REPORT
OF THE EIGHTH ANNUAL
ROUND TABLE MEETING
ON LINGUISTICS
AND LANGUAGE STUDIES

Research in Machine Translation

Edited by
Léon Dostert

GEORGETOWN UNIVERSITY PRESS
Washington, D.C.
FROM time to time The Institute of Languages and Linguistics, School of Foreign Service, Georgetown University, publishes monographs intended to contribute to the discipline of linguistics, and to the study and teaching of languages. Manuscripts should be addressed to

Léon Dostert, Editor
Monograph Series
on Languages and Linguistics
Georgetown University
Washington, D. C.
# Table of Contents

Foreword ................................................................. vii

INTRODUCTION
Welcoming Remarks, Reverend Frank L. Fadner, S.J., Regent, School of Foreign Service ............... 1
Introductory Comments, L. E. Dostert, Chairman of the 8th Round Table Meeting.
Brief History of Machine Translation ......... 3

I. PANEL I: SYSTEMS OF LOGIC IN MACHINE TRANSLATION
Meaning in Relation to Machine Translation .......... 13
Linguistic Analysis and Translation Analysis .......... 19
Language as Symbolic Logic ................................. 39
Discussion ......................................................... 45

II. CHARACTER SENSING AS AN INPUT TO MACHINE TRANSLATION ........................................... 53

III. PANEL II: LEXICAL PROBLEMS IN MACHINE TRANSLATION
The Rationale of the Micro-Glossary Techniques .... 63
Word Decomposition for Machine Translation ........ 71
The Requirements of Lexical Storage ...................... 79
German Prenominal Modifiers as Clues in Machine Translation .................................................. 89
Linguistics and Information Retrieval ................. 103
Discussion ......................................................... 112

IV. PANEL III. SCOPE OF SYNTACTIC ANALYSIS IN MACHINE TRANSLATION
Structure of Noun Phrases in German ..................... 125
Types of Russian Sentences ................................. 135
Some Ideas on Inter-structural Syntax ................... 143
The Use of SEAC in Syntactic Analysis .................. 151
Programming Aspects of Machine Translation ....... 163
Discussion .............................................. 167

V. PRACTICAL OBJECTIVES IN MACHINE
TRANSLATION RESEARCH .............................. 181

Appendix 1. Program of the Eighth Annual Round
Table Meeting ........................................... 187
Appendix 2. Participants .................................
Appendix 3. Index of Speakers ..........................
Foreword

The increasing interest in machine translation of languages and the growing scope of research in this field in several countries justified this year a departure from the established pattern of the Round Table Meetings on Linguistics and Languages Studies. Former meetings have presented individual studies under three related topics. This year all panels and papers are focused on diverse aspects of machine translation.

This was done in part to help disseminate more information in this growing area of linguistic and related research, in part to seek to bring up to date the published information now available.

Our thanks are due to those who have made these meetings possible — to The Very Reverend Edward B. Bunn, S. J., President of the University, and to the University Administration; to the many participants who have willingly given of their busy time to prepare and present the formal studies which constitute the core of the present Report, and to the National Science Foundation for its support of research in MT.

Our reverent recollection goes to the man whose vision created the Institute itself, the late Father Edmund A. Walsh, S. J.

Léon Dostert, Editor
ACKNOWLEDGMENTS

I should like to express my thanks to Miss Muriel Habel, Mrs. Alice B. Killen and Miss Chris Anne Montgomery for their assistance in the preparation of the manuscript of this monograph.

Léon Dostert
Editor
The Reverend Frank Fadner, S. J., Regent of the School of Foreign Service, made the following welcoming remarks:

Language is man's most primitive means of communication. And I don't suppose I exaggerate when I say that today, more than ever before in the world's history, there is greater need for effective communication among the peoples of this planet. If we are to have peace in our time, the rapid exchange of scientific data, the complete sharing of ideas by all the members of the polyglot human family is absolutely essential.

And yet, as you know — almost by way of vindicating the notion that human history is a record of action and reaction — in proportion as the need for international communication in the scientific world has grown, an obstacle that hinders and hamstrings such necessary exchange has presented itself to block the realization of one world of ideas. This is the blighting spirit of overweening nationalism — a form of provincial pride that glorifies the traditions and heritage of a given local bailiwick to the disastrous disadvantage of effective human communication. Thus, for example, quantities of scientific articles produced in the Russian world — works, which in another, more internationally minded age would have seen the light of day in the French or German languages — now hit a waiting world in the Russian language, and in most instances must undergo the laborious process of translation into the more commonly known world tongues. The day of Latin as an international Lingua Franca of the scientific world has long since passed.

In the latter days of the past century — as this deterioration and general breakdown became more and more noteworthy, certain linguistic pioneers sought to bridge the gap. Father Johann Schleyer invented Volapük in the 1870's; Louis Lazarus Zamenhof created Esperanto in the 1890's. Other attempts were subsequently made. These somewhat unrealistic efforts to educe an artificial means of international communication have met with but indifferent success.
We are now living in a distinctly mechanical age, and it is natural that we should seek to solve the problem by trying the means at hand. It is, I think, a source of justifiable pride that in the last half-dozen years these halls at Georgetown University's Institute of Languages and Linguistics have been the scene of much serious and painstaking effort — much conscientious experimentation — in the direction of making the dream of mechanical translation come true.

And that is why this particular Eighth Round Table meeting, which we are sponsoring — and which is devoted entirely to this timely and intriguing subject, is a fitting postlude to those efforts. That is why, too, as Regent of the School of Foreign Service I take extreme personal pleasure in welcoming you to these discussions. I assure you that it is the sincere and earnest intention of your hosts — the Administration and Faculty of the Institute of Languages and Linguistics of Georgetown University's School of Foreign Service — to wish you Godspeed and profit in the discussion of the subjects before you. The upshot cannot fail to be at least one more step on the way to a striking realization of the brotherhood of man.
Brief History Of Machine Translation Research
Leon Dostert
Georgetown University

Only five years ago the idea of using electronic computers to effect the translation of language seemed to many to be highly premature, if not actually fanciful. In the few short years that have elapsed since the first formal meeting on the subject in June of 1952 at the Massachusetts Institute of Technology, research in the field of machine translation has become widespread and has achieved academic and scientific respectability.

The story of the genesis of machine translation — the transference of meaning from one patterned set of signs occurring in a given culture into another set of patterned signs occurring in another related culture by means of an electronic computer — has been traced with care in the first compendium of essays on the subject entitled Machine Translation of Languages, edited by William Locke and A. Donald Booth.

From the outset it is understood that the type of machine to be used for translation is properly a logical machine. It must read the input text in the source language, manipulate the input translationally, and furnish a usable output in the target language. Since most digital computers operate with binary digits, the input operation has to include a transposition of the properly formulated source data into binary code, and the output operation includes the transposition of the binary result into more common symbols — decimal numerals or letters. The output can easily be equipped with a printer, resulting in legible printed text. The input of modern computers consists of previously prepared punched cards, punched tape, magnetic tape, or the like. For a translation program, this means that a human operator has to read the source text and, say, punch it on cards or tape, before it can be fed into the machine. In order to eliminate this preparatory human operation, the input of the machine would have to be equipped with an electronic scanning device capable of reading printed text and transposing it directly into binary code. Such a device has been the object of continued research for several years, and partial results have been attained.

Translation has to accomplish more than merely the one-by-one transfer of units from the source language into the target language. It must include some solution to the problems of choice implicit in
the fact that: A) a unit in the source language may have more than one possible equivalent in the target language, and B) that the order of source language units in the input may not be suitable for the output in the target language. The machine manipulation of the text fundamentally involves two types of computer operations: table look-up and algorithmic (that is, properly computative). The table look-up operation consists in matching the machine-read input against a set of data stored in the memory of the machine, and in delivering these stored-data to the arithmetic unit of the machine for algorithmic processing. The result of this processing is the translated output, which is then fed into the printer and delivered to the user.

The Georgetown-IBM experiment of January 1954, which succeeded for the first time in effecting machine translation from Russian into English on a limited basis gave, notwithstanding its critics, considerable impetus to research in the field. Admittedly, a number of the operations in the first trial were formulated on an ad hoc basis and the area of search and choice, as well as the extent of manipulation were strictly limited. However, as a result of this experiment, certain essential concepts were formulated which have largely remained valid. The fact that the experiment had attracted wide interest to the general problem was evidenced when in 1956 the Institute of Precision Mechanics and Computer Technology of the U.S.S.R. Academy of Sciences announced the successful performance of translation of English into Russian on their BESM computer, and acknowledged the relationship between their undertaking and the Georgetown-IBM experiment. This announcement was not unrelated to a renewal of interest and support for work in MT in the United States. In June of 1956 Georgetown University received a substantial grant from the National Science Foundation to undertake intensive research for the translation of Russian scientific materials into English. This grant has been renewed for a second year of continued research.

I shall give you a brief description of the work presently being conducted at the most important centers of organized machine translation research in the United States and abroad.

I have already mentioned the Academy of Sciences of the U.S.S.R., where a group of mathematicians and linguists have been working on various aspects of the problem, including the translation of limited segments of English scientific texts into Russian, and further, the investigation of machine translation of Chinese, French, and German into Russian. Two additional groups have been working on diversified
techniques both in Moscow and Leningrad. The information available is not complete, but it appears that the fundamental technique of these groups involves what they call the analysis of the source language and a synthesis (necessitated obviously by the inflectional character of Russian structure) process in respect to the target language. Their approach seems to utilize English inflectional suffixes and word order as cues to bring about the appropriate inflected forms in Russian. A second point which seems to emerge from published material is that their plan is to make Russian a sort of "core" language, so that machine translation from, for example, Chinese into French, would ultimately be effected through the medium of Russian. The Soviet group has apparently been successful in testing its translation program on computing equipment.

The approach of the Cambridge Language Research Unit in England differs from that of other groups by its emphasis on mathematical logic. Its plan of research involves the transfer of grammatical patterns from source to target language by means of Boolean algebra. This ensures the identification of translation units and their proper manipulation as wholes for translation purposes. Another feature of their approach is the use of a thesaurus routine to achieve semantic transfer from one language to another. That is to say, the semantic ranges of adjacent translation units are matched against each other by using coincident definitions in a thesaurus.

In this country, the group at the Massachusetts Institute of Technology, so far as is known, is concentrating on an exhaustive linguistic analysis of German on the one hand as source language, and English on the other as target language. The assumption of the MIT group appears to be that an analysis as complete as possible of the structure of the languages involved must first be made before the specific problems of meaning transference are approached.

On the West Coast, the International Telemeter Corporation of Los Angeles is now engaged in the planning of a translation machine of limited scope, with primary emphasis on storage capacity, and without envisioning too much complex manipulation between input and output. In conjunction with the development of this project, a group of linguists at the University of Washington is preparing a translation program which appears to be specifically geared to the characteristics or limitations of the machine under construction. The objective of their research is to investigate whether or not such a
deliberately circumscribed operation may not be adequate for certain practical purposes.

At U.C.L.A., Dr. Kenneth Harper has completed a preliminary project which goes considerably beyond Russian-English dictionary searching. Research is also presently being conducted at the following universities: Illinois, Michigan and Texas.

In October 1956 a meeting was held at the Massachusetts Institute of Technology, where American centers engaged in this work were represented, including groups from the University of Washington, UCLA, the International Telemeter Corporation, Harvard University, and Georgetown University, as well as MIT. Also present at this meeting were representatives of the Cambridge Language Research Unit in England. The U.S.S.R., though invited, was unable to send representatives to the meeting.

Next summer a seminar on machine translation will be held as part of the Linguistic Institute at the University of Michigan. It is supported by the National Science Foundation under the auspices of the American Council of Learned Societies.

Next August two reports on machine translation by members of the Georgetown project will be presented at the Eighth International Congress of Linguists at Oslo.

At Georgetown, after the 1954 experiment our research continued on a very limited scale until we received the grant from the National Science Foundation which enabled us in the fall of 1956 to engage in a full-scale project employing more than twenty senior and junior researchers. It was decided to focus on the translation of Russian texts in the field of organic chemistry.

The project is under the direction of a group of seven members of the regular faculty who represent varying types of linguistic specialization. Three are general linguists, two are Slavicists, one is an Arabist, and one is specialized in Germanic linguistics. Within the group competence in Romance and non-Indo-European linguistics is also found. While the focus is on the problem of translating Russian into English, the diversity of background of the group of linguists makes it possible for us to go ahead with the preliminary formulation of approaches to the problem of translation from and into other
languages on the basis of the techniques developed for the Russian-
English transfer.

The linguists work in close cooperation with nine research associ-
ates or assistants, all of them trained in linguistics, several of
them competent in Slavics, others in other linguistic fields, and a
few of them bilingual in Russian and English. Programming consult-
ants have been working with the group so that the emerging lin-
guistic solutions may remain within the limitations of programming
requirements for machine operations. Consultations with members of
the University's department of mathematics have been held.

Assisting the linguists and the associates is a group of three
bilingual translation analysts. Their task is to develop, with the
guidance of the linguists, a lexicon from an existing English trans-
lation of portions of the Soviet Journal of General Chemistry. The
particular text that has been chosen is concerned with experiments
in organic chemistry which have application to physical chemistry.
The English-language version was found to be inadequate for ma-
chine translation purposes and two persons were assigned the task
of preparing a standardized translation which would be free of styl-
istic idiosyncrasies and as consistent as possible. This is not to
be confused with pre-editing, or simplifying the text to any kind of
basic language. In this work, the analysis group has made use of
the services of a graduate fellow in organic chemistry to assure the
correct translation of technical terminology. It was found that the
problem involved in producing a translation especially suited to the
purposes of machine translation is largely one of consistency.

The staff is divided into working groups and assigned individual
topics. From the outset our policy has been not only to permit, but
to encourage diversity of approach.

Several members of the project are following and amplifying into
increasingly broader formulae the partly empirical and partly analyt-
cal technique of the 1954 Georgetown-IBM experiment. As a first
step, this group analyzed and reclassified the data of the first exper-
iment. As a result of this work, it was decided to focus on one set
of problems, namely, those involved in the translation of the Russian
noun phrase. These problems were approached by the Experimental
group in terms of the translation of the nominal and adjectival case
suffixes. At the present time, the Experimental group is engaged in
working out the code needed to present a considerably expanded and
more advanced sample suitable for a second experiment. The code now in preparation is intended to cover precisely and in a generalized way the translation of the noun phrase, within exactly stated tractability limits. They intend to apply this code to a normal technical text in Russian, and translate the entire text by giving, in addition to the precise solutions described above, intuitive, or ad hoc, solutions for the problems not so far included in the research, with the two types of solutions clearly differentiated in the sample.

Individual researchers have been given specific analytical problems, such as the formulation of a program for the insertion of the article in the English translation from Russian. Rules have been formulated for the prediction of the occurrence of *the* or "zero" article in English and have been demonstrated to have about 80% accuracy on the basis of the corpus tested. Russian prepositions have been analyzed in order to find the determiners of their variant translations into English and to study their correlation with the English prepositional pattern. An analysis of Russian noun and verb endings has been completed with a program consisting of some 500 statements of the type required for programming formulation. Specifically, a list was made of noun-adjective and verb suffixes, arranged in such a way as to allow for easy and economical storage in a glossary. After the preparation of the list, work was started with the consultant programmer, on a method by which the results obtained in the preliminary stage might be utilized in machine translation. The grammatical identification of Russian items can be achieved by matching stem with suffix, thus making it possible for the machine itself to recognize the grammatical function of a given item.

This identification program is directly connected with research on the correlation of syntactic operations in Russian and English. The underlying assumption is that syntactic analysis of successively included constructions within the sentence is essential for any method of translation that is to be reducible to mechanical procedures. The syntax function approach amounts to instructing the machine to proceed on the basis of syntactic analysis. This is achieved by approaching the translation analysis in terms of syntactic hierarchies and their sensing equivalents for a logical machine. The sentence must be handled in terms of its constituents in successive inclusion, so that the composition and order of smaller constituents can be adequately translated.
One member of the project has undertaken the study of a French chemical text and developed a strictly empirical technique based on the concept that the exhaustive translational analysis of a series of sentences within a continuous text would yield a number of formulations adequate to constitute the basis for a general program.

Finally, a research associate, with the assistance of one of the linguists, has undertaken the investigation of a technique in which a series of code affixes based on traditional grammatical categorization is added to Russian words so that ambiguity is resolved within a segment of the source language through a matching technique. The appropriate target item is then elicited by a second matching of the code affixes corresponding to identical affixes occurring as part of the target lexical items.

In order to ensure coordination among the several groups and individuals pursuing specific assignments, the group of linguists meets as required and special coordination meetings are held as circumstances warrant.

In addition, the entire staff participates in a weekly two-hour general seminar which is open to the public. As a result of the examination of specific problems through discussion, review, and comments by the other members of the project, certain conclusions are formulated which serve as orientation for further research. Individual members present papers summing up their own work or progress, or that of a group with which they are working. These papers are reproduced and distributed to those interested or working in the field.

One of the difficulties which has handicapped progress in the field of machine translation has been the lack of communication on the one hand, and the assumption of somewhat rigid positions on the other. To help to remedy the first difficulty, we hope that the publication and distribution of our seminar work papers will be one step toward greater communication among the different research groups in this country and in England. The publication of the proceedings of this meeting next fall will be another step in seeking broader communication of results and techniques. In respect to the second problem, it is somewhat more difficult to suggest remedies. It seems to me that since we are still barely past the threshold of our investigation, it would be both premature and unscientific to cling narrowly to a given hypothesis or theory as to the most efficient manner in
which the problem can be resolved. It can already be seen that the various techniques and approaches which have characterized the work of different individuals and groups are beginning to reach a point where results can be integrated. It is planned that as each technical approach reaches a certain level of formulation we shall conduct tests on existing computers which will give us a valid basis for determining which of the several approaches is the most efficient and on the basis of the results the possibility of a one-line approach will be considered.

On behalf of the faculty of The Institute and of my associates in MT research, I am happy to welcome you, and I hope that our exchange of ideas will prove of value to our common objective.
Panel I

Systems Of Logic

In Machine Translation
I

SYSTEMS OF LOGIC IN MACHINE TRANSLATION

Meaning In Relation To MT

MARTIN JOOS
University of Wisconsin

It is not my purpose here to compete with or to anticipate the contributions which other panel members will make with regard to such things as vocabulary size and storage, or even the problem of multiple meanings in present-day programming-research, for the bearing of my paper upon such things will be thoroughly abstract. It must be perfectly clear why this paper was placed in the lead-off position on the present conference: the management wanted to start with a clean slate. There just wasn't anybody else with less experience in MT, for mine is identically zero. I have had M experience — I have even built primitive computers — and I have done all kinds of T except only the M kind of T. Can I claim that I was doing MT when, in a certain national emergency, I spent dozens of hours translating Finnish into English with dictionary and grammar alone — that is, without knowing any Finnish? On second thought, I suppose that would be at least MT plus editing, for I never considered a sentence translated until I had made good sense out of it.

MT plus post-editing? No, not quite that either, for the editing didn't come after the MT. The two things were mixed, either done alternately (with several alternations within each sentence) or else simultaneously. I speak of 'simultaneous' MT and editing when you make the whole sentence (or as much of it as you can grasp) serve as your guide in choosing glosses or grammar formulas, so that you may even find yourself unable to use exactly what you find in the dictionary and grammar and are forced to 'guess in' something that fits. Thus a simultaneous MT and editing is not a logical addition, MT + editing, but instead a logical multiplication, MT x editing, with each of them limiting the other: logically quite a different thing.

Now in my translating from Finnish into English, the control of the process — I am using the word 'control' as a technical term: that which gives warning of blunders — the control of the process was identically semantics. The danger-signal was that the translation didn't make good enough sense. How good is good enough? How can
you, by the semantic control, ensure a good-enough translation? And are there useful implications for pure MT? These are not easy questions. But then, I understand that you don't expect perfect translation either, so that you may well be content with my approximate answers. I shall begin with exact semantic theory, and make the discussion finite by making it approximate.

The pattern of a language, and likewise the linguistic structure of any text, is that network of absolute restrictions (forbidding many or most random occurrences) which is called 'grammar' in the broadest sense of the term. This grammar may be described as the set of impossibilities of combination of linguistic units. On the standard probability scale from zero ('impossibility') to unity ('certainty' or 'strict implication'), we find that grammar uses only the two ends of the scale. For instance, in English it is 'impossible' to have this immediately followed by men; and the actually occurring open sequences of this and men 'necessarily, by strict implication' have something between, e.g. a comma or the sequence group of. Since impossibility is single, while this implication is a variable (a discrete variable, of which I have cited two values out of an indefinitely long list), it is cheaper to define grammar as a set of impossibilities than to describe it as a set of implications.

Some of the implied items are commoner than others; for instance, this group of men is commoner than this flock of men. I mean that one of these utterances is more often spoken or printed than the other; and I am not referring to another fact, of a different order, namely that groups of men are commoner than flocks of men — a separate question which might well occupy us elsewhere. We consider now the disparate or 'variable' commonness of the grammatically possible linguistic items, such as group of, flock of, etc., in the context this...men, and we give this whole phenomenon, proper to this and to all other contexts, a technical name. I call it the 'inside semantics' of the language; and the 'inside meaning' of each item is by definition the statistics of its occurrence in context with other items. Whether there is also something else, independent of this, which could be called the 'outside semantics' of the language, is a separate matter to be discussed later; but each 'content' item of a text has, by definition of the word 'content', a proper 'outside meaning' which is simply its referent, the real-world thing-and-event complex to which it refers in this occurrence.
Now the 'inside meaning' of each linguistic item in the text or in the language, and that system thereof which I call the 'inside semantics' of the language, are indisputable facts. I mean simply that people do not utter all grammatically licit utterances with comparable frequencies (oftennesses); rather, they seldom make such perfectly 'true' remarks as "I have never heard a green horse smoke a dozen oranges." The mathematics of this state of affairs, together with what was previously said about grammar, can be covered by a single statement, thus: 'The probability of given linguistic items occurring in given linguistic contexts is measured on the standard probability-scale running from zero to unity; the two ends of the scale define the grammar of the language, and the rest of the scale, the open interval between zero and unity, defines the inside semantics of the language.'

The objection has been raised that only a hopelessly long research program could work out the inside semantic system of a language. But that is merely a matter of degree. No statistician claims absolute precision anywhere. From a small body of data, a rough set of statistics will emerge; from a larger body of data, more precise statistics, and so on indefinitely. Impossibility of attaining absolute precision is not a legitimate argument against the existence of the phenomena being investigated. On the other hand, if the statistics do get more precise as the body of data expands, this is customarily taken as a valid argument in favor of the 'existence' of what is being statistically treated. Now the G. & C. Merriam Company finds it economically advantageous to expand its citation-file of word-occurrences and contexts indefinitely, while putting far less money into factual reference-books. I assume, therefore, that the existence of 'inside semantics' and of 'inside meanings' may safely be taken as established.

Let me return briefly, now, to the semantic control of my behavior in translating Finnish into English. When I had applied dictionary and grammar mechanically and thereby manufactured some English nonsense-text, what was the nature of the semantic control that told me it was nonsense? Was it (1) that the English words were strange bedfellows in the sentence; or was it (2) that the English sentence did not match the real world as I knew it? A case could be made out for either explanation, or for a combination of them in any proportion. I am going to try to make out a case for the first theory, the strange-bedfellows explanation.

What causes the hesitation between this and the real-world or outside-meaning explanation? I believe it is the fact that we keep passing knowledge back and forth between two containers, namely (1) language
and (2) sensation-and-manipulation. Very little knowledge, and that only of special kinds, is normally kept, even a little while, entirely inside one of those two containers. The extreme case of knowledge kept entirely inside language is: pure mathematics. The extreme case of knowledge kept entirely inside the other container, namely sensation-and-manipulation, is: the non-language arts, notably painting and sculpture. It is therefore easy to see why neither of those fields is dependent upon national languages for international currency. I have friends who read Russian mathematical publications without knowing a hundred words of Russian; and it is certain that a competent art teacher needs no language in his guidance of a pupil, just as we need none for enjoying a painting. But when mathematicians and artists communicate with others who are not of their own guild, in either case ordinary language always steps in. When pure mathematics is brought to bear upon engineering or atomic physics, ordinary language is always (I think necessarily) used as the mediator; and we all know that painting does get discussed interminably in ordinary language. Thus, as soon as society at large tries to profit from either extreme case, it gets assimilated to the intermediate, majority-party or normal cases. Therefore we can neglect both extremes here, and concentrate on the great majority of human fields of knowledge and action, where the general rule holds: all the socially significant knowledge continually gets passed back and forth between language on the one hand and sensation-and-manipulation on the other hand.

It is in the second container for knowledge, namely sensation-and-manipulation, that we would expect to find the previously mentioned 'outside semantics'. Although I would not deny its existence in the extreme case of painting or sculpture, or pure music or pure dance, I can nevertheless deny that this is 'outside semantics of the language' and proceed inductively to the conclusion that any 'outside semantics' as a system, independent of the system called 'inside semantics', does not belong to the language at all and therefore does not concern us as linguists or MT workers. Insofar as a semantic system exists which ties all 'outside meanings' together into a system, that system is inevitably isomorphic with the system already denominated 'inside semantics'. The isomorphism is maintained, and fostered as the culture and the inside semantics evolve, by the oscillation of knowledge between language and that non-language realm which I called sensation-and-manipulation. Every new thing, new sensation, or new manipulation, promptly gets named, and the discussion thereof gets standardized, with a standardization faithfully manifested in those contextual occurrence-statistics which I called the 'inside semantics'.
of the culture's language. It is only thus that the innovation can strike root and survive in the culture. Incidentally, this is surely the reason why new schools of painting nowadays change and fade and vanish and get supplanted in such a dizzy dance: deliberately devised to defy discussion in ordinary speech, they die for lack of the cultural survival-value of such discussion. Conversely, any new linguistic coinage which does not get firmly attached to either an old or a new sensation-and-manipulation item is forgotten in a few weeks or years: we call it slang. The exception here is pure mathematics: there the new formulations, devised from the beginning in such terms as make them independent of sensation-and-manipulation, can survive indefinitely and compensate us for the evanescence of slang and of artistic innovations.

With the extremes cleared out of the way, I can now concentrate on the overwhelming majority of human concerns, namely those in which the knowledge does get continually passed back and forth between language and sensation-and-manipulation. And I shall henceforth take it as certain that the 'outside semantics', which I allowed to be possibly autonomous in e.g. the graphic arts, is in these principal human concerns not autonomous at all, but necessarily isomorphic with the 'inside semantics' of ordinary language.

We would seem to have three technical terms left here: 'outside meaning' for one, 'inside meaning' for another, both proper to single linguistic items; and a single 'semantics'. The latter, originally defined as 'inside semantics' and as the system of 'inside meanings', appears now also to be the system of 'outside meanings'. Therefore, each 'outside meaning' is homologous (similarly placed in the system) to the 'inside meaning' of the same linguistic item; and thus we see that the distinction between inside and outside is otiose; they are equivalent and can be treated 'as if' identical, which is the same thing as treating them 'as' identical—even though originally defined quite separately. Thus the fact that windows are made of glass and are breakable and transparent if sufficiently clean—this outside fact does not need to be treated any differently, in linguistic or MT discussion, from the statistical fact concerning English utterances, that the word window occurs frequently in context with such words as glass, broken, and wash. Therefore, all sensation-and-manipulation facts can be built into an MT treatment as soon as they are sufficiently known, and it doesn't matter whether a detail of programming is based upon study of texts or study of the real world—as long as we confine ourselves to the open interval of the probability-scale, between zero and unity probability, for the two ends will govern a separate area of the
programming, namely the grammatical area. Originally I was tempted to claim that semantic programming should be based only upon statistical text-analysis, but now I see that it doesn’t matter.

Now semantic programming would seem to be the proper theme of my paper. I have no interest in describing an MT method which is essentially defective in that respect. But since my paper is abstract only, I won’t describe the non-defective machine either. I have already done that in another place: Lang. 32.296-7 (1956). Briefly, it is an imaginary machine which uses the presence of each content-word in a sentence as a guide in choosing among the possible renderings of each other word in the sentence, just as a human translator does.

But it may be worth while to add a few words here on 'denotation' and 'connotation'. In a decently written treatise, each content-word of a sentence has a definite reference as its outside meaning. Each occurrence of such a word, that is, means one certain 'thing' (in the broadest sense of 'thing'). But that same word in other occurrences (within or outside of this treatise) might just as well (though either more or less often) have one or more other outside meanings. Now the outside meaning which the word has in the sentence in question is its 'denotation' there. This denotation, being particular, is exempt from perturbation by the rest of the sentence. On the other hand, all the denotations which that same word could have in still other sentences are its 'connotations' in this sentence, to a first approximation. To a first approximation only, however, for these connotations (unlike its denotation here) are very much subject to perturbations from the denotations and the connotations of the other words in this sentence. It is one of the characteristics of skillful writing that this whole perturbation-field is carefully adjusted in two ways: so as to attenuate misleading connotations, and also so as to reinforce helpful connotations. It is not entirely outside the scope of the imagination to make a mental construction of a bilingual machine that would do some of this in MT work. But it is a problem of a different order of complexity than those previously envisaged. You see, the denotations are differently distributed over the vocabularies of the two languages. Therefore, the connotations cannot be simply carried over from S language to T language. Instead, within each sentence at least, the carefully adjusted complete perturbation-field in the S sentence has to be replaced by an adjusted perturbation-field in the T sentence, and this T field has to be composed mostly of connotations differently distributed among the denotive words. It is at present of course a desperate problem; but if the engineers of another millennium solve it, they can do MT of prose literature.
Linguistic Analysis And Translation Analysis
Paul L. Garvin
Institute of Languages and Linguistics

I have stated my theoretical position on MT in a recent paper in *For Roman Jakobson.*¹ I think that at the present stage of MT no useful purpose would be served by a further elaboration of my general ideas. What I would like to do instead is to discuss in detail some workaday problems of methodology on the basis of experience gained, rather than in terms of ideas as yet untested.

Concretely, I shall present a single Russian example and attempt to show how, and to what extent, certain morphological features and syntactic relations have been utilized to work out a translation procedure which, within its admittedly narrow limits, I consider to be sufficiently rigorous to be pragmatically valid, and yet sufficiently flexible to allow expansion beyond the limits as presently defined.

In doing this, I shall draw upon the procedures and results developed by the Experimental Group of the Georgetown University Project in MT Research² with whom I have been working.

In dealing with morphological features, these procedures are based on the assumption, which by now I consider validated, that the translation operation must proceed from the written morphs in the source language directly to the translation in the target language. In this, the assignment of the morph to its appropriate grammatical category is an essential preparatory step, but not one which has to be incorporated into the code by which the translation is effected, unless it is relevant translationally as well as grammatically.

The syntactic relations which have been considered in our procedures so far are those of agreement, government, and dependence. I shall attempt to show the extent to which these relations have been found to be relevant to, and have been utilized for, translation into English, within the limits of the present state of our research.

To define the relations of agreement and government for the present purpose, I shall utilize the traditional concept of grammatical

²Consisting of Dorita Lochak, Madeleine Mathiot, and Jane A. Pyne.
category (for Russian, the categories of case, number, gender, and person).

It can then be said that two or more units are in a relation of agreement if they share one or more grammatical categories (that is, agree as to number, gender, case and/or person).

Two units are in a relation of government if, irrespective of the grammatical category of unit A, unit B has to be assignable to a certain grammatical category. Unit A is then said to govern the grammatical category of unit B (thus, a preposition would govern the case of the following noun.)

The relation of dependence is a relation of presupposition: A is dependent on B, if A presupposes B for its occurrence, or for some other functional property.

The translation design under development by the Experimental Group first envisions a sensing and matching routine in which paradigmatic suffixes are separated from the preceding bases whenever translationally convenient; then undivided words, as well as base and suffix partials of subdivided words, are matched against corresponding entries in a stored glossary. The glossary is intended to contain the Russian items to be matched, together with their single or multiple translations, accompanied by a set of code symbols serving to effect the choice from among several alternatives in case of multiple translation, as well as other required translation steps.

The diacritic code thus carries the major translation load, since it is intended to handle those cases in which the translation is not one-to-one, that is, where it is not procedurally trivial. The function of the code is dual: it permits the recognition of decision points—items in the input which, because of possible multiple translation, require translation decisions; and it initiates a set of instructions designed to find the cue for the appropriate decision and to implement it.

---

Let me now discuss my example.

пи́рокате́хиновы́е эфиры́ триа́рилимети́лфосфи́новой кисло́ты

In the type of structure exemplified here, a relation of agreement can be said to exist within each of the two modifier-head constructions: пи́рокате́хиновы́е agrees with эфиры́, триа́рилимети́лфосфи́новой agrees with кисло́ты. A relation of dependence can in addition be said to exist within each of these constructions: the adjective is agreement-dependent on the noun, in the well-known sense that the grammatical categories of the modifying adjective vary with those of the modified nouns, and not vice versa. In addition, a relation of government can be said to exist between the two constructions, in that the first noun construction governs the genetíve of the second.

