

M E C H A N I C A L

P I D G I N

T R A N S L A T I O N

An estimate of the research value of "word-for-word" translation into a pidgin language, rather than into the full normal form of an output language.

BY

Margaret Masterman and Martin Kay

Cambridge Language Research Unit
Adie's Museum
20 Millington Road
Cambridge, England

July, 1960

SECTION I

Introduction: two long-run lines of research recommended for Machine Translation: one of them is translation obtained by the use of a "pidgin".

It has been tacitly presupposed, in our work, that the only form of Mechanical Translation which is worth trying for is what we, indirectly following Bar-Hillel, call F.A.R.I.M.T. ("Fully Automatised Reasonably Idiomatic Mechanical Translation"). It might be further inferred, from what we have said, that, believing as we do that the basic problem of determining the nature of semantic structure, - technically called the Problem of Multiple Meaning or of polysemy - has got to be faced and solved before F.A.R.I.M.T. becomes even in principle attainable, we do not think it worth while even to try Experimental Machine Translation, - that, as Oettinger has said about us (1), we are "quixotic" about taking the step of trying it.

This does not, however, represent the whole of the picture. We believe that there are two long-run lines of research which, with a Research Group of our composition and resources, are worth pursuing at the present time. One is interlingual F.A.R.I.M.T., done on a randomly chosen text, however short, with the aid of a very large thesaurus. The other is bilingual, word-for-word "pidgin" translation done with a large library of technical-word-and-phrase dictionaries.

In this "pidgin" translation, two conditions are fulfilled:

i) a different dictionary is used for each author (in literature) and each variant of each special subject (in science). With regard to the dictionary, therefore, the choice of the text is always non-random;

ii) the translation is made, not into full English, but into a much more primitive, though still comprehensible, language, the nature of which is discussed in this paper, and which we will call "pidgin".

(1) A. Oettinger: private communication to MM, June, 1959.

These two lines of Machine Translation Research have two features in common. They both set a low estimate, as opposed to the current high estimate, on the information-carrying value of grammar and syntax (see, however, the discussion of "piece-of-translation-information" below)(2). They also both set a high estimate, as opposed to the current low estimate, on the necessity of completely facing the multiple meaning problem before any kind of translation is attempted, no matter how. In the case of F.A.R.I.M.T., and according to the Cambridge Language Research Unit, multiple meaning problems must be attacked by finding a way of constructing a very large thesaurus.

In the case of "word-for-word" "pidgin" translation, they are attacked by the use of four devices. These are the following:

- i) the predominant use of phrases, rather than words, as the unit of the dictionary. (For the definition of "phrase", see below.)
- ii) by the antecedent multiplication of, and choice between, dictionaries.
- iii) by the deliberate use of two specially constructed types of symbol: a) widely ambiguous "pidgin" words, called here pidgin variables, which intuitively, the reader variously interprets, according to the context. b) by the use of a set of specially constructed grammatico-syntactic symbols, called here pidgin markers, which can be used, if desired, for the further transformation of the text,
- iv) by omitting altogether the translation of complex grammatical and semantic features of the text, an attempt to translate which cannot fail to cause a disproportionate complication of the translation program. The hope inspiring these omissions is, that, as it is these same features which cause complication within the thought-processes of the human translator, the text itself will be found to supply some alternative way (language being, as it is, 50 percent redundant) of making the information supplied by the complicating features available. Each of these

(2) CLRU ed. K. Spärck Jones, Semantic Patterns in Discourse. CLRU, ML112.

devices will be later exemplified.

It can without difficulty be inferred from the above that the whole possibility of improving pidgin translation, as such, is dependent upon achieving a state of greater clarity as to what a "piece of translation-information" is. We propose to discuss this point after the report has been given of a series of pidgin-translation trials (or, in a rough and ready sense, experiments) which have been carried out by CLRU. Meanwhile, there are one or two general introductory points which immediately need to be cleared up.

It might be thought that mechanical pidgin-translation, as widely defined by the characteristics given above, is just another way of describing the output of any form of Experimental Machine Translation, no matter what. For, (it might be said), in all Experimental Machine Translation as at present practiced, a special dictionary is used to translate a limited subject-matter; phrases proliferate; pidgin variables (e.e. Reifler's "HE/SHE/IT") which do not occur in full English, form part of the translation-output; special grammatical and/or syntax markers occur in the program, even if they are later eliminated; and some difficult grammatico-semantic features of English (e.g. the use of some auxiliary verbs, or of articles) are deliberately not accounted for by the program. This is true: and that it is true is an integral part of our argument. It is also true, however that the line of research which I here want both to recommend and to report on requires using the phrase "Mechanical Pidgin" in a more restricted sense. In the wider sense, all current M.T. output is Mechanical Pidgin. In the more restricted sense, the use of the five "pidginizing" devices given above does not fully characterize a Mechanical Pidgin. To be a Mechanical Pidgin generated by Mechanical Pidgin Research, as opposed to being an M.T. output produced as a first approximation to producing fuller English, a Mechanical Pidgin output must have the following characteristics, in addition to those which are outlined above:

i) It must be produced by a "word-for-word" procedure which does not allow of any choices being included in the output, between which anyone reading the output must find a way to choose. "The theory behind this rule is that a

reader is less confused by a text containing occasional vague equivalents than by one containing all the possible equivalents of every word."(3)

This procedure has two consequences, both favourable for this research. Firstly, it forces the research group who use it to sophisticate their dictionary-work, rather than to sophisticate their program. Secondly, it enables a sufficient quantity of Mechanical Pidgin to be easily generated for this to serve as raw material for analysis and further processing.

ii) The program must contain no provision for changing the word-order of the text. That is to say, the actual semantic sequence of units as they come into the output language must be what the user reading the output has to understand - somehow.

This forces research into semantic sequences in language, rather than into grammatical patterns in language. Another way of saying this, - which, however, encouraged contemporary grammarians to evade the issue, - is that this limitation pinpoints the importance of studying without cynicism what the older Latin grammarians called "the actual sequence of ideas"(4)

(3) G.W. King, Final Report on Computer Set AN/GSQ-16(XW-1), Vol.VI, Information Coding & Format (1959), p.1. Published by I.B.M. Research Center, Yorktown Heights, N.Y. This report, which will frequently be referred to in this paper, will be given the abbreviation IBM(N), "N" being the number of the vol.

