NATIONAL BUREAU OF STANDARDS REPORT

6295

A NEW APPROACH TO THE MECHANICAL TRANSLATION OF RUSSIAN

by

I. Rhodes

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

THE NATIONAL BUREAU OF STANDARDS

Functions and Activities

The functions of the National Bureau of Standards are set forth in the Act of Congress, March 3, 1901, as amended by Congress in Public Law 619, 1950. These include the development and maintenance of the national standards of measurement and the provision of means and methods for making measurements consistent with these standards; the determination of physical constants and properties of materials; the development of methods and instruments for testing materials, devices, and structures; advisory services to Government Agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; and the development of standard practices, codes, and specifications. The work includes basic and applied research, development, engineering, instrumentation, testing, evaluation, calibration services, and various consultation and information services. A major portion of the Bureau's work is performed for other Government Agencies, particularly the Department of Defense and the Atomic Energy Commission. The scope of activities is suggested by the listing of divisions and sections on the inside of the back cover.

Reports and Publications

The results of the Bureau's work take the form of either actual equipment and devices or published papers and reports. Reports are issued to the sponsoring agency of a particular project or program. Published papers appear either in the Bureau's own series of publications or in the journals of professional and scientific societies. The Bureau itself publishes three monthly periodicals, available from the Government Printing Office: The Journal of Research, which presents complete papers reporting technical investigations; the Technical News Bulletin, which presents summary and preliminary reports on work in progress; and Basic Radio Propagation Predictions, which provides data for determining the best frequencies to use for radio communications throughout the world. There are also five series of nonperiodical publications: The Applied Mathematics Series, Circulars, Handbooks, Building Materials and Structures Reports, and Miscellaneous Publications.

Information on the Bureau's publications can be found in NBS Circular 460, Publications of the National Bureau of Standards ($1.25) and its Supplement ($0.75), available from the Superintendent of Documents, Government Printing Office, Washington 25, D. C.

Inquiries regarding the Bureau's reports should be addressed to the Office of Technical Information, National Bureau of Standards, Washington 25, D. C.

NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT		NBS REPORT
1102-40-5213	February 6, 1959	6295

A NEW APPROACH TO THE MECHANICAL TRANSLATION OF RUSSIAN

by

I. Rhodes

APPLIED MATHEMATICS DIVISION

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

A NEW APPROACH TO THE MECHANICAL TRANSLATION OF RUSSIAN*

I. Rhodes

Applied Mathematics Division, National Bureau of Standards, U.S.A.

To characterize any transformation of a Source sentence into a Target sentence as a good translation is akin to characterizing a case of mayhem as a good crime, - in both cases the adjective is quite incongruous. Very likely, the source sentence, in itself, is but a poor translation of the intention of the utterer. Moreover, there is no certainty that this original thought has been fully or correctly shaped, before it was enunciated in the form of the source words. It would seem, therefore, that the only target at which a conscientious worker in the field of MT may properly aim is a practical translation.

If, for example, a translated article enables a scientist to reproduce an experiment described in the source paper, and to obtain the same results, such a translation may be regarded as a practical one. Perhaps it is not couched in elegant terms; here and there a target word may be followed by an alternate meaning; a phrase or two may appear as a mere transliteration of the original source words. Nevertheless, this translation has served its main purpose: it has made it possible for a scholar in one land to follow the work of his colleague in another. To this desideratum, we must add at least one other: MT must be effected at least as speedily and economically as is the chore of the human translator, else it would be hard to justify its use at all. At the present time we are all cognizant of the lack of suitable devices which would permit us to realize such an economy. However, engineering development in this field is so rapid that we can expect fairly adequate electronic systems to operate within the next half-decade. We can put the intervening period to good use by preparing a number of experimental programs which would not only guide our choice of an acceptable translation scheme but also throw clearer light upon our equipment needs.

---

*This work was sponsored by the Office of Ordnance Research, Department of the Army.

We started our experiments at the National Bureau of Standards a few months ago, and should like to submit our scheme to our far more experienced colleagues for consideration and criticism. Our program for translating Russian into English consists of two parts.