All of these relations are of equal significance from the standpoint of generating Russian utterances, such as is the case in speaking or writing, or translating into, Russian.

For the translation from Russian into English, the general design will have to be implemented as follows:

The sensing-and-matching routine will separate suffixes from bases and match the following items: пи́рокате́хинов-, -ые, эфи́р-, -ы, триа́рилимети́лфосфи́нов-, -ой, кисло́ты-, -ы.

Of the above, the following will be translationally unambiguous;

пи́рокате́хинов- - pyrocatechol

-ые - zero (Russian nominative/accusative suffixes require no English translation; since English modifiers do not agree with their heads, the grammatical plurality of the Russian suffix requires no English translation either.)

4 Taken from the chemical corpus of the Georgetown Project, Журнал Общей Химии, vol. XXII (1952).
The following right partials will require a choice from among several translational possibilities:

—ы, a grammatically ambiguous morph: assignable to a noun suffix for either genitive singular, or nominative or accusative plural;

—ои, a grammatically ambiguous morph: assignable either to an adjective suffix for nominative, genitive, dative, accusative, or instrumental singular, or to a noun suffix for instrumental singular.

Of the two remaining left partials, эфи́р— and кисло́т—, the former clearly requires a lexical choice from among two possibilities, ether or ester; the latter presents no lexical ambiguity: it translates as acid. Grammatically, they are both unambiguous: the first is a masculine noun base, the second a feminine noun base.

A grammatical ambiguity is introduced for these two bases, however, by the requirement of rendering the number meaning of the nouns as wholes. It has been found procedurally simpler to assign the number meaning to the translation of the base, and to retain, for the translation of the suffix, only the remaining elements of meaning, in order to reduce the coding and output load of the suffix entries in the glossary.

Thus, the glossary entry for эфи́р— will, in addition to the lexical ambiguity, contain a procedurally induced grammatical number ambiguity, requiring two sets of choices: ether versus ethers, ester versus esters, and once this choice of number is resolved, the remaining choice between the lexical meanings has to be effected (or vice versa). Likewise, the entry for кисло́т— will have to contain a choice of acid versus acids.

On the basis of the above, the right partials of my example will have to be entered into the glossary with following translations:

пи́рокатехино́в—  - pyro-catechol
—ые  - zero
— эфир—  — ether
          — ethers
          — ester
          — esters

— various translations based on the genetival, nominatival, or accusatival meaning of this morph (the corresponding number meaning will affect the translation of the preceding base).

триарилметилфосфинов—  — triarylmethylphosphinic

— various translations based on the nominatival, genetival, datival, accusatival or instrumental meaning of this morph.

кислот—  — acid
          — acids

Let me now discuss in series the routines developed to effect the appropriate choice from among several alternatives, where required, and the linguistic considerations that have entered into the routines.

эфир—  — ether
          — ethers
          — ester
          — esters

I shall first discuss the routine to cover the choice of number translation, since it has greater generality.

This routine is based on a classification of Russian noun bases in terms of the particular set of suffix morphs found to occur after each. The set of morphs particular to each type is the primary cue for the number translation of the base; if the morph unambiguously establishes the singularity or plurality of the translation, no further cueing is required.

—23—
Эфир—belongs to base type 209, which may be followed by the suffix morphs 0, -а, -у, -®, -e, -ом; -бы, -об, -ам, -бы, -ax, -ами. For type 209, any of these morphs constitutes an unambiguous and sufficient number-translation cue; if a type 209 base is followed by one of the morphs -0, -а, -у, -®, -e, -ом it will have to be given a singular translation; if it is followed by one of the morphs -бы, -об, -ам, -бы, -ax, -ами, it will have to be given a plural translation.

For this type, the following suffix morph will constitute an unambiguous cue to trigger off the selection of the appropriate number translation.

This translation routine is coded as follows:

The left partial Эфир— is a decision point (i.e., a point in the text at which a translation decision has to be made and an appropriate routine initiated to effect it). The glossary entries for all decision points will have to be provided with a suitable diacritic which will have to indicate: (a) that this is a decision point, and (b) which routine is to be initiated in order to effect the decision appropriate to this decision point in this or any context. The diacritic chosen for this particular decision point was P209, where P stands for "this is a decision point, initiate a routine here", and 209 stands for "the routine required here is that indicated by rule 209".

This is one of a set of singular/plural base-translation rules, numbered 201 up.

Since the cue to this particular decision is always contained in the suffix immediately following the base in question, the rule will be relatively simple to construct, as compared to rules based on more extensive and/or more distant cues.

For purpose of rule 209, therefore, it is enough if the glossary entries for all suffixes that may follow a base of type 209 are provided with a diacritic indicating whether this suffix is a cue for the singular translation or for the plural translation of the base. Specifically this diacritic will have to convey the following information: (a) "this is a cue"—indicated by C, (b) "this is a cue for the routine covered by rule 209"—indicated by 209, (c) "this is a cue only if located in the suffix attached to that same base (or, in MT parlance, in the other partial of the same subdivided item)"—indicated by s,
(d) "this is a cue to effect a translation by a singular or plural form"—indicated by $X$ or $Y$ respectively. Thus, in order to insure the consistent operation of rule 209, suffixes -a, -y, -e, -ом will have to carry a diacritic $C_{209}s_X$ (the zero suffixes are handled by a totally different routine not involving subdivision); suffixes -ы, -ов, -ам, -ах, -ами will have to carry a diacritic $C_{209}s_Y$. Furthermore, the translations of эфир—ether and ester will have to be marked $X$, the translations ethers and esters will have to be marked $Y$.

Rule 209 will then read as follows: when the glossary entry matched against the current input item contains diacritic $P_{209}$, check other partial of same subdivided item for the presence of either $C_{209}s_X$ or $C_{209}s_Y$. If $C_{209}s_X$ is present, select translation $X$ of current item (that is, singular); if $C_{209}s_Y$ is present, select translation $Y$ of current item (that is, plural).

At this stage of the procedure, the glossary entries for эфир— and -ы will look as follows:

<table>
<thead>
<tr>
<th>Эфир—</th>
<th>Ether $X$</th>
<th>P 209</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Etters $Y$</td>
<td>- Ester $X$</td>
<td></td>
</tr>
<tr>
<td>- Esters $Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ы</td>
<td>Various translations as stated above</td>
<td>$C_{209}s_Y$</td>
</tr>
</tbody>
</table>

If the routine by rule 209 is effected, the translation of the source item эфир—ы is plural: ethers, esters, rather than ether, ester.

Once the grammatical translation decision has been made, эфир— still remains an unresolved decision point for the lexical translation decision between ethers and esters. This decision will require a rule other than 209, since the cues are obviously not the same; a second $P$ diacritic is therefore required to indicate that a second routine is to be initiated to effect the appropriate lexical choice. The diacritic chosen for this particular decision point was $P_{1001}$, where 1001 stands for "the routine required here is that indicated by rule 1001". This is one of a set of lexical rules, numbered 1000 up.

-25-
In order to formulate rule 1001, all cards with эфир— were extracted from the lexical file of the Translation Analysis Group of the Georgetown MT Project, each of 93 cards showing one context in which эфир— has been found to occur in our corpus. The contexts were examined, to determine whether adjacent Russian words occurred with sufficient consistency to warrant serving as cues for translation by either ether/ethers or ester/esters, or neither. It was found that in 70 of the 93 cases the immediately adjacent word constituted a suitable cue, in 20 cases a cue was present within the sentence, but at a distance and in a relation to the current item such that they would require a very complex coding procedure which it was decided to defer to a later stage in the research, and in 3 cases no cue whatsoever could be found within the sentence.

The decision was made to incorporate the cues constituted by the adjacent words (73.5% of the cases) into rule 1001, and to leave the residual cases to be covered by an ambiguous translation. Translations ester and esters were designated by Β, translation ether and ethers were designated by Γ, ambiguous translations for those cases where cues would not be handled or were not present were designated by Δ, yielding the following glossary entry:

эфир—  
- ester Β  
- esters Β  
- ether Γ  
- ethers Γ  
- ester-ether Δ  
- esters-ethers Δ

Examination of adjacent cues revealed that the translation ester or esters applies whenever эфир— is preceded by certain adjectives and/or followed by certain nouns or adjective-noun constructions in the genitive, the translation ether or ethers applies when it is preceded by certain other adjectives and/or followed by certain other nouns or adjective-noun constructions in the genitive. Code diacritics thus had to be devised to be attached to the glossary entries for all potential preceding and following cues.
They would be decision cue diacritics, and thus carry the indication C, followed by the rule number 1001.

The location of the cue would be in either the preceding adjective or in the following noun or adjective (it was found that in adjective-noun constructions constituting cues for this lexical choice, the adjective—which is first in the sequence—was by itself sufficient to cue the decision).

Since, however, the sensing-and-matching routine subdivides all adjectives and all nouns in the genetive (with the reservation stated below) into left and right partials (base and suffix), the location of the cue had to be further specified.

For purpose of simplicity in stating this rule, I am here ignoring the feminine and neuter plural genitives, which have zero suffixes and therefore constitute complete items. The rule as actually formulated by our group does, however, take these forms into account.

In the case of preceding adjectives, the assumption could be made that (unless they were adjectives with nominal function, which would require a totally separate routine anyway) they would, in the overwhelming majority of cases, agree with the following эфир—in case and hence the right partial could be ignored as non-differential. Thus, the cue diacritic would have to be affixed to the glossary entry for the appropriate left partial. In the code of the Experimental Group, this location (left partial of immediately preceding subdivided item) is designated by L-1. This designation would be followed by the code letter indicating that the choice to be made on the basis of this cue is either ester and esters or ether and ethers, thus yielding diacritics C1001L-1B and C1001L-1Г.

Glossary entries for adjective bases like пирокатехинов— or алкилфосфорист— would thus be provided with diacritic C1001L-1В, and entries for adjectives like этилов— or изомерн— with C1001L-1Г, to yield translations ester and esters, ether and ethers, respectively.

For the following nouns and adjectives as potential cues, both partials would have to be provided with the appropriate diacritics in the glossary, since they function as cues only if they are in the genetive—that is, if the appropriate suffix is present. Thus, two cue locations have to be checked by this routine: left partial of
immediately following subdivided item, and also right partial of immediately following subdivided item. These locations are designated by \( l+1 \) and \( r+1 \) respectively. This designation would, for the left partial, be followed by the code letter indicating the choice to be made; for the right partial, the location designation would be followed by the code letters indicating both choices (since for either choice, the right partial would have to be constituted by one of the same set of genetive suffixes). The resulting diacritics are \( C1001l+1B \), \( C1001l+1\Gamma \), and \( C1001r+1\Gamma \).

Glossary entries for noun and adjective bases like кислот-, фосфорист- would then be provided with diacritic \( C1001l+1B \), and entries for noun and adjective bases like хлоргидрин- and иодоватист- with \( C1001l+1\Gamma \), to yield translations ester and esters, ether and ethers respectively, whenever the appropriate diacritic is present in the entry for the accompanying suffix. Thus, entries for all morphs assignable to the genetive would have to carry diacritic \( C1001r+1\Gamma \).

Rule 1001 will then read as follows: when the glossary entry matched against the current input item contains diacritic \( P1001 \), proceed to step 1.

Step 1. Check the entry for the left partial of the immediately preceding subdivided item for the presence of \( C1001l-1B \) and check the entries for both right and left partials of the immediately following subdivided item for the presence of both \( C1001l+1B \) in the left partial, and of \( C1001r+1\Gamma \) in the right partial; if either the preceding condition or both following conditions or all three conditions apply, select translation \( B \) of current item (that is, ester and esters); if not, proceed to step 2.

Step 2. Check the entry for the left partial of the immediately preceding subdivided item for the presence of \( C1001l-1\Gamma \), and check the entries for both left and right partials of the immediately following subdivided item for the presence of both \( C1001l+1\Gamma \) in the left partial, and of \( C1001r+1\Gamma \) in the right partial; if either the preceding condition or both following conditions or all three conditions apply, select translation \( \Gamma \) of current item (that is, ether and ethers); if not, proceed to step 3.

Step 3. Select translation \( D \) of current item (that is ester/ether and esters/ethers).
Step 3 of rule 1001 then covers those cases (26.5%) for which no appropriate cueing was worked out for the present, by putting out a multiple translation.

In our example, the entry for the left partial of the preceding subdivided item, пирокатехинон—, will carry С1001т-1Б, and the entries for the left partial of the following subdivided item, триарилметилфосфин—, and for the right partial —ои, will carry С1001т+1Б and С1001т+1БГ respectively, and thus, since the previous application of rule 209 has already narrowed down our choice to plurals only (esters, ethers, or esters/ethers), rule 1001 will now resolve our lexical ambiguity and will unequivocally translate 3(|)Wp— as esters.

—ы various translations based on the genetival, nominatival, or accusatival meaning of this morph.

The assignment of the morph —ы to the genetive, nominative, or accusative will affect its translation as follows: if assigned to nominative or accusative, its English translation has been found to be zero in all cases covered so far; if assigned to genetive, its translation will usually be of when preceded by a noun, it will be than when preceded by a one-word comparative form such as сильнее, and it will be zero in the remaining cases. A sampling of our chemical corpus and of some other Russian text has revealed that, at the present stage of our research, the above information is sufficient to determine the translation of this morph, and that an inspection of the base to which the morph is attached is unnecessary to effect the translation of the morph. Thus, as cues for the translation of this morph will function an immediately preceding noun, an immediately preceding one-word comparative form, or the absence of either.

The rule applicable to the morph —ы is 105, which is one of a set of suffix translation rules numbered 101 up, and the translation routine is coded as follows:

The entry for —ы, the decision point, gets a diacritic Р105.

Any noun in the language will, if it precedes a subdivided item the right partial of which is —ы, constitute a cue for its translation by of. Therefore, all nouns entered into the glossary will have to
carry an appropriate diacritic. This diacritic will be a cue diacritic for rule 105, hence $C105$ will be the initial code symbols.

Other information incorporated in the diacritic will be as follows:

Since some nouns will be entered in the glossary without requiring subdivision through the sensing-and-matching routine, that is, will be entered as complete items, and some nouns will be entered as subdivided items consisting of left-partial base and right-partial suffix, provisions for the location designation of the $C105$ diacritic will have to apply to both possibilities. Nouns entered as complete items will have to carry this diacritic, as well as the left-partial (base) entries for nouns entered as subdivided items; in either case, the location is to apply only if the complete or subdivided item immediately precedes the current subdivided item containing $-\text{bl}$. We mark this location by the code designation $c^l-1$, where $c$ stands for complete item, $l$ for left partial of other subdivided item and $-1$ for immediately preceding.

Finally, the diacritic attached to all nouns entered as complete items, and all noun bases entered as left partial, will carry the choice letter $A$, indicating that the translation to be chosen is of. The full diacritic will thus be $C105c^l-1A$.

Any one-word comparative in the language will, under the same circumstances, constitute a cue for the translation of $-\text{bl}$ by $\text{than}$. One-word comparatives will be entered in the glossary as complete items, hence the code diacritic will contain $c^l-1$ as its location designation, and it will contain $B$ as its choice designation, resulting in the diacritic sequence $C105c^l-1B$.

The glossary entry for $-\text{bl}$ will now be as follows:

$$-\text{bl} \quad \text{of} \quad A \quad P105 \quad C209sY$$

than $B$

- $C$

Rule 105 will then read as follows: when glossary item matched against the current input item contains diacritic $P105$, proceed to step 1.

Step 1. Check the entry for the immediately preceding complete item, or the entry for the left partial of the immediately preceding
subdivided item, for the presence of C105c\-1A; if it is present, select translation A of current right partial item (that is, of), and reverse the order of the translations of the right and left partials of the subdivided item of which the current item is part (this transposition is required, since while a Russian suffix follows its base, the preposition which may translate it will precede the English noun which translates the base); if C105c\-1A is not present, proceed to step 2.

**Step 2.** Check the entry for the immediately preceding complete item for the presence of C105c-1B; if it is present, select translation B (that is, than) and reverse the order of the translations of the right and left partials of the subdivided item of which the current item is part (the reasons for the transposition being similar to the above); if C105c-1B is not present, proceed to step 3.

**Step 3.** Select translation C of current item (that is, zero).

In our example the diacritics required for the first and second step of rule 105 are absent, therefore step 3 will be carried out which results in translation zero.

—\-0й various translations based on the nominatival, genetival, datival, accusatival, or instrumental meaning of this morph.

The assignment of this morph to the various case categories listed above depends on the class of the preceding base: if a noun base precedes, the morph will be assignable to the instrumental only; if an adjective base precedes, the morph will be assignable to all of the listed cases. The first step in the translation routine will thus be to determine the class of the preceding base, in order to proceed with a simpler routine if a noun base precedes, or with a more elaborate routine if an adjective base precedes. In our example, the latter is the case.

In this elaborate routine, it was found that the translation of the morph, if assigned to the nominative or accusative, is zero, as for the previously treated morph —\-у. In other instances, irrespective of its assignment to a particular case, its translation will usually be of when preceded by one group of nouns, thus both in \-дикой direction of such where —\-0й is genetive and управление \-дикой management of such where it is instrumental; it will be by when preceded by a second group of nouns and a group of verbs, with
when preceded by a third group of nouns and a second group of verbs, at when preceded by a fourth group of nouns and a third group of verbs; it will be than when preceded by a one-word comparative, as when preceded by a fourth group of verbs, to when preceded by a fifth group of nouns, a fifth group of verbs, or a noun in the accusative, and finally, when preceded by a sixth small group of verbs, its translation will be as when certain nouns follow and with when certain other nouns follow. Thus, as cues for the translation of this morph will function various immediately preceding verbs and nouns, an immediately preceding one-word comparative, an immediately preceding noun suffix morph with accusative function (we have not yet coded for accusatives with zero suffix in this context), and in a limited number of instances, an immediately following noun as well.

The rule applicable to the morph –əñ is 102, one of the suffix-translation sets numbered 101 up, and the translation routine is coded as follows:

The entry for –əñ, the decision point, gets a diacritic P102.

The entry for the preceding base will have to be the cue for the choice between the simple noun-suffix routine, called routine 1, and the more elaborate adjective-suffix routine, called routine 2. A C102 diacritic will thus have to be attached to all feminine hard noun bases, and another C102 diacritic will have to be attached to all hard adjective bases, the two sets of bases after which –əñ occurs. The location designation for both will be s, since we are here dealing with the other partial of the same subdivided item. The full diacritic attached to appropriate noun base entries will be C102s1, to indicate choice of routine 1, the full diacritic attached to adjective base entries will be C102s2, to indicate choice of routine 2.

To develop routine 2, which is the one of interest here, appropriate cue diacritics were selected and assigned to entries for the various groups of nouns (left partials and complete items) as well as verbs (complete items), one-word comparatives (complete items), and noun suffix morphs with accusative function (right partials), in a manner similar to the above, and to the procedure used in developing the translation for morph –bį.

I shall not present all detailed considerations leading to the formulation of rule 102, but confine myself to citing it.
The glossary entry for —0I will now be as follows:

—0I (routine 2 only) P 102

as A

with B

of C

than D

by E

to F

at G

— H

Rule 102 will read as follows:

When the glossary entry matched against the current input item contains diacritic P102, check other partial of same subdivided item for the presence of either Cl02s1 or Cl02s2. If Cl02s1 is present, go into routine 1; if Cl02s2 is present, go into routine 2.

Routine 2:

Step 1. Check the entry for the preceding complete item for presence of Cl02c-1AB; if this is present, then check the entry for the left partial of the immediately following subdivided item for the presence of either Cl02l+1A or Cl02l+1B; if Cl02l+1A is present, select translation A (as); if Cl02l+1B is present, select translation B (with); if these conditions do not apply, proceed to step 2.

Step 2. Check the entry for the immediately preceding complete item, or the entry for the left partial of the immediately preceding subdivided item, for the presence of Cl02c-1B or Cl02c-1C or Cl02c-1E or Cl02c-1G; if Cl02c-1B is present, select translation B (with); if Cl02c-1C is present, select translation C (of); if Cl02c-1E is present, select translation E (by); if Cl02c-1G is
present, select translation G \textit{(at)}; if none of the above conditions apply, proceed to step 3.

Step 3. Check the entry for the immediately preceding complete item for the presence of either C102c-1A or C102c-1D; if C102c-1A is present, select translation A \textit{(as)}; if C102c-1D is present, select translation D \textit{(than)}; if neither is present, proceed to step 4.

Step 4. Check the entry for the immediately preceding complete item, or the entry for either the left or the right partial of the immediately preceding subdivided item, for the presence of C102c\textit{F}-1F; if it is present, select translation F \textit{(to)}; if it is not present, proceed to step 5.

Steps 1—4. If a translation is selected from among A through G, reverse the order of the translations of the right and left partials of the subdivided item of which the current item is part.

Step 5. Select translation H \textit{(zero)}.

In our example, step 2 applies, yielding the translation \textit{of}.

\textbf{кислот-} \textit{acid}

\textit{acids}

This noun base entry, unlike the previously discussed \textit{эфир-}, requires only a number translation routine.

In line with our base classification procedure, \textbf{кислот-} is assigned to type 202, which may be followed by the suffix morphs \textit{−a, −ы, −е, −у, −е, −о́й; −ы, −ё, −ам, −ы, −ах, −ами}. If a type 202 base is followed by one of the morphs \textit{−a, −е, −у, −е, −о́й}, it will have to be given a singular translation; if it is followed by one of the morphs \textit{−ё, −а, −ах, −ами}, it will have to be given a plural translation. Any of these morphs, therefore, constitutes a sufficient number-translation cue. The remaining suffix morph \textit{−ы}, however, allows both a singular and a plural translation of a type 202 base, and additional cues must therefore be found whenever a type 202 base is followed by \textit{−ы}.

When a type 202 base is preceded by an adjective, the adjective suffix serves as a cue to resolve the number ambiguity: \textit{−о́й before
type 202 base followed by —бы will give singular translation for the base, —бы before type 202 base followed by —бы will give plural translation for the base.

When no adjective precedes, the cueing for the number-translation decision can be based on the relation between the number and the case meanings of morph —бы: when the morph has genetive case meaning, the number translation of the type 202 base will be singular; when the morph has nominative or accusative case meaning, the number translation of the type 202 base will be plural. Preceding nouns, prepositions, one-word comparatives, and various particles serving as cues for genetive identification are then utilized to translate the base as singular, absence of these serves to translate the base as plural.

The routine is then coded as follows:

The entry for кислот—, the decision point, gets a diacritic P202.

All nominal suffix morphs of the set found after type 202 bases will be given C202 diacritics, with location designation s, since they are other partials of the same subdivided items. Suffix morphs —а, —о, —у, —ый will be entered with diacritic C202sX, to effect singular translation; suffix morphs —ам, —ах, —ами with C202sY to effect plural translation, and morph —бы with C202sXY to indicate that the number ambiguity remains unresolved (the zero suffix is again handled by a totally different routine not involving subdivision).

For those cases where a type 202 base is followed by morph —бы, the additional cueing is provided as follows:

Suffix morph —ый, which will function as a cue for singular translation whenever it occurs as a right partial of the subdivided item immediately preceding the type 202 base, is assigned a diacritic C202r-1X; suffix morph —ые, which will function as a cue for plural translation whenever it occurs in the same location (unless it is attached to an adjective with nominal function, in which case a special routine is required), is assigned a diacritic C202r-1Y.

Finally, appropriate C202 diacritics are assigned to the nouns, prepositions, one-word comparatives and particles serving as cues when no utilizable adjective suffix is present.
Rule 202 will then read as follows:

When the glossary entry matched against the current input item contains diacritic P202, proceed to step 1.

Step 1. Check other partial of same subdivided item for the presence of either C202sX or C202sY or C202sXY; if C202sX is present, select translation X (singular); if C202sY is present, select translation Y (plural); if C202sXY is present, proceed to step 2.

Step 2. Check right partial of preceding subdivided item for the presence of either C202r-1X or C202r-1Y. If C202r-1X is present, select translation X (singular); if C202r-1Y is present, select translation Y (plural); if neither is present, proceed to step 3.

Steps 3–5 contain the routine for number translations based on the noun, preposition and other cues serving when the preceding steps are inapplicable. They are formulated in a manner similar to the comparable steps of rules 102 and 105.

In our example, step 2 applies, yielding the singular translation acid.

Providing the glossary entries for my example with the appropriate diacritics, and applying the rules as formulated above, the sequence pyrocatechol esters of triarylmethylphosphinic acid will be translated properly by pyrocatechol esters of triarylmethylphosphinic acid.

In conclusion, I would like to discuss in some detail in what way morphological features and syntactic relations enter into the four grammatical translation routines discussed above: the routines covered by rule 102 and 105 for the translation of suffix morphs -oJ and -uJ, and the routines covered by rules 202 and 209 for the number translation of noun bases.

It is clear from the above discussion that in translating the suffix morphs, their grammatical identification as to case was obviously taken into account, but the various steps of the routines were not designed to detour around case identification—they lead directly to translation. Cues were classified and provided with diacritics directly in terms of the English translation required, as for instance in
the translation of for morph —oil in terms of a preceding noun, irrespective of whether the morph itself represents a genitive or instrumental.

On the other hand, the case function of the suffix morph was specifically taken into account in the final steps of the number-translation routine for bases of type 202, by using cues grammatically applicable to the case relation, and not the number category.

The syntactic relation of agreement between adjective and noun in an adjective-noun construction is utilized by letting the first case suffix morph of the construction serve as the initiator of the proper translation of the case meaning of the entire construction. The reason is that any non-zero translation of a case morph will have to undergo rearrangement in order to precede the translation of the base, and to result in normal English, it will have to precede the translation of the first of the several bases contained in the construction. Subsequent case morphs in the construction will then be programmed to receive zero translation. Thus, rearrangement is held down to the necessary minimum. While the agreement relation itself is here utilized translationally, the agreement-dependence of the adjective and other modifiers on the noun within the construction is ignored.

The relation of government between two adjacent nouns or adjective-noun constructions is utilized extensively for the translation of case-suffix morphs, but without an intervening formally expressed grammatical identification routine, as can be seen in several steps of rules 102 and 105; this same relation, together with the government of noun or noun-adjective construction by a preceding preposition or particle is also utilized in the final steps of rule 202.

Note, however, that these morphological features and syntactic relations are not utilized whenever present, but only when required.

Thus, the nominal suffix property of morph —bl is ignored in rule 105, while the nominal and adjectival suffix properties respectively of morph —oil are utilized to initiate two separate routines in rule 102.

Likewise, the agreement between adjectival suffix morph —bie and nominal suffix morph —bl are ignored in arriving at a plural translation of ЭФИР—, while the agreement between adjectival suffix morph —oil and nominal suffix morph —bl is the basis for step 2 of rule 202 for putting out a singular translation of КИСЛОТ—.
Thus, structural linguistic information is utilized selectively in terms of its relevance to the translational result, which remains the final criterion.
Language As Symbolic Logic
William M. Austin
Institute of Languages and Linguistics

The symbolic logic of the last century, the "new logic" of Boole, De Morgan, Frege, Peano, Peirce, Russell, Carnap and others, must be understood as representing a system (or systems) partially based, but not consistently, on the Western Indo-European languages. However coherent and "logical" this system, or systems, may be, it will have to be regarded as arbitrary and its propositional solutions as non-unique. The clue to this non-uniqueness lies in mathematics itself where different systems, particularly of algebra and geometry, may each be valid for a particular frame. This is reinforced at the other end, if one may regard symbolic logic as the link between the disciplines of mathematics and linguistics, by languages themselves, where syntax exhibits logical structures, differing slightly between related languages, but much more widely between language families.

From the following elementary propositions (and they are elementary not only for the brevity of this paper, but also because of my own hesitancy in the field) we may see that languages have logical structures of their own which contain as valid propositions in the universe of discourse as the so-called "laws of thought" of symbolic logic. The "vagueness, imprecision and trickery" of "natural" languages, which logicians often repeat, disappears after a little examination and is replaced by the observation that system A differs from system B.

<table>
<thead>
<tr>
<th>English</th>
<th>Symbolic Logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barbarossa is Frederick I</td>
<td>Barbarossa = Frederick I</td>
</tr>
<tr>
<td>Barbarossa is a hero</td>
<td>Barbarossa ∈ hero</td>
</tr>
<tr>
<td>To sleep is to dream</td>
<td>To sleep ⊂ to dream</td>
</tr>
<tr>
<td>God is</td>
<td>E! God</td>
</tr>
</tbody>
</table>

Here = is "identical with"; ∈ (epsilon) is "is a member of a class, is a"; ⊂ is "entails" and E! is "exists". Only by such precise symbolism logicians maintain, can we bring logic out of language. Actually, the Western European languages, e.g. English, to which logicians purport to bring order, express the matter quite as precisely, but somewhat differently. The relation is always the same, the distinctions lie in the terms. "Barbarossa is Frederick I" might be expressed as A = B (where A and B are unit classes). Symbolic logic has a more technical way of indicating unit classes. Within its
calculus it is fairly easy to state that there is at least one member of a class, e.g.

\[(\exists x) : x \in A\]

which could be read as "There is at least one \(x\) where \(x\) is a member of \(A\)." For the logical concept that there is only one member of a class the formula is a little more complicated, that is,

\[(\exists x) (\forall y) : (x \in A) \land (y = x)\]

"There is at least one \(x\), such that, for any \(y\), \(x\) is an \(A\), and if \(y\) is an \(A\), then \(y\) is identical with \(x\)." Which is to say that the membership of class \(A\) is limited to one element, \(x\). English proper nouns, that is, nouns preceded by the zero allolog of the or nouns preceded by the are logically unique, that is, unit classes in the particular universe of discourse: "John is Mr. Smith" (and \(B = A\), "Mr. Smith is John") or "John is the king" ("The king is John", etc.). In "Barbarossa is a hero" the relation is not \(\epsilon\) "is a" as Peano so naively assumed, but again "is" while "a" is an indicator of a non-unit class (noun = class—so this can be symbolized by something like \(A = a\) "unit class—is—member of non-unit class"). The third proposition, "To sleep \(\subset\) to dream", where \(\subset\) presumably = "entails" again arises from an ignorance of "natural" language. Here the two terms, or classes, are, in extension, equal, both being infinitives. If one may quote here the motto of the Dominican order, *Laborare est orare*, "To labor is to pray" one may see at once that logical identity, not entailment, is involved. So the symbolic formula would read something like \(a = b\). Only in the last "God is" might a case be made for handling the relation "is" as something else. But logically this might be expressed as \(a = (A)\) where \((A)\) could equal an unexpressed first term.

One has to go only so slightly afield as Classical Latin and Russian to find similar propositions stated in a slightly different way. *Socrates est philosophus* would represent the two formulations of English \(A = B\) and \(A = a\) ("Socrates is the philosopher", "Socrates is a philosopher"). Russian would be the same as Classical Latin for the above Сократ философ but the Russian logicians would write "всё" between the two terms if they were of equal value (in this case, both animate), while if one were animate and the other inanimate, "Socrates is a rock". Сократ—камень they would use a dash. Here again the two presumed relators are actually a difference in the order of terms or classes.

---

Passing to other language families the situation is much stranger. In Mongol, for example, the whole problem of predication cannot arise, for "the sky is blue" is simply an equation of classes, "the sky blueness is". Here is resolved a long-standing problem of logicians, the "older logic" tending to rely on predication, while the "new logic" emphasizes the relations of terms. To go still further afield we have only to recall Whorf's remarks on Hopi, which has no nouns and hence no classes, at least in our sense. But a logician's nightmare would be a physicist's dream (viz. "it electrons").

Even the most classic and tradition-supported of all syllogisms is not immune to harassment by symbolic logicians and linguists alike.

All men are mortal  
Socrates is a man  
\[\therefore\] Socrates is mortal

In symbolic logic the relationships are expressed a little differently as,

men \(<\) mortals  
Socrates \(\epsilon\) man.

Here \(<\) means that all members of a class are included in the following class. But linguistically "is" and "are" are not a trick to confuse logicians, but rather express an identical relation. The difference is that Socrates and man are different classes of nouns. This syllogism, translated literally into Russian, results in the following:

Все люди смертны  
Сократ человек  
Сократ смертен

Here the syllogism, if it is retained in three sentences, and by definition a syllogism must, is an impossibility.

"All люди are mortal.  
Socrates is a человек  
\[\therefore\] Socrates is mortal."

---

Logicians forget the fact that люди and человек are in extension identical, one being a plural category of the other. So it seems that one has to add a fourth proposition stating that the two terms are equivalent.