(4)..."Always try to take the ideas in the order in which the Latin presents them. Read every word as if it were the last on the page and you had to turn over without being able to turn back. The mind soon becomes accustomed to the order of any language, as we see by the constant and almost unnoticed inversions of common speech and poetry. If, however, you are obliged to turn back, begin again at the beginning of the sentence and proceed as before. The greatest difficulty to a beginner is his ability to remember the first parts of a complex idea. This difficulty can often be lessened by jotting down, in a loose kind of English, the words as they come in the Latin. In this way it is often easy to see what a string of words must mean, though we should never say anything like it in English...the emphatic position of words plays a most important part in Latin writing...try to feel the emphasis position as you read. As the translation (i.e. word-for-word translation) is made expressly to bring out explicitly the force of order, it should not be taken as a model of desirable translation. Such a translation as is here given forces the emphasis on the attention more than is perhaps natural in English. The force is all present in the Latin, but in English, it may often be left to be brought out by the context, or by some kindred emphasis which the English substitutes. Caesar's Gallic War (Allen & Greenough's Edition), re-edited by J.B. Greenough, B.J. D'Ooge, & G. Daniell, London, 1888, pp. lvii-lviii. [Italics ours]. As this work will be frequently referred to in this paper, it will be given the abbreviation A&G.

The investigation of the nature of semantic sequences and of sequences of ideas is, however, so closely bound up with that of "pieces of information" that, in essence, the two have to be discussed together (see below),

iii) The pidgin must be treated and studied throughout as a homogeneous language with properties of its own (and this without consideration of the fact that specimens of it may be derived from different source languages); not as a defective version of some other language, nor yet as a set of disconnected outputs with no ascertainable properties in common.

Our object in defining a Mechanical Pidgin so restrictively can best be shown by comparing the research-line which is being reported on in this paper with that undertaken by the M.T. group at the University of Washington, Seattle, working under the direction of Professor Erwin Reifler (5). The project of preparing a large Russian-English word-and-phrase lexicon, to be encoded on a computer with photoscopic memory but almost no logic, has caused this group to make an intensive and sophisticated examination of the types of translation which can be achieved by this means. In their Report, under the sub-heading, "Vast Storage Capacity Versus Logical Limitation", they say, "...[The photoscopic] translation-system has a memory device with practically unlimited storage (permanent storage of 30×10^6 bits). It has also an exceedingly low access time (random access time on the order 0.05 seconds). But it does not yet have any logical equipment for linguistic purposes...We decided [therefore] to make full use of the vast storage capacity and to achieve an automatic solution of as many of our linguistic problems as possible through an optimum of lexicography [*italics Reifler's*]".(6) These terms of reference not only caused the Seattle group to consider, and to make the maximum use of, all the same translational devices which are also considered in this paper. It caused them also to make some highly general and cogent remarks, backed by tests, on the relevance of such devices to general problems of M.T. There is thus an analogy between the

(5) University of Washington, Seattle, Linguistic and Engineering Studies in the Automatic Translation of Scientific Russian into English. (Prepared for the Intelligence Laboratory, "Rome Air Development Center, Griffiss Air Force Base, N.Y. Contract AF30(602)-1566 & AF30(602)-1827). Project Director, Dr. Erwin Reifler, Professor of Chinese. As this Report will be frequently referred to in this paper, it will be given the abbreviation ATR.

(6) ATR, p. 7.

lexicographical work which was done by Professor Reifler and his associates and the lexicographical research which is being reported on here. But there is also a difference. In spite of the strong similarity between the M.T. programs used by both groups, and of the tendency of all such programs to produce a highly "pidginized" output, it is still the case that, for developmental reasons, Professor Reifler and his groups desire to minimize the "pidginness" of their Mechanical Pidgin; they desire to make their output as much like full English as possible. CLRU, on the contrary, have set out to expand and to explore the "pidginness" of pidgin; to test the whole Mechanical Pidgin idea to destruction, in order to see what can be done with it and what cannot; and to gain knowledge of, - and if possible, to utilise for M.T. purposes, - any general characteristics which a Mechanical Pidgin may turn out to have which are independent of its particular language of origin.

The body of this paper describes a minimal series of experiments designed to establish the general idea of a Mechanical Pidgin. The first of these was discussed by CLRU in the winter of 1955-56, but was not carried out until the winter of 1959-60; when the first thesaurus results were obtained, the whole idea of investigating Mechanical Pidgin per se was temporarily dropped. In 1958, however, a Latin--English Mechanical Pidgin of c. 700 entries was constructed and put on punched cards, and some dry-run outputs were obtained from it, in order to serve as a control for other forms of M.T. The maxim was that no form of Latin-English M.T. justified itself, unless it was noticeably better than the Mechanical Pidgin translation of the same passage. The extreme difficulty in doing better than the control revived CLRU's interest in the properties of Mechanical Pidgin as such; and in November-December 1959, after International Computers and Tabulators Ltd. had put a punched-card collator, sorter and reproducer-punch at our disposal, an actual pidgin-producing M.T. punched-card program was constructed and debugged. (This program is given in Appendix I). This program performed the same operations as the I.B.M. photoscopic translation system, except that there was no "Rho-stuffing" program.

(7) IBM(VI) , p.55, App. C. Rho-stuffing is a device which with minimal logic and housekeeping, enables the two separated parts of any chunked formula to be correctly rejoined.

It chunked words into sub-words, not by a "peeling off" method, (8) but by a method called by R. M. Needham, who devised it, "exhaustive extraction" (9). It had also a phrase-finding procedure; and performed a one-one dictionary match. It had no device for changing word-order, and has, as yet no mechanical device for actually printing out the output. Some output actually obtained by it is given in Appendix II.

As soon as it was clear that the punched-card word-for-word M.T. program could be debugged and was actually going to work, the series of pidgin-constructing and pidgin-analysing experiments which had been designed over four years earlier was rediscussed, redesigned and actually, carried out. (These are Experiments 1-5, given below). These experiments led to much further and much more basic discussion of the potentialities of Mechanical Pidgin, and of the nature of a "piece of information".