PART I

This part, which has been completed for the 704, deals merely with the Glossary look-up of individual source words. In order to minimize the time—and thus the cost—of look-up, we store our Glossary in a highly compacted form. Although at present we use, of course, only a sample glossary, we are able to estimate that an External Memory of 4×10^7 bits will be adequate for storing a Russian-English Glossary containing about 50,000 entries. Each entry corresponds to the stem of a Russian word, (i.e. the word deprived of its inflectional ending) accompanied by pertinent diacritical material and by one or more English correspondents, each of which carries its own diacritical notes.

When an Occurrence in a given Russian text is read into the machine—and we have reason to hope that this will soon be accomplished by a fully automatic device—this source material is subjected to the followingtreatment within the computer.

1. An Identification tag, t, is appended to the occurrence to indicate the page, sentence, and ordinal number. Its characters are counted and examined for various indications as to the type of the occurrence, and these are carefully recorded, in the space, S_t, allotted to the occurrence t.

2. If the given occurrence is a word, a search is made in a special—internally stored—List of some 1500 short and frequently used words. This search is carried out only within that portion of the list whose entries contain the number of characters possessed by the given word.

3. If the word is not found in the above list, it is decomposed into its pseudo-prefixes*, pseudo-root* (or roots), pseudo-suffixes* and ending, by means of Corresponding Lists, stored in the Internal Memory.

4. The ending is replaced by an address ϵ, accompanying its mate in the List of Endings, to be used in the second part of the Program.

5. Each pseudo-prefix and pseudo-suffix is replaced by a single character, consisting of 6 bits, and the combination of these characters (probably no more than 8) constitutes the Transform, τ, of the original source word. ϵ and τ

*The machine does not distinguish between true prefixes, roots, suffixes, and those which merely have the appearance of such.

are stored in S_t.

The entries in our externally stored Glossary consist of precisely such Transforms, arranged in groups according to the Pseudo-root which the words in their original form shared in common. Each group thus consists of what we call the Satellites of a certain Pseudo-root. These roots do not appear in the Glossary; as we already mentioned, they are stored in the Internal Memory, each followed by the location $\sim$, indicating where the group of Satellites begins in the Glossary.

6. To this location is attached the Identification tag, t; the combination, $\sim t$, is then intersorted with similar combinations corresponding to the source words processed previously. When all the internal space allotted for the sorted transforms is filled, a search is made throughout the entire Glossary for the corresponding entries. Since the time for such a transit throughout the glossary is formidable, and remains practically constant irrespective of the number of words to be looked up, it is obvious that an appreciable increase in internal storage space would result in a corresponding reduction in the look-up time per word. However, considering the high cost of internal storage media, it might be more expedient to utilize inexpensive non-erasible external storage media with clever buffering devices which allow for the simultaneous retrieval of information along several busses.

7. The last step of this part of the program consists in transferring all the information found in the Glossary corresponding to the stem of each source word, and in rearranging this information into the original order of the text occurrences.

As an example of the performance of Part I of the program, we shall offer the text word РАСПОЛОЖЕНИЕ containing 12 letters. In this case, the combinations РАС and ПО happen to be true prefixes. By referring to the stored list of Pseudo-prefixes, the routine would replace РАС by the letter V, and ПО by the letter R. Unable to discover more prefixes, the routine would isolate the ending ИЕ, and store a corresponding ϵ in the Internal Memory alongside the t of the given word. It would then identify ЕН as a Suffix and replace it by the letter K. Finding no more suffixes, the routine stores the numerals 2 and 1, to indicate the number of prefixes and suffixes, alongside the $t\epsilon$ in S_t. This is followed by the Transform, τ, which is VRK. The machine then enters the subroutine for identifying the Pseudo-root. In the present case, no difficulties will be encountered, as ЛОЖ will be located at once in the List of Pseudo-roots. In actual practice, a number of complications may arise. The given word may contain a polyroot; or what we assumed to be an ending may actually be part of the pseudo-root; or we may not be able to locate the root at all. The subroutine takes note of all these possibilities.

The root ЛОЖ is replaced by $\sim$; to this is attached the t of the given word; eventually, a transit will be made through the Glossary where VRK will be listed,

accompanied by all the material pertinent to the stem РАСПОЛОЖЕН . This material will be transferred to a set of cells in the Internal Memory which corresponds to the original order of the given Occurrence, i.e. in S_t.