Here, as elsewhere, many of the logicians' problems arise from a confusion of levels (and in this they are worse than linguists). They talk much of extension, the range of applicability of a term and intension, the content or meaning of a term. Logic is really possible only on the former level for only there can we deal with logical structure. In the realm of intension a logic, if attempted, would be basically trivial and, in a sense, would defeat its own purpose, which is the orderly arrangement of formal structures. All terms if understood as content or meaning would be different, and the best that could be achieved would be a sort of limited grammar of relationships. S. K. Langer (op. cit., p. 126) writes "the systematization of general propositions is the great contribution of logic to the concrete sciences. But general propositions, which are quantified propositional forms, always refer to members of a class, for it is only of such that we can say 'all' or 'some'. Obviously only propositions about extensions can be quantified." We have the identical situation in language where extension refers to the structural range of classes of morphemes, words and phrases. Intension would consist of "lexical meanings". It is this dichotomy of lexicon and structure that makes metaphor and paradox possible. To say "black is white" is, on the level of intension, impossible, but in extension where "black" and "white" are class equivalents, this is not only possible, but may even be considered profound, as are "the rich are poor" or "the weak are strong". This same extension also makes possible the grisly prospects of 1984, "war is peace" and "freedom is slavery". In short, when it comes to linguistic form we may say, to paraphrase Sapir, that ether and concrete are identical and that the cat walks with the hippogriff.

Only a fraction of the total problem, an elementary fraction at that, has been presented here. But enough has been said, I think, to indicate that symbolic logic has no stranglehold on truth, no monopoly in "the laws of thought", which, if they can be observed at all, can be observed only through language, "natural" or special (i.e. mathematics). If logicians must admit a multiplicity of systems, linguists, in their turn must, if language is a logical structure, learn to formulate their points of structure on a more mathematical basis.
than most have done in the past. Linguists could even learn from the syllogism. It is probably true, as Bloomfield said (Language, p. 170), that “each sentence is an independent linguistic form, not included by virtue of any grammatical construction in any larger linguistic form.” But there are relations between sentences and the relationship is in the extension of logic or the identity of form classes in language.

These statements are frankly Whorfian as all future statements on this important new frontier of science must be. (Cf. Whorf, op. cit., Languages and Logic, and Language, Mind and Reality, pp. 233-270). What is touched upon here is the specific concept of language as symbolic logic and the relationships of different logics to each other. Linguistics, suggesting different frames of logical structure and the consequently differing analyses of the universe, is at the very pivot of this new frontier. In one direction stretches philosophy, mathematics and logic, to which language is tied by extension, by the applicability of its points of structure. In the other direction language is bound by its intension or content to anthropology and sociology, which alone can furnish its “meaning”.
(Mr. Duncan Harkin, U. S. Government Consultant, was scheduled to present a paper on: "The Mathematics and Logic of Language." Due to illness he was unable to do so. - Ed.)
DOSTERT: I note in your paper, Mr. Joos, at the end, the same note of pessimism which was found in your review of Locke and Booth's collection of articles on MT. I don't know whether I am reading you correctly—or translating you correctly in terms of my concepts—but you seem to think that we are something in the nature of a millenium away from the solution of the problems of the translation of prose by machine processes. Do I interpret you correctly?

JOOS: The millenia have a way of getting shorter these days.

DOSTERT: You wouldn't care to indicate the rate of shrinkage in respect to machine translation, would you?

JOOS: As I remarked in a publication which will be out in a couple of weeks now, the Readings in Linguistics volume, the development of linguistics in this country from 1925 to 1955, 30 years—one generation—is about the same in extent as the development of mathematics from Newton to about 1850, or in chemistry from Dalton's atomic hypothesis of 1810 approximately down to nuclear fission, about two generations. So we have skipped from 200 years down to 60 years, and now down to 30. Thus the time scale keeps changing. The culture is developing and technology is developing more rapidly all the time. I do regard machine translation—which I would consider as being adequate as being in some sort of competition for what I can do myself in some of the languages I know—I do regard that as millenial. But considering the very swift progress in technology, possibly my grandchildren may witness it.

DOSTERT: It may be pertinent here in commenting on Professor Joos's remarks, to indicate the orientation of some aspects of our research. We have taken an existing English translation of texts in the field of organic chemistry. We have tried to standardize the English texts. That is to say, to remove from the translation some of the unnecessary, superfluous language and some of the idiosyncrasies which the human translator, who is usually frustrated in his creative processes when he translates, inserts into the translation in order to somehow manifest his personality in what he produces. After standardizing the text we proceed to analyze the material in terms of the transfer process; that is, the transfer of meaning from one set of symbols to the other. We do so by lexical and grammatical analysis. We have deliberately chosen what might be called a corpus in the 'inert' form
of language, because in the scientific treatises we are presented with a corpus of language relieved of many of the more elusive, and therefore more difficult to render in translation, forms of expression.

While I would go along with Professor Joos in saying that the translation of, say, War and Peace by mechanical process, it may well be that our grandchildren alone will see that sort of thing. In relatively accessible language, in the descriptive language of the sciences, we may be able to do something mechanically which will be acceptable within a reasonable number of years. In two-three years—we might have something adequate. We will have to retain human intervention. That is to say, the output text will have to be reviewed. I don’t mean retranslated, rather edited. We cannot expect the machine, at least in the foreseeable future, to come up with a text that will stand completely on its own feet. Our Russian opposite numbers are saying that they are already doing translation on fairly broad scope today, though they give little precise information on their technique.

GARVIN: I’d like to make just an extremely brief remark on this matter of literary vs. non-literary translation. The linguists in Czechoslovakia who have dealt with literature and translation in this respect have claimed—and I think they are correct—that in literary translation you really deal with more than translation—with a sort of "poetic re-creation", and obviously this cannot be expected to be done by a logical machine since it involves, in addition to logic, some of the other faculties of the mind. As far as technical translation goes, I think I would say that, as Mr. Dostert said, no more editorial work than is ordinarily required for any human translator would be required for the machine. Most translation agencies edit what they get from their translators and I don’t think that if we get a product two or three years from now it will be worse than what the human translator presents from the standpoint of requirements of editing.

JOOS: Concerning this Czech literary theory, I would like to use for their approach the old term "transcendentalism" and I expect such transcendentalism from the Czech school of literary critics. I am, as a more-or-less ordinary American linguist, rather on the "positivist" side against the transcendentalists. I hold that there is nothing ultimately transcendental in any literary document and I have done literary criticism work myself and also glosses of literary texts and annotating of them. My approach to this I think is adequately summed up in my long concluding paragraph on denotation and connotation which I reduced to a more elaborate programming of the kind that is
being worked on at present. It is particularly interesting to me to see how well Mr. Garvin's paper and mine meshed, specifically with his example of the two translations of эффект as "ether" and "ester." Where he finds that 75% of the cues are adjacent, 22% are local (by which I mean in the same sentence), only 3% can be called ambiguous in any way, and I think that the programming of the kind of machine-computer that they are working on now—a computer that learns by experience—has to include a pleasure principle, or a reward-and-punishment in the machine. If you include that, the 3% will shrink to the level of the carelessness of the original author. And then again, concerning the original author, when editing is spoken of here we generally mean post-editing, eliminating pre-editing, but I do not think that technical translation by machine on a large scale can afford to do without a certain modest kind of pre-editing which can be done by a monolingual editor—an editor who knows nothing about problems of translation, but operates quite mechanically. Such a person would have certain specific jobs to do. One, he must reduce the original text to sentences no more than twenty words long each. By hook or crook, they've got to be broken up in sentences each one of which is not more than twenty words long. Then this pre-editor is to introduce few stylistic simplifications. (That's not essential, but I think such an editor can be taught to introduce enough stylistic simplifications so that the machine can handle it more easily, and then this pre-editor will have to standardize abbreviations, expand some, expand all abbreviations that are not on the standard list. These three things can be done by a monolingual, and then the machine that I envisage for technical translation might be complete, I think, in a generation.

GARVIN: Would it be the job of the pre-editor to correct misprints?

JOOS: Oh, I assume that. This is a secretary's job anyhow. We expect the secretary to straighten out simple misspellings—simple misprints which can be called misspellings.

ALO RAUN (Indiana): In connection with connotation and denotation, would it not be possible to establish a linguistic norm like sememes as opposed to allophemes?

JOOS: I think those terms can be defined usefully. I have not been tempted to use them myself. In my reading where these terms are found, most authors have used them for begging various questions, so I don't use them myself. But I might have a use for them. I think
I could define them in a useful way. More or less as I attack the question of defining the two words denotation and connotation.

ZARECHNAK (Georgetown University): Denotation and connotation are not both properly the subject of manipulation by machine. My experience has been that the machine does not have to handle connotation, but if I have a good dictionary with the meanings in Russian I can translate adequately from Russian into English and from English into Russian. A Russian mathematician would understand terms which I would not. A machine has to translate the structures of the source language into the target.

JOOS: I agree that the normal or average reader, let us say the reader of the Reader's Digest, does not get a very large fraction of what I call connotations, and if you want to make your translation for a normal reader, then you don't need to worry particularly about connotations. I introduced the term as leading up to literary translations, and literature is so far away from complete grasp by any single reader that of course the problem of interpreting literature by college professors is a profession, at times lucrative.

CARLTON HODGE (Foreign Service Institute): I would like to ask if, granted that it will take some time to get anything that would translate literature, but in the process what linguistic byproducts can we expect from the research?

JOOS: The history of linguistic science in the last generation has proceeded in a way you might say from the small to the large. Phonology was well under control at the beginning of the second world war. Then, morphology and syntax, and we are now moving evidently in the direction of a linguistic semantics; surely the linguists' semantics will profit greatly from intelligent MT research (in meaning) transfer problems.

PAPER: (Michigan): I would like to add just one comment on what Mr. Joos just said. I think that in the areas he calls "inside meaning" since we now have at our disposal these complex and rapid data-processing machines, we can now get back to detailed study of the statistical frequencies of form classes, of words, of morphs, a project which we could not even dream of undertaking before because we didn't have the equipment at our disposal.
JOOS: Perfectly right. But when I start thinking about it, it seems to me that there would have to occur simultaneously substantially the same text in two languages, that is a good translation from one to the other. Feed them both in together in order that the machine may be informed, then study context distributions to see what semantic differences these context distributions are correlated with. I don't see quite how the machine can do it in one language, but if I retire into silence for a while, I may be able to figure that out later.

PAPER: Of course, as a linguist, I would be perfectly happy to get the statistical frequency information about the occurrence of morphs, for a particular language, regardless of the eventual feasibility of MT or not.

JOOS: As a linguist, or a semanticist, or literary critic, I compare the text with the imaginary picture of its reference as I build it up. I read a literary text and I build up a picture of a situation. That serves me in place of that other language. The machine needs it in the other language, or in Mr. Newman's type of analysis. For comparison, I use instead the real world for comparison of the text frequencies. Now, my concerns here have been mostly in the field of literature. For example, in Middle High German I find that the words for "eye", and "heart" tend to occur in context with each other, not necessarily in the same sentence. But if you take the same paragraph they are in context with each other better than three-quarters of the time, and that is certainly worth knowing for semantic research.
Character Sensing

As An Input To

Machine Translation
II

FIRST LUNCHEON PAPER

Character Sensing As An Input To Machine Translation

CLYDE C. HEASLY
Intelligent Machines Research Corporation

The need for automatic character sensing extends across the entire data processing field. There are numerous potential applications and, fortunately, a few present commercial applications which are proving the practicality of character sensing and are the source of our most useful experience. The most comprehensive of reading problems are those of reading for the blind and of reading for input to machine translation systems. These problems, in their ideal solution, require machines capable of reading a tremendous variety of material in terms of quality, style and format of the data.

It would seem that ideal solutions of either of these problems are not immediately forthcoming. I would like to describe to you what IMR is doing in some commercial applications and in some laboratory developments. This will give you an idea of what progress one company is making in character sensing. Against this background I will try to indicate what further steps remain in reaching machines of practical utility for machine translation.

In commercial applications, we have reached a goal which seemed far distant a few years ago. We have installed several machines which are now being relied upon for the daily processing of business data. These machines are reading documents of commercial significance and punching what they read into tabulating cards. They process documents at appreciable speeds creating punched cards at rates of from 100 to 180 per minute. A single machine can do the work of a large number of card punch operators; if the equipment fails, work piles up at an alarming rate. But our equipment is working with performance and failure rates that meet commercial operational requirements. One application has been field tested for over a year. Apparently it was not disillusioning, because the installation is being expanded system-wide as the principal means for processing this particular kind of data.
I would like to describe this installation in some detail. It is the application of character sensing to the reading of customer invoices in gasoline and oil credit sales. This equipment is far removed from the problem of machine translation, yet it demonstrates important principles. An important feature of this application is the advance control which was exercised over the nature of the data to be read.

The credit token which is issued to the customer when he is first approved for credit is a plastic Charga-Card* which is embossed with his account number at the oil company’s central office. The plastic Charga-Card was developed by Farrington Manufacturing Company who sponsored development of the character sensing equipment. Since the embossed image forms the type from which the account number is printed, any freedom in the design of the embossing can be used to advantage in improving readability of the imprinted invoice. Two steps are taken to improve the resulting image.

First, a maximum reliability type face is embossed into the card. The characters of this type face are formed of line segments which are deliberately made bold to reduce likelihood of broken lines due to light impression. The line segments are arranged to avoid small closed areas which might be filled in by smudge or heavy printing. Each digit of the set of ten digits differs from the other nine by at least two strokes. Since only one stroke-difference between each digit in the set would be sufficient to differentiate between them, the additional stroke-differences amount to a redundancy.

The machine utilizes this redundancy of information to reject digits which are defaced or imperfectly formed. It requires that all strokes which describe the character be present and that all strokes which do not describe the character be absent. If these requirements are not met, the character is rejected.

This use of redundant information is quite useful. The image which must be read is a carbon image which results from a pressure process. Heavy impression and subsequent handling can result in changing the apparent strokes. Light impressions can cause a stroke to be missed. In either case, the machine avoids a mistake by refusing to commit itself.

*Trademark Reg. U. S. Patent Office
At first appraisal, a rejected digit might appear to be as useless as a wrong digit, but this is not the case. The second preventive measure taken in preparation of the credit token makes it possible to restore the rejected digit if only one digit is rejected.

This second step amounts to adding a second kind of redundancy to the account number. An additional digit is included in the number. Thus, if nine digits are required to uniquely identify all customers, a tenth digit can be assigned in whatever manner will accomplish the most good. In this case the tenth digit is assigned according to a simple scheme, the IBM self-checking number scheme. This scheme will tell you the identity of one missing digit if all other digits have been correctly recognized.

The machine reads the account number. If a digit is rejected, the corresponding storage column is allowed to remain empty. The checking circuit responds to each correctly recognized digit. At completion of reading, the checking circuit will lack being satisfied by an amount related to the rejected digit. The checking circuit is interrogated to determine what digit is needed and this digit is inserted into the empty storage column. Thus the number finally punched will be the number which should have been read. As far as correct punching is concerned, a single rejected digit yields the same result as a number correctly read. It would be fortunate if all character sensing applications permitted the luxury of reading self-checking information, but such is not the case.

In other commercial operations, machines are operating successfully even though self-checking data is not available. In such systems the Reader occupies a position quite similar to the human operators it replaces. One such machine is reading the customer's telephone number—which is self-checking—and the amount paid on account—which is not self-checking. In the manual operation which the machine is supplanting, the punched cards are balanced against previously determined batch totals. A similar balancing operation is now used to prove the accuracy of automatic punching by the Reader. If card punch operators were perfect, it would be difficult indeed for an Analyzing Reader* to compete unless the material to be read were perfect. Of course, documents used in the real world are not perfect, but operators are not perfect either. This means that there must be

---

*Trademark Reg. U. S. Patent Office
some means to prove the accuracy of punching by either a card punch operator or a machine reading imperfectly printed documents. It usually works out that the great volume of correctly punched material which the machines can process more than compensates for the procedures necessary to prove the accuracy or isolate errors.

A more advanced machine which comes still closer to the requirements of machine translation input is used to read alphabetic as well as numeric characters. This machine, located at the Reader's Digest Condensed Book Club in New York, is used to read a card typed in their Order Department. The field to be read is the customer's name and address, which may have two, three, or four lines. From the first line the name is read and first and last initials selected. The house number is read from the middle line, if there is one, and the city and state are read from the last line.

This Analyzing Reader has difficulty on some material which can still be read by a human. But it does not get bored or tired, and it has "off days" only when components fail. It reads and punches cards at the rate of 150 per minute and this great capacity is particularly useful during the peak volume periods associated with sales promotions. So again, the use of character sensing is proven to be economically attractive, partially because of the ability of the machine and partially because of the intelligent procedures with which it is used.

Perhaps it would be appropriate at this point to discuss some of the means by which Analyzing Readers scan characters and identify what they see. While over five years of development work is behind us, the basic ideas which have evolved can be explained rather quickly.

The first element of the system is the scanner. The scanner can be likened to a television camera in its operation, although its structural details are quite different. A fixed line in space is scanned rapidly and repeatedly as the document to be read passes by. The line of scan, called a "reading station", is parallel to the vertical axis of characters and scanning is so rapid relative to document motion that the scans overlap.

If it is imagined that a capital letter "E" is passing the reading station, the following events will occur: while the vertical line on
the left side is being scanned, a pulse of long time duration will be emitted from the scanner. The duration of this pulse will be determined by the length of the line which caused it. Three pulses will occur during each scan across the three horizontal lines. Finally, after the character has completely passed the reading station, no pulses will result until the next character begins to be seen.

These pulses can be used to identify the character "E" if they are presented to a computer which has been programmed to look for and detect such things as "at least one long pulse", "three properly positioned pulses many times in succession" and "no pulses", provided, of course, that the computer which detects these pulse patterns can also keep track of where they are located, horizontally. This means that the computer can say that the "long vertical line" was on the left side, the "three crossings seen many times in proper vertical position" were in the middle and the "nothing" was on the right.

"End of the character" is used to tell the computer to make a decision on the basis of what it has seen. The things it has seen in this example have been the distinguishing properties of the letter "E", and it will be programmed to emit the signal for "E" every time that unique combination has been seen. In similar fashion, other characters are recognized by the unique combinations of strokes which define them.

The extent to which the details of the character play a part in recognition depends on the requirements of the problem. For example, in most styles of capital "E", the middle horizontal stroke is shorter than the top and bottom strokes, so that at the right end of the character three pulse-scans are followed by a few two pulse-scans. Yet this was not included in recognizing "E" because, in this example, it was not set up as a condition of interest. In similar fashion all other details will be ignored unless routines are set up to detect them. Whether or not the details are implemented will depend on whether or not they are needed in the individual reading problem. In some cases, it may be useful to require that characters be of exactly prescribed size and that certain prominent details peculiar to the type face be seen. In other cases it may be useful to allow considerable latitude as to how long is "long" in a vertical line and to accept three crossings with some variation in spacing. The general method is to instrument whatever characteristics need to be
recognized in order to uniquely identify the characters as found on the actual documents being read.

It becomes evident from the preceding remarks that the machine actually identifies the characters by recognizing the strokes of characters as a code. It is, of course, true that characters are a code and the application of the technique to real problems involves determining how minutely the code must be examined in order to successfully recognize each character in the presence of the noise (smudge, overstamp, light printing, etc.) which accompanies the code.

This reading of code in the presence of noise is related very closely to the question of whether, or perhaps more accurately, when IMR can build an Analyzing Reader for MT input purposes. It would appear that many of the documents of interest for automatic translation will carry a great deal of noise. The characters will not be of uniform quality, and quality cannot be improved because the documents are generated by sources over which no control can be exercised.

However, this does not mean that character sensing cannot eventually be applied to MT problems. There are other techniques which can be interposed between scanning and final decoding which greatly enhance the ability to read through noise and through variations in the characters themselves (i.e., subtle changes to the input code). These techniques increase the flexibility with which a machine can analyze a particular stroke and recognize it even though the stroke is imperfect.

Some of these techniques are being applied to a machine now being built which will read the alphabetic capitals, the numeric characters, fourteen punctuation marks and two hand-drawn edit symbols. This machine will read from page copy and punch out what it reads in teletype tape at the rate of 60 characters per second.

The parameters of the problem thus far described are all necessary for machine translation input. The machine falls short of MT requirements in that the material it is to read is prepared under reasonably well controlled conditions on a well maintained printer. It does not read lower case characters, either, but this machine is very much closer to what is needed than the first machine which was described as reading stylized numeric characters only. The simplified
conditions under which the alphameric Reader will perform are perhaps at the halfway point between the simple problems being solved commercially and the more difficult problems of machine translation input.

The use of simplified conditions points to an important facet of IMR's methods. We look at each reading machine in terms of its place in a system. We try to make each element of the system contribute as much as it can to overall performance. Improvements of a minor nature often have profound effects on the output quality of printers and typewriters. Once the other elements have been controlled as well as is practical, we then build the Analyzing Reader of least complexity which will still handle the resulting data.

This approach has led us into problems bearing high economic impetus, where the desire to reduce costs causes minor improvements in document-generating equipment to appear attractive in the light of the end in view. We have at the same time tried to include important innovations in each new machine so that as we proceed from problem to problem our techniques improve and the field of our competence expands.

We believe this to be the method which will cause the fastest progress. Perhaps the steps which we will take on the next few machines will bring character sensing and machine translation closer together. Just as we started reading under restricted conditions and then expanded step by step, so apparently are you beginning your translation efforts in well defined, restricted areas of the source language. As we each expand our scope, the day should arrive when a Reader of less than ideal ability can be combined with a translator of less than ideal ability to solve a practical problem of real and useful significance.
Ill

Panel II

Lexical Problems In

Machine Translation
LEXICAL PROBLEMS IN MACHINE TRANSLATION

The Rationale Of The Idioglossary Technique

VICTOR A. OSWALD, JR.
University of California, Los Angeles

Upon reviewing recent linguistic publications in the field of Machine Translation, one cannot fail to be struck by the extent to which the principle of translation via idioglossaries, which met with wide acceptance at the first conference on Mechanical Translation (Massachusetts Institute of Technology, June 17-20, 1952), has either fallen into abeyance or is being applied unconsciously rather than with a full awareness of its implications. To cite specific instances, the most elaborate recent attempt to resolve lexical problems, Miss Masterson's chunk-lattice-thesaurus procedure¹, is riddled with needless complications introduced by the failure to isolate scientific discourse from the tangle of general language; and Mr. Gode's procedure of translation via Interlingua², which is to date the most wide-sweeping venture toward a solution of problems of word-order and syntax, incorporates, but does not acknowledge, translation by the use of an idioglossary. Evidently the proponents of the idioglossary principle³, instead of concentrating upon a demonstration of how to compile and apply a specific idioglossary, should have elaborated at greater length upon the rationale of the procedure.

It ought to be taken as axiomatic that each and every realm of scientific discourse requires for the expression of ideas peculiar to itself a vocabulary which is, at least in part, peculiar to itself. To be precise, the substantives which "name" specific objects or established concepts in a particular realm of discourse, together with the verbs that "describe" the operations carried out with these

¹Margaret Masterson, "The Potentialities of a Mechanical Thesaurus," Cambridge Language Research Unit, 1956 (typescript).
objects or these concepts, constitute a unique jargon intelligible only to the initiated. This is true not only of scientific discourse, but of the communication of trades and handicrafts and hobbies as well. Let us descend for a moment from the austere domain of the sciences and examine a brief, but graphic, sample on the level of the *lingua populi*.

Not many years ago a gifted cartoonist, the late H. T. Webster, capitalized on the existence of popular "private" languages by running off a series of cartoons entitled "They Don't Talk Our Language." A hypothetical reconstruction of such a cartoon strip might run somewhat as follows (in all of the panels a husband and wife are represented as sitting at ease in the living-room after dinner).

**Panel One**

Husband: I checked the rig in the lab this afternoon. The new electrostatic tweeter brings the frequency way up.

Wife: But doesn't it get out of phase with the old woofer?

**Panel Two**

Husband: Well, I think we can compensate by putting in a new preamp. I'm much more worried about a tendency in the arm toward lateral inertia.

Wife: I thought something was wrong with the tracking, but I assumed it was vertical inertia. Maybe a hydraulic arm would be best, after all.

And so for at least two more panels. The last panel carries the punch line.

**Last Panel**

Wife: Oh, by the way, May Johnson called just before you came home. She and Phil want us to come over tonight.

Husband: Heavens, no. They'd want to go over the slides they took last week with that new camera of theirs, and I wouldn't understand a word all night. *They don't talk our language.*

Please note that this specimen of "hi-fi" jargon displays two quite different kinds of impediments to intelligibility. One is the
use of coined words like "tweeter" and "woofer", which simply mean nothing outside of hi-fi context. The other—and the more formidable impediment to intelligibility—is the use of what we may call analogical conversions, the adaptation of established words to new and more or less inscrutable uses. The "arm" (a conversion) is, of course, the "pick-up (another conversion) arm" which holds the "cartridge" (a conversion), and the cartridge in turn has in it the "stylus" (a conversion) whose "tracking" (a conversion), that is, its undeviating conformation to the "grooves" (a conversion) of the "disc" (a conversion), is vital to the attainment of "high fidelity" (a conversion).

Fortunately, for those who hope to see mechanical translation realized, the processes of coinage and analogical conversion are used by men everywhere in the world to create the unique terminology of any particular "mystery," with the happy result that the vital terms of the jargon of, say, a German-speaking biologist, will have unique equivalents in the jargon of an English-speaking biologist. An idioglossary is simply a bilingual list of such unique equivalents within a given realm of scientific discourse.

Coinages, whether those of recent origin or those of long-established usage, are likely to have unique equivalents not only, say, between a jargon of English and a jargon of German, but even within the entire scope of these two languages. To roam about for a moment in the territory with which I am most familiar, English linguistic terms of recent origin, such as "phoneme," "morpheme," "taxeme," or "sememe" will turn up in any foreign language with a unique equivalent or no equivalent at all. It would appear to be an unconscionable waste of machine time—no matter how rapidly machine analysis may proceed—to put terms like these through anything so complicated as a chunk-lattice-thesaurus process. On the other hand, analogical conversions probably cannot be latticed because they would frequently prove inextricable from the maze of general language. To put the case on a very homely level again, it would take more ingenious programming than I can imagine to extract from the jargon of plumbing the fact that in this context the foreign language equivalent for "snake" would have to be a word naming a long section of flexible metal used to dislodge foreign matter from a drain pipe; or that "plumber's friend" would require a word designating a rubber suction cup with a wooden handle. If these examples seem unfair or extreme, consider the case of "set" in modern psychology, or "crotchet" in musicology, or "stop" and "click" and "stress" in phonetics. How
is a lattice going to separate "articulate" as used in anatomy from "articulate" as used in linguistics? One can only feel torn between admiration for the flexibility of the human mind and despair at the thought of trying to reproduce such flexibility in any mechanical device. It will be much easier to operate with a rigid comparison of terms which have unique equivalents.

One cannot but disagree with Miss Masterson that the chunk-lattice-thesaurus process "feels like a model of what we ourselves do when we translate". It must rather be said that what we do, when we ourselves translate, at least when we translate scientific discourse, is to run the foreign language input through a sorting mechanism in our personal translating machine, a process in which we refer all terms from any special realm of discourse to a special repository of target language output equivalents. We do not scan the context for clues to help us in resolving lexical ambiguities. In any special realm of discourse there are no lexical ambiguities. It is for this reason that Mr. Gode can say that, in translating "atrial fibrillation and flutter" into French: "There are but four equations involved which yield the French words "auriculaire, fibrillation, et, flutter." Mr. Gode knows in advance that this example is from an "actual medical text," and he automatically turns to his personal, built-in medical idioglossary for his French equivalents. If he had not done so, he would have found that "atrial" has not one, but at least three possible equivalents: in the domain of archeology "atrial" would require a translation meaning "pertaining to the central main room of an ancient Roman private house"; in architecture it would require an equivalent meaning "pertaining to a courtyard, usually surrounded by colonnades, in front of early Christian or medieval churches." Only in his medical idioglossary can Mr. Gode find a one-to-one equivalent, meaning "pertaining to an auricle of the heart," i.e., French "auriculaire."

I have focused my comments on the work of Miss Masterson and Mr. Gode, not because I do not respect them, but on the contrary, because I think they have made, from the standpoint of linguistics, the most original propositions of recent years. Mr. Gode's translation via Interlingua appears capable of solving, at one stroke, many of the tiresomely complicated problems of syntactic resolution. His

4Op. cit., p. 9
5Op. cit., p. 56
procedure represents, to be sure, pre-editing on a massive scale; but one cannot fail to be impressed with the fact that it also represents the first demonstration of how mechanical translation might operate on a multilingual scale. If his process is to be mechanized, however, he will have to provide the machine with the idioglossaries he himself uses. Miss Masterson's chunk-lattice-thesaurus operation can be used to save time in programming access to idioglossaries, but it cannot dispense with them. When she chose an Italian article on botany to carry out her demonstration with the "chunk" PIANT-, she was actually working with an idioglossary that was only in part cluttered up with general language terms and idioms.

I do not wish to belabor my point much further, but it might be well to point out that another recent proposition of considerable ingenuity, Mr. Reifler's "Mechanical Determination of the Constituents of German Substantive Compounds," will likewise fail to work unless it is used with an idioglossary. I am persuaded that he can mechanically separate the components of compounds like "Säuge-tier" and "Säuge-tier-blume," and that for "Säuge-tier" he can eventually arrive at the English equivalent, "mammal"; but where, save in a botanical idioglossary, is one to find the information that a "Säuge-tier-blume" is neither a flower that feeds like a mammal, such as the Venus flytrap, nor some such incredible monstrosity as a flower that suckles its young, but simply a plant that depends upon mammals to carry out the process of pollenization? Or that a "Vogel-blume" is a plant with an analogous dependency on birds, and not a plant that looks like a bird, such as the flower we call "bird of paradise," or, perchance, a flower that flies like a bird?

The existence of a private jargon for every domain of scientific discourse and the need for idioglossaries to translate the jargon of one language into the jargon of another has hitherto seemed to me to be so obvious as not to require blunt and homely demonstration. Perhaps Mr. Lawson and I are to blame for not having made our point more forcefully or more colorfully. In 1951-53 we were intent upon finding out how to construct an idioglossary of no more than 5,000 terms in a foreign language (German) together with the same number of equivalent terms in a target language (English)—the hypothetical limit of a

---

magnetic drum which could serve as a "mechanical dictionary." To assure optimal precision within such a limited scope we compiled our own idioglossary—to the best of my knowledge, the first idioglossary ever compiled—from articles on brain surgery. By undertaking our task as we did we may have made it appear more formidable than it really is, and we certainly appear to have failed to make the point that the use of idioglossaries provides the only immediately practical solution for the major lexical problems of machine translation.

We did, at least, demonstrate the effectiveness of translation via an idioglossary, and we also made the important discovery that, along with their inevitable private jargon, scientists use a very limited range of the general language at their disposal.

In view of recent engineering developments that have provided programming devices of wider scope and storage devices with more rapid access time and vastly greater capacity than anyone would have dared hypothesize five years ago, there is no reason why idioglossaries cannot now be compiled from dictionaries for such large but essentially self-contained fields as Biology, Chemistry, Physics, or Medicine. The process of mechanical translation would then proceed in no more than four steps. First, the input material would be scanned as a segment from a special field of scientific discourse and the output material held "in storage." This step would isolate those items of the input (primarily nouns and verbs, but also a certain proportion of adverbs and adjectives) which are peculiar to the particular jargon in question. Second, the input would be scanned for "general language" items of the same categories, a not very formidable proceeding in view of the fact that scientists characteristically restrict themselves to a rather narrow range of expression. Thirdly, the input material would be scanned for language constants: articles, pronouns, prepositions, and conjunctions which are used with the same frequency and in the same way in any segment of language. Finally, the input material would be submitted to whatever programming might be required to resolve the syntactical patterns of the input language into those of the output language.

I do not wish to sound more sanguine than I really am. The small-scale experiment with syntactical rearrangement that Fletcher and I carried out in 1950-51, and the small-scale experiment that

Lawson and I performed with our brain-surgery idioglossary in 1952-53, appear to indicate that the maximal equivalence to be attained mechanically—and at that a very crude equivalence—could not be higher than about 80 per cent. I cannot pretend to know what results would be obtained from large-scale operations. I can only reiterate that the procedure outlined above appears to be the only wholly mechanical procedure likely to attain results of any kind. It is time for someone to produce the facilities to arrange for an investigation on a scale large enough to provide definitive results.
Word Decomposition for Machine Translation

R. H. Richens and M. A. K. Halliday

Cambridge Language Research Unit, Cambridge, England
Presented by R. A. Crossland, University of Durham, England

All feasible systems of machine translation are based on a unit smaller, in a great many cases, than the word. This unit, which provides the source-language entries in a mechanical dictionary, is conveniently termed a "chunk" so as to avoid confusion with other linguistic categories. There are, however, a number of ways in which words may be decomposed for machine-translation purposes and the following remarks deal with some of the principles that are involved.

Linguistic structure

By applying the well-known method of testing for linguistic commutability, it is possible to establish classes of chunks each characterized by its internal commutation relations. All the chunks considered below satisfy commutation tests; but since in many cases this technique gives a number of possible word decompositions, it is necessary to consider which possibilities should be adopted. In addition, some systems at least of machine translation require principles of word decomposition based on quite other criteria than those used in structural linguistic analysis.