Meanwhile, in March 1960, and concurrently with all this discussion, we had received from Dr. G. W. King of I.B.M. some word-for-word output from the photoscopic translator. This output, which was approximately 6,000 words in length, had been obtained by mechanically translating articles from Pravda and Izvestia into English, and had not been post-edited in anyway. Since the I.B.M. photoscopic translator produces this type of information faster than any other method available, we decided to transfer to this output the set of pidgin-improvement devices which we had used in the smaller scale tests; the result of doing this is shown in the account given of Experiment VIII. The result of this experiment, together with the discussion of the nature of "translation-information" to which it led, seems to us to show that research into the nature of the "pidginness" of pidgin, though it brings up fundamental difficulties, is worth continuing; and some suggestions as to how it should be continued conclude this paper.

(8) E. Reifler, "The Mechanical Determination of German Substantive Composition", Studies in Mechanical Translation No.7, Dept. of Far Eastern Languages & Literature, University of Washington, Seattle.

(9) M.Kay & T.R.McKinnon Wood, A Flexible Punched-card Procedure for Word Decomposition, CLRU, ML119.

Two introductory points remain to be cleared up. It will be argued, of course, by nearly all linguists, that analysis of Mechanical Pidgin per se will not be scientific at all, but artificial, since it is only the set of devices for generating it, namely its mechanicalness, which give it its apparent homogeneity. Such linguists will argue that since there is nothing in common in the texts in different languages from which the pidgin is produced, except the fact that they are all being used for translation into Mechanical Pidgin, any general characteristics of Mechanical Pidgin which may emerge will be either characteristics of the English language which are retained in the pidgin, or else characteristics which are directly inserted into the output by the use of the various "pidginizing" devices, and which therefore do not belong to any language at all.

It can be shown, however, we think, that such considerations are not only implausible, but also unlinguistic. The devices which are used to produce a Mechanical Pidgin were not invented to mislead logicians and linguists. They were framed to enable a pidgin-language to translate as effectively as possible. Moreover, as has just been established above, these devices are characteristic, to a greater or less degree, of all M.T. outputs, whatever dictionary or program is used to generate them, and from whatever language they initially come. Why, then, should they not be presupposed to be endemic to language, and brought out into the open and examined as such?

As to the accusation that other apparently general characteristics of Mechanical Pidgin are merely unconsciously retained characteristics of fuller English, not only is it the case that such characteristics of English can be identified (see, once more, the discussion of a "piece of information"), but it is also the case that, in linguistic studies also, characteristics of the language in which the analysis is made can intrude into the language of which the analysis is made, and that precautions have to be taken against this. Moreover, in linguistic studies also, if the grammar of a language is to be the end-product, heterogeneities within the language due to time-difference, regional differences and

and differences of authorship have to be disregarded, if a homogeneous grammar is to be obtained (10), and these are no different, if formally considered, from the heterogeneities which appear in Mechanical Pidgin outputs when these are obtained from different source languages.(11)

One final point: to say, "We believe that interlingual thesaurus-research, and Mechanical Pidgin research are the most promising long-run lines of research on Machine Translation (note the words underlined) is not to be construed as an implied depreciation either of current experimental work being done on bi-lingual Machine Translation, or of the increasing current use of computers in linguistic textual analysis. What this paper desires to draw attention to is not a defect in current research, but a gap in it. Knowledge is accumulating about how to program a computer to handle translation-material. In time, we are sure, an international auto-coded library of M.T. sub-routines will be built up, and this, in itself, will be a very great gain in all fields of data-processing. At last, also, as linguists become more

(10) There is some interesting discussion of both these problems in Late Archaic Chinese, a Grammatical Study, by W.A.C.H. Dobson, (Toronto, 1959), in which he makes a distinction between analysis and statement. He there says (Introduction, pp. xv-xvi)... "No category [in the analysis] has been recognised which is identifiable only by exterior criteria (e.g. by prior translation into a 'reference language') ... [But] an attempt [has been made] to resolve the problem in statement (though not in analysis) of describing the 'source language' (Late Archaic Chinese) in terms of the 'target language' (English). Hence, while analysis is purely formal, statement takes account of certain linguistic features of the language of description..."

On the homogeneity question, he says, ... "Late Archaic Chinese is an abstraction. The term is a convenient descriptive label, nothing more. It represents a hypothetical norm for the literary language in use in North China in the fourth and third centuries B.C. ... Furthermore, the statement only claims to account exhaustively for the features of four samples taken from [four] authors, each of some two thousand 'characters' in length. The statement has, however, been tested over a wide range of authors and material of the period and found to be generally valid. In this description of LAC no account is taken of possible dialectical, regional or social stratifications of the language, though the presence of such features is hinted in the material itself..."

(11) See below, for examples of Mechanical Pidgin obtained from different source languages.

computer-minded, more of the exactly-obtained structural linguistic descriptions of languages made from adequate samples, and which we have so long desired, will begin to be available to serious scholars. All this is fine. Nevertheless, consideration of the good things which we are going to get one day does not excuse us for not looking harder at what we have got now, especially those of us whose desired end-product is not general routines for data-processing, nor uni-lingual linguistic description, but Mechanical Translation itself. What is coming out of the machines at the minute is a pidgin, whose characteristics per se are never investigated nor their implications followed up. Either the samples of this pidgin are immediately post-edited into a more ordinary form of English: or it is explained away as "low-level M.T.", or "rough M.T;" (12) or some vague, euphoric remark is made to the effect that pidgin M.T. is all right for most purposes (13), which covers up the fact that no investigation is being made to discover what it is. It comes to this: either the "pidginness" of current output is hastily forgotten about, or it is taken for granted. The suggestion made in this paper is that it should be hauled out into the light; and that this, together with thesaurus-research, - or indeed, of any other research which assists, instead of evading, down-to-the-bone consideration of polysemy, - are the long-run ways through to increasing our knowledge of Mechanical Translation, as opposed to increasing our knowledge of general data-processing, or computer-aided linguistic research.

