterization of this word is not available. This precludes the possibility of early systematic treatment of inflectional processes, other than by handling each inflected form as a unique entity.

The second approach may be called the “fish net” method. Available dictionaries are dragged for words useful according to some reasonable criterion. Anything caught in the net is retained. The chief advantage of this procedure
is that a large selection of useful words is obtained very rapidly. The major disadvantage is that the resulting dictionary is only as good as the criterion of selection or as the dictionaries from which it was selected. Moreover, words without current utility may well be caught in the net.

Neither method is guaranteed to yield a vocabulary precisely suitable for the first texts to which a dictionary is applied, but either lends itself to the addition of new words or new forms whenever these appear in a text. How rapidly the ratio of text words not in the dictionary to those in it will diminish to a satisfactory level is still open to conjecture. For that matter, so is a satisfactory definition of this level.

The procedure described here is a combination of the two methods. An initial set of nearly ten thousand words was selected, in part, from a general dictionary (2) to obtain words of common currency and, in part, from a specialized electronics dictionary (3) to obtain as complete as possible a coverage of technical terms in this area. In cases of doubt about the utility of words, they were usually included. It seems more efficient to carry a few doubtful words through routine compilation procedures and to provide for their automatic removal (based on a criterion of frequency of use), once operating experience has accumulated, than to spend valuable personnel time on intricate and inconclusive selection procedures. Once the dictionary is functioning, all new words encountered in texts submitted for translation, but not in the dictionary, will be printed as a by-product of dictionary operation. These new words, from text sources, will eventually replace the initial stock as the raw material for an up-dating procedure almost identical to the compilation process. For up-dating, therefore, the fish net process is replaced by a modified gold panning technique, where the gravel is the major object of processing, and the gold a valuable, but easily obtained by-product; prior to further processing, each new form found in a text is reduced to the form normally listed in dictionaries. The accumulation of frequency data, the analysis of contexts and other analytic procedures can be carried out on texts already recorded in automatically readable form because of their interest as objects of translation.

2. The form of dictionary entries

In ordinary dictionaries, the paradigm of a word is conventionally represented by a standard or “canonical” form, e.g., the nominative singular for nouns, the nominative masculine singular for adjectives, and the infinitive for verbs. This lexicographic device presupposes on the part of a person using the dictionary the ability to perform the grammatical analysis necessary to reduce an
inflected form of a word found in a text to the canonical form listed in the dictionary. As conventionally practiced, this grammatical analysis requires a fairly thorough acquaintance with the inflection system of the language in question, as well as a certain amount of imagination.

An automatic dictionary may provide for the treatment of inflected forms either (a) by providing a distinct entry for each distinct inflected form or (b) by providing only a single canonical form as an entry for each word, together with an algorithm for transforming other forms to the canonical one. A variety of compromises between these two extremes is also possible. Whichever type of entry is used, if words are selected from existing dictionaries by the fishing technique, an algorithm capable of generating all distinct inflected forms of every word given in the conventional canonical form is essential. The need is obvious if a distinct entry is to be made for each distinct inflected form. Under circumstances described further on, the generation of all distinct inflected forms is useful even if only the canonical forms are to be used as dictionary entries. The pure gold panning procedure by-passes these difficulties, but at the price of commitment to the system of distinct entries for distinct inflected forms.

If experimental work is to be carried out on existing machines, facilities sufficiently ample and economical to store a large dictionary are presently available only in the form of magnetic tapes or punched cards. Because using distinct inflected form entries would increase storage requirements and search time by an order of magnitude, this method will be practicable only when large capacity internal memory devices with sufficiently low cost and access time become available (4,5). Certainly for the present, and probably for the future, the use of canonical form entries seems indicated.

The canonical form chosen for our purposes is best described as a “stem,” because its definition is very close to that of stems as commonly defined by linguists. The precise definition of stems and a description of the algorithm used for reducing arbitrary members of a paradigm to this canonical form have been given earlier (6). The reduction algorithm or “inverse inflection algorithm” cuts forms into two parts, one, the stem, the other, an ending usually identical with or at least very similar to the usual Russian inflectional desinences. When the dictionary is in operation, each form occurring in a text is split by the inverse inflection algorithm into a stem and an ending. The stem is used as the key for search in the dictionary, and the ending is retained for eventual use in syntactic analysis. To insure that the stems produced by splitting text forms will be identical with stems entered in the dictionary, the latter are obtained by applying the inverse inflection algorithm to all members of the paradigm of any word about to be entered in the dictionary.
3. The generation of paradigms