Invariant words

There is no particular difficulty with invariant words such as English prepositions. Here the word is a chunk. Yet it may be that such a chunk forms a semantically irreducible compound with some other chunk or chunk class. Thus *up against* and *have (has, had, having) up* need to be treated as couplets whose meaning cannot be inferred from the normal range of meaning of the component chunks. It is simplest, however, to treat *up* in each case as a chunk whose meaning can only be elucidated by comparisons with neighbouring chunks. This comparison may have to await syntactic analysis as in *hurry him up* where *hurry* and *up* form an irreducible but disjunct semantic compound.

Affixation

A single affix likewise presents few difficulties. Both linguists and machine translators will decompose *dogs* into *dog-s* and *unkind* into *un-kind*. The notion of affixation, however, contains implications
that must be investigated further, in particular the implied distinc-
tion between stem and affix. It is certainly the case that this dis-
tinction is commonly made on semantic grounds; the stem is regarded
perhaps as an argument, operated on by its affix, or at least as
having some sort of semantic priority to it. The distinction could be
maintained purely on formal linguistic criteria with reference to the
extension of the commutation classes, but it is probable that most
methods of machine translation utilize the stem-affix distinction and
it does not greatly matter for machine-translation purposes whether
the distinction is purely formal or formal-semantic.

The question then arises whether a segment which can be either
a stem or an affix, such as or and -or, is to be regarded as one chunk
or two. In the case of two-chunk words, the position of the space
bounding the word can be used to provide a basis for distinction;
but this becomes more difficult with three-chunk words such as
possess-or-s. It could be maintained on structural grounds that -or
and -or- are one chunk and or another; alternatively -or and -or-
could be regarded as different chunks. The position is even more
complex with regard to segments such as en, en- and -en which can
function as stem, prefix or suffix. Moreover, both en- and -en can
be infixed as in dis-en-thrall and moist-en-s.

To avoid merely verbal dispute, it is convenient to regard all
segments composed of the same letters in the same order as one and
the same chunk. However, there is no reason why the stem and
affix distinction or distinction into different classes of affixes
should not be applied within the chunk if the system of machine
translation being used requires it. Thus the -en- in disenthrall can
be regarded as (1) a prefix, (2) an infixed prefix or (3) an infix,
each being a subcategory of the affix category of the en chunk. It
is possible also to classify an affix by numbering its distances
either from the beginning or end of the word or from the stem. Thus
the -en- of disenthrall is in position 2 from the beginning, 2 from
the end, and 1 before the stem. Any of these methods can be justi-
fied on structural analytic grounds; the one adopted will depend on
the machine translation method used, and different methods might
well require different ways of classifying affixes.

Multiple affixation involves further problems. Commutability
considerations would permit disenthrall to be divided either as dis-
en-thrall, disen-thrall or dis-enthrall. It is not even necessary for
all parts of a chunk to be contiguous. Something is to be gained by
decomposing the German *abgeschrieben* into *ab-* , *ge-* -en, *srieb* ;
and similarly in the Semitic verb. However, it is probably simpler
to regard disjunct affix couplets as consisting of two different
chunks which form an irreducible but disjunct semantic compound as
discussed above. It is doubtful whether any consideration based on
formal-linguistic analysis alone can be adduced for any particular
system of decomposition when multiple affixation occurs. The deci-
sion has in fact to be made on the various other criteria mentioned
below.

**Mutable stems**

The simplest instance is exemplified by initial mutation in
Welsh where mutation occurs without affixation, e.g. *pen,* *ben,* *mhen,*
*phen.* There are two possibilities here, either to regard these var-
iants as four different chunks or to divide into *p-en,* *b-en,* *mh-en,*
*ph-en.* Since most Welsh words are affected by initial mutations,
the increase in the size of the mechanical dictionary if all variants
were entered as separate chunks would be very serious. On the
other hand, if *en* is regarded as a unitary chunk, it is necessary to
be able to distinguish the above series from *c-en,* *g-en,* *ngb-en,*
*ch-en*; moreover *g-en* is not only a variant in the *c-en* series, but may
be a root form of the series *g-en,* *en*; moreover *en* is itself a root
form. It is clear that the removal of the mutable initial from the stem
is only possible if the root form is recoverable at a subsequent stage
in the mechanical-translation procedure. This is easily done, as
pointed out by Richens (1956) by treating 9 Welsh initial mutation
series as flexional classes, adding this information to the mecha-
nical dictionary and then comparing stem and initial letters for flex-
ional class. The mode of decomposition adopted must, therefore,
depend on the flexional system set up.

**Mutation plus affixation**

Combined mutation and affixation is more complicated. An
example is provided by *half,* *halves.* Obvious possibilities for
decomposition are as follows:

<table>
<thead>
<tr>
<th>half</th>
<th>halves</th>
</tr>
</thead>
<tbody>
<tr>
<td>half</td>
<td>balv-es</td>
</tr>
<tr>
<td>bal-f</td>
<td>bal-ve-s</td>
</tr>
<tr>
<td>bal-f</td>
<td>bal-ve-s</td>
</tr>
</tbody>
</table>
Here again, economy of mechanical index space will tell against the first two possibilities. The choice between the others will depend on the subsequent choice of a flexional system. The last form, \textit{hal-f}, \textit{hal-v-es}, has the advantage that it can be integrated easily with the verbal set.

\textit{hal-v-e}, \textit{hal-v-es}, \textit{hal-v-ing}, \textit{hal-v-ed}

Reduplication

Reduplication plus affixation is very similar to the preceding. Thus the series

\begin{align*}
\text{lop} & \quad \text{lops} & \text{lopping} & \text{lopped} \\
\text{lop} & \quad \text{lop-s} & \text{lopp-ing} & \text{lopp-ed} \\
\text{lop} & \quad \text{lop-s} & \text{lopp-ing} & \text{lopp-ed} \\
\text{lo-p} & \quad \text{lo-ps} & \text{lopp-ing} & \text{lopp-ed} \\
\text{lo-p} & \quad \text{lo-ps} & \text{lopp-ing} & \text{lopp-ed}
\end{align*}

As before, the first possibility is likely to be discarded since it involves two stem entries in the dictionary. Choice between the other forms will depend on the flexional system used subsequently. The third possibility is probably the best for a number of systems of machine translation.

Semantic requirements

In the foregoing, word decomposition has been treated almost entirely as an exercise in linguistic analysis. For machine translation, however, it is frequently necessary to sacrifice a commutatively possible division if there is no semantic parallelism. Thus while \textit{in-exact} and \textit{in-excusuable} are usable decompositions, \textit{in-famous} or \textit{in-fam-ous} is not, since \textit{infamous} does not mean \textit{not famous}. Prepositional prefixes to verbs in Teutonic and Romance languages offer many other instances. It is simplest to treat \textit{infamous} as a single chunk.

Flexional system

It has been noted already that the system of decomposition depends on the flexional system. The following table illustrates two of the many possibilities of treating a constellation of words containing the segment \textit{cop}.

—74—
<table>
<thead>
<tr>
<th>Word</th>
<th>System I Division</th>
<th>Flexional class</th>
<th>System II Division</th>
<th>Flexional class</th>
</tr>
</thead>
<tbody>
<tr>
<td>cop</td>
<td>cop</td>
<td>b</td>
<td>cop</td>
<td>p</td>
</tr>
<tr>
<td>cops</td>
<td>cop-s</td>
<td>b</td>
<td>cop-s</td>
<td>p</td>
</tr>
<tr>
<td>copping</td>
<td>copp-ing</td>
<td>b</td>
<td>cop-p-ing</td>
<td>p</td>
</tr>
<tr>
<td>copped</td>
<td>copp-ed</td>
<td>b</td>
<td>cop-p-ed</td>
<td>p</td>
</tr>
<tr>
<td>copper</td>
<td>copper</td>
<td>b</td>
<td>copper</td>
<td>b</td>
</tr>
<tr>
<td>coppers</td>
<td>copper-s</td>
<td>b</td>
<td>copper-s</td>
<td>b</td>
</tr>
<tr>
<td>cope</td>
<td>cope</td>
<td>b</td>
<td>cop-e</td>
<td>e</td>
</tr>
<tr>
<td>copes</td>
<td>cope-s</td>
<td>b</td>
<td>cop-es</td>
<td>e</td>
</tr>
<tr>
<td>coping</td>
<td>cop-ing</td>
<td>b</td>
<td>cop-ing</td>
<td>e</td>
</tr>
<tr>
<td>coped</td>
<td>cop-ed</td>
<td>b</td>
<td>cop-ed</td>
<td>e</td>
</tr>
<tr>
<td>copious</td>
<td>copious</td>
<td>a</td>
<td>copious</td>
<td>a</td>
</tr>
<tr>
<td>copy</td>
<td>copy</td>
<td>a</td>
<td>cop-y</td>
<td>y</td>
</tr>
<tr>
<td>copies</td>
<td>copies</td>
<td>a</td>
<td>copies</td>
<td>y</td>
</tr>
</tbody>
</table>

Total stems: 7  
Total affixes: 3  
Total flexional classes: 2

a = invariant stems; b = stems with simple affixation; c = stems affixing -e; p = stems infixing -p; y = stems affixing -y.

The main advantage of system I lies in its economy in affixes and the small number of flexional systems. A serious drawback is the large number of stems, more than double that in System II. The second system is economic in stems but at the cost of 8 affixes and 5 flexional systems. However, since the number of affixes in any system is limited while the number of stems is roughly proportional to vocabulary size, system II is probably far more economic of mechanical-dictionary space for a large vocabulary. If dictionary size is not important, system I may be simpler to manipulate.

**Chunk identification**

In addition to the limitations on decomposition set by the conjugation system, further limitations may be set by the mechanical matching technique. Most methods of machine translation envisage comparison of the words of the source passage with the chunks in a mechanical dictionary, and since a large number of words contain several chunks, techniques have to be devised to identify these...
chucks correctly. There are doubtless many ways of doing this, but it is probable that most methods impose restrictions on certain semantically and structurally permissible decompositions where otherwise misidentifications will occur.

An example will make this clear. One of the simplest techniques of matching (cf. Richens and Booth, 1955) is to match each word, beginning at the front end, against a mechanical dictionary of chunks arranged in alphabetic order but with the longer words preceding the shorter. Then, when a match is made, i.e. a chunk in the mechanical dictionary corresponds exactly with an initial segment in the word, the remaining segments, if any, are rematched. This method is simple and works in a great many cases.

thus *disloyalty* will be decomposed as follows:

\[
\text{disloyalty} \quad \text{dis} \quad \text{loyalty}
\]

\[
\text{loyalty} \quad \text{loa} \quad \text{al} \quad \text{ty}
\]

\[
\text{ty} \quad \text{ty}
\]

However *discontent* by this method would be liable to yield:

\[
\text{discontent} \quad \text{disc} \quad \text{ontent}
\]

\[
\text{ontent} \quad \text{on} \quad \text{tent}
\]

\[
\text{tent} \quad \text{tent}
\]

This can be prevented by applying the standard solution of so many machine-translation problems, namely by putting the cause of trouble in the mechanical dictionary. Thus, *discontent*, though semantically separable into *dis-content*, is treated as the unitary chunk *discontent*.

It is possible to avoid trouble with this particular word by using a different matching technique, but it is likely that any comparatively simple technique will result in misdivision in some cases. This is of no consequence if it is clearly recognized that any awkward word is to be treated as a unitary chunk.

Translation field

Machine translation schedules may be classified into 9 categories, according to whether they go from or to the particular, comparative or universal (cf. Halliday, 1957); that is whether there are one, several or all source languages catered for, and one, several or all target languages. Thus, a scheme which applies only to English-Italian is an example of one-one translation; a general programme to render any language into a Romance language would be an all-several translation.
Word decomposition for machine translation may depend to some extent on the translation field. In translation between related languages, it may be possible to utilize nonsemantic parallelisms due to common origin or borrowing. Thus the infix -iz- in English has a number of quite different meanings, e.g.

\[
\begin{align*}
sympath-iz-e & \quad \text{to manifest sympathy} \\
pulver-iz-e & \quad \text{to bring to powder} \\
mechan-iz-e & \quad \text{to do by machine}
\end{align*}
\]

Parallels to these words exist in French, Italian and Rumanian:

\[
\begin{align*}
sympath-iz-e & \quad \text{sympath-is-er} & \quad \text{simpat-izz-are} & \quad \text{simpat-iz-a} \\
pulver-iz-e & \quad \text{pulver-is-er} & \quad \text{polver-izz-are} & \quad \text{pulver-iz-a} \\
mechan-iz-e & \quad \text{mecan-is-er} & \quad \text{meccan-izz-are} & \quad \text{mecan-iz-a}
\end{align*}
\]

The parallel uses in these four languages are such that it is feasible to some extent to decompose as above and to translate English -iz- by Italian -izz- etc.; even though these infixes vary widely in meaning within each language. On the other hand, in translating any of the above into Japanese, this type of decomposition would be less appropriate as the divergent significance of the infix in the examples quoted requires a different rendering in each case.

The above illustration represents a relatively clear instance of a chunk which may be translatable comparatively but not universally. It can be maintained as axiomatic that no affix is universally, one-all translatable; that is, in linguistic terms, there can be no universal identification for translation purposes of any grammatical category. For this reason especially, there are obvious advantages in one form of decomposition for any one source language, whatever the target language or the translation field. If there is no possibility of the translation of a given segment as a chunk it is probably best handled at a later stage in the machine-translation programme. For example, there is a category of "plural* in part of the Chinese noun-system, but it cannot be arrived at by simple translation from the English plural, i.e. translation of the English -s chunk; this does not necessarily mean that in an English-Chinese translation programme the -s plural should not be handled as a chunk, but that it would not be represented directly in the Chinese, and the Chinese category of plural would have to be introduced by other means.
Conclusion

The object of the preceding note is to show that the range of possible word decompositions can be established by commutation tests as carried out by normal structural linguistic analysis. The actual decomposition appropriate to any particular situation can only be decided by additional criteria. Of these, semantic requirements, flexional system, chunk identification technique and translation field are the ones that have been considered; but other criteria are not excluded.
Lexical Search

In recent studies of Machine Translation a good deal of attention has been paid to translation, but very little to machine. There seems to be a feeling the machine will be more or less like existing computers. Such an assumption must be taken with caution.

There are two ways to carry out computations on a machine. One is to construct the required result by algorithms; for example, the quantity $\sin x$ can be calculated by a repetitive formula equivalent to a power series. The other is to rely heavily on table look-up. In present-day computers the latter method is almost extinct, and in Mechanical Translation we must strive as much as possible toward algorithmic methods.

Inasmuch as it seems impossible to construct the meaning of a word from its spelling or phonemes, except in the few cases of onomatopoeia, Mechanical Translation must always rely heavily on table look-up rather than algorithmic methods. Furthermore, a word not only has its dictionary meaning, but also the adhesion of a great deal of psychological and unexpressed descriptive material. A "sack" and a "coffin" are both "containers", but it would take a paragraph to modify the word "container" to make it mean either "sack" or "coffin". Thus in order for the machine to choose the most appropriate word of this category, we must store away additional material with each word to aid, to the degree of sophistication required, in the ultimate selection.

So although we expect to look up meanings associated with words, we do not wish to have an automatic dictionary, but to de-emphasize this approach, and try to introduce as many algorithmic techniques making use of context as possible. Thus we should consider "lexical search" rather than "dictionary look-up".

Magnitude of the Search Problem

The extent of the lexical search is determined not only by the theory, but by practical limitations. We now know that Mechanical Translation is possible, probably to as high a degree of refinement as we wish, so what are our objectives now?
Are we to pursue Mechanical Translation as an academic stunt? Do we expect to turn out useful translations, but presume they will always be crude and inelegant? Are we to provide a means to translate a specific field such as science or technology, or all types of literature?

The first is not enough, the last beyond our capabilities now. But the second is possible in 1958. In fact our objective should be to translate scientific or technical material in accurate readable form, with one proviso. Such an effort would be of great value to the nation, only if it can be done as fast as foreign presses print the material. The problem of lexical search is what is known in the computer field as a "real-time problem".

No hardware yet exists to carry out Mechanical Translation in real time. The current output of the leading nations is of the order of $3 \times 10^6$ pages per year, or $10^9$ words per year. In the next year or two we may expect text readers to be developed which will be able to read printed material at the rate of 1000 characters/sec. With $10^7$ sec in a working year and 6 characters/word, this amounts to $1.5 \times 10^9$ words/year, of the order of magnitude of the rate of publication. Thus we can expect the rate of input to the machine to be adequate.

Storage for Lexical Search

The corresponding rate at which the lexical search must be carried on is $10^9$ words/year or 100 words/sec. Thus the first requirement on this memory unit of the machine (which we shall call Store I) is that it must have 10 millisec random access time to every entry. It will take one-fifth of a second to look up all the words in an average sentence.

The size of the store will depend on the number of words in a language, and the amount of lexical material to be associated with each word in an entry. There are some $6 \times 10^4$ words in a dictionary. However, at the present state of translation theory we can hardly afford to neglect the clues offered by inflexional forms, so the total number of source words which must be in the store will be more like $10^6$. At present we average about 250 bits (6 bits to define a character) in an entry, and more sophisticated translators will require about $10^3$ bits. Thus the second requirement of the store is that its capacity must ultimately be about $10^9$ bits.
There is a third design parameter of the store which must be established to make the translating system efficient. Access to an entry has been established at 10 milliseconds; the size of the entry at $10^3$ bits. This material must be read out in a reasonable fraction, say 10%, of the access time. Thus the third requirement of the store is that its read-out rate must be $10^6$ bits/sec.

**Storage for Logical Processing**

The problem of lexical storage involves more than the mere storage and access to lexical material. A good translation also involves the interrelation of the lexical material found on the basis of syntax and semantics. The first disgorgement of the store is only raw material, on which a logical unit of the machine has to work. (Here "logic" means that mathematical or symbolic logic which can actually be done with a computer. Some "logical" operations are purely housekeeping details of the mechanical operations of the computer.) Rough estimates based on current theories of Mechanical Translation would indicate that some $10^4$ logical operations may be required per sentence to straighten out the disgorged material into a good translation. Even if only a fraction of these operations were requested for further look-up (as many theories demand), the restriction of producing output as fast as material is fed in makes it imperative that no further look-ups in the large store are permitted during logical processing.

This means that the disgorged material on the first look-up (from Store I) should be necessary and sufficient for analysis of the sentence (or paragraph). In other words, the output of the first look-up operation creates a "microglossary" sufficient for the analysis of the sentence. This selection from Store I should be dumped in a fast memory (called Store II) for logical processing. With 20 words per sentence on the average, $10^3$ bits output/word the requirement on capacity for the intermediate memory is of the order of $10^5$ bits (100 thousand-bit computer "words").

We have seen that the rate of flow, from source through input equipment and in table look-up in Store I, are all well matched at 0.2 sec per 20-word sentence. High-speed memories of $10^5$ bits capacity are currently available with a 10 microsecond random access. Hence $0.2 \div 10^{-5}$ or $2 \times 10^4$ logical operations (computer-type) may be made with the microglossary. Since some 20 computer-type housekeeping operations are normally required for one purely logical operation (e.g. a comparison of endings), about $10^3$ of the
latter are permitted per sentence. Of these perhaps $10^2$ may be further table look-ups (in the fast memory). This facility seems adequate for current Mechanical Translation theories.

**Nature of the Logical Processing**

This scheme of setting up a microglossary for each sentence imposes not only the above physical requirements on the intermediate memory, but also begins to define the logical elements necessary in the entries.

At this point in the machine it is actually unnecessary, and is in fact premature, to have any translation into the target language.

We may cover most of the theoretical approaches by defining the contents of the entry in Store I as clues. That is, given a sequence of words in the source material

$$S_1, S_2, \ldots, S_i, \ldots$$

the first operation is to look up in Store I lexical information concerning the words $S_i$ (or word sequences $S_{i}$, $S_{i+1}$, \ldots $S_{i+k}$). The output will be a sequence of terms

$$S_1A_1B_1, S_2A_2B_2, \ldots, S_iA_iB_i, \ldots$$

where $A_i$ refers to characters (possibly binary) giving syntactical information (such as the part of speech), and $B_i$ refers to characters giving semantic information, e.g. "this is a word from physics"; but not the translation.

The sequences $(S_iA_iB_i)$ form an expanded sentence, and form the microglossary in the intermediate fast memory of the logic portion of the computer. Here the sequences $A_i$, $B_i$ are examined and a new set of characters $C_i$ are constructed and assigned to each $S_i$. Note that the $i$'th $C_i$, assigned to $S_i$, is in fact a function of all the preceding and succeeding $A_i$'s and $B_i$'s (called the "local" or "minor" context). The determiners $(A_i's$ and $B_i's$) for the $C_i$'s may not only be in the sentence, but possibly (especially for pronouns) lie in previous sentences, or even of the title (field), called the "Major Context".

These logical operations will consist of two groups. The first will be a syntactical analysis of the sequence $A_1, A_2, \ldots A_i$ (without the $S_i$'s or $B_i$'s). This is like a schoolboy's diagramming of the sentence,
which finds the relations between words. According to the Cambridge Language Research Group this analysis can be made by algebraic lattice theory, which is highly algorithmic.

To give a very elementary example from French let all nouns have an \( A=a \), all verbs an \( A=\beta \) and the word \( S=\text{le} \) have an \( A=a+\beta \). Here the plus symbol is the logical "or" operation. There will be as many terms in \( A \) as there are multiple meanings for the \( S \). Then the syntactical analysis of "\( \text{le} \)" followed by a noun would involve the Boolean multiplication, which is easy to mechanize,

\[
(a+\beta)(a) = a.
\]

The result, \( a \), would constitute a character of the \( C \) for "\( \text{le} \)", so that the output \( SC \) for "\( \text{le} \)" would be \( \text{lea} \). The augmented word \( \text{lea} \) has a unique meaning "the". Note the actual meanings of the augmented words \( S,C_i \) are not yet at hand, and are to be found by a third operation in the machine, to be described below.

In the case of "\( \text{le} \)" followed by a verb, the multiplication is

\[
(a+\beta)(\beta) = \beta
\]

and the output \( SC \) for "\( \text{le} \)" would now be \( \text{le}\beta \). The augmented word "\( \text{le} \)" has the unique meaning "it". (The other meaning "him" would be assigned to "\( \text{lea} \)", derived from other \( A \)'s.)

A more complicated example would be the phrase

"... penetree d'abord de ..."

in which the logical operations on the \( A \)'s for the four words (d'abord not being treated here as an idiom) should show that "\( \text{de} \)" not "\( \text{d}' \)" is modified by the "penetree" (and is ultimately to be translated as "by" not "of").

According to the MIT group these operations will require another series of table look-ups. The storage involved probably does not require high capacity, but will require fast access, and be similar to Store II.

When the permissible connections between words has been established by purely syntactical analysis, by means of the \( A \)'s, a second
series of operation, involving the B’s, is carried out. Consider the following three elementary examples in French.

1) . . . le livre est à lui . . .
2) . . . il est pour travailler . . .
3) . . . pour . . .

In sentence 1) “est” has a specific meaning “belongs”, the clue for this selection being “à”, which has its normal meaning “to”. In sentence 2) “est” has its most probable meaning “is”, but “pour” is to mean “about to” here. In sentence 3) “pour” is to have its most probable meaning “for”. We shall not complicate matters by giving sentences where “à” is controlled by other words giving it meanings other than “to”, but remember this in the formulation. To handle the multiple meanings for the three words est, à and pour, whose clues are specific words elsewhere in the sentence, rather than purely syntactical, we assign to these words B’s which are logical sums of characters a, b, c . . ., one for each possible meaning. Thus for

\[ S = \text{est} \quad B = a + b + \ldots \]
\[ S = \text{à} \quad B = b + c + \ldots \]
\[ S = \text{pour} \quad B = a + d + \ldots \]

(Subscripts to the S’s, used previously, giving the position of the word are omitted.)

Then in the first sentence we would have

\[ (a + b + \ldots)(b + c + \ldots) = b \]

to get estb and àb.

For the second

\[ (a + b + \ldots)(a + d + \ldots) = a \]

to get esta and poura. For the third we would get only poura, and for sentences where “à” has other meanings àc etc. The corresponding augmented words have the following unique meanings

\[ \text{esta} = \text{is} \]
\[ \text{estb} = \text{belongs} \]
The output of the logical unit is then a sequence of

\[ S_{i1}C_1, S_{i2}C_2 \ldots S_{iC_i} \ldots \]

The point here is that we are no longer concerned with raw words \( S_i \) of the source language, but augmented words \( S_iC_i \), and these augmented words, if our method of construction of the \( C_i \)'s is adequate, have a unique meaning.

At this point in the machine we should have then solved the multiple-meaning problem with the aid of the syntactical and semantic context.

**Output Store**

We now come to the final stage of the machine, which again is a memory look-up operation. We enter with the individual augmented words \( S_iC_i \) and find a single target equivalent \( T_i \). (Note \( S_iC_i \) may stand for a string of words \( S_k \ldots S_m \), from which some \( S_kC_k \) have no target equivalent.)

The statement that the machine has a second look-up in a large store for each word does not violate our precept that time does not permit more than one look-up, because this operation is on another store, and can be done in the interval when the preparatory look-up for the next sentence is going on. (It is reasonable to suppose the intermediate Store II is flexible enough to be accepting \( S_iA_iB_i \) from the first memory for the following sentence while simultaneously supplying \( S_iC_i \) for look-up in the last memory for the sentence in hand.)

Nevertheless the speed of the last large memory (Store III) must be such as not to delay the overall flow of information through the system.

Since the logical operations have only made a one-to-one correspondence between the \( S_iA_iB_i \) and \( S_iC_i \), the number of look-ups for the sentence remains the same. Thus the requirement on the last store in regard to access time is the same as for Store I.
In simple Mechanical Translation theories, on the average there are 3 multiple meanings for each source language word, the number of entries in Store III will be three times that of Store I. Further the length of the address, \( S_iC_i \), will be about twice the length of the address \( S_i \) used in Store I. On the other hand the information sought is only a simple target equivalent, averaging 6 characters, or less than 50 bits. The length of an entry will thus be about 150 bits. The total capacity of Store III will be about one-third that of Store I. Nevertheless, in view of the rudimentary state of the theory, for the following reasons one should consider Store III as having essentially the same capacity as Store I.

**The Thesaurus**

It seems that a more advanced theory of Mechanical Translation, or more accurately, of mechanically understanding the written word, could be developed along these lines. The semantic information \( B_i \) associated with each input word \( S_i \) in the first lexical search, could be elaborated in great detail; so much so that the output of the logical unit could dispense with the symbols, \( S_i \), of the source words, and be merely a string of \( C_i \)'s,

\[ C_1C_2 \ldots C_i \ldots \]

This presupposes that the \( B_i \)'s, and the analysis of relationships by means of the \( A_i \)'s, are sufficiently detailed that the sequence of \( C_i \)'s has retained all the content and relationships the whole idea, in some coded form related to symbolic logic. In this event Store III would be a kind of thesaurus, for which the input is a sequence of symbols, \( C_i \), associating in a Boolean function a large number of ideas and relations which must be stated in the output, as determined from the initial contextual analysis; and for it we wish the machine to choose the most appropriate word. This word is not necessarily the one we would find in a dictionary, nor is it a synonym, but a particularly cogent word for the idea in the particular context. In passing we remark that the \( C_i \)'s themselves constitute a language analysis to symbolic logic or the proposed "ruly English", but are unsatisfactory output in themselves as they do not convey the richness and desirable ambiguity (after Empson) which makes ordinary languages sophisticated means of communication. In short the thesaurus reattaches to the primitive \( C_i \)'s the psychological content and background description that makes languages.
In order to point out that the effort spent on both the theory and hardware for Mechanical Translation is of value not only in itself, but for the larger problem of information retrieval, we may point out that in the above system the output $T_i$ from Store III may indeed be the same language as the input $S$, so that the machine translates English into better English. Or $T_i$ may be the more primitive English used by librarians and indexers, so that the system could be used for classifying, indexing and abstracting.

**Incomplete Matching**

There is an important point in imagining the construction of local context and introduction of the thesaurus in contrast with a dictionary.

Inasmuch as the $C_i$'s are determined from the local context, which, if the material is worth translating, should have some novel combinations of ideas, we cannot expect all possible $C_i$'s to be listed with an $S$ in Store III. That is we do not necessarily have unique addresses to the entries of Store III. Hence we must arrange to locate not necessarily a specific $C_i$, but a best match. There are various ways of defining "best"; one is, recognizing $C_i$ to be essentially a Boolean function, to find a $C_i$ which dominates $C_i$ in the sense of lattice theory, i.e.

$$C'_i \geq C_i \geq T_i$$

A system such as this will have to be introduced even in simpler Mechanical Translation schemes, to handle typographical errors and grammatical errors on the part of the original author.

**Summary**

The Mechanical Translation system consists then of three parts, first a high capacity millisecond-access store of lexical information concerning the source language; second, a low-capacity microsecond-access store for logical processing of lexical information into augmented words for selection; and third, another high-capacity millisecond-access store of thesaural information concerning the target language. The whole system must operate in real time.
From items that precede a noun in a German noun phrase (NP) we obtain a great amount of information. Much of the information is redundant, much of it seems irrelevant within the area of English morphology. But what is superfluous morphologically, may become significant on the syntactic level.

Nouns can be preceded by adjectives. These may be primary adjectives with so-called strong endings, some of which are ambiguous, others specific. Or they may be secondary adjectives featuring the so-called weak endings, which even more frequently are ambiguous. In the case of adjectives with secondary endings there must be another item preceding which belongs to a group that we may call the der-words, or limiting words. Among the der-words we list all those that cause secondary endings in a following adjective, such as der, die, das, dieser, jener, eine, keine, mancher, solcher, welcher and their inflected cases. The der-words, however, may also be the only items preceding a noun. Among primary adjectives we list all those that take endings themselves, but do not cause secondary endings in a following adjective, i.e. the usual comparable adjectives as well as mehrere, viele, wenige, einige, andere, etliche verschiedene.

Moreover, nouns can be preceded by items with zero-endings, like ein, kein, welch, solch, manch, also the possessives and numerals. This group may occur alone before a noun, or preceding another adjective without causing secondary endings.

From the noun itself we obtain information about number, gender, and case. Number is of immediate relevance for translation into English. If we cannot obtain information on number from the form of the German noun, i.e. if the plural morpheme is zero (Fenster, Schüler), the identification becomes more complicated and search for clues outside the NP may become necessary. Gender is not rendered in English, but case is syntactically relevant, and it may be that case becomes identifiable only by reference to gender. For example, the NP der Fenster is identified as a gen.pl. by reference to the fact that Fenster is a neuter noun.
The information, then, offered by the German NP for translation into English concerns number and case. We obtain such data by a combination of clues which in isolation are quite often unspecific. Even with all the NP data combined, the identification has often to be sought elsewhere, i.e. in the verb, the word order, etc.

The greatest amount of information is yielded by the noun itself. The noun must be coded in the glossary with all the information that it supplies. Some of this information is specific, some is not. *Mannes* is specific as to gender, number, and case. *Frau* is specific in gender and number, but not in case. *Schüler* is specific in gender only.

As regards other items in the NP, there is only one specific primary adjective ending, namely *-em*, which denotes sg.dat. The fact that it may be masc. or neuter does not impair its specificity for translation. In the category of limiting words, specific information is carried by the ending *-em* and by the word *des* (not by the ending *-es*, since in *dieses, jenes*, etc. there is ambiguity). Through secondary adjectives alone, no unambiguous identification is possible.

Since the nouns carry at least some specific information regularly, it is felt that the search for complete data should begin there and proceed backwards through the NP.

The following is an attempt to outline the steps that have to be taken in machine translation to identify the NP as to case, and where the plural morpheme is zero, as to number. It will become clear in the process where the NP does not yield sufficient data for complete identification.

A preceding preposition may have to be referred to for identification of case. It is not included here in the NP because it is often linked with a verb or adjective, in which case it may call for a lexical correspondence in English that does not contain a preposition (e.g. *sich erinnern an*—to remember).

Upon encountering a *masc. noun in the sg.*, look for preceding adjective.
If yes, look for *-e* ending
  if yes, NP is nom. (Include preceding der-word) der kleine Herr

—90—
if no, look for -en ending
if yes, look for preceding des-word
  if yes, NP is gen. des kleinen Herrn
  if no, look for dem-word
    if yes, NP is dat. dem kleinen Herrn
  if no, look for den-word
    if yes, NP is acc. den kleinen Herrn
  if no, look for gen. suffix in the noun
    if yes, NP is gen. guten Mutes
    if no, NP is acc. guten Mut
if no, look for -er ending
if yes, NP is nom. (Include preceding zero-word, if any) (ein) kleiner Herr
if no, look for -em ending
if yes, NP is dat. grossem Fleiss

If no, look for preceding der-word
if yes, NP is nom. der Herr
if no, look for des-word
if yes, NP is gen. des Herrn
if no, look for dem-word
if yes, NP is dat. dem Herrn
if no, look for den-word
if yes, NP is acc. den Herrn
if no, look for preceding zero-word

If yes, NP is nom. ein, kein, mein, unser Herr
If no, look for preceding preposition
If yes, NP is non-nominative case mit Mut, durch Fleiss
If no, clue is outside NP Mann, Mut, Fleiss (nom. or acc.)