(12) Personal communication to MM from D. W. Davies, National Physical Laboratory, Teddington, Middlesex, June, 1959.

(13) ..."The translation system is capable of deriving adequate English equivalents for a high percentage of Russian inputs... I.B.M. (I), p. 12. ..."useful and meaningful translations can be performed automatically in most cases..." (I.B.M.(I), p. 20. ..."The translation that follows each Russian entry is restricted to as few meanings as possible. For the most part, only one meaning is given... For these Russian words that have a common as well as a technical meaning, the dictionary lists only the technical English equivalent. We have found that our dictionary is sufficient for the purpose of providing comprehensible translations..." I.B.M.(VI), p. 13.

Experiment VIII: Sophistication, on principles
developed in Experiments I-VII,
of Raw-Pidgin output obtained
from I.B.M.

The design of this experiment was taken from that of Experiment V, but a longer text was produced and we had the benefit of a bi-lingual Russian consultant, Mr. G. Trapp, of the Department of Slavonic Studies, Cambridge University. The improvement as between the raw-pidgin and the sophisticated pidgin output was therefore much more marked.

This improvement was achieved above all by the prodigal use of phrases, and by the correct translating of proper names. With regard to the first, the definition of a phrase (or cliché) as established by Mr. Trapp, was, "Any combination of words which occurs twice or more in any given type of literature". This is the widest definition of a "phrase" which could possibly be given, and 42 such actual phrases were inserted into the text.

The totality of possible phrases required to get this information over, in all its forms, is however, as computed by Mr. Trapp as 170 (about). It is worth remarking that Mr. Trapp insisted of his own accord on attaching minimal "x-values", based on strictly grammatical consideration, to each phrase, in spite of the fact that CLRU had by then definitely established that no such x-values could ever be computed.

Examination of the words which the I.B.M. Russian-English dictionary had left untranslated revealed a depressing fact, relevant to all M.T., to which, in our view, too little notice as up to now been given. This is that, in any text for translation, understanding of a new word, (coined by the author for the purpose of constructing his argument), or of one rare word, (which almost certainly will not be in the dictionary) or of some key proper names, (which, if the input is Russian, the dictionary will almost try to translate, not knowing that they are proper names) is essential if the output is not to be garbage. At present, we see no short-term way round this difficulty.

We were both horrified, as were the rest of CLRU, to discover, under the pithy guidance of Mr. Trapp, to what

extent we had misunderstood the I.B.M. raw-pidgin output without realizing that we had done so; i.e. the extent to which the raw-pidgin output was garbage. The raw-pidgin mistranslations of creation (hundredththief), defeat(disease), orthodoxy(rightglory), calendar(country) and belonging to (consisting in), together with proper names and occasional mis-chunkings greatly contributed to this.

We both wish to thank Dr. G. W. King, on behalf of CLRU, for making available to us, immediately on request, a large and assorted quantity of the I.B.M. output, even though he was warned of the use which would be made of it.

The experiment was done by Masterman, Kay, Mr. Trapp and L. Braithwaite with punched-cards, not with paper slips, but the only use made of them was for 2-pocket sorting, to produce the lists.

The list of documents relevant to the experiment, and which immediately follow, in order, is given below:

List of documents used for or prepared in Experiment VIII:

- a) Photostatted Russian input text (from Pravda).
- b) Photostatted raw-pidgin output, supplied by I.B.M.
- c) Input text in list form, giving:
 - i) where Rho-stuffing was used in the experiment.
 - ii) English phrases created in the experiment.
 - iii) Russian phrases created in the experiment, with phrase-increase "x-values".
 - iv) "Scientific-technical" special pidgin dictionary words and phrases used in text.
 - v) "Science-and-religion" special pidgin-dictionary words and phrases used in text.
 - vi) "Religious" special pidgin-dictionary words and phrases used in the text.
 - vii) "Literary Vocabulary" special pidgin dictionary words used in text.
- d) Sophisticated pidgin-output, annotated.
- e) Idiomatic English translation of text, prepared by Mr. Trapp and L. Braithwaite.

I.B.M. OUTPUT

wda - november 3, 1959

article no. 2

page 1

by pagean magazines

Science against religion

In arsenal means scientific-ateisticheskoy propaganda, materialistic training worker appeared new weapon - magazine "Science and religion". Issued in light it first number. This collection contained kh, various, with/from interest reading matter evily, sharp that directed against religious severly and prejudice.

Magazine publish All-Union society by propagation political and scientific knowledge - mass organization advance Soviet intelligentsia, prizvannoy active help party in realization decision/solution XXI congress CPSU, in forming man communist society. Accomplishment this noble problem assume all surmounted religious ideology, propaganda materialism and ateizma.

"Peoples education, propagation scientific knowledge, study laws nature, - talk comrade N.S. Khrushchev, - not leave place for believe in ga". Material magazine illustrate this position convincing examples out of our socialist life, practice communist building in USSR, success in the area of astronomy, atomic physics/physicists, rocket technology/technician, aircraft building, agrobiolgy, such prominent creations human reason, how creation and start Soviet artificial satellites earth and cosmic rocket.

"Man - power nature" - thus is called one of divisions magazine. It contain article academician V.A. Ambartsumyana "Science about Universe and religion ya" and converse with/from doctor physicist atenaticheskikh sciences O V. Kukarkinyn about start second Soviet cosmic rocket on Moon. From these materials reader will see that since Minesrnik threw call church by authority in explanation nature, astronomy became bed conquer one position for/after other, expeling ga out of all sections material world/peace. Idea existence ga, idea hundredth thief world/peace suffered full disease. Every shag in development-studies about Universe all greater convince us in right those materials in and falsity religious world view.

The same thought underline in article academician A. I. Parina "By istokov life", that answer on question: how

appeared life, as well as whence appeared on Earth numerous animal and plant, how happened man? Scientific cognition, talk in article, decide/solveing problem origin life, finally developed chalo religious idea of hundredththief living essence nificheskim)gon.

PuOkished in magazine article A. Valyatynova "About ratio Sovietgo state to religion and church" and L. Mitrokhira "Contemporary rightglory" arm reader correct understanding policy Communist party and Sovietgo government in respect to religion and church. They explain yayut one of article Constitution USSR, talking about that that for the purpose of support for/after Soviet grazhdanani freedom sovesti church in our country department from state and school - from church that freedom sending religious kul'tov and freedom antireligious gropagandy recognize for/after all grazhdanani.