Generating all inflected forms belonging to the paradigms of ten thousand canonical forms obtained from standard dictionaries is a task not to be undertaken lightly. As a purely manual operation, this task would be staggering, and the probability of errors extremely high. The standard Russian noun paradigm has twelve forms, the adjective paradigm twenty-four or twenty-eight forms, and verbs have well over one hundred forms, if participles are counted. Because not all members of every paradigm are distinct (e.g., the nominative singular form of some nouns is identical with the accusative singular form), it is sufficient to generate a “condensed paradigm,” containing only distinct inflected forms. However, even the number of distinct inflected forms is enormous.

In Russian, inflected forms are characterized by their desinences, so that the specification of a paradigm is tantamount to specifying a type of arrangement of inflectional desinences. While the morphological differentiation of word stems is great, that of inflectional desinences is relatively small. The inflection pattern of one word can therefore be used to obtain the inflected forms of all other words characterized by the same nature and disposition of inflectional desinences. This suggests the possibility of developing an algorithm for automatic direct inflection, which will treat alike all words identified as having identical or sufficiently similar inflection patterns.

A new system of classification was therefore developed (7) such that, once a word has been identified as belonging to a class whose members are inflected in a certain way, it can be inflected automatically. Classification systems given in existing grammars were found to be both incomplete and often incompatible with our requirements. Our system of classification is based on the assumption that the identification of the inflectional pattern of a word must, for the time being at least, remain a manual function, while the actual generation of distinct inflected forms can be an almost completely automatic process. Therefore, ease and accuracy of identification must be promoted by using any readily obtainable data meaningful to a person, while the generation must be based strictly on explicit orthographic data recognizable by a machine.

Some of the criteria used in defining classes may be illustrated with reference to Fig. 1. The first column of Fig. 1 describes the full paradigms of the Russian words дама, комната, and граница. Condensed paradigms, where only distinct inflected forms appear, are illustrated in the second column. It will be noted that дама and комната have been included in the same class, even though they differ in the accusative plural, because their condensed paradigms
can be generated by identical procedures. The justification for including граница in this class is twofold: first, differentiating граница from комната is inefficient at this stage and is accomplished more readily at a stage described in Section 4. Second, the only price to be paid for simplifying classification in this way is the addition of a spurious inflected form to the paradigm of each member of this class. This addition is harmless. If a dictionary of distinct inflected forms is being compiled, the spurious forms will never be consulted and therefore will eventually be eliminated by virtue of their zero frequency of use. In a dictionary of canonical form stem entries, a spurious form usually leaves no traces, since it leads to a stem identical to those obtained from the other distinct inflected forms. This procedure is very much like one used quite frequently in approximating mathematical functions over a given range: any convenient function may be used that suitably approximates the desired function in the specified range; its behavior outside of this range is of no consequence. Other expedients of this kind are also proving their worth in terms

FIGURE 1. Condensation of paradigms (examples drawn from class N4).
of greater systematization than would be possible without them. As those experienced in automatic data processing well know, the handling of a single exceptional case often proves more costly and more time-consuming than the routine handling of a few extra elements.

The definitions of inflectional classes and the rules for forming distinct inflected forms of words belonging to these classes are illustrated in Fig. 2. The

N4. дама, лампа, игла, служба
I. A class ending in -а preceded by any consonant except:
   (i) г, ж, к, х, ч, ш, й,
   (ii) п, whenever preceded by another consonant;
   (iii) the majority of cases when the consonant is л, н,
   (iv) some cases when the consonant is м, р,
   (v) a few cases when the consonant is б,
   preceded by another consonant
II. Formation roles for distinct inflected forms:
   (i) Generating stem = canonical form - last letter;
   (ii) Distinct inflected forms:
      (a) canonical form
      (b) generating stem + ы
      (c) + е
      (d) + у
      (e) + ой
      (f) generating stem + ей
      (g) + #
      (h) + ам
      (i) + амн
      (j) + ах

N4.31 вьшлавка, кишкá
I. A “reappearing о” class embracing the nouns ending in -а preceded by:
   (i) any consonant (except й, ж, ш, ч, ц), not followed by ш, + к;
   (ii) ж, ш, ч, + к, whenever the stress falls upon the ultima of the word;
II. Formation rules for distinct inflected forms:
   (i) Generating stem = canonical form - last letter, but an о must be inserted between
      the penult and the ultima of the generating stem in (g);
   (ii) Distinct inflected forms as in N4, II (ii), except -ы for -ы in (b).