Upon encountering a masc. noun in the pl., look for preceding adj.
If yes, look for -e ending
  if yes, look for preceding acc.prep.
    if yes, NP is acc. (Include zero-word, if any) für (zwei) alte Herren
    if no, clue is outside NP (zwei) alte Herren (n. or acc.)
  if no, look for -er ending
  if yes, NP is gen. (Incl. zero-word, if any)
    (zwei) alter Herren
if no, look for *-en ending
if yes, look for preceding limiting word
  if yes, look for die-word
    if yes, look for acc. prep.
      if yes, NP is acc.
      if no, clue is outside NP
        für die alten Herren
die alten Herren (nom. or acc.)
  if no, look for der-word
    if yes, NP is gen.
    if no, look for den-word
      if yes, NP is dat.
      if no, NP is dat. (Incl. zero-word, if any)
        (zwei) alten Herren

If no, look for preceding limiting word
  if yes, look for die-word
    if yes, look for acc. prep.
      if yes, NP is acc.
      if no, clue is outside NP
        für die Herren
die Herren (n. or acc.)
  if no, look for der-word
    if yes, NP is gen.
    if no, look for den-word
      if yes, NP is dat.
      if no, look for preceding preposition
        (Incl. zero-word, if any)
          für, wegen, mit (zwei) Herren
        (zwei) Herren (nom. or acc.)

Upon encountering a masc. noun with no indication of number
(i.e. with zero or ambiguous plural morpheme), look for preceding adj.
If yes, look for *-e ending
  if yes, look for preceding der-word
    if yes, NP is nom. sg.
    if no, look for acc. prep.
      if yes, NP is acc. pl.
      if no, NP is pl., but clue
        for case is outside NP
          der kleine Schüler
        für kleine Schüler
        kleine Schüler (nom. or acc.)
if no, look for -en ending
if yes, look for limiting word
    if yes, look for des-word
        if yes, NP is gen.sg. des kleinen Schülers
        if no, look for dem-word
        if yes, NP is dat.sg. dem kleinen Schüler
        if no, look for die-word
        if yes, look for acc.prep.
            if yes, NP is acc.pl. für die kleinen Schüler
            if no, NP is pl., but clue
                for case is outside NP die kleinen Schüler (nom. or acc.)
        if no, look for der-word
        if yes, NP is gen.pl. der kleinen Schüler
        if no, look for den-word
            if yes, look for dat.pl. suffix in the noun
                if yes, NP is dat.pl. den kleinen Schülern
                if no, look for preceding acc. prep.
                if yes, NP is acc.sg. für den kleinen Wagen
                if no, look for dat.prep.
                if yes, NP is dat.pl. bei den kleinen Wagen
            if no, clue is outside NP den kleinen Wagen (acc.sg. or dat.pl.)
        if no limiting word, look for gen. suffix in noun
            if yes, NP is gen.sg. guten Willens
            if no, look for dat.pl. suffix in noun
            if yes, NP is dat.pl. kleinen Schülern
            if no, look for preceding acc.prep.
            if yes, NP is acc.sg. ohne kleinen Wagen
            if no, look for dat.prep.
            if yes, NP is dat.pl. mit kleinen Wagen
            if no, clue is outside NP kleinen Wagen (acc.sg. or dat.pl.)
        if no -en ending, look for -er ending
            if yes, look for preceding zero-word
                if yes, NP is nom.sg. ein, mein, kleiner Fehler
                if no, look for gen.prep.
                if yes, NP is gen.pl. wegen kleiner Fehler
                if no, clue is outside NP kleiner Fehler (nom.sg. or gen.pl.)
if no, look for -em ending
if yes, NP is dat.sg.

kleinem Fehler

If no preceding adj., look for limiting word
If yes, look for der-word
  if yes, look for preposition
    if yes, NP is gen.pl.
      wegen der Fehler
    if no, clue is outside NP
      der Fehler (n.sg. or g.pl.)
  if no, look for des-word
    if yes, NP is gen.sg.
      des Fehlers, des
      Studenten
  if no, look for dem-word
    if yes, NP is dat.sg.
      dem Fehler
  if no, look for den-word
    if yes, look for dat.pl. suffix in noun
      if yes, NP is dat.pl.
        den Fehlern
      if no, look for prec.dat.prep.
      if yes, NP is dat.pl.
        mit den Wagen
      if no, look for acc.prep.
      if yes, NP is acc.sg.
        ohne den Wagen
      if no, clue is outside NP
        den Wagen (acc.sg. or 
        dat.pl.)

if no, look for die-word
if yes, look for preposition
  if yes, NP is acc.pl.
    ohne die Fehler
  if no, clue is outside NP
    die Fehler (nom. or 
    acc.pl.)

If no limiting word, look for word with zero-suffix
If yes, look for numeral
  if yes, look for preposition
    if yes, NP is non-nominative pl.
      für, wegen, mit zwei
      Wagen
    if no, NP is pl., but clue for 
    case is outside NP
      zwei Wagen (nom. or acc.)
  if no, look for ein-word
    if yes, NP is nom.sg.
      ein, mein, unser Wagen 
      Wagen, Schüler, Fehler
    (nom.acc.sg.pl.)
If no zero-word, clue is outside NP

 Upon encountering a fem. noun in the sg., look for preceding adj.
If yes, look for -e ending
if yes, include preceding *die*-word, if any, and
look for acc. preposition
  if yes, NP is acc.
  if no, clue is outside NP

if no, look for -*en* ending
if yes, include preceding *der*-word and
look for gen. prep.
  if yes, NP is gen.
  if no, look for dat.prep.
  if yes, NP is dat.
  if no, clue is outside NP

if no, look for -*er* ending
if yes, look for gen.prep.
  if yes, NP is gen.
  if no, look for dat.prep.
  if yes, NP is dat.
  if no, clue is outside NP

If no adj., look for preceding limiting word
If yes, look for *die*-word
  if yes, look for acc.prep.
    if yes, NP is acc.
    if no, clue is outside NP
  if no, look for gen.prep.
    if yes, NP is gen.
    if no, look for dat.prep.
    if yes, NP is dat.
    if no, clue is outside NP

if no, look for *der*-word
If yes, look for gen.prep.
  if yes, NP is gen.
  if no, look for dat.prep.
  if yes, NP is dat.
  if no, clue is outside NP

If no limiting word, look for prep.
If yes, NP is in non-nominative case
  mit Not, für Mutter
  Not, Mutter, Frau
  (nom., gen., dat., acc.)
If no, clue is outside NP
Upon encountering a *fem. noun in the pl.*, look for preceding adj.
If yes, look for -e ending
  if yes, (include preceding numeral, if any, and look for acc. preposition)
    if yes, NP is acc.                     für (drei) junge Frauen
    if no, clue is outside NP               (drei) junge Frauen
  if no, look for -en ending
  if yes, look for preceding limiting word
    if yes, look for die-word
      if yes, look for acc. prep.
        if yes, NP is acc.                  für die (diese, manche, etc.) jungen Frauen
        if no, clue is outside NP           die jungen Frauen
      if no, look for der-word
        if yes, NP is gen.                 der (dieser, etc.) jungen Frauen

    if no, look for den-word
      if yes, NP is dat.                  den (diesen, etc.) jungen Frauen

  if no limiting word, NP is dat.
    (incl. preceding numeral, if any)     (drei) jungen Frauen
    if no, look for -er ending
    if yes, NP is gen.                    (drei) junger Frauen

If no preceding adj., look for limiting word
If yes, look for die-word
  if yes, look for acc. prep.
    if yes, NP is acc.                  für die (diese, etc.) Frauen
    if no, clue is outside NP            die (diese, etc.) Frauen

    if no, look for der-word
      if yes, NP is gen.                 der (dieser, etc.) Frauen
    if no, look for den-word
      if yes, NP is dat.                  den (diesen, etc.) Frauen

If no limiting word, look for prep.
If yes, NP is in non-nominative case
  (incl. numeral, if any)             mit, für (drei) Frauen
If no, look for dat. pl. suffix
If yes, NP is dat.
  (incl. numeral, if any)  (drei) Müttern, Töchtern
If no, clue is outside NP (drei) Frauen (nom.,
  dat., acc.)

Upon encountering a neuter noun in the sg., look for preceding adj.
If yes, look for -es ending
  if yes, (incl. preceding zero-word if any, and)
    look for acc. preposition
    if yes, NP is acc.
    ohne (ein, mein, solch, etc.) kleines Kind
    if no, clue is outside NP
    (ein, etc.) kleines Kind
  if no, look for -e ending
    if yes, (incl. preceding das-word and) look
      for acc. prep.
      if yes, NP is acc.
      für das (welches, dieses) kleine Kind
      if no, clue is outside NP
      das (welches, etc.) kleine Kind
      (nom. or acc.)
    if no, look for -en ending
      if yes, look for preceding limiting word
      if yes, look for des-word
      if yes, NP is gen.
      des (dieses) kleinen Kindes
      if no, look for dem-word
      if yes, NP is dat.
      dem (diesem) kleinen Kind (e)
      if no limiting word, NP is gen.
      letzten Endes
    if no, look for -em ending
    if yes, NP is dat.
    gutem Geld

If no adj., look for limiting word
If yes, look for das-word
  if yes, look for acc. prep.
  if yes, NP is acc.
  für das (welches, dieses) Kind
  if no, look for gen. suffix in the noun
  if yes, NP is gen.
  welches, dieses Kindes
  if no, clue is outside NP
  das (welches, dieses) Kind (nom. or acc.)
if no, look for *des*-word
if yes, NP is gen.

if no, look for *dem*-word
if yes, NP is dat.

If no limiting word, look for zero-word
If yes, look for acc. preposition
  if yes, NP is acc.
  if no, clue is outside NP

If no zero-word, look for prep.
If yes, NP is in non-nominative case
If no, clue is outside NP

Upon encountering a *neuter noun in the pl.*, look for preceding adj.
If yes, look for -e ending
  if yes, (incl. preceding numeral, if any, and)
    look for acc. preposition
      if yes, NP is acc.
      if no, clue is outside NP
  if no, look for -er ending
  if yes, NP is gen. (incl. numeral, if any)
  if no, look for -en ending
  if yes, look for preceding limiting word
    if yes, look for *die*-word
      if yes, look for acc. prep.
        if yes, NP is acc.
        if no, clue is outside NP
    if no, look for *der*-word
      if yes, NP is gen.
      if no, look for *den*-word
        if yes, NP is dat.
        if no limiting word, NP is dat.
          (include numeral, if any)
If no adjective, look for limiting word
If yes, look for die-word
   if yes, look for acc. preposition
      if yes, NP is acc.
         von
         für (manche) Kinder
      if no, clue is outside NP
         die (diese, manche, etc.) Kinder (nom. or acc.)
   if no, look for der-word
      if yes, NP is gen.
         von
         der (dieser, etc.) Kinder
      if no, look for den-word
      if yes, NP is dat.
         von
         den (diesen, etc.) Augen
If no limiting word, look for preposition
If yes, NP is in non-nominative case
   (incl. numeral, if any)
      von
      für, mit, wegen (zwei) Augen
If no preposition, look for dat. pl. suffix in the noun
If yes, NP is dat. (incl. numeral, if any)
   (zwei) Kindern
If no, clue is outside NP
   (zwei) Augen (nom., dat., acc.)

Upon encountering a neuter noun with no indication of number (i.e. with zero or ambiguous plural morpheme), look for preceding adj.
If yes, look for -e ending
   if yes, look for preceding das-word
      if yes, NP is sg.
         von
         durch das (dieses, welches) kleine Fenster
      if no, clue for case is outside NP
         das (dieses, etc.) kleine Fenster (nom. or acc.)
   if no das-word, NP is pl.
      look for acc. prep.
      if yes, NP is acc. pl.
         von
         durch kleine Fenster
      if no, clue for case is outside NP
         kleine Fenster (nom. or acc.)
   if no, look for -er ending
   if yes, NP is gen. pl. (include numeral, if any)
   if no, look for -es ending
   if yes, NP is sg. (incl. preceding zero-word, if any, and)
      look for acc. preposition
      if yes, NP is acc.
         von
         ohne (ein, solch) kleines Fenster

—99—
if no, clue for case is outside NP

if no, look for -em ending
if yes, NP is sg. dat.
if no, look for -en ending
if yes, look for limiting word
if yes, look for des-word
if yes, NP is sg. gen.
if no, look for dem-word
if yes, NP is sg. dat.
if no, look for die-word
if yes, NP is pl.
look for acc. prep.
if yes, NP is acc.
if no, clue for case is outside NP

if no, look for der-word
if yes, NP is pl. gen.
if no, look for den-word
if yes, NP is pl. dat.
if no limiting word, look for gen. sg.
ending in noun
if yes, NP is gen. sg.
if no, NP is dat. pl.

If no adj., look for preceding limiting word
If yes, look for das-word
if yes, NP is sg.
look for acc. prep.
if yes, NP is acc.
if no, clue for case is outside NP

if no, look for des-word
if yes, NP is sg. gen
if no, look for dem-word
if yes, NP is sg. dat.
if no, look for der-word

(ein, solch, welch) kleines Fenster (nom. or acc.)
kleinem Mädchen

kleinen Mädchens
dem kleinen Mädchen

ohne die kleinen Mädchen
die (diese, etc.) kleinen Mädchen (nom. or acc.)
der kleinen Mädchen
den kleinen Mädchen

kleinen Mädchens
kleinen Mädchen

für das (dieses, etc.) Mädchen
das (dieses, etc.) Mädchen (nom. or acc.)
des (dieses, eines) Mädchen
dem (diesem, einem) Mädchen
if yes, NP is pl. gen.  
der (dieser, mancher)  
Mädchen

if no, look for *den-word*
if yes, NP is pl. dat.  
den (diesen, manchen)  
Mädchen

if no, look for *die-word*
if yes, NP is pl.  
look for acc. prep.  
if yes, NP is acc.  
für die Mädchen  
die (diese, etc.)  
Mädchen (nom. or acc.)

if no, clue is outside NP

If no limiting word, look for zero-word
If yes, look for numeral
if yes, NP is pl.  
look for preposition  
if yes, NP is in non-nominative case  
für, wegen, mit zwei  
Mädchen  
zwei Mädchen (nom. or acc.)

if no, clue for case is outside NP

if no, look for *ein-word*
if yes, NP is sg.  
look for acc. prep.  
if yes, NP is acc.  
für ein (mein, kein)  
Mädchen  
ein (mein, kein) Mädchen (nom. or acc.)

if no, clue for case is outside NP

If no zero-word, look for preceding preposition
If yes, NP is non-nominative pl.  
für, wegen, mit Mädchen

If no, clue for number and case is outside NP

Mädchen (nom., dat., acc. sg.; nom., gen., dat., acc. pl.)
(Mr. Jacob Rabinow, of the Rabinow Engineering Company, at this 
time gave an oral presentation on a new large-capacity optic data 
storage device. He did not submit a manuscript and his presentation 
is therefore omitted from this Report.—Ed.)
Retrieval of information from technical literature as practiced in the Patent Office is much more concerned with the specific *interrelations* of things than with the specificity of the things themselves.

In this literature, any two documents which refer to the same or to closely similar phenomena will ordinarily express these phenomena in two widely different sets of verbal expressions. A patent examiner who may be searching for a similar phenomenon in formulating a retrieval request will probably utilize a still different, but accurate, verbal expression. Hence, in order to retrieve the documents, both the request and the documents must be transliterated into a common linguistic denominator.

We have been coining terms for an "unambiguous" or *Ruly English*, (as distinguished from "unruly",) after the terminology of Professor Dodd¹.

Our method of setting forth a disclosure, including the relationships found therein, has been termed *itemization*. In this system, each thing is listed as a separate item. Additional *descriptors* which state other aspects of each thing are listed in the same item. Relationship between items is shown by means of *interrelational concepts*, coined in mirror-image form, one of the images being *distributed* (listed) with each of the things being related. Actions are also distributed with each of the things involved. Distributed terms are identified by identical numbers, called *interfixes*.

We are coining Ruly English *roots* to be used as descriptors. These roots can be *modulated* to indicate such notions as process, thing processed, thing processing, condition and number. Ruly roots of action words are coined to encompass changes in all the things taking part in the action. Ruly roots of *qualifiers* likewise are coined to encompass complementary terms. A simple solution to the *quantification* of the qualifiers is proposed.
The Linguistic Approach

One of the approaches taken by our Office in its research on storage and retrieval of information in technical literature is linguistic in nature. This linguistic investigation was undertaken for two reasons. Most scientific and technical literature is already verbally expressed, and it is clear that other references in the form of drawings, tables of figures, photographs, models, working machinery, etc., could be transliterated into their verbal counterparts. It is also quite clear that no matter what final scheme or method of information retrieval is ultimately developed, an "unambiguous" metalanguage will be found to be of value.

Since conventional linguistic analysis did not prove to be helpful in the solution of our technical problems, we found it necessary to undertake steps to begin the creation of a metalanguage in which each unit will have one, and only one meaning, and in which each meaning may be expressed with one, and only one unit. This metalanguage and the method proposed for its use undoubtedly will appear unconventional to trained linguists; however experimentation to date points to the eventual solution of some Patent Office problems in information retrieval.

We have named this metalanguage Ruly English after the terminology of Professor Dodd, who has pointed out that English is quite "unruly". Such a metalanguage will have at least two mutually exclusive uses in information retrieval.

Mechanized Information Retrieval

In our work at the Patent Office we are much more interested in the interrelation of things than we are in the specific details of the individual things themselves. Any two technical documents which refer to the same or to closely similar phenomena will ordinarily express these phenomena in two widely different sets of verbal expressions, both of which accurately convey the same information to the human mind. In using a mechanized searching system, a patent examiner might formulate his request for the retrieval of this same information by utilizing a still different, but accurate verbal expression. If his search is to yield both documents, the request and the documents must each be transliterated into a metalinguistic common denominator. This metalanguage for the first time will make possible the conversion of the many complex and interrelated notions in a particular document into single unique forms. Such a scheme as this is fundamental in the development of any successful retrieval system.
In an automatic assembly program system for accepting search request data, these same unique terms may be used to encode such data in a form suitable for machine instruction or command. The utilization of such systems is being widely developed by the machine industry at this moment.

Once the terminology and definition of the language units have been created, the encoding of search questions appears to be merely a problem of program development.

It also appears to be quite feasible to program a data processing machine to communicate with its operator during a search, in order to inform him of the trends of the results, so that he may vary the questioning program, or may substitute a different form of program or question before completing the search. This same metalanguage, thus, would serve as the means by which such intercommunication could be effected.

Basic Elements Of Ruly English

The basic elements of Ruly English are itemization, distribution and interfixing. Itemization consists of assembling all the descriptors of a single thing, each of these descriptors describing the thing from a different aspect, and grouping them as one item in a numbered list. Such numbers are used only for identification. Distribution is applicable to notions of interaction or interrelation between two or more items. Such notions are expressed by means of two or more cognate descriptors, one of which is placed with each related item. Thus distributed, the notion expresses the relation between the items. Interfixing is a method of tying together the cognate descriptors of distributed notions by affixing to each descriptor the same arbitrary number.

The Interrelational Concept

The notions of interrelations between items which are conveyed in English by prepositions were found to be quite troublesome. We therefore went to the list of Basic English words and took twenty-five which we recognized as prepositions. We collected all the phrases or sentences that we could find which utilized these prepositions, and we attempted to separate them into their various meanings. Fig. 1 is a list of the different meanings of from.
1. "differ" from - We distinguish day \textit{from} night.
2. "means" from - The ball hangs \textit{from} a string.
3. "cause" from - We get hives \textit{from} berries.
4. "while" from - He throws \textit{from} a standing position.
5. "start" from - She makes a cake \textit{from} flour.
6. "reject" from - We shall appeal \textit{from} a decision.
7. "exclude" from - We saved him \textit{from} injury.
8. "away" from - They live far \textit{from} a city.
9. "whence" from - The train came \textit{from} New York; he took a penny \textit{from} his pocket; the words are separated \textit{from} context.

also

10. "whence" from . . to - He was measured \textit{from} head to foot.

"FROM"S

Next, we collated these phrases and sentences and collected those containing prepositions having equivalent meanings.

As shown in Fig. 2, we found that one meaning of each of \textit{from},

<table>
<thead>
<tr>
<th>Preposition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>as \textit{from} in</td>
<td>The train came \textit{from} New York; he took a penny \textit{from} his pocket; the words are separated \textit{from} context.</td>
</tr>
<tr>
<td>as \textit{on} in</td>
<td>He drew a check \textit{on} the bank account; fighting the attack made an inroad \textit{on} the supplies.</td>
</tr>
<tr>
<td>as \textit{off} in</td>
<td>He ate \textit{off} the plate; he cut the end \textit{off} the stick.</td>
</tr>
<tr>
<td>as \textit{of} in</td>
<td>He came \textit{of} a noble family; wines \textit{of} France are well known; art is bought \textit{of} a dealer.</td>
</tr>
<tr>
<td>as \textit{from whence} in</td>
<td>Back to the dust \textit{from whence} he came.</td>
</tr>
</tbody>
</table>

INTERRELATIONAL CONCEPT: WHENCEFROM(FROMWHENCE)

Fig. 2.
on, off, of, and the complex from whence all conveyed the same notion, which we arbitrarily named WHENCEFROM. Since this interrelational concept has a direction or polarity, in order to point out the direction of action, we created a mirror image of this concept, termed FROMWHENCE. For example, the train (of example 1, Fig. 2) is FROMWHENCE, and New York is WHENCEFROM. Not all interrelational concepts have polarity, e.g., the mirror image of AMONG is identical with its object term.

Itemization

Let us take the simple sentence, The stale water is emptied from the china pitcher, and note one way in which its meaning may be unambiguously presented. Fig. 3 shows the itemization of this sentence.

<table>
<thead>
<tr>
<th>Item #</th>
<th>Unruly root</th>
<th>Ruly interrelational concept</th>
<th>Interfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>water</td>
<td>FROMWHENCE</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>stale</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>empty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>pitcher</td>
<td>WHENCEFROM</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>china</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>empty</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

THE STALE WATER IS EMPTIED FROM THE CHINA PITCHER

Fig. 3

Item 11 is the water and the additional descriptor stale is included. Item 12 is the pitcher, and the additional descriptor china is included. The action empty has been distributed between both items, and the interrelated concept WHENCEFROM(FROMWHENCE) shows the direction of the emptying. The interfix 15 ties the two distributed cognate notions together.

This same idea could have been stated The china pitcher is emptied of its stale water. Using this same technique, we derive the itemization of Fig. 4 which is identical with Fig. 3, except for those factors which do not influence meaning, i.e., the order of the items, their numbers and the interfix number.
We then decided that if we could reduce the elemental Ruly English terms to Roots⁴, we could modify the meaning by adding a modulant. Fig. 5 gives a partial list of modulants. For example, if the Root for

\[
\begin{align*}
=\text{NT} & \quad \text{process} \\
=\text{W} & \quad \text{work} \\
=\text{M} & \quad \text{made from or out of} \\
=\text{E} & \quad \text{condition} \\
\end{align*}
\]

\begin{itemize}
\item \text{Numerical} \\
\item \(=X\) for or more \\
\item \(=Y\) for exactly \\
\item \(=Z\) for or less \\
\item \(=B\) for as an ordinal
\end{itemize}

MODULANTS

Fig. 5

the unruly notion empty is DISPEN (from the verb dispense) the action or process of emptying is DISPEN=NT, the condition of being emptied is DISPEN=E, etc. Numbers may also be modulated, e.g., 3=X for three or more; 3=Y for exactly three; and 3=B for third.
In choosing terms for Roots, many of the common names of things must be avoided. Such names often suggest a mere accidental use to which the thing is put, or some property incidental to its use. A tray may be inverted and become a cover. A plastic water glass or a piece of melamine dinner china are contradictions in themselves. From common usage we have chosen only such terms as ring or sphere which connote structural shape.

Modulants

As we create terms for our Ruly vocabulary, we must define them in part in unruly English, and, if possible, give examples to help interpret each term. For instance, CONFORM (from conformation) may be defined as a geometric shape or figure in zero, one, two or three dimensions; e.g., a point, a line, a surface or a volume. As more terms are defined, we can go back over our unruly definitions and substitute Ruly terms. Fig. 6 illustrates the Ruly definition and its unruly transliteration. The Root chosen is POLAR and it is defined in terms of its condition modulant, POLAR=E.

<table>
<thead>
<tr>
<th>Ruly definition</th>
<th>Unruly transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLAR=E is __________ The condition of polarity is given by a contrasting_________ given by a contrasting characteristic of __________ characteristic of 1=Y ELEMENT or PORTION____ one or more elements or portions of a CONFORM__________ of a conformation distinguishing it from________ distinguishing it from other ELEMENTs or PORTIONs____ other elements or portions of the same CONFORM________ of the same conformation e.g., the front of________ e.g., the front of a piece of furniture________ a piece of furniture</td>
<td></td>
</tr>
</tbody>
</table>

POLAR (from polarity)

Fig. 6

Dual Aspect Roots

Action words often must be radically changed. For example, consider the situation in which one heats a room by cooling water - the conventional hot water heating system. This process is dual-aspect
in that the temperature of the air in the room is raised while that of the water is cooled. For this situation we coin the Root HEATCOOL, and as a ground rule we apply all interrelations to the first part of the compound Root, i.e., HEAT-.

This situation has been itemized in Fig. 7, in which we note that

<table>
<thead>
<tr>
<th>Item #</th>
<th>Root and modulant</th>
<th>Interrelational concept and interfix</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>room HEATCOOL=NT</td>
<td>MORE-31</td>
</tr>
<tr>
<td>16</td>
<td>water HEAT=E</td>
<td>LESS-31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HEATING A ROOM BY COOLING HOT WATER</td>
</tr>
</tbody>
</table>

the room increases in temperature because the process HEATCOOL=NT has the cognate image MORE of the interrelated concept MORE-(LESS). Conversely the HEAT=E (heated) water is cooled because the cognate image LESS applies to the HEATCOOL=NT.

**Qualification of Quantifiers**

Qualifying terms used in technical documents are ambiguous and troublesome. Both the balance spring of a watch and an enormous bridge girder may be designated as flexible. Actually analysis will show that all things are located somewhere on a scale between rigid and flexible, and none are found at either end of the scale. Hence we coin a *dual qualifier* such as RESILRIG to describe the condition, and we quantify it by one of two terms, slightly (SLI') or substantially (SUB*). Again we use the ground rule and apply the quantifier to the first part of the compound Root, RESIL-. Hence the bridge girder is SLI'RESilRIG and the balance spring is SUB'RESilRIG. If quantification can be measured along some scale, e.g., the Brinell standard of hardness, some predetermination will be made as to what portions are slightly hard or substantially hard, and it may be necessary to use both descriptors for values in the middle of the scale.
The one phase of research and development work expressed in this brief summary should provide insight into the linguistic problem which confronts the Patent Office. It follows that current research in the methodology of machine translation is, by extension, a contribution to the furtherance of our project. In proposing to adapt some of the techniques of machine translation to our specific uses, we become allies of your cause. Can we reciprocate in some measure with efforts that are useful in your research?

REFERENCES

2. In such metalanguages as Interglossa, on the other hand, several units may convey the same meaning.
3. All Ruly English terms will be written in upper case letters.
4. Ruly English roots will be denoted Roots, i.e., with an upper case R.
OSWALD (to KING): I'd like to start the ball rolling by asking Mr. King what kind of language program he thinks it would take to make this computer work.

KING: You mean, how do we get the machine loaded? One of the points I wanted to make is that you have to have some idea of the kind of equipment you're going to have to put this on when you prepare your information. I think also that once some theoretical basis has been established much of this can be done mechanically, too. I think we ought to give some thought to using the machines themselves to manipulate the lexicon of the source material.

DOSTERT: In your paper, Mr. Oswald, you seem to deplore the fact that your advocacy in 1952-53 about what you now call idio-glossaries had been overlooked. It has not really, at least not by us here, since we have decided to start with one of the fields of science, namely, chemistry—and in that field in organic chemistry. We are indeed following the tracks you were tracing several years ago, accumulating a glossary of what we call "general language", that is, the items that will recur in any given text, regardless of the technical field area in which we are operating, and then the special glossary, peculiar either in form or in usage to that particular field. We have in mind, if our research continues, to investigate texts in the field of physics, let us say, or mathematics and medicine, so that in due course we shall accumulate a series of specialized glossaries or dictionaries for several technical fields.

RABINOW: I would like to comment on this. It is very easy to take a section of one of these plates and make one of your idio-glossaries just by having a particular area,—as you see, the capacity is very large. Also, I should have said that you don't need a binary input. With each plate with twenty-six positions you can put in letters directly and come out with letters in the alphabet. The mechanics get more complicated, if there are 26, and it's not as efficient area-wise as binary.

OSWALD: May I ask one question about your system, Mr. Rabinow? What prevents friction in this mechanical operation?

RABINOW: The plates don't touch. They are about 1/16" apart.

-112-
OSWALD: No friction on the wires hanging down?

RABINOW: They just swing. The wire is a spring and—as a matter of fact if you use the right materials and don't stress them too much, their life is infinite. The plates are very cheap, by the way. Etched completely in small quantities these would cost $18.00 apiece; in large quantities a dollar or two. So you can throw away the whole set of plates if you like. It means, of course, that you have to repunch the whole set of holes, but this can be done by machinery too.

GARVIN: I would like to comment on Mr. King's paper. I was just wondering about the storage routine, because in the 1954 experiment and in our present work we envision only one permanent storage, plus one temporary storage. Now I see that in King's routine there is a permanent storage, then it goes into temporary storage, then it goes back into another storage and then out. I am not quite clear as to why there has to be a double lookup. That is to say, on the one hand a lookup for identification of morphological and semantic elements, and on the other hand a separate one for identification of translations.

KING: I think in the early model of the system there is no particular reason to have the two memories. One could make do with both of them, but this complicates the intermediate high-speed storage. If we try to make a fairly sophisticated translating system the dual would be just more economical and convenient. This means, for instance, if we were looking up a word like "DE" in French, which means almost anything, if we try to get all meanings (I counted 54 in my short investigations) we would impose a tremendous amount of material on our operations.

NEWMAN: First of all, I wanted to say that Jack Rabinow's statement of thinking in terms of a picture may possibly clear up some of the thinking as to the way we are going to use our "Ruly English". We are going to take the thoughts and put them in the form of a picture and then verbalize that picture. To get back to Mr. King's use of 'container', I don't think the container is going to be one of our ruly words, because as we pointed out in one of our early reports, many nouns and many name-things are given to items because of the use to which they are put. We are going to look for more structural terminology. "Sack", for instance, could be 'paper sack', + 'closed tube', + 'coffee container'.

-113-
KING: But I think you are throwing away the wonderful thing about the English language. When I say "sack" I don't have to say all these other words.

NEWMAN: I know, but if you're going to the question of information retrieval, we have to identify, whether you call it a sack, or a poke, or a bag—no matter what the word is, we have to take that thought and put it in terms of a picture, and verbalize that picture in such a way that no matter how you verbalize it it will be the same language. In that way, you can take these different things, like a coffee poke, a coffee bag, and a coffee sack, and get, when you itemize it, the same item with the series of descriptives which will include the same terms.

KING: Yes, but you can take the word "sack" and blow it up with all this descriptive material that I don't have to say. Then you're going to take a coffin and say this is a container with such-and-such properties!

NEWMAN: But we don't plan to use the word container. We're going to use a whole series of words to describe specific concepts or things.

KING: But I'm just saying that you make life harder by 'translating' one word into many words.

NEWMAN: Well, this part of the feature may not help you in machine translation. The emphasis that I put here in my talk was on the interrelational concept and the idea of distributing your action words between the items that were being described. And I think that in this area machine translation can be helped by using that approach.

OSWALD (to HEASLY): Is there any possibility in the foreseeable future that your Intelligent Machines will be able to scan any kind of print line by line?

HEASLY: Well, they're scanning print now. Scanning line-by-line is what we are doing on a machine used by a government agency where we're reading teletype line-by-line, recognizing alphabetic and numerical characters and some hand-drawn edit marks. One of the things we are not interested in at all is handwritten texts in English or foreign languages. But I don't think that is a problem for you either.
OSWALD: No, I'm only concerned with the printed texts. Is there any hope that the machine will be able to recognize a reasonably limited variety of printed characters?