V. I. the Lena in own time indicateed, how much "is important use those book and pamphlet, that contain much concrete facts and comparison, showing connection classs interests and classs organization contemporary bourgeoisie with/from organizations religious founded and religious propaganda". In this meaning very urgent located in magazine article about first entsiklike (message) chapter Roman-Catholic church papy Ioanna XXIII with/from that it inverted at the end June present year to episkopam, sacredkan and believeing, consisting in Catholic church. Article expose authentic nature papskoy entsikliki, penetrate unnavist'yu to Communism, to progress, show inseparable connection Cotton with/from international reaction.

Greater interest represent napechatannaya in magazine ateisticheskaya work known contemporary - English scientist and philosopher T)rtrana Disperseed "Introduceed whether religion useful contribution in civilization?", and also not published till now letter remarkable Gernango thinker biggest philosopher-materialist Peopleviga Feyerbakha, that contain series idea, and ponyne preserveing own value in struggle against religion.

Significant place in magazine occup division "Relate, fel'eton, pamphlet". In it expose prestupnaya activity forbidden in our country sekt, noseshchikh izuverskiy nature, - iegovistov, fiftiethkov and other, show, how church and sectarian attempt support, ozhivit' -

religious experienceki by separate Soviet people. One of materials this division draw dirt appearance rogue Shavrova, that debt/duty year evidently disregarded laws and morals our society, deceived and developprashchal believeing, createed bend make.

Valuable that already with/from first number magazine strive establish close contact with own readers, publish them/their letter. Interesting letter ural'skogo worker A. S. Shark and answer to it old greatervichki, member CPSU with/from 1903 year, E. City of Lefttskoy, in which goes conversation about that, is/eat whether fate. Instructive also letter worker Stalinburnkoy G()r()e()with/from A. Zaytseva, relateing about that, how it pulled out of pautiny sectariango studies.

Useful information reader pocherpnet out of "Country's red-hotgift", devoteded glorious fighters against religious obscurantism past century's Russian thinker, writer-satiriku Antiokhu Kanteniru and prominentiya XVIII century Deni Didro. Contents magazine varioustookzya t and such division, how "Ateisty for/after work", "In world/peace book". "Out of last mail", "Our encyclopedia".

New monthly popular science ateisticheskiy magazine "Science and religion" calculateed

on massgo reader, and also on lektorov and propagandist, occupying scientific-ateisticheskoy propaganda. In this connection wanted would wish in order to it material, especially devoteded achievements science, were expounded yet popular and accessible.

Let with pages new magazine sound passion, scientific argumentrovannoe word militant ateizma, reaching it/then clear from religious prejudice, calling to active struggle for/after celebrate Communism itselfgo bettergo, itselfgo true society!

Russian Input

On the following pages, the Russian text together with information relevant to the pidgin translation, is given. The key to the columns is as follows:

Column 1:

The number before the period refers to the line in the original Pravda article on which the word begins and following the period is the number of the word in that line.

Column 2:

Russian input words. All punctuation marks count as separate words as does the symbol "§" marking the beginning of a paragraph.

Column 3:

The letter "p" in this column indicates that "congruence matching" was used to identify the parts of the word in the dictionary.

Column 4:

The letter "e" indicates that this word is translated by a phrase in the output.

Column 5:

The letter "r" indicates that this is the first word of a phrase in the Russian. The following number shows how many words make up the phrase. Thus, if "r3" appears against a word, that and the following two words make up a phrase.

Column 6:

The following numbers refer to the special pidgin-dictionary word and phrases used in the text:

- 1 Scientific-technical
- 2 Science and religion
- 3 Religious
- 4 Political-social
- 5 Literary.

* "Rho-stuffing".

In arsenal - - of methods of scientific atheistic propaganda, of the materialistic training of the workers appeared new weapon-journal - - "Science and religion". It was published that one first number - - . This comprehensive collection of, various, with interest which are read material-s-ism, point* which is directed against religious superstitions - - s-ism and prejudice-s-ism .

Journal - - published is by the all-union society for the propagation of political and scientific knowledge, mass independent* organization-by forward* Soviet intelligentsia-'s, called* active to help the Party to carry out the decisions XXI congress-ism CPSU, in forming a member* of the Communist society . Accomplishment this noble problem-'s assume-s full overcoming* religious ideology-? , propaganda - - materialism-ism and atheism*-ism .

"People's education - - , propagation - - of scientific knowledge , say-s comrade N.S. Khrushchev , - not leave - place-'s for belief in God*" . Material-s journal-ism illustrate - - this position - - convincing example-s-by out of our socialist life-? , practice-'s Communist building-'s in USSR , success-s-by in the region* astronomy-? , of atomic physics , of rocket technology , aircraft construction-'s , microbiology-'s , such prominent creation-s-by human reason-ism how/as creation - - and setting off - - of Soviet artificial earth satellites and cosmic rocket-s-ism .

"Man - - master* - - nature-'s" - thus called+is one out of division-s-ism journal-ism . That one contains article - - academician -? V.A. Ambartsumyana "Science - - about Universe-? and religion - - " and discussion with doctor of mathematical* and physical sciences B.V. Kukarkinym about setting off - - second Soviet cosmic rocket-'s on the moon - - . Out of these material-s-ism reader - - will see, that since that time , how/as Copernicus* challenged* church authority-ward in explanation-? nature-'s, astronomy - - became gradual* to conquer one position - - for/after another , expelling God*-? out of all section-s-ism of the material world . Idea - - of the existence God*-? , idea - - of the creation* of the world suffered full defeat* . Every step - development*-? teaching-? about Universe-? all greater convince-s us in rightness-? materialism-ism and falsity-? religious world-view-'s .

The same thought - - underlined-is in article-? .
 academician-? A.I. Oparina "at source*-s-ish life-?" ,
 which* answer-s on question-s: how/as arise-ed life - - ,
as+well+as whence appear-ed on Earth-? numerous animal-s
 and plant-'s , how/as happened man - - ? Scientific
 cognition - - , said-is in article-? , having+solved+the+
problem origin-'s life-? , final debunk-ed religious idea - -
 on creation-? of+living+beings mythical* God-by .