PART II

This part of the Program consists of an iterative process which attempts to assemble the individual target words, corresponding to those of a source sentence, into a practical, meaningful, translation. In order to accomplish this, we have introduced such aids as

1. Foresight Pool, containing future Expectations
2. Hindsight Pool, storing previously encountered Conflicts
3. Linkage Numbers
4. Chain Sequences
5. The Morphology of Inflectional Endings.

These are utilized in connection with the following material, gleaned—for each text word—from the Source Sentence by means of the routine in Part I:

1. The relative address, ϵ, of the ending.

2A. The material abstracted from the Glossary, if it contains the current stem, consisting of:

   a. Statistical information regarding the stem, such as

      1) The number of cells occupied by the entry
      2) The number of parts of speech, P, represented by the stem
      3) The number of correspondents for each P.

   b. For each P of the stem

      1) Its Morphology
      2) Predictions as to the Structures, S, it is likely to govern
      3) For each of its English correspondents, E.

         a) A Category Number
         b) A Locution Number
         c) The English meaning in a greatly compacted form, consisting of a Transform (similar to the one described for a Russian word) and a "streamlined" root.

2B. The transliteration, T, of the stem, if the text word is not located in the Glossary.

In order to illustrate the basic principles of our scheme, we shall at first eschew all complications by using an exceedingly simple Russian sentence, culled from an article by Chebyshev. For the present, the source sentence will not be disclosed, in order to highlight the work of the machine. As a result of the routine in Part I, the given sentence would yield the following information.

Occurrence Number	SOURCE					TARGET		
	Morphology of Stem	Predictions of Stem			Ending of word	Cat.	Loc.	English Correspondent
		Case	Urgency	Prepos				
1	[Not in the List of Pseudo-roots]				ИЯ	--	--	T_1
2	Nn, fem, Cl. 6	--	--	--	ЕЙ	C_2	L_2	E_2
3	Vb. Cl. 2 tr. rfl. act. prd.	dat. ins.	1 1	-- as	ИТ	C_3	L_3	E_3
4	Nn. fem. Cl. 1	dat.	1	to	ОЮ	C_4	L_4	E_4
5	[In List of Frequent Usage] Adj. mas. or neut. ins. sing. Adj. mas. or neut. dat. plur.	--	--	--	--	C_5	L_5	E_5
6	Nn. neut. Cl. 3	--	--	--	ИЯМ	C_6	L_6	E_6
7	[The punctuation mark, period.]				--	--	--	.

The iterative process proceeds as follows:

A. Preliminary Housekeeping at the start of a sentence.

1. The Hindsight Pool is cleared
2. The following two initial Expectations are stored in the Foresight Pool:

Predictor No.	Structure	Urgency No.
0	Subject	2
0	Predicate	2,

where an Urgency of 2 indicates an absolute necessity.

3. The Linkage Number is set to unity.

B. Steps executed for <u>each</u> occurrence in the sentence.

1. If the occurrence is a word not belonging to the List of Frequent Usage, the relative address, ϵ , of the Ending will indicate where its morphology starts; this will, as a rule, contain several possibilities.
2. If the stem is contained in the Glossary, its morphology usually serves to eliminate a number of possibilities found in B.1.
3. The remaining one or more possibilities constitute the Temporary Choices for the word's syntax.
4. The machine consults the Foresight Pool. The first Expectation (if any) which is in agreement with any of the Temporary Choices will be accepted as the Final Choice during the current iterative cycle. If no suitable Expectation is found, the first Temporary Choice is accepted as final.

Note: For word no. 1 in our illustration, the Final Choice will be as follows

Predictor No.	S	P	Gdr.	Case	No.
0	Subject	Nn.	fem.	nom.	sing.

5. A fulfilled Expectation is eliminated from the Foresight Pool and the juxtaposition of the Predictor number and the Number of the current occurrence will constitute the current Chain Sequence. The Linkage number remains unchanged. If the Foresight Pool contains no suitable Expectation, a large fixed number (indicating infinity) is used for the Predictor number, and the Linkage Number is raised by one, <u>unless</u> the Current Occurrence is one of the following:

a. A Preposition
b. A Conjuction
c. An Adverb
d. A Punctuation Mark.