HEASLY: Yes, I think, in view of our present techniques, if the quality of character is pretty good, we may not care too much about the variants, such as between one type face and another, or between one type size and another, or variations in pitch due to justification and various other typographical characteristics. On the other hand, if the material is not well printed, if for example it is something from a poorly inked mimeograph machine, we would have trouble. Applying it to a particular problem is another question. But in general I think that there is hope, and before a millennium! We have read already varieties of type-size of the order of three to one and varieties of type style from a fine Gothic to a rather fancy bold Roman, all in the same program. But this was good material, and in this particular case it was restricted to numeric digits. I think this will give you an idea of the scope of our outlook at this time. We are doing some experimenting on sorting mail. In that operation we just look for some things. We take the position that there is much redundancy, and place names are probably more redundant than anything. We just look for some characteristics and we don't even require all of the characteristics that we know would normally be necessary to recognize a particular character. We just look for some things that are easy to detect and we say, if we get this characteristic and a couple of characters later we get something else, then we can recognize the name of a city accurately, even though we may not have recognized all of the individual characters which composed it.

RABINOW: I'd like to make a comment on this. If you tie one of the memories to your reader rather than to your translator, you could always give perfect copy if your error rate was less than say 1%. The reader could then say, "This word makes sense and this one doesn't". With words, this makes perfectly good sense, except in very special cases.

HEASLY: Unfortunately most of the work has been on information that is really random in nature. The teletype copy we are reading is not plain language.

RABINOW: Could I ask an embarrassing question? What are you reading teletype for?
HEASLY: The answer is that in this particular application they tried using other techniques, such as using a perforated tape which also had the characters typed on it. Then it had the editors splice the tape and that sort of thing, and they found that they just got no production out of the editors at all. What they do now in this operation is to have the material come in on a digital teletype, the editors edit it with their red pencils and give it to the key punch operator to put it back in digital form again.

KING: What kind of an input have you thought of for this set-up?

HEASLY: A scanner. We say scanning first, and then this incredible whirl of the machines.

DOSTERT: I have one general comment to make about the papers read by Mr. Crossland and Mr. Mueller on the one hand and the papers by Mr. King and Mr. Rabinow. When we listen to the papers on the linguistic side, the paper on word decomposition (we have called it "splitting" here) and the syntactic or grammatical analysis in Mr. Mueller's paper, one is struck with the complexity of the operations described and wonders at the volume of explanation required to instruct the machine to perform the complex operation described in decomposition and in syntactic grammatical analysis. On the other hand, the papers by Mr. King and Mr. Rabinow suggest that the storage possibilities and rapidity of search are almost limitless in their promises. It seems to me that the time may not be too far away for the practical consideration of the economy of the two basic techniques of multiple entries—increasing the number of one-to-one equivalents—as against analytical program instructions. We have had considerable discussion on this problem in our seminar. We have not come to any formal or definite conclusion yet. It seems to me that this is a basic problem in the research, whether or not the linguists should strive to formulate exhaustively the syntactic and lexical operations for programing instructions, or whether they should not rather be prepared to accept the more simple technique of storage. This is not, of course, the time to resolve the question. I am merely raising it as a fundamental point in the whole problem.

RABINOW: I think this is a case of the tail and the dog wagging each other. If the tail is the memory, and it is big enough, it may very well wag the dog. If we can give you any word you like with any combination of any other word—we could even put two words in simultaneously, so that BIG CAT appears, or BLUE CAT—we could
then give you groups of words if you like. It is just a question of economics. Which is better—to combine them in the logic or combine them in the machine. If the machines get powerful enough, and I know that International Telemeter makes some very powerful optical machines, then it has much better sense to do it by brute force—look up the whole phrase—the whole sentence, if it's a common enough sentence.

**KING:** To make really good idiomatic output I think the emphasis is going to have to be on elaborating the answers, rather than cutting them down. We should not give too much thought to that problem now. In these early days we do want to have simple, single meanings to the fullest extent possible.

**HODGE (to NEWMAN):** This is really not a question, but I am just curious to see the application of this to a patent application.

**NEWMAN:** I can assure you, we would too. This is one of the milleniums we hope will not be forever. As a matter of fact, one of the projects I have in mind is to take a patent specification and do this. In order to do it, though, we have to create a vocabulary, which is very slow in its creation. I might state here that possibly I am flying under false colors. I am not a linguist. I am an engineer and I have been thrown into this thing, and possibly I am cutting across a lot of things that a linguist might worry about because I know not where I tread. However, I do feel that in this business of getting a single word unit to mean a specific thing, that we can clarify a great many things that we speak of very ambiguously in a normal way, and allow the context to tell us really what we are saying.

**RABINOW:** I disagree quite entirely with your approach to the problem of simplified speech. Whatever the duties of the Patent Office, which very often covers more than an inventor thought of, even more than the attorney thought of, many years later you discover the importance of a thing that was done. I'd hate very much to have you change my language, even though it might be better, because it might not be the thing I thought of, and I think that with your simplified speech you won't get any place. Actually, if anything, speech is not accurate enough—it needs more words. And I don't think you could define things precisely. To give your example back to you about the stiffness of a beam vs. the stiffness of a watch spring, I disagree first technically very much that there is any difference. It may very well be that the beam is stressed much more than the watch
spring and that it is actually much more resilient in any technical term. In any case, we take my words and you start to change my meaning. I am very skeptical that much can be done to improve the language by using your special words.

**NEWMAN:** Well, let me give you two sentences which occurred to me earlier, because I thought something like this would come up. Take two ideas: "To get hives from berries" or "to get seeds from berries". Now those two *froms* are entirely different *froms*. To get *hives* from berries is a sort of a cause-result "*from"", and to get the seeds from berries is a *whencefrom-fromwhence type of from*. Now, I understand, at least from one translator, they said the ideal translation was the ability to tie the ambiguities in the source language into another ambiguity in the target language. If this is the case, this is not what we're trying to hit. But what we're trying to do is, if you will, to pre-edit the source language and transliterate it into a target language in which its meaning is clear.

**LEHMANN (to GARVIN):** I am curious about the MT notation that you propose here. We have had the problem, as you know, in linguistics of a variety of notations. Is any effort being made to develop a uniform notation for the analysis of linguistic material for mechanical translation?

**GARVIN:** Actually, we have no notation for the linguistic analysis because it is not to be incorporated into the actual translation code. What we have is notation for what we expect a translating program to accomplish and this is, as I tried to say, not linguistic analysis, although it is *based* on linguistic analysis.

**LEHMANN:** I am referring to the former days when linguistic terminology was being built up, after which there was considerable revision of terminology.

**GARVIN:** We have found that since there aren't any complete descriptions of Russian except in traditional grammars and in a few other sources that we have available, we use traditional terms. We say, "instrumental" and "genitive" and "singular" and "plural" and let it go at that, if this is what you mean. For our purposes this is adequate because we are dealing with written Russian anyhow and what is a flaw in traditional Russian grammar, namely, to give the suffixes as written, is for our purposes an advantage.
LEHMANN: May I go on then to amplify this? Now, why did you choose a "P" then, for your decision point? Say, your symbols?

GARVIN: Point starts with the letter "P", and cue starts with the letter "C", so I use "P" and "C".

LEHMANN: Do you suppose Cambridge uses the same notation?

GARVIN: They don't have the same. I don't know whether any existing translation research group has yet come to the point where they actually have recognized what their decision cues are and what code they are using to translate from a Russian text into English, therefore I couldn't answer your question. I have a feeling that most machine translation discussion has been on a theoretical level and therefore has had no use for this. The Cambridge people are handling it on the basis of a primarily logical approach and they conduct various mathematical operations for which my symbols would be unsuitable. Now several other groups have worked in MT translation, but they have not gone into any extensive routine to resolve the choice problem; therefore they don't need the "P"'s and "C"'s. Within our particular research group there are several other sub-groups in addition to the one that I have been working with, and they use somewhat different symbols simply because they formulate operations of a different nature with different symbols. I don't think anybody has done exactly this on a general basis and therefore the problem of standardization of symbols has not yet arisen.

DOSTERT: This is in amplification of what Mr. Garvin just said in response to Mr. Lehmann. I think it would be premature now to think of standardizing codes, as they now emerge, because there is no uniformity even of theory among the several groups in different localities. There is not even theoretical uniformity within our own sub-groups here at Georgetown. We have deliberately encouraged diversity of approach, rather than to harness research to a pre-judged theory. Now we have three approaches. One calls itself the "experimental group", which, speaking in general terms, proceeds empirically from the specific pattern to the general formulation. From a semi-analytical and semi-empirical approach, these staff members expect to formulate reasonably soon a series of generalizations which can be turned into machine instructions. Another group has developed a machine technique based on grammatical analysis of Russian, to resolve what may be called 'internal structural ambiguity', before proceeding with transfer based on diacritics matched with code.
symbols on the target language. This is the “code-matching” tech-
nique. A third hypothesis was advanced by one of the members of
the staff which, so far as now known, seems a purely empirical
approach, i.e., sentence by sentence. The argument here is that
exhaustive analysis of a series of sentences in a continuous text
will yield general rules. Given the divergences in theoretical ap-
proach, at least as far as our own experience shows, it would appear
premature to think of a standardized coding of the results of the
investigations of the several groups.

JESSE MANN (to AUSTIN): I would like to make just one short
note on a syllogism that Mr. Austin correctly cites as a classic, and
indicate a misconception that he quoted. It is frequently thought to
be an Aristotelian syllogism. Actually, it has been long since
pointed out that Aristotle always uses symbols (A., B., C, etc.)
and that it was a mistake in the traditional thought that Aristotle
formulated syllogisms of this type in which he would use a singular
term in the minor premise, which was evidently far from his theory
of logic. I would like to point out that at least Beech, of the Univer-
sity of Indiana, has written a book which he has called specifically
“Intentional Logic”. I would like to get some amplification on the
meaning of that statement.

AUSTIN: Well, logicians sometimes make a statement that the
passengers of the Mayflower are the founders of Plymouth. Now,
by intention they admit this all falls apart because passengers and
founders are different, but by extension they can be made to be the
same. I have not seen this book.

MANN: All I had in mind was that to frame a syllogism such as
“to speak a language is good; to speak French is to speak a lan-
guage; therefore to speak French is good”, but there you have no
expressed reference. (I hasten to qualify that hastily with regard to
any extension.) Or you could say in your slogan there, “To work is
to play, therefore to play is to work”. I am just trying to make ex-
PLICIT the point that some intentional ends, it seems, will be neces-
sary even in a logic. Even in the most abstract of formulae most of
the formula-makers would not be willing to concede to your putting
zeros in place of all the arguments. Perhaps I misunderstand. I
understand that there are considerable difficulties in assigning some
kinds of meanings, and I think you brought out very well the fact
that you sometimes mistake sentences which have meaning for

---120---
sentences which don't. I don't mean that at all. I just mean that we can't say that any logic which seeks intension is childish, and that is because we have to have some intention.

**AUSTIN:** I still don't see the point of intention at all. Scientists construct logical existence and then find contents for the existence, and the contents will be intention.

**MANN:** The only point I'm making now is that I don't think the logician who forms systems would be content to substitute zeros for all elements in the system. He would hope, at least, that there was some intelligible content.

**AUSTIN:** Language would be like an algebra—a system. Then you would put in the meanings—the content of this system. I should perhaps not say "childish"; let me just say "difficult".

**GARVIN:** I just wanted to make a comment as a transcendentalist-positivist in the sense that I think the direction has been at least in linguistics to use little terminological tricks—to go from transcendentalism to positivism. For instance, we now no longer speak of "meaning" when we are very positivist; we speak of "context classes", and such as that. And likewise, I feel that this whole argument about intentional logic may perhaps be resolved by saying that in some way content is accommodated by different class membership, because if you look at the members of the class *passengers* as restricted by the class *Mayflower*, and on the other hand the membership of the class *founders* and restriction by membership in the class of our country, etc., and you look at this in terms of possible members included in each class, then you have formulated your intention without using the word "content" and you have then shown that this is after all quite right. What I am driving at is that I feel that if you have a need for handling something in a formal way, then you do this, whether or not this is overtly formal or not. In other words, if I want to handle meaning, and if for some purpose or other I do not wish to be transcendent and say that this is a property of the mind, then I simply develop a formal system for handling meaning. For instance, I will say that meaning is equal pragmatically to either translation or paraphrase, and the general meaning of a given unit is that which all translations of this particular unit have in common and all paraphrases of this particular unit have in common. Or I will say that meaning is the class of all those units that can be substituted for the current unit, or I will say that meaning is that
set of relations into which this unit can be entered for purposes of whatever manipulation I want to do to it. And so to me this long argument about whether you are being this, that or the other thing, epistemologically, is rather trivial, because when you want to do things you have to be more positivist and when you want to explain things, I think you have to be more transcendentalist. Then you have to make up your mind on what you have to devote yourself to, and so I am a transcendentalist-positivist!

MANN: I merely wanted to insist, though, that even at the beginning of your process where you assign certain starting points, you are assigning meaning.

DOSTERT: Before we close, I want to correct an impression I may have given a moment ago when I said that in our own groups we had three different approaches to the problem of machine translation. This is deliberate and in any field of scientific investigation various hypotheses must be given complete freedom to be expressed and tested. When we can move from hypothesis to theory, and then to established facts, we are proceeding along perfectly valid methodology. In due course the data gathered by the various approaches may come together to some extent; some phases of it will not be susceptible of integration, but that is the methodology we are following. Diversity of approach to the solution of MT problems does create a measure of confusion, admittedly. Out of mutual challenges will come, we believe, the more effective formulation. Freedom of approach to problems is still the key to scientific solutions.
IV

Panel III

Scope Of Syntactic Analysis

In Machine Translation
Analyzing German materials with a view to mechanical translation in accordance with current linguistic practices, we find the noun phrase as one of the commonest constituents. The noun phrase is an immediate constituent of most subject-predicate sentences, and frequently an immediate constituent of a predicate as well, as may be illustrated with a fairly frequent type of sentence (1): 1

Jede sprachliche Veränderung und mithin auch die Entstehung jeder dialektischen 'Eigentümlichkeit hat ihre besondere Geschichte. 'Every linguistic change and consequently also the origin of every dialectal peculiarity has its special history.'

By our first analysis of this sentence we determine one immediate constituent, which consists of the material preceding hat, to be a compound noun phrase. (Since the position of the finite verb is rigidly circumscribed, analysis of most German clauses into two immediate constituents is a simple matter.) By our second analysis, we determine another noun phrase, the material after hat. While analysis and translation into English of this sentence should cause no difficulty for man or machine, those who have consulted materials written in technical German are aware that interpretation of noun phrases occasions their greatest difficulty for man; one might therefore expect that it would also occasion difficult problems in machine translation. Principles involved in its analysis will be discussed here as well as its structure. For the structure of noun phrases can fairly readily be summarized.

We may classify noun phrases in four groups: I. those consisting of nouns preceded by modifiers; II. those consisting of adjectives functioning as nouns, preceded by modifiers; III. those consisting of nouns or adjectives, that is, phrases of types I and II, followed

1 Among materials analyzed for this sketch were sections from H. Paul's *Prinzipien der Sprachgeschichte* (5th ed. Halle, 1920), N. Trubetzkoy's *Grundzüge der Phonologie* (Prague, 1939) and Helmut Rehder's "Karl Jaspers" in *Philosophen des 20. Jahrhunderts* (Stuttgart, 1957). German examples cited may be located in these texts: (1) Paul, p. 42; (2) Rehder, p. 739; (3) Trubetzkoy, p. 15; (5) Paul, p. 8; (6) Paul, p. 14.
by modifiers; IV. those consisting of phrases of types I and II accompanied by positionals, usually prepositionals. The following is a brief summary of this classification, with examples.²

I. Nouns preceded by modifiers

A. φ No

B. Aj No

a. Li No a. DE No

β. KE No

γ. Adj No

b. De No

(De De ...) (Cnj)

Das Bier 'the beer'

sein Bier 'his beer'

Pauls Bier 'Paul’s beer'

dunkles Bier 'dark beer'

Das dunkle Bier 'the dark beer'

sein dunkles Bier 'his dark beer'

cf. recht dunkles Bier

'really dark beer'

Das recht dunkle Bier 'the quite dark beer'

sein recht dunkles Bier 'his quite dark beer'

Das recht dunkle Bier 'the quite dark beer'

sein recht dunkles Bier 'his quite dark beer'

(Adv) (Adv)De

(cf. sentence (5) below

²The abbreviations used here have been chosen for mnemonic purposes. Abbreviations composed of two letters represent inflectional elements; of three, invariant elements; of four, immediate constituents of larger constructions. When all members of the abbreviation are capitalized, the abbreviation refers to a class which is named after one of its members. Statements on morphophonemic/morphographemic variation are not dealt with in this paper.

The simple sentence consists of Subj { Pred Verb Objt }

Abbreviations used are:

Adv Adverbial

Adj Adjectival

Aj Adjective

Cnj Conjunctional

De Descriptive Aj

DE der-Li Objt Object

KE kein-Li Ppn Postpositional

Li Limiting Aj

Adj PpP Prepositional

or I. B. a. y

Subj Subject

Though I. A., II. A, etc. consist of but one word, they are included in the roster of phrases to provide a complete analysis of nouns which may occur in the subject position. Further, certain sub-classes of nouns are marked by the absence of ein, i.e. by an A rather than one form of B.a. constructions.
II. Adjectives preceded by modifiers
A. φ De Dunkles 'dark (beer)'
B. Li De sein Dunkles 'his dark (beer)'
C. Li Adv De das recht Dunkle 'the quite dark (beer)'

III. I and II followed by modifiers
A. I + (Li) (De) No das dunkle Bier dieser Brauerei 'the dark beer of this brewery'
B. I + Adv ein reichliches Material zum Beleg 'copious material as evidence'

IV. I and II with positionals; such phrases are Adv
A. Prp + I mit einem solchen Bier 'with such a beer'
B. I + Ppn diesem Gasthaus gegenüber 'opposite this inn'
C. Prp + I + Ppn von dieser Grundlage aus 'from this foundation'

Other classifications of noun phrases might be made. This one is based on morphological differentiation of the head of the phrases and immediate constituent analysis of sentences. Each of these four types of phrases is of course found with various complexities, as is indicated after I.B.b and I.C.c; but these do not present any essentially new features. Moreover, each component of II, III and IV phrases must be analyzed in greater detail, as in the analyses after I labeled a.a, a.β, etc. For the sake of brevity such further analysis has been omitted here; but an example of such analysis is provided in the third last paragraph of this paper. Given a summary like that presented here, various procedures for mechanical translation could be devised, or various types of notation. Some of these will be touched on as further aspects of noun phrases are discussed.

At the present stage of machine translation, analysis of materials to be translated must be undertaken on the graphemic level, the morphological level, and the syntactic level, if we leave aside here lexical problems. Moreover, it is restricted to one style of German, since machine translation work will be directed at technical rather than at literary or colloquial German. While with Joos one may be dubious about the possibility of achieving machine translation from
the imperfect representation that is provided by the written language, our imperfect knowledge of the signals of spoken German and the stage of our translation equipment suggest that we must first attempt to achieve a suitable method of dealing with written texts, subsequently with spoken materials. If machine translation of written texts is impossible of achievement, other than in the dubious form of the so-called pragmatic approach, preliminary linguistic work will be useful for the machine translation of spoken texts and it will contribute to a description of German. For the treatments of the noun phrase in our extant grammars of German need some modification, based as they are on a traditional morphological approach. Useful descriptions of some types of phrases however have been published and others are utilized, though not systematically presented, in some pedagogical works on the German language; previous discussions of the machine translation of German also include valuable material on the noun phrase. These differ from the present study in dealing essentially with morphological data, while a syntactic analysis is presented here.

Although written German may be deficient in marking some of the signals of spoken German, it provides us with graphemic markers which may in part compensate for the absent phonological markers. The first of these is the sentence-final marker, normally the period, which is equivalent to phonemically significant final junctures. (Another, the interrogation point, also has a distinctive function, though different in extent from the German sentence-final junctures; and the exclamation point again has a significant function, though this may be redundant with morphological and syntactic markers of the imperative construction.) From these, with the help of morphological and syntactic markers, we can readily determine clauses, which as noted above can be broken into constituents. Two other

---


graphemic markers are of great significance for the analysis of the noun phrase: the comma and the capital letter. Since commas stand before any new clause in German, they assist greatly in determining the constructions which must be subjected to immediate constituent analysis, in distinguishing homonymic pronouns from adjectives, and so on. We may note at once, however, that graphemic markers will not of themselves permit analysis of all clauses. A simple clause, determined by a comma and a period, like (2): *welche mögliche Vollkommenheit simulieren.* 'which simulate possible perfection' can only be analyzed on the basis of morphological data outside the noun phrase, i.e. by the morphological datum that a noun phrase consisting of *welche* followed by the *e* of *mögliche* before *Vollkommenheit* would require final *t* on *simulieren* if we were dealing with a unit noun phrase *welche mögliche Vollkommenheit*; the syntactic datum that *simulieren* stands at the end of the clause requires us to make our immediate constituent analysis after *welche*. Moreover, the graphemically significant capital letters in German provide information unavailable in the spoken language, for they mark nouns, as may be illustrated by the following sentences: *Er geht zu messen.* 'He is going for the purpose of measuring.' and *Er geht zu Messen.* 'He goes to fairs.' It might be assumed then that the final extent of noun phrases is marked essentially by capital letters, and that the chief machine translation problem in the determination of noun phrases is one of order, not of extent. For noun phrases of type I this assumption holds. Graphemic devices, capital letters, are instrumental in determining the final limit of immediate constituents in successive cuts, marking nouns; syntactic and morphological sub-classes mark their prior limit. Accordingly in describing in greater detail noun phrases of this type, we must determine the classes and sub-classes of their components.

Homonymity of phrases like *der Wagen* 'the car' and *der Wagen* 'of the scales' suggests that we cannot make adequate analyses from the sub-class of the noun modifiers alone, but that we also must classify nouns into various sub-classes—indicating the sub-class by the form of the noun we choose to list. Through comparison of minimally contrasting clauses like: *Er hat den Wagen gekauft.* 'He

---

5Note however the qualification discussed by Reifler, *Machine Translation of Languages*, 150-51; since capital letters are required at the beginnings of sentences, nouns in this position are not distinctively marked.
bought the car.' and *Er hat die Wagen gekauft.* 'He bought the cars.',$ where the contrast between the noun modifiers alone marks the difference between singular and plural, it is clear that distinguishing data may further be presented in the sub-classes of modifiers. Distinctions marked in this way for noun phrases of type I.B. (except for a,y) number seven for nouns of masculine gender, six for nouns of neuter gender, five for nouns of feminine gender.

Although we make provision for these five to seven forms, we will still need to rely on syntactic, or semantic, markers to distinguish between homonymic nominal phrases like *den alten Wagen* 'the old car (ace.)' and *den alten Wagen* '(in) the old cars' or *die neue Bank, -e* 'the new bench' and *die neue Bank, -en* 'the new bank.' Homonyms like *die Wagen* 'cars (nom.)' and *'cars (acc.)'* will be distinguished on the basis of our initial cut of the sentence into two immediate constituents; if the homonymous nominative and accusative plural falls in the constituent accompanying the verb, it usually stands in the accusative. This procedure will also distinguish the homonymous nominative and accusative singular of feminine nouns. Though listings of nouns and their modifiers in their five to seven possible forms will then account for all of the information provided by the morphological markers of phrases of type I.B, additional distinctions are indicated by syntactic markers. For type I.A too, we have to rely heavily on syntactic markers. Type I.C requires on the other hand even further syntactic analysis, for in this sub-type lie many of the difficulties met in technical German.

Simple examples of type I.C are contained in sentence (1): *jede sprachliche Veränderung* and *jeder dialektischen Eigentümlichkeit.* *jed-* belongs to a syntactic sub-class, the DE-words, which marks the prior limit of a noun phrase. Mutually exclusive with the definite article, it contains information which defines the function of a following noun. But while noun phrases of type I.C are usually bounded by adjectives, generally limiting, and nouns, the elements contained within these bounds must be analyzed for their form; for the endings of the enclosed elements disclose the proper analysis of noun phrases longer in extent than two words. If in accordance with certain rules, the enclosed elements end in -e/-en, the phrase can be equated to corresponding English noun phrases and managed
as simply as are the noun phrases of our first sentence. If, however, the enclosed elements end in different graphemes or if they belong to sub-classes incapable of forming type I.C noun phrases, we are generally confronted with a I.C.c construction. The following sentence (3) illustrates an impossible grapheme.

vielmehr kommt es auf die mit ihrer Hilfe zu erlangenden allgemeinen Begriffe an. 'rather, it depends on the general concepts which may be obtained with its aid.'

mit signals as impossible analysis of the segment from die to Begriffe as a simple noun phrase; the -en ending on allgemeinen, however, suggests assumption of the intervening material before erlangenden as adverbial. Just as a descriptive adjective ending in -e or -en without an immediately preceding limiting adjective generally marks this construction, so do two limiting adjectives—unless they indicate that one of the apparent limiting adjectives is a pronoun, as in the following sentence (4):

Ihre Betonung stimmte mit der Genetive der konsonantischen Stämme überein. 'Their accentuation coincided with that of the genitives of the consonant stems.'

While constructions of type I.C.c by their variety and complexity often provide man with the sole diversion he derives from reading technical German, their basis structure is that of the examples after I.C.c on the chart, as we may illustrate from the following moderately complex example (5):

[Es werden] mit Hilfe des menschlichen Leibes bearbeitete oder auch nur von dem Orte ihrer Entstehung zu irgend einem Dienste verrückte Naturgegenstände ...[übertragen] 'natural objects handled with the aid of the human body or merely moved from the place of their origin for some use...'

The material of this example which is not enclosed in brackets consists of the construction given after I.C.c, with two adjectives

---

6On the other hand, an -e or -en ending after Li may not indicate De, as in the following examples: Paul 17 ...von den oben Seite acht angegebenen vier Kategorien... 'Of the four categories given above on page eight'; Paul 18 ...die aus ihnen bestehenden Sätze... 'the sentences consisting of them...'
modifying Naturgegenstände, bearbeitete and verrückte. Each adjective is modified by an adverbial. The adverbial preceding bearbeitete is a type IV noun phrase, mit Hilfe, accompanied by a III, (Hilfe) des menschlichen Leibes; the material preceding verrückte is a type IV accompanied by a type III, followed by a type IV, von dem Orte ihrer Entstehung zu irgend einem Dienste. Each of these phrases in turn incorporates or is a type described after I: (mit) Hilfe a type I.A; des menschlichen Leibes a type I.C.a; dem Orte a type I.B.a.a: ihrer Entstehung and irgend einem Dienste a type I.B.a.β. If one wished to proceed in reverse, to apply successive rules for the expansion of a given symbol (as has been suggested by some workers in machine translation) this could readily be done. One can handle such complex phrases in a variety of ways, by treating them as transformations or by using whatever other analytical device seems most economical.

The other noun phrases may similarly be analyzed with the procedures outlined above. Simple examples of II (which conforms to I, with the difference that the nominal retains the inflections of the descriptive adjective) were presented on the chart. A variant is found in the following sentence (6) ending in entsprungenen:

Um die einer in ihr selbst entsprungenen entsprechende Vorstellungsverbindung in einer anderen Seele hervorzurufen,...
‘In order to produce in another mind the association of ideas which corresponds to one which has arisen in oneself,...’

Examples of III and IV were given in the chart, and further illustrated in sentence (5) above.

A complete description of the noun phrase in German would be much more explicit with regard to sub-types of these four types, to variations in forms, and to frequency of the various types. It would also deal with problems such as the order of adjectivals, especially that of descriptive adjectives. Their order presents little difficulty when one moves from German to English, as is evident from the following extensive noun phrase of the type beloved by linguists:

All the nice little new two-story white brick houses...
All die netten kleinen neuen weißen zwei-stöckigen
Backsteinhäuser...
The order of the modifiers is virtually identical in the two languages; the propensity of German to make compound nouns as illustrated in this phrase would be a problem outside the scope of analysis of the noun phrase, one in the sphere of morphological analysis. Further problems such as the non-correspondance of limiting adjectives in some German and English noun phrases, methods of distinguishing between adjectives like *der, jeder, seiner* and pronouns homonymous with them would also require full description in a complete analysis of the noun phrase in German.
In relation to problems of MT from Russian to English one must first of all consider the nucleus-centers of the source language, because it is only on the basis of these that one can obtain proper equivalents in the target language.

It seems appropriate to make a few preliminary remarks which will clarify some technical aspects of this paper. First among them is my belief that a sentence definition should be positive. We need to know what a sentence is and not what it is not.

Since our formulation is to be applied to computers, it is advisable to express it in logical and concise terms. For that reason, I have included such terms as independent and dependent variable in my definition of the Russian sentence, and to the greatest extent possible I have reduced the complexity of Russian structural types to symbolic formulas.

One necessary condition for MT research is the knowledge of morphophonemic process both in Russian and in English. In Russian, an example of this would be variation within pronominal forms. Examples:

1. получена фракция
   a fraction was obtained

2. выход почти количественный
   the yield is almost quantitative

We believe that Russian sentence types can be described only on the assumption that Russian speakers, while generating sentences, are guided by a general relational meaning of two types:

1) the actor-action
2) the actor-action goal.

This is especially true of morphemic types where there is no graphemic difference between nominative and accusative case. In such instances word order in Russian becomes functional.
Definition Of Russian Sentence

For the field worker who watches how words behave within Russian structure, it becomes clear that there are usually two nucleus-forms:

1) the nominative case of the noun, personal pronoun, cardinal numeral, or the series of words which can assume the function of the nominative case (infinitives, adverbs, etc.); here the nominative function is expressed in writing by a delimiting punctuation mark " ", and in speech by a special intonation;

2) the personal verbal indicator or the form which can replace it. (In speaking of substitutes, we have in mind short pronominal forms.)

Thus, from a positive point of view, the Russian sentence may be expressed as a function of the nominative case and the personal indicator, the nominative case being the independent, and the personal indicator the dependent variable. The verbal form is predictable on the basis of the nominative case, but not vice versa.

Such a definition of the Russian sentence overcomes the difficulties arising from a situation in which there is interdependence of whole sentences or separate members thereof. Examples: Russian personal pronouns ОН, ОНА, ОН, etc., cannot be transferred as ОН—he, ОНА—she, etc., since ОН may be rendered in English by he, she or it depending on the source item to which ОН refers. The details on this problem were discussed in a separate Georgetown University work paper on MT. The definition of the Russian sentence is satisfactory in all such instances.

Since we consider the largest unit for syntactic search to be a sentence, the whole Russian syntactic field can be thought of as consisting of sentence types which are not subject to further breaking down, and of transformational operations which can be effected in terms of these irreducible Russian sentence types. This refers to the syntactic transformation and kernel theory of Harris.

Graphically, one can describe the syntactic field as a circle divided into the nominative case and the personal indicator with their interrelationships.

—136—
From combinations of those two, we can produce the following sentence types:

1) Both parts are explicitly present;
2) one of them is lacking;
3) both are lacking.

Type One is of highest frequency; Type Two is next highest; Type Three is of lowest frequency.

If we start to analyze Type 3, in which both components are absent, we would see that here belong titles, headlines, slogans, inscriptions, formulas, etc. From the structural point of view, they represent a government structure, whose determiners can be listed. Within the government structure an agreement restriction may occur. The latter does not hinder the string operation\(^1\) of such a structure if it is taken for granted that selection and arrangement problems must be dealt with, anyway.

Examples: 3. On the spread of culture in countries situated in the African continent
О распространении культуры в странах, расположенных на африканском материке.

4. Concerning science in the USSR
О положении науки в СССР

5. Concerning the action of triarylbro methanes on alkylpyrocatechol esters of phosphorous acid.
О действии триарилброметанов на алкилпирокатехиновые эфиры фосфористой кислоты

6. Experimental Part
Экспериментальная часть

7. Hydrolysis of the ester
Омыление эфира

---
\(^1\) By the string operation we understand the set of operations being performed after the syntactic transfer was affected; i.e. H and P were transferred. In a string operation we transfer
a) lexical meaning which may involve a lexical choice
b) desinences which may cause a rearrangement or insertion of a preposition before certain nouns
c) rearrangement of post-modifiers, etc.
Type No. 2 in which one component is lacking, comprises impersonal and nominal structures in Russian. The nominal type is characterized by the absence of a verbal indicator, the impersonal type by the absence of the nominative case. There are special transformations within the Russian sentence indicating which item assumes the function of the nominative case. Examples:

8. I have a headache  У меня голова болит
9. I hate to look at him  Мне противно на него смотреть
10. The wind carried away the boat  Лодку унесло ветром
11. Not a word about it  Ни слова больше об этом

In comparison with Russian, the English translation shows that the English equivalents of the Russian non-nominative items expressed in genitive, dative, prepositional, and instrumental cases, are the nuclei of the noun phrase. The Russian speaker would never look for a preposition before the nucleus of the English noun phrase. Example:

12. a new paper—  новая работа

If in a Russian sentence the verbal indicators are missing, then we have nominal phrases. For example:

13. The question of mutual relations between the West and East  Вопрос о взаимоотношениях между востоком и западом

14. Melting of bromide and iodide systems  Плавление бромидных и йодидных систем

From the point of view of translation it is necessary to point out that in this category the verbal nouns when they do occur are determiners for the string operation and will have in the immediate vicinity a non-nominative case, usually genitive. For example:

15. The obtaining of the ethylpyrocatechol ester of phosphorous acid.  Получение этилпирокатехиноового эфира фосфористой кислоты

One has to remember that quite frequently the zero verb can be a problem if the cues for its transfer are not properly identified. In this case, we depart from the pure linguistic point of view. We attack this problem not structurally, but rather in terms of transfer.
Since the copula of the complex predicate equals zero, we take the nominal part of it as a cue and effect a transfer either as present/past and singular versus plural with insertion of "IS/ARE" or "WAS/WERE" depending on the source cues.

