A STUDY FOR THE DESIGN OF AN AUTOMATIC DICTIONARY

A thesis presented

by

Anthony G. Oettinger

to

The Division of Applied Science

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Mathematics

Harvard University

Cambridge, Massachusetts

April, 1954

PREFACE

An automatic dictionary is the fundamental component of an automatic translator. It may be used independently to produce rough translations of technical texts for direct use by specialists in the subject matter of the texts. This thesis is a report on the first steps in the design of an automatic Russian-English technical dictionary.

The author wishes to thank Professor Howard H. Aiken for proposing this study, for continuing encouragement and support, and for many valuable suggestions. He is indebted to Professor Joshua Whatmough and to Professor Roman Jakobson for many helpful discussions and suggestions regarding the linguistic aspects of this study. The responsibility for the conclusions reached lies, of course, entirely with the author.

The collaboration of those members of the staff of this laboratory studying the theory of switching has been highly instructive and is greatly appreciated. The advice of Mr. Robert Ashenhurst, Dr. Kenneth Iverson, Dr. Robert Minnick, and Mr. Mark Pivovonsky is especially acknowledged. Special thanks are due to those who liberally gave of their time and effort to the experiment reported in Chapter 3, and to Mr. John Bilitz for the extended loan of a Russian typewriter.

The author wishes to express particular thanks to Miss Jacquelin Sanborn, who typed the plates for printing, for cheerfully and ably mastering a keyboard in a foreign language. The photographic work 7;as done by Mr. Paul Donaldson of Cruft Laboratory, and the drawings by Miss Carmela Ciampa. Mrs. Joan Allen, Miss Joan Boyle, Mr. Robert Burns, and Mrs. Sandra Grass assisted in the preparation of the manuscript.

CHAPTER 1 AN AUTOMATIC RUSSIAN-ENGLISH TECHNICAL DICTIONARY

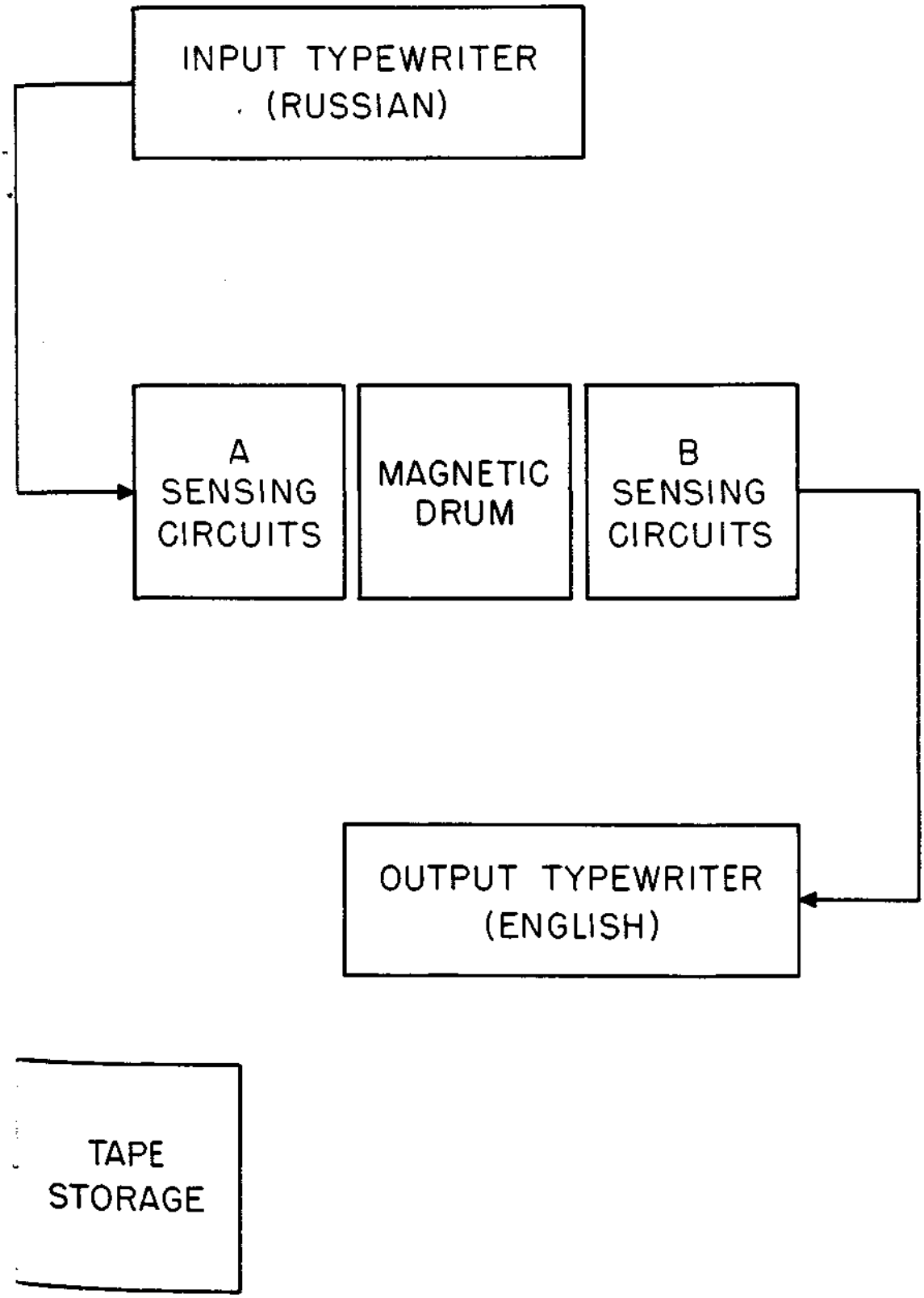
1. The Automatic Dictionary and its Applications.