* The canonical form here is that used in standard dictionaries, not the “stem” canonical form used in the Automatic Dictionary.

FIGURE 2. Rules for class identification and inflection.
to the distinct inflected forms, and the resulting canonical stems are not always identical with the generating stems. For example, the past passive participial forms of some verbs are constructed most readily by adding -енный to a generating stem. The canonical stem of the corresponding paradigm includes the letters -енн.

Because the assignment of words to classes is intended, for the present, to be a manual task, any classification criterion that is easily recognized by persons can be used. For example, stress distinctions, which cannot be used in defining formation rules, serve as a means of class identification. In addition, significant examples, lists of exceptions, and so forth, have been given wherever possible.

Our system also embraces a large number of words whose formation is “irregular.” For example, the class N4.31 comprises words in which the vowel “о” is introduced in one inflected form. The formation rules for this class (Fig. 2) therefore specify that an “о” must be inserted between the last and next to the last letters of the generating stem before constructing the form (g). The class N4.31 is further distinguished from the class N4 by the use of the letter и for the letter ю in the inflected form (b). Whenever the formation rules for one class deviate only slightly from those for another class, they have been stated as exceptions to the rules for this other class. Considerable economies in programming inflection are achieved as a result.

Our system of classification comprises eight classes of adjectives, thirty-eight classes of nouns, and forty-six classes of verbs. Indeclinable words are assigned to a special class. The system of classification is sufficiently comprehensive to include all but a few unproductive classes with highly atypical paradigms; it is completed by the definition of a class, labeled Z99.99, to which all words not falling into any of the other classes are assigned. These words are eventually inflected by hand. Of a total of seventy-six hundred words classified to date, all but twenty-four were assigned to genuine classes. The distribution of these words among the major groupings is indicated in Fig. 3.

Total number of:

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives</td>
<td>2477</td>
<td>32.59</td>
</tr>
<tr>
<td>Nouns</td>
<td>3972</td>
<td>52.26</td>
</tr>
<tr>
<td>Invariables</td>
<td>74</td>
<td>0.97</td>
</tr>
<tr>
<td>Unclassified nouns and adjectives (Z99.99)</td>
<td>1053</td>
<td>13.86</td>
</tr>
<tr>
<td>Verbs</td>
<td>11</td>
<td>0.14</td>
</tr>
<tr>
<td>Total</td>
<td>7600</td>
<td>99.99</td>
</tr>
</tbody>
</table>

FIGURE 3. Distribution of words among major groups.

The class definitions as given in Fig. 2 are not in a form that readily lends itself to rapid recognition. To make classifying as easy as possible, the rules
were expressed in a tabular form (3) illustrated in Fig. 4, where the complete classification table for adjectives is displayed. The vertical lines in Fig. 4 divide endings on the right from significant terminal stem letters on the left. The special symbols C, V, and C_bC are interpreted as follows: C denotes any consonant not specified earlier in the group between horizontal lines, and V the same for vowels. The combination C_bC signifies any consonant, whether; followed or not by the soft sign, and not specified earlier within the group. For example, an adjective ending in -ий preceded by к may be assigned to any of the classes A6, A7, or A8, depending on the letter preceding к. If the letter preceding к is one of the three indicated next to A7, the adjective is assigned to this class, if the letter preceding к is any consonant not in the list belonging to A7, the adjective is assigned to the class A6, and finally, if the letter preceding к is any vowel, the adjective is assigned to the class A8. Vowel changes are marked by the sign >. On the left side of the sign is the vowel before the change, and on the right side the vowel after the change. For example, in the third group in Fig. 4 adjectives ending in -ый preceded by л or п are normally assigned to the class A3, unless the vowel e is inserted as the second letter from the end in the masculine predicative form of the adjective.

Through the use of these charts, assigning to classes becomes a routine task which can be done by a person with relatively little knowledge of Russian, although occasionally dictionaries must be consulted in the process. Experience has shown that the amount of dictionary consultation at this stage is negligible. With this one exception, all inevitable dictionary consultation is concentrated at a single stage of the process, namely, when English correspondents are assigned and grammatical codes are added to the stems. Assigning words to classes has been successfully done at an average rate of approximately one thousand words per day per person.