Examples:

16. The yield is almost quantitative
17. My father is a physician
18. The physician is my father
19. It is not very good
20. It's bad
21. The system of fusibility was studied
22. The data are given

Among the type in which both components are present, we differentiate two subgroups:

1) non-third person indicator plus a noun or third person pronoun;
2) first or second person indicator plus the personal pronoun of the 1st or 2nd person.

In the latter case the pronoun may be missing, since it is redundant.

Example: Иду. Я иду. Люблю грозу в начале мая..."

One has to consider forms where the lack of a morphemic signal identifying the accusative and nominative cases causes the word order to become functional. (We will not go into that question since it will be discussed separately in this morning's session.)

Thus we can list the Russian kernels with corresponding English kernels in the following manner:

<table>
<thead>
<tr>
<th>Russian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Реакция должна идти</td>
<td>The reaction should proceed.</td>
</tr>
</tbody>
</table>
2) N is N
   a) The yield is 10 g.
   b) The product of hydrolysis are pyrocatechol and triarylmethylphosphinic acid.

3) N is A
   The esters ... are resistant
   The formulas ... are given
   The yield ... is high

4) N is D
   The pencil is here

Versus those where the relational meaning of the actor-action-goal structure is expressed like in:

5) NVD (NV+N is PN)
   The reaction actually proceeds according to the above mentioned equation

6) NVN
   (Their) behavior ... confirmed ... the correctness...

7) N is PN
   The dimethylaniline is in ether solution

and finally a typical Russian structure which has no English equivalent, i.e. impersonal structures such as:

8) It is A to V
   It is necessary to note ...
In terms of spoken Russian, one has to remember that a Russian sentence can be described further in terms of the open or closed intonation contours. These are reflected in the writing system by two categories of Russian punctuation marks, namely, separators and delimitators. (These categories will be discussed in a separate paper.) Conjunctions, specifically coordinating conjunctions, are used for the same purpose.
Some Ideas On Inter-structural Syntax
Jane A. Pyne
Institute of Languages and Linguistics

Many theories and applications of syntax are discussed in the literature of machine translation. Various linguists' utilization of syntactic analysis in MT research represents diverse points of view, indicating far-reaching interest in syntax since the latter is perhaps the most essential tool for reducing the translation process to mechanical procedures. In his paper on the Georgetown-IBM Experiment included in the Locke and Booth collection,¹ L. E. Dostert suggested the possibility of developing a coded general or core syntax, common to several languages, into which the syntax of these natural languages would be programmed for purposes of multi-lingual translation by machine. This concept has been further revised and developed by him during the current research project at Georgetown University.

It should be noted in this connection that the question of multi-lingual core syntax has particular relevance to a future stage in the research, that is, when we hopefully will have come to a consideration of machine translation of more than two languages. At the present time our focus of study in the Georgetown project is one-directional and bilingual only. Consequently this paper deals with certain problems of structural transfer from Russian to English, and is based on the research in syntactic analysis being carried out by M. Zarechnak and myself.

We are attempting to develop a mechanical procedure for effecting structural transfer from Russian chemical discourse to its English translation. The sensing of functional units is considered essential for the machine to be able to transfer meaning adequately, in that translation is defined as the transfer of meaning from the linguistic pattern of one language to that of the other. Functional units, in turn, are defined as structural or syntactic units of language, as opposed to the linear commutativeness of individual words or morphemes. I will refer to these functional units later in connection with a definition of translation units at various levels, from the lexical to phrase level. In an article entitled "Transfer Grammar",

Z. S. Harris states, and I quote: "translating the morphemes ("word-by-word") is in any case not enough for translation, since the grammatical interrelation of the morphemes in each language is a matter of the subdivision of the sentence into constituents (in successive inclusion) which will often differ in the two languages; and the order of the morphemes within each constituent will often differ. The analysis of a sentence into successively included constituents, and the composition and order of smaller constituents (down to morpheme classes) is therefore necessary for any method of translation that is to be reducible to mechanical procedures."  

Although any method of translation, whether human or mechanical, requires the substitution of the morphemes of one language for those of the other, the nature of linguistic structure precludes linear substitution. The morphemes or words of English cannot be linearly substituted for the morphemes or words of Russian because the grammatical interrelationships are not identical. Furthermore, a given Russian morpheme or word may have more than one possible equivalent in English, and thus be translationally ambiguous either lexically or grammatically or both. The problem of linguistic analysis for MT consists in separating from the start those translations which are free variants of transfer from those which are positional variants. Positional variants are those where the choice is determinable from some lexical or grammatical item in the determining environment. These can be called positional variants of transfer, and constitute the data of MT, as opposed to the unimportant choice between two translations which are free variants for the transfer. Free variants refer to those whose selection in any environment is a matter of style or individual preference.

Because not all choices among given positional variants are cued by the presence of some item within a predictable and definable distance from the ambiguous item, and because a linear method of translation does not solve problems of rearrangement—such as when the Russian verb precedes the noun, or may even be zeroed, and English structure requires different order—for these reasons, it is necessary to view the transfer operation in terms of a machine-programmable analysis and transfer of successively included constituents within the sentence. The goal of our research is to prove

---

that the sentence can be handled by the machine in terms of its constituents in successive inclusion, so that the composition and order of smaller constituents (down to morpheme classes) can be adequately translated.

The Russian sentence can be defined, as previously discussed in this morning's session, as a bicomponential function, where H is the independent variable in the nominative case and P is the dependent variable, a verbal form or its substitutes. Once the H and P as nuclei are handled, it is relatively straightforward to translate the elements surrounding the nuclei since the majority of these elements are in direct dependence relation to the nuclei. All Russian sentence types are expressible in terms of H, P and three features of grammatical relation: agreement, government and apposition.

Since we are attempting to achieve mechanical translation from Russian to English, not the reverse, any approach to English syntax is in terms of its minimum difference and maximum similarity to Russian. This difference, determined through comparison of Russian structural types and their English translations, can be defined as the number and content of grammatical instructions needed to generate the English sentences out of the Russian. To refer again to Harris' conceptual framework of transfer grammar, we use the criterion of translation as equivalence, and postulate a transfer relation between each sentence of Russian and its English translation, and then construct transfers between paired items within the sentence. Detailed examples of such pairs are given in a separate Georgetown Work Paper on MT. (MT-38)

The English synthesis part of the research is based on the syntactic theory of the construction, transformation and kernel, in that all sentence structures are combinations and/or transformations of just a few simple sentence structures, the kernels of the grammar. We are particularly concerned with non-linear transfer, that is, cases in which a given Russian unit cannot be translated directly or component-by-component into English. The problem is to establish and describe the regularities in the transforming operations needed to obtain the proper transfer for all Russian structural units.

The sheet which you were given in connection with Mr. Zarechnak's paper shows the basic comparison of Russian and English kernels. Languages are in general much more similar to each other in their kernel sentences than in their final resultant sentences.
The factorization of Russian chemical discourse and its English translation into kernels and transformations has enabled us to establish the regularities of insertion and rearrangement operations necessary for English in contrast to the Russian original.

It is interesting to note that the kernel analysis of a given Russian sentence is remarkably similar to that of its English translation. Where the languages differ is largely in respect to the transformations employed, just as this factor causes vast stylistic variation within one language itself.

The approach in question here, however, is not the mechanical factorization and translation of kernels (although this has wide interest for information retrieval procedures) but rather the transfer of Russian structure in its original transformed (or put-together) state into the corresponding English translation.

Machine operations depend on the sensing or delimitation of translation units in the source language and their transfer into the target. Translation units are describable at three levels: lexical, syntagmatic, and syntactic. On the lexical level a translation unit may be a single morpheme or several words. On the syntagmatic (or sub-sentence) level a translation unit may be a suffix function or a prepositional phrase and the like. On the syntactic level there are only two basic translation units, the noun-phrase and the verb-phrase.

Any translation unit may be subject to selection and/or arrangement, which in the transfer procedure means the following three operations may be involved in translating any translation unit: 1) choice between positional variants 2) insertion and 3) rearrangement. All translation units must be delimited by the machine, which, in view of the fact that they overlap—because they are successively included—presents considerable but not insurmountable problems. The lexical and syntagmatic translation units may consist of one unit sensed (a word between spaces) or part of one or more than one. The search area for delimiting these units rarely constitutes the complete stretch of input, the sentence. Conversely, the syntactic translation unit is delimited by examining a search area including the entire input, the complete sentence including punctuation.

See Z. S. Harris, "Transformations Manual", mimeographed booklet, not yet published. (University of Pennsylvania)

---

See Z. S. Harris, "Transformations Manual", mimeographed booklet, not yet published. (University of Pennsylvania)
We have concentrated to date mostly on the problem of the sensing, delimitation and transfer of syntactic translation units, and have arrived at a procedure for machine transfer (at least on paper) of the basic structure or kernel of any Russian sentence into English.

The procedure followed in the research was to reduce to symbolic formulas a very large sample of Russian chemical discourse and its English translation. Compression techniques were applied to each sentence, that is, a method was devised to express any sentence type in terms of its particular transfer features. Thus a given HP relation in Russian is transferred into English structure according to a specific operation. Although the scope of this paper does not permit a detailed explanation of the techniques of Russian analysis and English synthesis, a brief summary of the transfer syntax procedure would seem to be pertinent at this point. The procedure involves a cut of each minimal Russian sentence into two parts, verbal and nominal. These are in turn handled first in terms of their head words, the H and P respectively, and then the rest of the components fall into a string-type operation. Altogether there are three basic transfer instruction operations: H, P and S. H operation extracts the head of the noun-phrase, and P operation the head of the verb-phrase. Their relative morphological composition and order determines the particular transfer instruction for English structure. The S operation (meaning string, or chain) completes the transfer and is directly dependent on the first two. Locating the H and P is necessary for the delimitation of the unit boundaries.

We have formulated search sequences for extracting the H and P which are the structural nuclei, and the next crucial step is to discover the quantitative relation of H and P, namely, to discover how many H's and P's are in the stretch. The result of this tally determines the particular type of operation to be employed for cutting the stretch into its components and for performing the transfer instructions. The standard, or nominal sentence in Russian has one noun-phrase and one verb-phrase, each having a head word whose relative position and morphology determine the transfer instruction for a given structural type. This we label a 1-1 type. There may be no H and/or no P, and these cases take a particular transfer instruction procedure. Where there is more than one H and P, (called a 2-2 type) the search types do not (and note that this is a departure from the usual classification) follow the sentence groupings of simple, complex, dependent, etc. Instead, as a result of the gathering of a large corpus and of extensive testing, we found that the crucial
feature separating one search type from another was simply the relative number of H's and P's. Only after a 2-2 type is discovered, for example, is it necessary to separate conjoined from complex structure.

The complete series of questions for the component distribution search (the title we give to the process of finding out how many H's and P's are found in the stretch) and all the series of operations for the various search types and transfer instructions, will be presented in a forthcoming Seminar Work Paper.

In conclusion, however, I will touch briefly on the types of questions which are "asked" in sequence to discover the number of head words, namely nouns in the nominative case and verbal forms or their substitutes.

Punctuation helps to mark structural breaks. Separaters, such as ; : , function as positional signals and enable one to make cuts within which a certain search should be initiated. For example, the search stops immediately at a semi-colon. Delimiters mark off inserted structures. I refer to parenthesis, brackets, quotations and the like. These have no positional function in terms of extracting the H and P components, and whatever is found to be an inserted structure is functionally another and separate search area.

Nouns, pronouns, and numerals in the nominative case function as H. They can be pulled out simultaneously, whereas verbal forms or their substitutes must be looked for in sequential order, namely: verbs with person markers first, then short participles or short adjectives, then full adjectives or adverbs which have predicative function.

Next comes the count of H's and P's, and then the corresponding operation, either 0-1, 0-0, 1-1, 1-2, 2-2 and so forth. These differ markedly in the search sequences applied to make the cut between the noun-phrase and verb-phrase, as well as in the instructions for transfer. The first problem is to discover the structural type, as in 0-1, where H is lacking, which must be separated for transfer according to whether it translates as NV or simply as V. 1-0, on the other hand, where P is lacking is not subject to any modification beyond the syntagmatic level. 1-2 may be either NV plus NV (conjoined sentences) or NV plus V. 1-1 is perhaps the most basic transfer operation, representing the minimal sentence in Russian. It is basic because many more-than-one types are reducible to 1-1, whether
they turn out to be conjoined or complex structures. After the structural type is determined, the instructions for rearrangement and insertion go into operation.

Again let me apologize for the rather summary character of these remarks by referring to our forthcoming work paper on the transfer syntax procedure, which will clarify and amplify these statements.
The Use Of SEAC In Syntactic Analysis
Richard B. Thomas
National Bureau of Standards

The purpose of this structure search is to determine whether the syntactical patterns of English sentences, expressed symbolically, show that a relatively small number of such patterns represents a significantly large number of sentences.

The routines written for SEAC (National Bureau of Standards Electronic Automatic Computer) examine the structure of sentences within a corpus of expository prose chosen from scientific and technical writings. SEAC performs several functions in the search: (a) accurate high-speed tabulation; (b) precise comparison of data; (c) compression of coded data in terms of syntactical equivalence relationships. Preparation of the material includes coding the elements of the source sentences functionally according to the scheme shown in Figure One. This is but one of many feasible analytical schemes, and it represents a rather gross cut made according to traditional grammar. Each functional element to be coded is given a notation comprising one number and one letter, as shown in Figure Two. The notations (maximum 9) of each sentence then become SEAC input.

Primary Search

Except for tabulation, the program rejects the letters in the codes. (The letters are included in the raw data so that they may be available for other programs.) The computer reduces the first incoming sentence code to its numerical pattern (e.g., 56416500000) and stores the pattern. Each subsequent candidate sentence is reduced in the same way and its numerical pattern is compared with all stored patterns. If a candidate pattern is identical with a stored pattern, a recurrence tally of 1 is added into the least-significant place of the stored pattern and the candidate pattern is rejected. But if the candidate pattern proves to be unique, it is stored along with the others. Check routines are included in the programming of this preliminary phase to reject data erroneously prepared.

When all sentences have been processed, the resultant unique primary patterns, with their tallies, emerge from SEAC via high-speed magnetic wire, along with tallies showing (a) the number of sentences processed, (b) the number of patterns stored as unique, (c) the number of sentences rejected because of errors in preparation or inscription, and (d) the number of patterns having 1 digit, 2 digits, ....9 digits.
Each time 50 sentences have been processed, SEAC prints out the number of unique patterns being held in storage.

**Compressed Search**

The unique primary patterns are then fed back into the computer for compression and a 1 is added to every pattern tally so that each tally will show the actual number of occurrences. The routine first rejects all-but-one of any digit that is contiguously repeated within a pattern. For example, 444166655 becomes 4165, but 414656 remains unchanged. The basis for compression in this manner is the assertion of equivalence relationships whereby "The little red hen clucks" (44416) is here considered to function syntactically as "The hen clucks" (416) or "The hen will cluck" (4166), etc. A compressed pattern is thus construed as a basic form from which all corresponding primary patterns could be developed by regular structural transformation.

Since a number of patterns whose primary forms were different would probably be identical in their compressed form, the compressed patterns are then compared, each with all others. When SEAC finds that two compressed patterns are identical, their respective tallies are added together; the sum is stored in the least-significant places of the first pattern in question and the other pattern is cleared to zeros. At the end of this operation, the unique compressed patterns with (some new) tallies are printed out, together with (a) the number of compressed patterns stored as unique and (b) the number of unique compressed patterns having 1 digit, 2 digits, ..., 9 digits. Separate routines are then employed to list the patterns in numerical order, as shown in Figure Three.

**Results of the Search**

The original corpus of 1002 sentences is a very small sample. The curve of cumulative occurrences (Figure Seven) shows little tendency to reach zero slope, as unique primary patterns were still occurring at a nearly fixed rate. The distribution of patterns according to the number of digits (notations) per pattern is shown in Figure Four. The components of the raw and processed data are shown in Figure Five. Figures Six and Seven express the rate of occurrence of new patterns.

The 1002 sentences yielded 541 unique primary patterns. Of these, the five most common are listed here in sequence, followed by the number of occurrences (in parentheses) and a sample sentence for each:
The dog has run across the street.
The dog with floppy ears has run across the street.
Dogs eat bones.
Fido is the dog with floppy ears.
Fido ran across the street.
Fido has run across the street.

The total number of sentences (95) shared by these patterns is less than 10 percent of the corpus.

Compression of the 541 primaries yielded 247 unique compressed patterns. The five most common are listed here, as above; the parenthetical number is the number of original sentences represented by the compressed form:

4165 (80) The dog ran across the street.
165 (62) Fido ran across the street.
41465 (56) The dog with floppy ears ran across the street.
54165 (35) Finally the dog ran across the street.
416424 (28) The dog ate the bone which he had dug up.

The total number of sentences (261) shared by these compressed patterns represents only 26 percent of the corpus. The first four compressed patterns listed above contain at least one adverbial element in every case and lack objects or predicative nominatives; these patterns represent 233 sentences or 23 percent of the corpus.

The results must be viewed as specifically inconclusive because the corpus is small. But the technique of compression appears valid and useful for examining possible "base" or "kernel" forms of syntax.

The computer is admirably suited to this type of search. To duplicate the search on other types of equipment would require considerably more time and more complex operations, especially in the comparison and compression phases. The total SEAC running time for this program was about 40 minutes.
FIGURE ONE. Coding Scheme for Syntactic Analysis

<table>
<thead>
<tr>
<th>LEXICAL UNITS</th>
<th>PHRASES</th>
<th>DEPENDENT CLAUSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) 1, 2, 3 - Noun, Pronoun</td>
<td>Noun { Infinitive, Gerundive, Prepositional }</td>
<td>Noun</td>
</tr>
<tr>
<td>1 A - Subject</td>
<td>1 B - Subject</td>
<td>1 C - Subject</td>
</tr>
<tr>
<td>2 A - Object (d.&amp;i.)</td>
<td>2 B - Object (d.&amp;i.)</td>
<td>2 C - Object (d.&amp;i.)</td>
</tr>
<tr>
<td>3 A - Pred. Nom.</td>
<td>3 B - Pred. Nom.</td>
<td>3 C - Pred. Nom.</td>
</tr>
<tr>
<td>(B) 4 A - Adjective</td>
<td>4 B - Adj. { Infinitive, Participial, Prepositional }</td>
<td>4 C - Adjective</td>
</tr>
<tr>
<td>5 A - Adverb</td>
<td>5 B - Adverb { Infin., Prep. }</td>
<td>5 C - Adverb</td>
</tr>
<tr>
<td>6 A - Main verb</td>
<td>6 B - Auxiliary</td>
<td>6 C - Modal Auxil.</td>
</tr>
</tbody>
</table>

NOT CODED: Connectives (relative pronouns, coordinating conjunctions, subordinating conjunctions, conjunctive adverbs); absolutes; appositives; interjections; non-functional expletives; internal structure of phrases and dependent clauses; and elements which modify portions of "B" or "C" structures (except verbs).

Such structures have been omitted from the coding because they do not affect the basic structure of the independent clause.
FIGURE TWO. SAMPLES OF CODES AND PATTERNS

5-A 6-B 4-A 1-A 6-A 5-A
1. When does the balloon go up? 5A6B4A1A6A0
--- 2-C --- 1-A 6-A
2. Whatever Lola wants, Lola gets. 2C1A6A00000-
1-A 6-C 6-B 6-B 6-A 5-A 5-A
3. It could have been solved more simply. 1A6C6B6B6A0
6-A 4-A 2-A 5-A
4. Polly, put the kettle on. 6A4A2A5A000-
4-A 4-A 4-A 4-A 1-A
5. The electronic automatic digital computer 4A4A4A4A1A0
   dropped a bit. 6A4A2A00000-

Primary patterns formed from the foregoing sentences:

1. 5 6 4 1 6 5 0 0 0 0 0
2. 2 1 6 0 0 0 0 0 0 0
3. 1 6 6 6 6 5 0 0 0 0
4. 6 4 2 5 0 0 0 0 0 0
5. 4 4 4 4 1 6 4 2 0 0 0

Compressed patterns formed from the primary patterns:

1. 5 6 4 1 6 5 0 0 0 0 0
2. 2 1 6 0 0 0 0 0 0 0
3. 1 6 5 0 0 0 0 0 0 0
4. 6 4 2 5 0 0 0 0 0 0
5. 4 1 6 4 2 0 0 0 0 0

(SEAC Input)
**FIGURE THREE**  
Unique Compressed Patterns - Listed - Last Two Digit Places

Show Number of Occurrences

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Number of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>16000000</td>
<td>14</td>
</tr>
<tr>
<td>62000000</td>
<td>09</td>
</tr>
<tr>
<td>65000000</td>
<td>01</td>
</tr>
<tr>
<td>51460000</td>
<td>01</td>
</tr>
<tr>
<td>14600000</td>
<td>02</td>
</tr>
<tr>
<td>16200000</td>
<td>23</td>
</tr>
<tr>
<td>16300000</td>
<td>07</td>
</tr>
<tr>
<td>16400000</td>
<td>11</td>
</tr>
<tr>
<td>16500000</td>
<td>62</td>
</tr>
<tr>
<td>21600000</td>
<td>02</td>
</tr>
<tr>
<td>41600000</td>
<td>03</td>
</tr>
<tr>
<td>51600000</td>
<td>06</td>
</tr>
<tr>
<td>56200000</td>
<td>01</td>
</tr>
<tr>
<td>61400000</td>
<td>02</td>
</tr>
<tr>
<td>62500000</td>
<td>06</td>
</tr>
<tr>
<td>64100000</td>
<td>02</td>
</tr>
<tr>
<td>64200000</td>
<td>01</td>
</tr>
<tr>
<td>65100000</td>
<td>01</td>
</tr>
<tr>
<td>65420000</td>
<td>01</td>
</tr>
<tr>
<td>14250000</td>
<td>01</td>
</tr>
<tr>
<td>14630000</td>
<td>02</td>
</tr>
<tr>
<td>14650000</td>
<td>09</td>
</tr>
<tr>
<td>15620000</td>
<td>02</td>
</tr>
<tr>
<td>15650000</td>
<td>03</td>
</tr>
<tr>
<td>16240000</td>
<td>04</td>
</tr>
<tr>
<td>16250000</td>
<td>05</td>
</tr>
<tr>
<td>16340000</td>
<td>02</td>
</tr>
<tr>
<td>16350000</td>
<td>01</td>
</tr>
<tr>
<td>16420000</td>
<td>07</td>
</tr>
<tr>
<td>16430000</td>
<td>07</td>
</tr>
<tr>
<td>16450000</td>
<td>06</td>
</tr>
<tr>
<td>16520000</td>
<td>03</td>
</tr>
<tr>
<td>16530000</td>
<td>01</td>
</tr>
<tr>
<td>16540000</td>
<td>05</td>
</tr>
<tr>
<td>16560000</td>
<td>04</td>
</tr>
<tr>
<td>21620000</td>
<td>01</td>
</tr>
<tr>
<td>41450000</td>
<td>01</td>
</tr>
<tr>
<td>41460000</td>
<td>03</td>
</tr>
<tr>
<td>41620000</td>
<td>05</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>41456420 01</td>
<td>41565650 01</td>
</tr>
<tr>
<td>41456540 01</td>
<td>41642450 01</td>
</tr>
<tr>
<td>41456560 01</td>
<td>41645420 01</td>
</tr>
<tr>
<td>41462450 01</td>
<td>41653450 01</td>
</tr>
<tr>
<td>41464240 11</td>
<td>41654240 01</td>
</tr>
<tr>
<td>41464250 02</td>
<td>41654340 02</td>
</tr>
<tr>
<td>41464340 07</td>
<td>41656420 02</td>
</tr>
<tr>
<td>41465450 02</td>
<td>41656430 01</td>
</tr>
<tr>
<td>41465640 01</td>
<td>45436410 01</td>
</tr>
<tr>
<td>41465650 07</td>
<td>51456540 01</td>
</tr>
<tr>
<td>41562540 01</td>
<td>51465450 01</td>
</tr>
<tr>
<td>41564240 01</td>
<td>51654240 01</td>
</tr>
</tbody>
</table>
FIGURE FOUR. Distribution of Patterns according to Number of Digits

Number of Digits

1 2 3 4 5 6 7 8 9

Number of Patterns

UNIQUE PRIMARY PATTERNS

Number of Digits

1 2 3 4 5 6 7 8 9

Number of Patterns

UNIQUE COMPRESSED PATTERNS
FIGURE FIVE. Components of Raw and Processed Data

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>970</td>
<td>10</td>
<td>15</td>
<td>995</td>
</tr>
<tr>
<td>(2)</td>
<td>279</td>
<td>23</td>
<td>48</td>
<td>350</td>
</tr>
<tr>
<td>(3)</td>
<td>98</td>
<td>9</td>
<td>12</td>
<td>119</td>
</tr>
<tr>
<td>(4)</td>
<td>1368</td>
<td>435</td>
<td>54</td>
<td>1857</td>
</tr>
<tr>
<td>(5)</td>
<td>305</td>
<td>645</td>
<td>147</td>
<td>1097</td>
</tr>
<tr>
<td>(6)</td>
<td>1002</td>
<td>432</td>
<td>154</td>
<td>1588</td>
</tr>
<tr>
<td>Total</td>
<td>4022</td>
<td>1554</td>
<td>430</td>
<td></td>
</tr>
</tbody>
</table>

COMPONENTS OF RAW DATA

Components of Unique Primary Patterns:
(1) 529   (4) 1162
(2) 211   (5) 676
(3) 69    (6) 844

Components of Unique Compressed Patterns:
(1) 243   (4) 414
(2) 112   (5) 281
(3) 39    (6) 300
FIGURE SIX. Percentage of Unique Primary Patterns in Each Increment of 100 Sentences Processed
FIGURE SEVEN. Cumulative Number of Unique Primary Patterns.
Programming Aspects Of MT
Dan A. Belmore
Consultant Programmer
Institute of Languages and Linguistics

The general steps in MT research prior to programming might be outlined thus:

Step 1: Linguistic analysis of the source and target languages.

Step 2: Expository description of behavior patterns of linguistic components of the languages.

Step 3: Description by symbolic logic statements (Formulation).

For example:

Source sentence: стакан царапает стол.

From glossary: glass scratch table

From Russian "stem" diacritics: inanim. transitive inanim. noun verb noun

From inflectional endings: nom. or 3rd sing. nom. or acc. present acc.

Expository description: If in a Russian sentence the "current" item is an inanimate noun having an ambiguous nom. or acc. inflectional ending and the "immediately subsequent to current" item is a transitive verb and the item immediately following this verb is an inanimate noun having ambiguous nom. or acc. inflectional ending, then the "current" item is nominative (i.e., the subject) and the item following the "immediately subsequent to current" item is accusative (i.e., the object).

Description by symbolic logic:
Let the symbol: "aa" mean "the item is a noun"
"ab" mean "the item is a pronoun"
"ac" mean "the item is animate"
"ad" mean "the item is nominative"
"ae" mean "the item is accusative"
"af" mean "the item is a verb"
Let the symbol: 

- "ag" mean "the item is a transitive"
- "i* signify the i^th item in the sentence (where "i* can have the values 1, 2, ..., n; where "n" is the number of items in the sentence.)
- "." mean "and"
- "v" mean "or"
- "⊃" mean "if . . . then"
- "¬" mean "not"

Then the symbolic logic statement is:

\[
[a_\text{a}, a_\text{c} . (a\text{d} \lor a\text{e})]_{i} \cdot (a_{f} \cdot a_{g})_{i+1} \supset a_{d_{i}} \cdot a_{e_{i+2}}
\]

The flow-chart which follows is intended to show in a very general and very simplified form some of the proposals of contributors to MT and how such proposals might fit into an overall plan of MT.

The outputs of the Key-Punch (at the top of the chart) are labeled in parentheses with subscripts "H". Information so labeled is in a coded form corresponding to a pattern of punched holes. As the information emerges from the subsequent programs it is in a binary coded form.

Two types of glossary entries have been proposed.

Non-split entry:
Each paradigm form with its handling cue (diacritic) and with its English equivalent would constitute one entry.

Split entry:
The Russian "stem" with its handling cue (diacritic) and with its English equivalent would constitute one entry.

The same outputs of a "Look-Up" program are possible whether the glossary contains only split entries or only non-split entries or assorted split and non-split entries.
input (Russian corpus) → KEY-PUNCH (Russian) → MECHANICAL ASSEMBLY PROGRAM → MECHANICAL PRELIMINARIES PROGRAM (Russian) → LOOK-UP PROGRAM (suffix) (diacritic) (stem) → ALTERNATIVE "A" PROGRAM (English translation) → MECHANICAL TERMINARIES PROGRAM (English translation) → PRINTER output (English translation)
More details of Alternative "A" program:

First step:
Test for Russian words with suffix "-a"; rearrange and/or resolve lexical ambiguities and idioms and translate.

Subsequent steps:
Test for Russian words with the other suffixes and proceed as above.

More details of Alternative "B" program:

"Idiom" recognizer

Russian Grammatical Ambiguity Resolver

Identification of Russian Structure Type
(Russian Syntax Analysis)

Rearrangement for English Structure
(English Syntax Synthesis)

Lexical Ambiguity Resolver

English Grammatical Identifier

Look-Up
(includes English morphological synthesis)
DISCUSSION — Panel III, Saturday morning

OSWALD: May I ask, Mr. Belmore, what the speed of the key punch operation is?

BELMORE: This is done gradually. It is similar to a typewriter. The material to be translated could be punched days before actually putting it on the machine. All the punched material for one Russian book or journal could be put on the machine and then run in a few hours.

OSWALD: Have I called the correct time on how long it’s going to take to do the key punch, whoever does it? 60-75 w.p.m.?

BELMORE: It would depend upon the speed of the typist. It is strictly a manual operation.

GARVIN: The key punch will not necessarily remain part of this operation. It could be replaced by one of these reading instruments when perfected.

BELMORE: Exactly. If we use a reading instrument of the type mentioned yesterday,....when perfected.

MUELLER: I have a question to ask Mr. Lehmann. I would like you to elaborate a little on these categories. There are limiting adjectives and descriptive adjectives based on analyzing their formal behavior. We should list them somewhere else. Under II-B and II-C you have "sein Dunkles" and "das recht Dunkle". You have there the first item, the "sein" and "das" listed as limiting adjectives, but they behave quite differently. Why don’t you categorize them differently?

LEHMANN: Well, the term "limitation" is one that would include both the DE and the KE. Let us take that as a cover term.

MUELLER: But then you have to subdivide them. Somehow or other you have to indicate that one is a limiter that conditions a secondary meaning and the other is a limiter that does not.

LEHMANN: This may be a bit too free, but that is what was done up under I-B. Then you go over to a, then you go over to alpha and beta. That same subdivision is presumed. You see, a is limiting, then alpha and beta are the DE and the KE words, and that would
be assumed to carry all the way through the chart. The chart would become rather long, of course.

MUELLER: I see now. It wasn’t completely clear to me from your listing here.

BUCKLAND (USAF): About this question of print reading and the idea that we shouldn’t worry about it now and that we won’t need it until we have a translator. I’m not sure I go along with this, but I think that you linguists sooner or later are going to have to use machines to look through a lot of running text and I don’t feel that looking through a million words of running text is out of the question. However, if you exclude the factor of speed, key punching 1,000,000 words of text costs about $200,000. I was wondering if there wouldn’t be some possibility of using monotype or some of these foreign type-setting machines to provide this source of data. I take it that monotype is something similar to linotype machines except that they produce this type from punched tape, and then it would be simply a matter of converting this into code, into a form suitable for use with our own machine. Does anyone know anything about this?

THOMAS: I believe McGraw-Hill make paper punch tapes of all the materials they print. This is for their own use. It would be very nice if everybody did this. It would solve some problems.

KIRSCH (Bureau of Standards): I think it is important to emphasize here that in going to the trouble of preparing a lot of data in mechanical form, — in other words, in typing a lot of data, — it costs no more to type it in such a form that it is ultimately usable by a computer— that is, typed and produced on punched paper tape copy at the same time, — and I think that the concern over what form this data should be in punched paper tape form, needn’t worry you for the time being, because when you finally have something ready that you can feed into a machine, this large corpus on punched tape will be usable by most machines, or certainly convertible by mechanical procedures to be used by large machines. Consequently, I think that Mr. Buckland’s point that it is worthwhile to prepare this data in such a form that it is feedable into some machine eventually is a particularly important one, — one that will permit considerable saving.