Modern magnetic drums are capable of storing approximately 10° binary digits. This storage capacity is sufficient to hold nearly five thousand Russian words, together with their most frequently used English correspondents.

Consider therefore the system represented schematically in Fig. 1-1, consisting of a magnetic drum with associated sensing circuits, and of two typewriters. The input typewriter is provided with a Cyrillic Russian) keyboard, and with key-operated contacts which control the drum sensing circuits A. The output typewriter is an ordinary English model with solenoid-operated keys, controlled by the drum read-out sensing circuits B.

The system outlined above would be operated as follows: A Russian text would be copied by an operator using the input typewriter. the depression of the space bar, signalling the end of a word, would be used to start the drum sensing circuits A on a search for this word, which would have been read into a register in the sensing circuits as it was being typed. Once the Russian word would have been found, its English correspondents would be transferred from the drum into a register in the drum sensing circuits B. This register in turn would control the typing of these correspondents by the English output typewriter.

As the result of the operation of this system any sequence of Russian words entered on the input typewriter would be matched on a word-'or-word basis by a corresponding sequence of the most frequently employed



An Automatic Dictionary

English correspondents appearing on the output typewriter. Apparatus performing in accordance with the above specifications may aptly be called an AUTOMATIC DICTIONARY.

Several problems connected with the design of an automatic dictionary are immediately apparent:

1. A storage capacity of five thousand words is adequate for only about ten percent of the words held in a standard desk type dictionary.* A careful selection of the entries must therefore be made. The roost useful selection is likely to be that obtained by the restriction of entries to the most frequent words pertaining to a specific subject, e.g., electronics, aeronautics, or calculating machinery. The above estimate is based on the assumption that means can be found to refer words differing from one another only by case and number, or by tense, person, and number to a single dictionary entry. This entry itself should be as short as is possible consistent with the requirement that it identify the word uniquely. These restrictions are not serious because it is simple to provide auxiliary storage on magnetic tapes. Each tape could store the vocabulary specific to some particu lar field, and the contents of the appropriate tape be trans ferred to the drum as required.