4. The preparation of stem entries

Because our dictionary is intended for operation on the Univac I computer at the Harvard Computation Laboratory, some details of the preparation of stem entries and of other phases of compiling apply directly only to this machine. However, very similar procedures can be readily used with other types of contemporary large-scale computers.

Words selected from the two dictionaries mentioned earlier were originally transcribed onto file cards. An inflectional class marker, assigned in the manner outlined in Section 3, was then written on each card. The card file is used only in the initial compilation process, since new words found in texts as a by-product of dictionary operation will be made available automatically in a
<table>
<thead>
<tr>
<th>ADJECTIVES</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{ш, ж, ч, щ} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ны} )</td>
</tr>
<tr>
<td>(\text{н} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{г, х} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{C} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{C} )</td>
<td>(\text{k} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{V} )</td>
<td>(\text{k} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{жк, } \text{йк, } \text{ьк} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{C} )</td>
<td>(\text{k} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{V} )</td>
<td>(\text{k} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{жк, } \text{йк, } \text{ьк} )</td>
<td>(\text{ий} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
<tr>
<td>(\text{C} )</td>
<td>(\text{k} )</td>
<td>(\text{ой} )</td>
<td>(\text{ый} )</td>
</tr>
</tbody>
</table>

**FIGURE 4.** Classification table for adjectives.
The Design of New Systems

printed format suitable as raw material for the periodic dictionary up-dating.

Classified words are next recorded on magnetic tapes by means of a standard Unityper. Because this input device is designed to handle normally only numerals, the Roman alphabet, and a few special characters, some adaptation was necessary to use it for typing Cyrillic material. The set of characters used for this purpose is shown in column 1 of Fig. 5. It is a fairly simple matter to place over the normal keys of any typewriter special keytops engraved with any desired alphabet. We chose to arrange the keyboard in one of the standard Cyrillic layouts, to ease the work of typists already familiar with it. The correspondence so established between Cyrillic and machine characters is described by columns 1 and 2 of Fig. 5. Although this correspondence preserves a

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>0</td>
<td>Z</td>
</tr>
<tr>
<td>1</td>
<td>&amp;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>r</td>
<td>2</td>
<td>Й</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>К</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>4</td>
<td>Л</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>М</td>
</tr>
<tr>
<td>6</td>
<td>@</td>
<td>6</td>
<td>Н</td>
</tr>
<tr>
<td>7</td>
<td>t</td>
<td>7</td>
<td>О</td>
</tr>
<tr>
<td>8</td>
<td>/</td>
<td>8</td>
<td>П</td>
</tr>
<tr>
<td>9</td>
<td>J</td>
<td>9</td>
<td>Р</td>
</tr>
<tr>
<td>A</td>
<td>F</td>
<td>1</td>
<td>А</td>
</tr>
<tr>
<td>B</td>
<td>,</td>
<td>2</td>
<td>В</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>V</td>
<td>У</td>
</tr>
<tr>
<td>Y</td>
<td>4</td>
<td>Г</td>
<td>Ф</td>
</tr>
<tr>
<td>Д</td>
<td>5</td>
<td>Д</td>
<td>Х</td>
</tr>
<tr>
<td>Е</td>
<td>R</td>
<td>6</td>
<td>Ц</td>
</tr>
<tr>
<td>Ж</td>
<td>;</td>
<td>7</td>
<td>Z</td>
</tr>
</tbody>
</table>

1: Cyrillic Available on Keyboard
2: Typewriter Code
3: Ranked Code
4: Transliteration Code

FIGURE 5.
corded in the ranked code obviously cannot be easily read, so that material which must be read quickly is subjected to still another code conversion run in which the character strings given in column 4 of fig. 5 are substituted for their correspondents in column 3. Eventually, the conversion from the code of column 1 to that of column 3 will be made simultaneous with typing.

Two lists are prepared from the tape recorded on the typewriter. The first, illustrated in Fig. 6, presents the words in the order in which they were typed.
and is used for catching errors made in typing the words; it will eventually be replaced by hard copy made by the typewriter itself. The other, shown in Fig. 7, is alphabetized by the last letters of words rather than by the first as usual, and justified to the right to bring all words with like endings together. It is used to check the classification of words. Because the criteria for classification are based largely on the configuration of the last letters of each word, words with the same class marker tend to be brought together on this list.