6. In the light of the Final Choice the remaining Expectations in the Foresight Pool may be amended, and New Expectations—bearing the number of the Current Occurrence as Predictor Numbers—are culled from the following sources:

   a. The Predictions of the Stem, found in the Glossary, are copied out.
   b. The Grammar Section in the Routine dealing with New Expectations may yield several items.

   Note: For word no. 1 of our illustration, this Section will yield two new Expectations.

Predictor No.	S	Urgency No.	Gdr	Case	No	Prepos.
1	Object	0	-	gen.	-	of
1	Modifier	0	fem	nom	sing	-

   where a zero urgency alerts the machine to erase the Expectation, if it is not fulfilled after the next occurrence.

7. All Expectations in the Foresight Pool are sorted first by Predictor number and then by Urgency number within the Predictor.

8. The Hindsight Pool is consulted in order to ascertain whether the current Choice can throw any light on the Conflicts which have been previously recorded. If any of the indicated Conflicts can now be resolved, the proper amendations, rearrangements, and consequent revisions are carried out, and the notation in question is eliminated from the Hindsight Pool.

9. If the Current Choice gives rise to new Conflicts, they are added to the Hindsight Pool with the Number of the Current Occurrence attached.

10. The following Lexicological Sections of the Routine, are consulted:

   a. The Polysemy Subroutine, if more than one correspondent exists. (This subroutine has not been fully thought out by us. We are relying on a careful assignment of Category and Locution numbers to resolve most of the conflicts. Where they fail to do so, we shall be forced to print out all available target meanings).
   b. The Pidgin Endings Subroutine. In order to save precious Memory space, no attempt will be made, at present, to achieve elegant English; it will be assumed that all words are regular.
   c. The Insertion Subroutine. giving crude rules for placing certain articles or prepositions before, or after, the chosen target word (or words).

Note: If the Current Occurrence is not a word, the appropriate Routine Section is consulted, e.g., one dealing with either Numerals, or Formulas, or Punctuation Marks, et. al.

C. When a period is encountered, a special Subroutine is consulted to establish whether the context of this punctuation mark indicates that the end of the sentence has been reached. If such be the case, it also marks the end of the current iteration, and the following criteria are applied:

1. Does the Foresight Pool Still Contain Expectation with an Urgency of 2?
2. Does the Hindsight Pool Still Contain notations of Unresolved Conflicts?
3. Is the Linkage Number unduly high?
(We have not yet established what value for this variable should be considered acceptable. For the present, we arbitrarily place the maximum at 10% of the number of words in a sentence.)

If the answer to any of the above questions is in the affirmative, we compare the current Chain Sequence Pattern to those obtained in previous iterations and retained in the Memory. If it is identical with one of them, we assume that we have got ourselves into a hopeless "loop" and jump to the "Failure Subroutine". If the Pattern is different, we ascertain whether the maximum allowable number of iterations (at present ten) has been executed. If such is not the case, we raise the count and enter upon another iteration guiding ourselves, inasfar as it is possible, by the aggregate of information gained during the previous iterations. If, however, the maximum number of iterations had not produced satisfactory results, the Failure Subroutine instructs the machine to print out the following:

1. A signal indicating failure
2. All the information about the sentence obtained at the end of Part I
3. The faulty target sentence obtained during the last iteration.

The failures, as well as the supposed successes, will be studied for clues which we hope will lead to the discovery of improved techniques. The Source Sentence of our illustration is

ТЕОРИЯ ВЕРОЯТНОСТЕЙ СЛУЖИТ ОПОРОЮ ВСЕМ ЗНАНИЯМ

The Target Sentence after one iteration, becomes

(THE) [TEOR] (OF) PROBABILITY-S SERV-S (AS) AID (TO) ALL SCIENCE-S.

Notes on the Target Sentence:

a. Words within parentheses indicate additions by the Insertion Subroutines.
b. Material within brackets indicates a transliterated stem. (Our Glossary does not contain stems of non-Slavic origin.)
c. The misspelled target stem "serv" is the result of the curtailment of superfluous letters in the roots of Correspondents.
d. English Prefixes and Suffixes, originally replaced in the Glossary by single letters, are reintroduced before printing.