2. A given Russian word may have several English correspondents in which case only an analysis of the context can lead to a correct choice. The number of alternatives can be reduced by the restriction of the dictionary to a specific subject. In some instances one of the alternatives will yield the others by an obvious metaphorical extension, and hence may be treated as a unique correspondent of the given Russian word. Whenever such multiple correspondences exist it is also possible to leave the choice to the reader, rather than to the machine, by requiring the machine to type all of the alternatives, possibly in the order of their frequency.

3. The most serious problem concerns the question of how well a reader could interpret the output of the machine. In this connection, it should be noted that even an expert translator with a profound knowledge of Russian cannot be expected to translate accurately texts dealing with subjects outside the range of his knowledge. It is to be expected, however, that if the automatic dictionary were restricted to the processing of terms dealing with a specific subject, and experts in this

*e.g., A. I. Smirnitskij, Russko-Anglijskij Slovar', Moskva (1949).

subject were to make use of the results, their knowledge of their specialty would greatly assist them in extracting meaningful and useful information from the output of the dictionary.*

The very least one can expect from the use of an automatic dictionary is that a person well versed in the subject matter being treated should be able to assign the output of the machine to one of the following three categories:

1. those "translated" articles from which he can abstract sufficient information to meet his needs;

2. those articles which, while they cannot be fully understood, are clear enough to reveal their irrelevance, in which case they may be ignored;

3. those articles which, although not fully understandable, show sufficient potential interest and importance to warrant their being translated in full in the ordinary manner.

It is believed, therefore, that the automatic dictionary, far from becoming a substitute for the human translator, can become a device which will assist him and those who use his work in obtaining a greater output of useful information with much less wasted effort.

*The following comment by an expert translator is of great interest in this connection:

"Die Kenntnis des Sachverhaltes ist fur den Dolmetscher tatsachlich eine unerlassliche Vorbedingung. Im Laufe der Jahre bin ich auf Grund meiner Erfahrung immer mehr zu der Uberzeugung gelangt, dass ein guter diplomatischer Dolmetscher drei Eigenschaften besitzen muss: er muss in allererster Linie, so paradox es auch klingen mag, schweigen konnen, zweitens muss er selbst in gewissem Ausmass Sachverstandiger in den Fragen sein, um die es sich bei Seinen Ubersetzungen handelt, und erst an dritter Stelle kommt eigenartigerweise die Beherrschung der Sprache. Ohne Sachkenntnis genugen auch die besten Sprachkenntnisse nlcht. Ein zweisprachiger Laie wird die Ausfurungen eines Cheinie-professors niemals ubersetzen konnen, aber ein Chemiestudent, der sich etwas eingehender mit fremden Sprachen befasst hat, kann Sich einem auslandischen Chemiker gegenuber ohne weiteres verstandlich machen."

Paul Schmidt, Statist Auf Diplomatischer Buhne, Athenaum-Verlag, Bonn, 1949, pp. 18-19.

There is, at present, considerable interest in the development of "translating machines." If by a "translating machine" one is to understand "a machine which will present a well-written English version of Russian input text," there is little doubt but that the great difficulties standing in the way of designers will make such machines unavailable for some time to come. It should be clear, however, that an automatic dictionary is one component essential to a translating machine. It is strongly believed, therefore, that the construction of such a dictionary and its application to useful work will lead to a new insight into the problem of translation. It is likely that the final development of full-scale translating machinery should be achieved sooner by this modest approach, than would be possible were the basic plans more ambitious.

2. The Concept of Translation.

> When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode it."
>
> —Warren Weaver

In telecommunication, as in cryptography, the arbitrary nature of any given symbolic representation of information has long been recognized, Consider, for example, a telegraph system. The telegraphist's raw material is a text written or printed using conventional symbols for letters, "a,b,c, etc." The nature of the transmitting medium requires the telegraphist to translate systematically from the "a,b,c, etc.," notation to the familiar ".—, —...,' —.—., etc.," notation. He proceeds to manipulate his key, depressing it briefly to transmit a dot, longer to transmit a dash. The information is now represented by electric impulses traveling along a wire. At the receiving end, these impulses may be used to actuate

a device which punches holes in a paper tape according to a pattern defined by the sequence of pulses and by their duration. The information is now represented as holes in a paper tape. Finally, the paper tape may be used to control a printing device, which reproduces the information in "a,b,c, etc.," form once again.