FIGURE 7. End alphabetized word list.
Checking class markers is facilitated by the presence of runs of identical class markers, in which the occurrence of an odd marker shows up clearly. Errors detected on these lists are not corrected on tape. It is easier to delete the affected items prior to assigning correspondents to them, and then to process them again in routine fashion with another batch. Therefore, items found in error are simply marked for deletion at the later stage.

Throughout compilation (8), Russian words represented in the ranked code are imbedded in “items” consisting of five machine words of twelve characters each. The first three machine words are used to store the Russian word, allowing a maximum of thirty-six letters. The first six character positions of the fourth word of each item are reserved for the inflectional class marker as shown in Figs. 6 and 7. Eight positions in the fifth word are used for an identification number which designates the batch in which the word was originally processed, and the serial number of the word within this batch. This identification number accompanies all forms of a word throughout compilation, to facilitate the identification of Russian words represented in the ranked code, and the tracing of errors. Of the last three digits of the fifth machine word, the two low order ones specify the character position within a machine word at which the last letter of the Russian word occurs, while the high order digit specifies the machine word (0, 1, or 2) in which the last letter of the Russian word occurs. These numbers are computed during the initial code conversion, and are used to control shifting and extracting operations when the Russian word undergoes later transformations.

After the lists of Figs. 6 and 7 have been checked for errors, the Russian word tapes are ready for automatic inflection (9). This process is illustrated in Fig. 8, with the verb писать as an example. This verb is shown with its inflectional class marker V5, assigned as described in Section 3. The forms generated after one inflection run are indicated in column α of Fig. 8. It will be noted that adjectival inflectional class markers are given next to the participial standard canonical forms. These forms are generated by the inflector routine, together with their class markers. Therefore, the inflection of the participles is entirely automatic, without need for further manual classification.

The inverse inflection algorithm is then automatically applied to the set of distinct inflected forms, splitting each into an ending and a potential stem canonical form. The resulting set of stems is condensed to yield a list where only one representative of each distinct type of stem is retained. As in the example of Fig. 1, and for the same reasons, some artificial forms are generated which do not properly belong to the paradigm of the verb писать. These forms are marked on the diagram by the letter I. In addition, several so-called academic forms are generated. These are distinguished from the artificial
forms in that they are morphologically acceptable members of the paradigm of пиарить but are very unlikely to occur in actual texts. Stems obtained from artificial or academic forms can be identified and deleted, if desired, at the time that English correspondents are assigned. In case of doubt they are retained, to be deleted when their zero frequency of use after a long period of operation automatically indicates that they should be.

FIGURE 8. The generation of paradigms and canonical stems.
In practice, automatic inflection is separated into three major runs. Words marked with nominal and verbal class markers are inflected before those with adjectival class markers. This means that, on the last run, those adjectival forms generated together with their class markers as a result of verb inflection can be inflected with the other adjectives. A list of inflected forms obtained from the inflector runs is shown in Fig. 9. A similar list, showing forms after splitting of their endings, is given in Fig. 10. Adjectival forms generated as a result of verb inflection are distinguished by the presence of one of the letters A, B, C,
or D in their serial number, in the position where a hyphen occurs in all other words. On the magnetic tapes used for further processing, only a stem is present in the first three machine words. The split ending is stored in the last five character positions of the fourth machine word of the item, where it may be seen in ranked code. Of these five characters, three are reserved for normal endings, the last two are spaces except for verbs ending in -ся or -сь. For ease in reading transliterated lists of split forms, the endings stored in these
positions are brought back into the first three machine words during the transliteration run. The hyphen is inserted to mark the position of the split.

The list of distinct stems obtained by automatically condensing a list of the type illustrated in Fig. 10 is printed with the layout shown in Fig. 11. This layout is designed to guide the manual inscription of English correspondents and of grammatical coding associated with the stem canonical forms. The paper is ruled to show clearly the divisions between machine words; exactly
two characters must be written in the spaces defined by the vertical dashed lines. Heavy black horizontal lines delimit the space allowed for an entry. Dictionary entries are items consisting of thirty machine words, of which the first five make up a standard Russian item. The last twenty-five machine words are devoted to correspondents and to grammatical information. The English correspondents are written immediately after the Russian item. The last four of the twenty-five words are reserved for coded information. Distinct correspondents are numbered, the last correspondent being marked by the use of a percent sign in place of a numeral.