Which+are+published in journal - - article-'s A.
 Valyatinova "About relation*-? Soviet state-'s to religion-?
 and church-?" and L. Mitrokhina "Contemporary orthodoxy* - -"
 arm - - reader-? correct understanding-by of+the+Communist+
party+policy and Soviet* government-'s with+respect+to
 religion-? and church-? . They explain - - one out+of
 article-s-ish Constitution-? USSR , say-ing about this that
for+the+purpose+of maintenance-? for/after Soviet citizen*-s-by
freedom+of+conscience church - - in our country-? separated+
 is from state-'s and school - - --- from church-? , that
 freedom - - of+religious+practice* and freedom - - anti-
 religious propaganda*-'s recognized+is for/after all
 citizen*-s-by .

V.I. Lenin* in own time - - indicate-ed , how+much
 "is+important utilization - - those of+books and of+pamphlets ,
 which* contain - - much concrete fact-s-ish and comparison-s-
 ish , showing connection - - of+class+interests and of+class+
organization contemporary bourgeoisie-? with/from organization-s-
 by religious institution-s-ish and religious propaganda-'s."
From+this+point+of+view* very timely*-is which+is+located in
 journal - - article - - about first encyclical-? (message-?)
 head*-'s Roman-Catholic church-? pope*-'s John* XXIII
 with/from which that+one address-ed at+the+end June-ish
 present year-'s to bishop*-s-ward priest*-s-ward and to+
 believers* , belonging+to* Catholic church-? . Article - -
 expose-s authentic nature - - of+the+papal+encyclicals* ,
 penetrated hatred*-by Communism-ward , to progress-ward , show-s
 inseparable connection - - of the Vatican* with/from inter-
 national reaction-by .

Great interest - - represent - - which+is+printed* in
 journal - - atheistic* work - - contemporary English scientist-?
 and philosopher-? of+Bertrand+Russell*"introduce-ed whether
 religion - - useful contribution - - in civilization - - ?" ,

and also not which+is+published till+now letter* - - remarkable German thinker-? biggest philosopher-? -- materialist-? of+Ludwig+Feuerbach* , which contain-s series - - idea-s-ish , and up+to+now* preserve-ing own significance* - - in struggle - - against religion-? .

Significant place - - in journal - - occupy-s division - - "Story* , serial* , lampoon*" . In that+one exposed+is criminal* activity - - forbidden in our country-? of+sects* , carrying* barbarous* nature/ (omission: see below) , shown-is now/as churchgoer-s and sectarian-s attempt - - to+support , to+revive* religious anachronism-s at individual Soviet people-? . One out+of material-s-ish this division-ish depict*-s dirty character* - - rogue-? Shavrova , which for+many*+years open* disregard-ed law-s-by and morale-by our society-'s , deceive-ed and corrupt*-ed believer-s? , create-ed vile* deed*-'s .

Valuable+is , that already with/from first number-'s journal - - strive-s to+establish close connection*-? with/from own reader-s-by , publish-s them/their letter-'s . Interesting+are letter - - Ural* worker-? A.S. Akulina* and answer - - on that+one old Bolshevik+female*-'s , member-? CPSU with/from 1903 year-'s , E. Levitskoe* in which go-s conversation - - about this , is/eat whether fate . In-structive also letter - - worker-? Stalinogorsk* hydro-electric+power+station* A. Zaytseva , relate-ing about this , how/as that+one extricated*-is out+of web*-'s sectar-ian teaching-'s .

Useful information-'s reader - - will+get* out+of "Page-little-s-ish calendar*-ish" , which+are+devoted glorious fighter-s-ward against religious obscurantism-ish past century-s-ish - Russian thinker-ward , writer-ward-satirist*-ward to+Anti+ch+Kantemir* and prominent figure - - French education-'s XVIII century-'s Denis+Diderot , contents - - journal-ish add+variety*+to and such division-s , how/as "Atheist*-s at*+work" , "In world/peace - - of+books" , "Out+of last mail-'s" , Our encyclopedia" .

New monthly popular scientific atheistic* journal - - "Science+and+religion" calculated-has+been on mass reader-? and also on lecturer*-s? and propagandist-s? occupy-ing-self by+scientific+atheistic*+propaganda , In this connection-? wanted would to+wish+that* that+one material-s , especial

(Omission) the+followers+of+Jehovah* , the+liberal+movement+of+the+50's* and other, shown-is.....

devot-ed achievement-s-ward science-'s, expounded-would+
be yet more+popular and more+accessible .

Let from/with of+pages new journal-ish resound-s
passionate* scientific argued* word - - militant atheism*-ish ,
reaching minds-ish and heart-s-ish people-? , help-ing them-
ward to+free+self* from religious prejudice-s-ish , call-ing
to active struggle-? for/after triumph* - - Communism-ish -
most* best , most* just society-'s .

Notes:

- 1) Russian phrases underlined.
- 2) All English phrases marked with crosses.+ .
- 3) Punctuation and capital letters are as in the Russian.
- 4) Untranslated or wrongly translated in I.B.M. output
marked with asterisks*.

Endings: No adjective has been given an English ending.
Nouns and adverbs which could possibly be adjectives have
been treated as adjectives. The Russian noun-endings have
been translated into English as follows:

<u>Singular</u>	<u>Plural</u>
nom acc. - -	nom. acc. - -
gen. -ish	gen. -s-ish
dat. -ward	dat. -s-ward
prepositional - -	prepositional - -
instrumental - by	instrumental - by
ambiguous endings -?	ambiguous but definitely plural -s?
present tense of verbs - -	plural subject -s sing. subject.

Ambiguity between nom. plur. and gen. sing. is denoted by -'s.
Certain gen. plur. consist of stem with no endings. These
would appear as separate dictionary entries. They become
therefore "of+sects" etc.

Mistakes in I.B.M. output about 1/2 not in his dictionary.
Quite a few through attempting to chunk proper names owing
to not distinguishing capital letters.