The Routine in Part II is far more sophisticated than the above illustration would lead one to believe. For example, it takes into account all types of subordinate clauses and has a special Subroutine dealing with Coordinative conjunctions. In brief, this subroutine does the following:

1. A signal is placed in Hindsight to indicate the occurrence of such a conjunction, and the next occurrence is processed. When step 8, above, is reached in the processing of the first word following the conjunction, the Hindsight signal will cause the Routine to revert to the Conjunction Subroutine.

2. The choices of each of the previous words (in reverse order) will be examined for "coordinateness" with the current word, in order to find out what parts of the sentence the conjunction is trying to link together: it could be

   a. two modifiers
   b. two structures
   c. two phrases, etc.

In any given iteration, the Subroutine will accept the first "coordinate" situation, encountered in its backward glance, as the correct one.

It is impossible, of course, to describe in detail the myriad subroutines we plan to include in Part II. The amount of allowable internal space will determine which subroutines will have to be jettisoned for the present.

U. S. DEPARTMENT OF COMMERCE
Lewis L. Strauss, *Secretary*

NATIONAL BUREAU OF STANDARDS
A. V. Astin, *Director*

THE NATIONAL BUREAU OF STANDARDS

The scope of activities of the National Bureau of Standards at its headquarters in Washington, D. C., and its major laboratories in Boulder, Colo., is suggested in the following listing of the divisions and sections engaged in technical work. In general, each section carries out specialized research, development, and engineering in the field indicated by its title. A brief description of the activities, and of the resultant publications, appears on the inside front cover.

WASHINGTON, D. C.

Electricity and Electronics. Resistance and Reactance. Electron Devices. Electrical Instruments. Magnetic Measurements. Dielectrics. Engineering Electronics. Electronic Instrumentation. Electrochemistry.

Optics and Metrology. Photometry and Colorimetry. Optical Instruments. Photographic Technology. Length. Engineering Metrology.

Heat. Temperature Physics. Thermodynamics. Cryogenic Physics. Rheology. Engine Fuels. Free Radicals Research.

Atomic and Radiation Physics. Spectroscopy. Radiometry. Mass Spectrometry. Solid State Physics. Electron Physics. Atomic Physics. Neutron Physics. Radiation Theory. Radioactivity. X-rays. High Energy Radiation. Nucleonic Instrumentation. Radiological Equipment.

Chemistry. Organic Coatings. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Molecular Structure and Properties of Gases. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

Mechanics. Sound. Mechanical Instruments. Fluid Mechanics. Engineering Mechanics. Mass and Scale. Capacity, Density, and Fluid Meters. Combustion Controls.

Organic and Fibrous Materials. Rubber. Textiles. Paper. Leather. Testing and Specifications. Polymer Structure. Plastics. Dental Research.

Metallurgy. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion. Metal Physics.

Mineral Products. Engineering Ceramics. Glass. Refractories. Enameled Metals. Concreting Materials. Constitution and Microstructure.

Building Technology. Structural Engineering. Fire Protection. Air Conditioning, Heating, and Refrigeration. Floor, Roof, and Wall Coverings. Codes and Safety Standards. Heat Transfer.

Applied Mathematics. Numerical Analysis. Computation. Statistical Engineering. Mathematical Physics.

Data Processing Systems. SEAC Engineering Group. Components and Techniques. Digital Circuitry. Digital Systems. Anolog Systems. Application Engineering.

• Office of Basic Instrumentation.

• Office of Weights and Measures.

BOULDER, COLORADO

Cryogenic Engineering. Cryogenic Equipment. Cryogenic Processes. Properties of Materials. Gas Liquefaction.

Radio Propagation Physics. Upper Atmosphere Research. Ionospheric Research. Regular Propagation Services. Sun-Earth Relationships. VHF Research. Ionospheric Communication Systems.

Radio Propagation Engineering. Data Reduction Instrumentation. Modulation Systems. Navigation Systems. Radio Noise. Tropospheric Measurements. Tropospheric Analysis. Radio Systems Application Engineering. Radio-Meteorology.

Radio Standards. High Frequency Electrical Standards. Radio Broadcast Service. High Frequency Impedance Standards. Electronic Calibration Center. Microwave Physics. Microwave Circuit Standards.


NBS