In the course of this whole process, the same information has been translated several times. These translations happen to be of an extremely simple kind for several reasons. First, they are all one-to-one; each symbol in either one of a pair of codes has a unique correspondent in the other. Second, the translation is done at the most elementary level, i.e., that of letters. It is possible for an operator to memorize the translation rules, and hence to manipulate the key as he reads the printed page. When whole sentences are translated into a single combination of dots and dashes as in the use of commercial codes, the rules become sufficiently complicated to require frequent reference to a code book. Finally, and most important, the rules are completely known, having been defined for this very purpose.

Another interesting illustration of the same process is provided by the translation from decimal to binary numbers required in many automatic calculators. In "coded decimal" machines each individual decimal digit of a number is translated into a corresponding group of binary digits. The number as a whole remains expressed in the radix 10. Consequently, a correspondence need be established between only ten pairs of symbols, and the translation is readily accomplished by simple electric circuits. In "binary" machines, each number as a whole is translated from an expression of radix 10 to one of radix 2. A machine capable of handling numbers of n decimal digits must therefore be able to establish corre-spondences between 10^n pairs of symbols. It is fortunate that the structure

of the number system allows the formulation of relatively simple algorithms for the determination of the binary equivalent of any decimal number. If this were not the case, a table with 10^n pairs of entries would be required.

It is not difficult to see the similarity between the elementary translations described above and translation from one language into another. In either case the process is the recoding of a message into a different set of symbols. Ideally the symbols of the new code must be in correspondence with those of the old code in such a fashion that the message itself remains invariant.

This ideal can be attained without difficulty when one is free to choose the codes. This is not the case with language translation where two codes are given, and the problem is to find the correspondences. It would be tempting to think of changing the codes, if not for the example of the sad history of "artificial" languages. It seems more reasonable to accept the necessity of working with natural languages as a constraint imposed on the translation problem. No algorithms for the translation of natural languages are known. One-to-one correspondences are the exception, not the rule. Correspondences on the level of letters or phonemes are unfruitful, those on the level of propositions thoroughly unmanageable in view of their incalculable number, and those on the level of words are notoriously multi-valued.

3. Technical Russian and Technical English.

The pair of natural languages which will be studied in this thesis are English and Russian. In the absence of one-to-one correspondence the direction of translation must also be specified. Russian to English will be chosen as the direction more useful to English readers, and, it

turns out, as the more tractable one. It was suggested in Section 1 that storage limitations require restricting the scope of the machine to something narrower than all possible discourse. Such a restriction is also in keeping with a desire to define a manageable problem. At least for the present, therefore, Russian and English will be taken to mean technical Russian and technical English. The range of the technical literature to be considered will be restricted, also for the present, to that pertaining to "electronics." These restrictions are not as narrow as it might appear at first glance. If Russian and English have any uniform structure at all, then adapting a machine designed for one area of discourse for use in another area should require at most a change in glossary. In addition, it is in the translation of the vast volumes of technical literature now inaccessible to the majority of American scientists that automatic devises are likely to be most useful.

From the theoretical point of view, the technical literature is likely to have a sufficient degree of statistical homogeneity to make possible the design of adequate machines on the basis of a study of a number of samples of this literature. There would be no point in designing machinery to perform a certain task if the whole task had to be done first in order to design the machinery. It is this consideration which, coupled with respect for esthetic sensibilities, rules out the application of machines to literary works of art, since these often shine by virtue of their originality, or, from a different point of view, by virtue of their deviation from the statistical norm.

Most international technical communication proceeds through the medium of scientific journals, textbooks, and treatises. Indeed; this is **true** of scientific communication in general. It is therefore written or **more** precisely printed English and printed Russian which are the proper

subjects for study. The assumption that the input to the machine is printed Russian text will consequently hold throughout the following chapters. This assumption must always be kept clearly in mind, especially when reference is made to the present literature of linguistics. Phonetic criteria play an important part in linguistic research, but are inadmissible in the present context.