A look at the journals.

Science and Religion.

Among the mediums for propagating scientific atheism, for the materialistic education of the workers, a new weapon has appeared - the journal "Science and religion". The first number has been published. It is a comprehensive collection of interesting and varied material, whose main thrust is directed against religious superstition and prejudice.

This journal is published by the national society for the propagation of political and scientific knowledge, an independent organization run by the more outstanding Soviet intellectuals for the benefit of the general public; they were called upon to put to practical effect the decisions of the 21st congress of the Soviet Communist party, namely the decision to form an ideal member of Communist society. The execution of this noble task implies the complete abolition of religious modes of thought and the propagation of materialism and atheism.

"The education of the general public, the propagation of scientific knowledge, and the study of the laws of nature," says comrade N. S. Krushev, "result in their being no place for belief in God." The contents of this journal illustrate this position, both by convincing examples drawn from our socialist life, of the building-up in practice of Communism in USSR; and by the successes in the fields of astronomy, atomic physics, rocket technology, the construction of aircraft, and agrobiolgy, and by such outstanding creations of the human mind as the development and launching of Soviet artificial Earth satellites and interplanetary rockets.

"Man - the master of nature", this is the title of one of the sections of the journal. It contains an article by the academician V. A. Ambartsumyan "Religion and the study of the Universe", and a talk by doctor of science B. V. Kukarkin on the launching of the second Soviet interplanetary rocket to the Moon. From these articles the reader will see that since the time that Copernicus challenged the authority of the Church by his explanation of nature, astronomy gradually conquered one position after another, chasing away the concept of God from all aspects of the material world. The ideas of the existence of God and the creation of the world suffered complete defeat. Every stage in the development of the study of the Universe convinces us more and more of the truth of materialism and the falsity of the religious Weltanschauung.

The same line of thought is emphasised in an article by the academician A. I. Oparin "the origins of life", which answers the question - how did life arise? - as well as from where the numerous species of animals and plants appeared on Earth and how man originated. Scientific knowledge, the article says, by solving the problem of the origins of life has finally debunked the religious idea of the creation of living beings by a mythical god.

Two articles published in the journal, by A. Valyatinov "The Soviet state's attitude to religion and the Church" and by L. Mitrokhin "Orthodoxy in our time", arm the reader with a correct understanding of the policies of the Communist party

and the Soviet government, with respect to religion and the Church. They make clear one of the clauses of the USSR Constitution, which states that for the purpose of ensuring freedom of conscience for the Soviet citizen the Church in our country has been separated from the State, and education from the Church, and that freedom of religious practice and the freedom to propagate anti-religious ideas is recognised for all citizens.

Lenin in his time indicated how "important it is to use those books and pamphlets which contain many concrete facts and comparisons showing the connection between the class interests and the class organizations of the present-day bourgeoisie, and their religious institutions and propaganda!" The article in the journal about the first encyclical (message) issued by the head of the Roman-Catholic church, Pope John XXIII, which he addressed at the end of June this year to the bishops, priests and faithful of the Catholic church, is very appropriate in this context. The article exposes the authentic nature of the papal encyclicals, impregnated with hatred of Communism and of progress, and shows the inseparable connection between the Vatican and the forces of International Reaction.

Published in the journal and of great interest are an atheistic treatise by the well-known contemporary English scientist and philosopher Bertrand Russell "Has religion been a useful contribution to civilization?", and also a letter, hitherto unpublished, by the most remarkable and important German thinker and materialist-philosopher Ludwig Feuerbach, containing a number of ideas which are still relevant in the struggle against religion.

An important section of the journal is "Story, Serial, Lamoon". In it are exposed the criminal activity of the barbarous sects forbidden in our country, the followers of Jehovah, the liberal movement of the 50's and others. This section shows how the churchgoers and the members of the sects endeavour to maintain and to revive religious anachronisms among individual Soviet people. One of the items of this section depicts that dirty rogue Shavrov, who for many years disregarded the laws and customs of our society, deceived and corrupted believers, and performed infamous deeds.

What is valuable is the journal's attempt - even from its first number - to establish close ties with its readers and to publish their letters. The following letters are interesting, one from a Ural worker A. S. Akulin and the reply to it from an old Bolshevik woman, a member of the Communist party since 1903, E. G. Levitskaya, in which they discuss the question of free will. Also instructive is a letter from a worker in the Stalinogorsk hydro-electric power station, A. Zaytsev, describing how he extricated himself from the web of sectarian teaching.

The reader will obtain useful information from the section called "the pages of the calendar", which is devoted to the glorious fighters against religious obscurantism in past centuries - to the Russian thinker and satirist Antioch Kantemir and to the prominent French 18th century man of letters Denis Diderot. Such sections as "Atheists at work", "in the world of books", "From our latest mail", and "Our Encyclopedia" add variety to the contents of the journal.

The new popular monthly scientific and atheistic journal - "Science and Religion" - has been designed for the general reader and also for lecturers and those concerned with the propagation for scientific atheism. In this connection one would wish that

the contents, especially those devoted to the achievements of science, were expounded in a yet more general and accessible manner.

May the passionate and scientifically argued ideas of militant atheism ring out from the pages of this new journal, reaching the hearts and minds of all people, helping them to free themselves from religious prejudice, calling them to fight actively for the triumph of Communism, the best and most just of societies.

Conclusion and Summary

The series of experiments reported in this paper were performed with the following objects, practical and analytic, in view:

A. Practical Objectives

i) To provide, by using a strict word-for-word procedure, control-translations for certain pieces of text which were also being mechanically translated or mechanically analyzed by other means. See especially on this Experiments VI and VII,

We shall continue to use this procedure for obtaining word-for-word control-translations, as and when we need them; but it is intended that, before we use the program again, we shall add to it the developments, and, among these, especially the Rho-stuffing device, listed under ii) below.

ii) To use punched-card machinery for Experimental M.T. See especially, on this, Appendix I of this paper, and also the following sections of the long photo-lithoed CLRU Report, currently being issued: A Flexible Punched-Card Procedure for Word Decomposition, by M. Kay and T. R. McKinnon Wood (now ready) and Notes on the Presentation of Punched-Card Programs, by M. Kay and C. Wordley (in press).