An effort is made to rank the correspondents in the order of their likely frequency of use. Initially, this ranking is necessarily somewhat arbitrary. It is expected that, as a major by-product of automatic dictionary operation, the sets of correspondents and their ordering will gradually be adjusted in accordance with the experience of the technical experts who are the ultimate users of translations and the best judges of their value.

In the first of the four words reserved for coded information, the significance of a character depends on its position within the word. This word is therefore called the “organized word.” In the other three words, the significance of characters is independent of their position within the set of three words. These words are therefore called “semiorganized words.” Because the inflectional class markers have chiefly formal significance, one of the major functions of the coded information in the organized word is to identify the functional role of the entry. This may be illustrated by the following two examples: a word declined like an adjective may function as a noun and will be coded as such in the organized word; the set of words classified as formally indeclinable includes, along with prepositions and conjunctions, some words functioning as nouns, and these are distinguished by an appropriate notation. Functionally distinct paradigms lumped into one inflectional class to simplify classification and automatic inflection are also distinguished by means of a notation in the organized word. For example, the distinction between animate and inanimate nouns is of no consequence so far as the generation of distinct inflected forms is concerned, but it is vital to the interpretation of the functional significance of endings. That distinction is therefore made by means of a symbol in the organized word. Such information as the names of dictionaries or texts consulted in preparing the entry is mentioned in the semiorganized word. In addition, important grammatical data, whose range of application is insufficient to justify their inclusion in the organized word, are introduced in the semiorganized words.

Although every effort is made to provide for coded information of wide scope, it is clearly not possible to foresee all contingencies likely to be met in
experimental work. Therefore, the space between the percent sign marking the end of the last English correspondent and the organized word is left free for the insertion of ad lib English prose comments. In this way, information of significance whose need was not foreseen in planning the layout of the organized and semiorganized words, or likely to apply to too few entries to warrant inclusion in these words, may be recorded and retrieved automatically at a later date. The systematic inclusion of such information into the organized words is possible whenever large enough classes of similar comments are found.

After the correspondents and coded information have been written as shown in Fig. 11, the handwritten material is transcribed onto a magnetic tape. This tape is then merged with that containing the Russian stems, and the complete dictionary entries are recorded on a new tape. The markings in the left-hand margin of Fig. 11 are used to control in part the process of dictionary assembly. A zero in the left-hand margin indicates that the stem is to be deleted and not to appear in the final dictionary. This is the means for deleting spurious or academic forms, or errors detected earlier but not then corrected.

The paradigms of nouns and adjectives where vowel insertion occurs, and of many verbs, may be represented in the list of Fig. 11 by more than one stem. This problem of multiple stems has been discussed in detail elsewhere (6). Whenever two or more stems belong to the same paradigm, they are identified by the same serial number. The first stem in such a set is always that split from the standard dictionary canonical form, and is marked by the letter F in the seventh column of the fourth machine word of the Russian item. Unless it is to be deleted, this stem is usually given a left-hand margin marker 1. If other stems with the same serial number require precisely the same English correspondents and coded information, repeated writing of this information is not necessary. It is sufficient to write the symbol 1R in the left-hand margin for such entries. Where one or more of a set of stems with the same serial number require a set of correspondents distinct from that assigned to the F form, the margin markers 2, 2R, 3, 3R, etc., may be used.

In transcribing the English and coded material, each entry is identified simply by its left-hand margin marker and the serial number. The whole Russian entry need not be copied.

Once the correspondents have been transcribed onto magnetic tapes, they are automatically assembled with their stems. A section of assembled dictionary is shown in Fig. 12. It should be pointed out again in this connection that Russian word forms in dictionary entries are stored in the ranked code on magnetic tape and are represented by stems only. Endings are introduced during the process of transliteration only to facilitate human recognition of the printed entries.
FIGURE 12. Assembled dictionary.
ACKNOWLEDGMENTS

The authors are indebted to Professors Howard Aiken, Roman Jakobson, and Joshua Whatmough for their interest in this research and for many valuable suggestions. The work has been supported in part by the Harvard Foundation for Advanced Study and Research, the Ohio Oil Company, the United States Air Force, and the National Science Foundation.

REFERENCES