On pp. 26-27 of the former paper three improvements which it is intended to make to the program are described in outline. The first of these is to keep separate in the dictionary a sub-dictionary of initial chunks. The second is to have a new character in the alphabetic code marking the end of a word. (This has the same effect upon the program as inserting into the dictionary a separate dictionary of final chunks, but because of the nature of the hardware, a different technique has to be used.) The third is, effectively, Rho-stuffing as used by I.B.M. and which is called by CLRU congruence matching.

All these three additions can be inserted into the program with minimal trouble. In addition, survey-work will be done to see if it is worth while to add to the program a routine for finding discontinuous phrases occurring in the input text. (The program already finds continuous phrases, as can be seen by looking at Operations 5 and 6, in Appendix I).

iii) To enable any required number of Experimental M.T. trials to be carried through quickly and cheaply and with simple machinery, so that subsequent programs, and together with large pidgin dictionaries, prepared for big digital computers could be drawn up with less wastage of time and labour than is currently accepted as inevitable in Experimental M.T. work.

We are at present not sure how many more such punched-card experiments we shall do in the immediate future (see on this the discussion of theoretic objectives below). What we are clear about is that the more we use punched-card machinery for this purpose, the better we like it. It is simple to use, unpretentious, and yet completely determinate; it drastically cuts the cost and labour of M.T. experimentation, while yet encouraging enterprise and imagination in it. It also encourages a certain light-heartedness; the cost of failure, when produced by this method, is not so high as to prevent experiments being done which, like Experiment III, look at first sight plain silly. Thus, by using this technique, M.T. Experiments become really experiments, (though sometimes, like Experiment V, trivial and futile). On this technique, M.T. experiment is sharply distinguishable from M.T. demonstration.

B. Theoretic Objectives

The deeper, theoretic objectives of the exercise were the following:

i) To analyse the whole concept of Mechanical Pidgin as it stands; given that some form of Mechanical Pidgin is what is in fact being currently generated by all M.T. programs of all kinds, although it is usually immediately post-edited into fuller English. See especially for a discussion on this, Section I, The Introduction.

ii) To see whether it is possible to evolve any principles of design, for a Mechanical Pidgin, so as to facilitate and speed-up the whole process of experimental M.T. dictionary-making.

To date, this effort has been only partially successful (see on this Experiments II, IV and VIII), since we have not yet tackled the basic problem of getting down the rate of phrase-increase, which alone (as Experiments V and VIII

show) checks the emergence of the phenomenon of M.T. garbage-production.

Nevertheless, we are glad that we made the effort, and feel that we know something more of the pidgin-basis of language as a result of it. In particular, we are interested in exploring further Zipf's Group I-Group II distinction as brought out by Experiment III.

To analyse this distinction on sufficiently large samples to be interesting requires a character-recognition machine. When a great many kinds of information are to be extracted from one text, it is indeed worthwhile, as many people, headed by Oettinger, have pointed out, to key-punch it. When, however, only one kind of information is to be extracted from many texts, the use of a character-recognition machine becomes imperative.

iii) To see if there can be found any reliable inter-lingual basis, which could be used to guide the design of any Mechanical Pidgin. See especially on this Experiment IV.

Although this objective may sound far-reaching, we are concerned to pursue it; and the further set of suggested experiments which are given below would be, in our view, very well worth doing, though we do not, at present, see ourselves having time in the immediate future to do them.

Further suggested experiments to develop a generally applicable Mechanical Pidgin;

1) Take sets of sentences comparable to those used in these experiments, but from other sets of 20 languages. Compare the analyses of the markers of the resulting "languages" and analyse for overlap of function.

2) Take 20 paragraphs, instead of 20 sentences, from the set of 20 languages chosen by Richens. Compare the analysis of the resulting "language" with the analysis obtained by these experiments.

3) Repeat Experiment IV, using other languages to form the basis of the output pidgin: e.g. Russian, Hungarian, Chinese. Compare the resulting sets of markers for trans-lateability.

In other words, there is a research future, in our view, for pidgin-marker research, if only in the hope of obtaining further light than we have now on the role of prepositions (and their post-positional translation-analogues), and of post-verbs (and their pre-positional translation-analogues) in fuller languages.

For the senses in which "pre-positional" and "post-positional" are used here, see Experiment IV.

iv) To examine the whole phenomenon of phrase-increase in dictionaries.

To do this completely would be, in our view, to test word-for-word M.T. to destruction. This being so, it will be evident that both our Experiments (V & VIII) and our analysis of the way to treat the phenomenon are inadequate. See especially, on the latter, the unsatisfactory discussion of the desirability, and impossibility, of computing what are there unsatisfactorily called "phrase-increase factors" for any M.T. dictionary.

Nevertheless, we are glad to have made the attempt. Making it, and especially, performing Experiment VIII, has brought the phenomenon of M.T. garbage-production to our notice with a new urgency. And if the only way to correct it, in word-for-word M.T., is by a huge-scale phrase-increase, then it is clear, too, that the basic and urgent research-problem, for this kind of M.T., is to learn how first to get insight into, and then control, the present unbridgeable activity of M.T. phrase-making.

We are already starting to tackle this problem by making a sample thesaurus of phrases, giving "pieces of information", these "pieces of information" being interlingually defined and interrelated by means of an interlingua with two connectives and of the order of 100 elements. By means of this we can define, for the first time, "low-level translation" in such a way as to separate all forms of it from "garbage"[†]. But even this sample thesaurus is not nearly finished yet.

*This phenomenon was first pin-pointed by Ida Rhodes at the M.T. conference at Los Angeles, held in February, 1960,
[†]At present, "low-level M.T." and "garbage M.T." are formally indistinguishable, and the fact that they are indistinguishable constitutes the impasse in which Experimental M.T. is increasingly landing itself.

For the grisly fact remains, - and it is equally relevant to all forms of M.T., that in order to get any technical phrase or cliché out of an M.T. dictionary or thesaurus, you have first to put it into it. And there are a great many technical phrases and clichés in language.

Margaret Masterman

Martin Kay

Cambridge Language Research Unit

July 13th 1960