PAPER: I would like to bring up a small terminological point with respect to Mr. Zarechnak’s paper. I wonder if a term something like “morphographemic” wouldn’t be better than “morphophonemic”, since
happily at this stage we are not interested in spoken input, but rather simply printed input. Of course, the morphographemic shape reflects a morphophonemic difference.

ZARECHNAK: As you know, I have used morphographemic in all my papers, but only in one place one has to be acquainted with the morphophonemic processes in Russian to qualify and classify properly morphographemic symbols. And there you have to be sure you are doing it by morphophonemic processes before you classify your morphographemic signals. I agree with your statement.

PAPER: There is one general question I have. I wonder if the people who have been involved in MT research have come up with a specific definition of the term "idiom" for their purposes?

DOSTERT: I think I can formulate one which may be acceptable. An idiom is a combination of separate words, each of which loses its individual signification in the context. For example, in "right away", in which "right" and "away" no longer mean "right" as "right" and "away" as "away", then each of the components has lost its original signification to acquire a new one when combined. Now, the number of items that can be so combined is variable, obviously. When in French we say "tout à l'heure" or "tout de suite", we don't think "tout", "à", "l'" and "heure", we think "tout à l'heure" as a single sign signifying a single concept. Does Mr. Joos agree with that?

JOOS: Yes. That is a perfectly reasonable way to start the discussion. One wonders just how one can draw the line between the point when we are going to know that we have finished finding all the idioms. I mean to say, how are you going to draw the line between idiom and construction?

GARVIN: I will mount my positivist horse and say that whenever I can translate an item in a context individually, it's not an idiom. If the translation of each individual piece gives us gibberish, it is an idiom. And then you will find out how many idioms you have when you have covered a large enough corpus.

JOOS: You could put it this way then: An idiom would be whatever your machine is not prepared to handle by its primary routines.

DOSTERT: We will have to store idioms as lexical units. In other words, you will have to take the complex which we call an idiom and
store it as a single unit, giving it its 'idiomatic' equivalent in the largest language and handling the whole as a single lexical manipulation rather than an analytical one. I think that has been fairly well agreed.

**JOOS:** I think it's a perfectly sensible engineering approach. If you want to give time to a linguistic discussion, you could find out an opposite possibility. You could say that it is possible to define an idiom within one language at a time; only then you won't find so very many. For English I have found very few. For example: "Get your own breakfast", which doesn't mean "get your own breakfast", but get it without help.

**DOSTERT:** In a case like that, I think you would have to cue the word "own" when it operates in the manner you have indicated.

**JOOS:** The way it operates in this particular idiom is in contradiction to the general rules of English grammar. Here is another contradiction: "I can't seem to find it". According to the rules, that ought to mean, "I can't give the appearance of finding it". We have here a syntactic-semantic shift resulting in another thing that I would call an idiom without comparison with another language. However, once you do make that kind of survey and find these idioms in English like the two I have just mentioned, you will find that you have only a tiny fraction of what needs to be taken care of as idiom in a translation program.

**YNGVE:** I would like to say that the papers by Mr. Zarechnak and Miss Pyne come very close to the approach that we have been using for German. What they are apparently trying to do for Russian is very nearly the same thing that we're trying to do for German. Naturally, I agree wholeheartedly with this.

**DOSTERT:** In this connection, I would like to address a question to Mr. Belmore. He has put down as Step No. 1 on Page 1 of his handout the linguistic analysis of the source and target languages. Now, that is a very broad statement indeed. I have struggled with the idea that it may not be necessary to go through a complete analysis of the source language as such and of the target language as such, on the theory that there are some parts of the analysis that are not pertinent to the translation process. It seems to me that there is an area of analysis which you might call your "transfer area" between your S and T languages. The question is, before you can focus on the transfer, must you do an exhaustive analysis of S and then an exhaustive
analysis of T, and then look for the transfer pattern? That seems to be the trend with your group. Mr. Yngve, we would welcome comments on this, because it seems to be rather basic to the whole problem of MT.

JOOS: I wouldn't mind commenting on that. I think that if you do it in the MIT fashion, analyzing the two languages in question exhaustively first, you will find that for your MT purposes most of your results are trivial. I think that if you seriously want to do this as an engineering proposition, the empiric approach of Georgetown looks more reasonable to me, and should not be very disappointing to a linguist either, because by this empirical approach he will now get the complexities of each language in order of their importance, from at least one viewpoint. It may be that if he does this with a number of languages he will get them in order of their importance generally, which would be a very nice thing for a linguist to know.

YNGVE: I wouldn't want to be pushed all the way to the very wall and say we are going to do an exhaustive—completely exhaustive—grammar of English and German first, but I will say this much, we certainly ought to have a more exhaustive analysis of German and English than we have at present before we go and look at the specific coding problems.

DOSTERT: Isn't it also not only a matter of the extent of the analysis, but also the type of analysis which is specifically suited for the transfer process? It seems to me that the formulation of the results of your analysis should constantly bear in mind that you are faced with the problem of meaning transfer between one set of signs and the other, and you are interested in the internal structure of each set of signs only to the extent that they relate to the job of transfer.

YNGVE: Yes, but even this we don't have yet.

DOSTERT: You mean that we don't have enough now to start looking at the transfer?

YNGVE: No, I'm talking about the structure of English or German or Russian as of now.

DOSTERT: Well, we can do it with segments, and out of the segment analysis and the cumulative results we can arrive at broader formulations.
PYNE: I would like to ask Mr. Joos if in referring to the empirical approach at Georgetown he referred to an approach to the languages in terms of their relationships with each other. That is, English as viewed only in terms of its difference or similarity to Russian?

JOOS: When I said empirical I meant to use the word in the sense that Mr. Dostert introduced it in yesterday. Namely, you work out a routine for translating the first sentence that you randomly encounter from one language into another, then you work out your routine for the second sentence and usually you're lucky enough to be able to use some of what you used in the first, and so on, thus gradually accumulating notations, procedures, etc. That is what I meant by the empirical approach. I think that's what Mr. Dostert meant.

DOSTERT: That is one of the techniques that one of our co-workers has developed, working on French. He took a corpus of chemical text in French and worked on it sentence by sentence. In other words, his idea was eventually to look at the forest by looking at enough trees, one after the other.

JOOS: I think I'd like to add another comment here on the other side. It may well be that the result of this kind of empiric approach, if it ever does result in a successful machine translation, will be a set of routines so complex that they will cry aloud for simplification by symbolic logic methods, and they may have gotten it so complicated that it would take much labor to simplify them.

OSWALD: I just wanted to go on record as agreeing with the statement you made, Mr. Dostert. There are three kinds of analysis involved, really. There is an analysis of English, let's say, as a source language; an analysis of, let's say, German as a target language; and the third body of information concerns entirely the relation of these two languages and constitutes a new realm of discourse, in which the formulations can be quite different from those normally used to analyze the languages as such. This I take it is what we are all agreed upon as meant by the empiric approach. It does constitute a new sort of formulation, a new series of statements, in a quite different realm from the normal type of linguistic analysis.

DOSTERT: I don't think, Mr. Oswald, it means exactly this, and I wouldn't try to put the tag "empirical approach" on the process I tried to describe and with which you agree. That I would call the analysis of the transfer pattern, rather than the empirical. Now let me reassure
Mr. Joos. Our group is working with three different techniques, the empirical, that is the one you described; the other is Mr. Garvin’s method where he tries to find through an analytical process the formal cues within a segment and tries to develop a theory from an analysis of a given number of sentences. It is not as empirical as the one of Mr. Brown. It is more a combination of the analytical and empirical, with a focus on formal cues. Then the third is an attempt to formulate a rather broad theory—that is what Mr. Zarechnak, Mr. Pacak and Miss Pyne are doing—and then to actually test it on a text. A fourth group has followed what we have called the internal coding system, or internal coding plus bilingual matching of code to arrive at translation. One word of clarification is needed. As evidenced by the paper given by Mr. Belmore, we are trying as we formulate our linguistic steps, or our steps in terms of linguistic analysis, to bring the discipline of symbolic analysis to their rendition so that we will not run the danger of being confused to a point beyond retrieval. The reformulation would become a very complicated task.

*GARVIN:* I just wanted to elaborate on this last point. I think that in the procedure of developing rules and getting a large body of rules there will have to be intermittent periods of stock-taking and compression. What I mean is that, for instance, at the present time we find that in our glossary there are recurrent sets of entries which have each the same ten diacritics. In a case of that sort, provided we have a large enough body to do this with a degree of statistical and other validity, you can then replace this set of ten diacritics with a single one thus to reduce the problem of size of special entries. Another thing I visualize as one of the future operations in compressing the corollary codes is that equivalent steps in various rules can be spelled out more simply and by a simple reference, rather than to have to start out each time with the instructions. In other words, Rule 22, Instruction A, B, and C, and then Rule 23, Instruction C, B, and A, and so on down the line. This is another form of compression. The third thing is that we are already beginning to see that by itemizing one rule per situation we have a tremendous amount of redundancy because it turns out, for instance, that out of our 100 rules, a set of rules covering suffix translation, there are so far about a half a dozen of them which have exactly the same steps and exactly the same key entries. Now if this is borne out by a larger body of data, you could then take these six rules and lump them into one, and use the same set of diacritics for six times in these situations instead of six different sets of diacritics, and so on down the line. I think that if this is not
done as you go along, then confusion will result. I agree that some awareness of this is present.

LEHMANN: It seems to me that Mr. Thomas' paper illustrated very well that we will need a better analysis of English such as Mr. Yngve suggested before. If, instead of using the analysis you did, you would begin with a linguistic analysis based on immediate constituents, I think you would get far fewer types of English, and I think you indicated that in your talk as well.

ZARECHNAK: We are approaching here the same problem on a different level. Still we have one common denominator, namely, we do not forget the data which we are facing. We believe that if the same data are approached from different points of view, we would get information which we would mutually exploit later on. One comment more: When we are explaining or taking some attitude, we never can forget transfer from source to target; this transfer is actually information on two codes, and specifically what is common in between. Here is the place where we so appreciate cooperation with programmers.

DOSTERT: Yes, the fact that after all we are all working with the same data means that there will have to be a measure of concordance in the results, even though in the initial phase there seems to be considerable divergence. The fact that the data with which we are working are finite and systematic is another encouraging factor. Another thing I think that should be pointed out is that languages of relatively similar cultures and of the same family present much less difficulty than languages of different families and of widely divergent cultures. Since at the beginning we are focussing our efforts on languages in relatively similar cultures and relatively similar systems, we are trying to tackle the more modest problems before the big ones.

I had asked Mr. Crossland yesterday if he would not be willing to take the floor this morning to explain to us the method and objectives of the Cambridge Language Research Unit. Mr. Oswald took Mrs. Masterman somewhat to task yesterday for her formulation. I understand that a further exchange of views between Mr. Oswald and Mr. Crossland rectified what may have been originally an erroneous impression. I would not want the group to be uninformed about the Cambridge Unit's efforts, and since Mr. Crossland is with us, I would like to ask him if he would come up here and tell us about some of the things they are doing.
CROSSLAND: I don’t think I wish to give a long statement and explanation about what we’re doing, because I think that was well put forward at the MIT conference in October. I did feel, though, that Mr. Oswald had considerable uneasiness about Mrs. Braithwaite’s close reliance on a thesaurus, and for the sake of informing this meeting, I thought I would mention our method in a broader sense. That is, we are prepared to deal with specialized vocabularies, idioglossaries if you wish to call them so, though perhaps we think that is a little too precise a term, to explain that we have to cut up the vocabulary into technical vocabularies, but we are certainly prepared to play with the idea of specialized vocabularies, although we now think the thesaurus method is the most promising for solving this problem of language and exact semantic correspondence between words in different languages. Another point I wanted to make is that our experience has not been that the amount of general vocabulary of ambiguous terms is very wide, from random experience on each side. We are at the moment working with restricted language. We are trying out our technique on a certain amount of botanical literature in Italian. So in a sense we are doing perhaps what you suggested in an early stage, although relying on the thesaurus method, or giving it a try for solving problems of this sort for semantic correspondence.

OSWALD: It may be that Cambridge linguists working with Italian and botany have had a different experience from myself and Lawson working with German and brain surgery, because of the temperamental differences between the Italians and the Germans on the one hand, and also because of the fact that botany, I should think, probably lends itself to a more metaphorical way of expression than brain surgery. As Mr. Dostert has suggested, Mr. Crossland and I have had no difficulty in ironing out our differences. I would like to add that I spoke to Mr. King yesterday and asked him whether in his super-gigantic millenial computer he could dispose of idioglossaries—whether this meant that the whole principle of idioglossaries could not be discarded as obsolete and belonging to the primitive stages of our investigation—and he said “no”. He felt quite definitely that no matter how large his storage system would be, the idioglossaries would have to be in there in some form, and suggested only that the question of where they should go—he put it quite bluntly—we should leave that to the hardware people and just produce them and not worry about where they would go in a given system.

DOSTERT: The fact that they should be produced is not controverted at all.
AUSTIN: I'd like to ask Mr. Crossland if the thesaurus method he has in mind is anything like that outlined by Mr. Parker-Rhodes in a recent paper?

CROSSLAND: I would give a qualified yes to that.

AUSTIN: Well, there were two categories in this. As I remember there were around fifteen or twenty that he listed, and one of these categories was "objects", and another was "things to think about". Can you tell me any possible use for those categories?

CROSSLAND: I think it's a good thing to try out, that's all.

AUSTIN: Wouldn't this about cover the vocabulary of any language?

CROSSLAND: That could be taken in a broad sense, I suppose.

KING: As I understood that paper of Parker-Rhodes, this was only a trial idea more or less along the lines of "Twenty Questions", and how you organize a thesaurus. Well, we have Roget's, but that isn't necessarily directed toward mechanical translation.

GARVIN: I'd like to change the subject back to idioglossaries and their use in translation. I think that a compilation of idioglossaries is at least a necessary preparatory step. They will either have to be left each separate or lumped together in one big glossary depending on the particular storage method and what-have-you, that a given scheme contemplates. But I do think that this is merely a preparatory step because you still have the problem of choice in a text, namely, whether to go to a general glossary or to an idioglossary, and the problem of ambiguities within a technical terminology. The point is that even a chemist in a chemical article occasionally departs and uses a word from some other field or uses a chemical term in a non-chemical sense.

OSWALD: I was very careful to say yesterday that the totality of my experience with idioglossaries, which is not very great, nevertheless indicated an equivalence only up to about 80%. I haven't the faintest idea how you solve the other 20% of ambiguities, of unpredictable words, and brain surgeons suddenly using unpredictable adjectives. This clearly is something that has to be gone into and it is precisely in this area of choice, as Mr. Garvin calls it, that that 20% falls.
GARVIN: Well, I think there is another problem here. That is, whether or not the choices were necessary, because one of the things that came out in the very, very small piece that we have done so far, is that apparently you don't have to solve every single problem, in order to get a viable translation. And from the linguistic point of view, this relevance criterion which Joos brought up I think is tremendously significant to the whole problem. In the "transcendentalist" Prague School tradition, it used to be called functional load, and now "functional yield", according to Martinet. And I think that this is one way of getting at it in a purely numerical, pragmatic and precise way. At the same time, of course, we can make the decision on the following basis: if I don't bother to translate this properly, what will happen to the rest of the text? Will it be unintelligible, or will it be intelligible enough and elegant enough even though I have left out this property? For 23½% of the instances you would have to have a very extensive routine to resolve the ambiguity between ether and ester. After that Father Sohon came to me and said, "Well, this is really not very important anyhow, because as a chemist I have the impression that the two terms are roughly equivalent and that it is just a stylistic difference". Now if this is the case, then I will leave my 23½% sit there and put out a translation ETHER/ESTER or ETHER/ESTERS and not worry about it and say to myself that the labor I have saved might better be used for resolving an ambiguity without the resolution of which the machine is going to blow a fuse, and so on down the line. So that you can get down to these engineering criteria if you want, or you can use some other criterion of efficiency, and it's just a continuous process of decision-making in detail. Those little details each require a very specific decision. Now, for instance, you have eight possible translations of the suffix "oi". Now at this point we would theoretically include all eight of them and devise a routine for it, and it's extremely extensive and very complicated. I assume that after we have gone through 65 or 100 or 200 sentences, if we find that out of those eight translations, the one requiring the most complicated routine occurs once in a text of I-don't-know-what magnitude, then we might just drop this altogether, and say when we get to that particular point we will simply take translation A. Or we might say we have already developed a machine that reads fast, and it costs only three bits to leave this in there, so we leave it in. But this sort of continuous bread-and-butter decision-making, I think, is what has to be kept in mind in order to solve your problem of choice.

-177-
Practical Objectives

In

Machine Translation Research
Practical Objectives Of Machine Translation Research
Closing Luncheon Address
Léon Dostert

I think it is fitting that the sessions of our Eighth Annual Round Table Meeting on Linguistics and Language Studies—focused as they have been on the theoretical, linguistic, and technical aspects of machine translation—should now come to a close on a practical note. This topic, "Practical Objectives in Machine Translation Research", was to have been discussed by a staff member of the National Science Foundation, the institution to which we owe the financial support making our work possible. I am sorry to have to tell you that Dr. Alberto Thompson, Head of the Foundation's Office of Scientific Information, the speaker originally scheduled for this closing luncheon, has been hospitalized since last week. I wish him, in the name of the University and of all of us, prompt and complete recovery. I will do my best to deal with the subject on which he had planned to speak today.

It may well be that, after the scientific approach which has marked the presentation and discussion of the topics listed on our three panels, an attempt to come to more mundane and practical considerations will be looked upon by some of our participants and perhaps by certain of our future readers as being out of tune with the purpose of a gathering such as this. Still, clearly, we must be prepared to recognize that the financial support, without which the research of the increasingly numerous groups now seeking to formulate solutions to the problems of language translation by machine, will not be forthcoming unless our activities are in fact related to practical objectives.

One may indeed wonder whether there is such a thing as pure research. For after all, research is exploration—it seeks to push back ever farther the frontiers of the unknown. It is dedicated to the discovery and formulation of truths for the purpose of serving the ends of man. When the great explorers set out on their journeys into uncharted lands or seas, they may not have been motivated by the desire to claim for their sponsoring governments the new lands to be found. The late Admiral Byrd was, if you will, a pure explorer, but who will gainsay the practical purpose of his voyages in the icy polar seas? We have to live on our world and in our time. A century ago when Pasteur was working in his garret laboratory, with pitifully limited means, he was driven by the will to know, yes; but also, by the desire to protect his fellow man from disease. A century later, the disciplined research
which went into the discovery of antibiotics or Salk's vaccine had a practical end in view. And it is right that this should be true also among the day-to-day workers that we are in the field of machine translation, obscure though we remain. Unless we can and do relate the results of our work to some meaningful and useful end, we shall fall short of one of the true and valid objectives of research, which is to have a purpose—the purpose of serving the ends of man, of improving his lot as an individual or as a member of a cultural group, or of men as members of the related communities that make up the world we live in.

One may, then, quite properly ask, "What are the practical objectives of research in the field of translation of languages by machines?" There may be some who would answer that they are primarily, if not exclusively, interested in what may be called the linguistic results of the investigations—that is, in the inter-lexical, inter-morphological, and inter-syntactic relationships among the languages under scrutiny. Worthy as that objective is as part of the total purpose of our research, to identify it with the final results that are sought would be confusing the means with the end. Unquestionably, the study of the meaning transfer process between languages, between systems of signalization of experience, looked at from the point of view of its mechanical formulations, will yield significant data for the science of linguistics. I would look upon this as an important corollary gain, rather than as a major objective. There may be others who, and quite properly, because of their concern with adding to the versatility of electronic computers, will look upon research in the field of machine translation as a way of discovering just how much the machine can do. Again, this would be confusing instrumentality with objectives. Doubtless the challenge to the art of electronics posed by the problem of machine translation will lead to significant development in the operations, techniques and construction of computers. But again this ought not to be considered as an essential objective of our research. Indeed, there is the danger that concern with early tangible results in machine translation may lead us to lowering our sights to the process of essentially lexical search which, though not without value, remains far from actual language translation.

Our work has very definite and significant practical objectives. First, the aim of machine translation is to permit ready, undelayed access to scientific information written in the languages of the several scientifically creative cultures of our day. This means that scientists
engaged in diverse fields of research and development will, when we are successful in resolving the problems of machine translation, have more readily available to them the results of the research of their co-workers in other countries with different languages. The lapse of months, indeed of years, which now exists between the date of publication of the findings of, let us say, a Swedish physicist in his own language and the time of their accessibility to, for example, German, French, Russian, Japanese, or American scholars will have been, practically speaking, eliminated. The span of time between discovery and formulation by some, and awareness and sharing by others will have been significantly reduced. The immediate practical aim of machine translation thus is to accelerate communication among scientific workers, regardless of the language in which their findings may be expressed.

But this is only one of the practical objectives of our work. Were our research to serve this one purpose alone, it would be amply justified. The purpose of our work goes beyond this attainment, which is now well within our reach. Let us pause for a moment to consider the economic implications of machine translation. Clearly, if the capital investment required to achieve translation of languages by machine is to be greater than the traditional way of doing the job by employing translators, then no case can ultimately be made for machine translation. No matter what the economic structure of a given culture may be, machine translation will have to justify itself economically. Our Soviet colleagues, who since 1954, after the first demonstration of the feasibility of machine translation in which we participated in association with the International Business Machines Corporation, have been diligently pursuing research in the field, will have, even as we will, ultimately to face the test of cost accounting.

I am well aware that to intrude with these rather earthly preoccupations may displease some of the scientifically-minded among us, but still this very practical problem has to be faced. The time would seem to be at hand to undertake a study of the economics of language translation as now performed through the exclusive use of human talents as against the likely cost of translation by machine. I mentioned earlier that one of the anticipated results of our work will be to reduce significantly the time lag between publication of scientific data and their general availability in other languages, but if this is to entail excessive expenditures, then obviously the tendency will be to put up with the delay. It is my considered opinion that even as the use of computation instruments has proved economically sound in other fields of
data processing, this will be shown to be true in respect to the language translation processes. There is no evidence except by analogy to substantiate this statement, and I hope that the study of relative cost which I propose here will not be postponed.

Parenthetically, this leads to the consideration of another aspect of the general problem. When in January, 1954 the Georgetown-IBM experiment received what some critics have considered undue attention, some of my former associates in the Language Division of the United Nations asked whether this new venture would put them among the unemployed by automation. Obviously, in this field, as in many others, the use of the machine will only serve to liberate human talents for their proper end. I remember how countless man hours were harnessed to the production in four languages of the written proceedings of the Nuremburg Trials. When I recall the long hours of drudgery that gifted translators had to spend in the purely mechanical phase of translation, I am gratified at the prospect that creative human talents will be released for their true purpose in this field by the use of machines. It cannot be envisaged that what we have in our jargon come to call the "output" of the machine will be such as to stand completely on its own. This is where the capabilities of translators will find full expression as revisers or editors of the texts produced mechanically. Even as the development of simultaneous oral interpretation increased opportunities for competent interpreters, so will the greater volume of translation effected by machine provide openings for creative work for capable translators.

So far, then, we have seen that two of the important practical aims of our research are to facilitate scientific communication and to reduce the cost of translation and to free translators for more creative work. There is, I believe, a third area of practical significance in machine translation which is seldom discussed and which has not been widely recognized. The major concern so far has been to move toward the translation of scientific materials from other languages into English. It seems to me that the reverse process can be of equal, if not greater, significance. We know that one of the major obstacles to the dissemination of knowledge, techniques, and skills in the less developed areas of the world is the inaccessibility of technical literature.

I will mention two specific cases with which I am directly familiar. The government's Aid Program in Yugoslavia and Turkey found itself
seriously handicapped by the fact that many technicians who were to be trained in their several fields to higher degrees of competence in modern techniques could not receive their training because of the obstacle of language. It was obviously out of the question to think at that time of the practicality of making the technical literatures in such fields as textiles, animal husbandry, or soil drainage available to them in their native idioms through translations. So a program was initiated to train those technicians in the English language, at very considerable cost and time investment, so that they could receive their advanced training and have access to the current materials in their specializations published in English.

Suppose that it should become practical through machine translation to make available economically the basic reference materials and the current literature produced in the major scientifically creative areas of the world in the local languages of the less developed regions. One of the major objectives of our Foreign Aid Program is to disseminate knowledge in technical fields to ensure improvement of the economy and the strengthening and progress of the social structure. Obviously, if the basic technical literature can be translated in volume and economically, and placed on the library shelves of the schools, technical institutions, colleges and universities of these areas, the sharing of knowledge and spreading of information will be greatly facilitated and broadened. Now that we are witnessing the emergence of national states so rapidly in the hitherto called "backward" areas of the world, it is of the utmost importance to provide accessibility to the kind of information and know-how which will protect these developing peoples from the snares and blandishments of totalitarian ideologies.

We are thus allowed the hope—indeed, the expectation—that the results of our work will have many practical as well as scientific results. Research in machine translation will make a significant contribution to the science of linguistics; it will encourage inter-discipline exchange between the fields of mathematics and symbolic logic on the one hand and linguistics on the other; it will contribute to the increase of the capabilities of electronic data-searching processes. It has recently become quite evident, as a result of our contacts with the patent office researchers in the field of information retrieval and with the reports and statistical service of the Veterans Administration, that there will undoubtedly develop an increasing connection between machine translation on the one hand and the general data-searching and information retrieval techniques on the other. These are less tangible, but very significant objectives. More practically, our work
may contribute to the more rapid and wider dissemination of scientific information; it is likely to increase the scope of data accessibility to researchers in many fields; it may well make volume translation economically feasible; it can be of help to disseminate basic information in many technical fields in those areas of the world where it is urgently needed. These, then, are the practical objectives of machine translation research.

And now let me conclude by expressing our thanks for the support which our several groups have received from the National Science Foundation and agencies of the Department of Defense. On behalf of the University, let me thank you for the valuable contributions you have made to this meeting through the formal papers presented and the discussions which followed them. May I be allowed to give special recognition to the participation in our meeting of a representative of the Cambridge Language Research Unit in England and of his contribution. Perhaps next year it will be possible for research groups from other countries also to join with us in the sharing and discussions of the results of our common endeavors.

It is particularly gratifying that this reunion, following the pioneering work initiated at the Massachusetts Institute of Technology meeting of June 1952, and so fruitfully continued in the conference of October 1956 at the same institution, has brought together representatives of the several disciplines which must pool their efforts in the formulation of solutions to the problem of machine translation. The presence on the panels, in a joint effort, of linguists, computer engineers, data-searching specialists, programmers, and mathematicians is an indication of the increasing recognition that the solution of the problem of machine translation is by no means exclusively a question of linguistic research.

Thanks to your contributions, the next monograph will be a rather up-to-date statement of the status of work in the field of machine translation—at least as adequate a statement as is now available. The distribution which will be given to the proceedings of this meeting will, it is your hope and mine, stimulate increasing interest in the promising field of machine translation research.

I am looking forward to the spring of 1958, when I trust all of us and others can meet to review and discuss the results of our work between now and then.

On behalf of Georgetown University, let me thank you most sincerely.
APPENDIX 1

Program of the
Eighth Annual Round Table Meeting
on
Linguistics and Language Studies

FIRST SESSION—Friday, April 12, 9:30 a.m:

WELCOMING REMARKS:

Reverend Frank L. Fadner, S. J.
(Regent, School of Foreign Service)

INTRODUCTORY COMMENTS:

L. E. Dostert, Chairman, Eighth Annual Round Table Meeting
Brief Review of the History of Machine Translation Research

PANEL I: Systems of Logic in Machine Translation

Chairman: Victor H. Yngve, Massachusetts Institute of Technology

Martin Joos, University of Wisconsin
Meaning in Relation to Machine Translation

A. C. Reynolds, Stromberg Time Corporation
Linguistic Analysis and Mechanical Formulation

Paul L. Garvin, Georgetown University
Linguistic Analysis and Translation Analysis

William M. Austin, Georgetown University
Language as Symbolic Logic

Duncan Harkin, U. S. Government Consultant
The Mathematics and Logic of Language

(Discussion)
LUNCHEON MEETING—1:00 p.m., Hotel Dupont Plaza

Speaker: Clyde C. Heasly, Jr., Director of Planning, Intelligent Machines Research Corporation

Character Sensing as an Input to Machine Translation

SECOND SESSION—Friday, April 12, 2:30 p.m.

PANEL II: Lexical Problems in Machine Translation

Chairman: Victor A. Oswald, Jr., University of California, Los Angeles

The Rationale of the Micro-Glossary Techniques

R. A. Crossland, King's College, University of Durham, Newcastle upon Tyne, England

Word Decomposition for Machine Translation, by R. Richens, with the collaboration of M. A. K. Halliday, Cambridge Language Research Unit

Gilbert King, International Telemeter Corporation

The Problem of Lexical Storage

Hugo Mueller, Georgetown University

German Prenominal Modifiers as Clues in Machine Translation

Jacob Rabinow, Rabinow Engineering Company

New Large-capacity Optic Data Storage Device


Linguistics and Information Retrieval (Discussion)

RECEPTION—5:30 p.m., Faculty Lounge

Reception offered by Georgetown University

THIRD SESSION—Saturday, April 13, 9:30 a.m.

PANEL III: Scope of Syntactic Analysis in Machine Translation

Chairman: L. E. Dostert, Georgetown University
W. P. Lehmann, University of Texas
Structure of Noun Phrases in German

Michael Zarechnak, Georgetown University
Types of Russian Sentences

Jane A. Pyne, Georgetown University
Some Ideas on Inter-structural Syntax

Richard B. Thomas, National Bureau of Standards
The Use of SEAC in Syntactic Analysis

Daniel A. Belmore, Consultant Programmer, Washington, D. C.
Programming Aspects of Machine Translation

(Discussion)

CLOSING LUNCHEON—1:15 p.m., Hotel Dupont Plaza

Speaker: L. E. Dostert, Director, Institute of Languages and Linguistics, Georgetown University
Practical Objectives in Machine Translation Research

Luncheon offered by Georgetown University

APPENDIX 2

Membership of the Eighth Annual Round Table Meeting on Linguistics and Language Studies
12 to 13 April, 1957

Allen, Rolfe L. Department of the Army
Atsal, Ülkü Institute of Languages and Linguistics
Bagnall, J. J. U. S. Government
Bargin, Germaine Department of Agriculture Graduate School

Barry-Perroud, J. Department of the Army
Baydalakoff, Alexandra Institute of Languages and Linguistics
Berry, Madeline M. National Science Foundation
Beverly, Helen F. D. C. Teachers' College
Beym, Richard Foreign Service Institute

-189-
Paine, S. H.
Paper, Herbert H.
Pavia, Mario N.
Plavsky, Regina K.
Pyne, Jane A.
Rabinow, Jacob
Rashid, David O.
Raun, Alo
Rauscher, Dorothy
Roy, Joseph A.
Rubin, Joan
Salcines, Dagmar
Schramm, Joseph W.
Schweder, William H., S. J.
Segreda, Guillermo
Sbnitnikoff, Boris
Singer, J. R.
Sister Joan Margaret
Sister Mary Victor, O.S.B.
Sohn, Frederick W., S. J.
Sokoloff, Iris
Spainbour, Mary
Streadbeck, Arval L.
Sushko, Michael
Thomas, Richard B.
Thompson, Marie
Treviño, S. N.
Uçer, Perran
Van Norden, Linda
Vangül, Emin
Varga, Enüd von Kibédi
Vohs, Ralph H.
Waters, Halsey
Watkins, Mark Hanna
Wilhelm, Maria
Williams, Jacobia
Williams, Thyllis
Yngve, Victor H.
Young, John
Zarechnak, Michael

Washington, D. C.
University of Michigan
Georgetown University
Department of the Army
Institute of Languages and Linguistics
Rabinow Engineering Company
Institute of Languages and Linguistics
Indiana University
Foreign Service Institute
Institute of Languages and Linguistics
Institute of Languages and Linguistics
Institute of Languages and Linguistics
Washington, D. C.
Georgetown University
Foreign Service Institute
Institute of Languages and Linguistics
National Scientific Laboratories
Catholic University of America
Catholic University of America
Georgetown University
Cornell University
Institute of Languages and Linguistics
University of Utah
Institute of Languages and Linguistics
National Bureau of Standards
D. C. Teachers' College
Foreign Service Institute
Institute of Languages and Linguistics
University of California
Institute of Languages and Linguistics
Institute of Languages and Linguistics
Department of the Army
Institute of Languages and Linguistics
Howard University
Washington, D. C.
Institute of Languages and Linguistics
Washington, D. C.
Massachusetts Institute of Technology
Institute of Languages and Linguistics
Institute of Languages and Linguistics

-192-
## APPENDIX 3

### Index of Speakers

<table>
<thead>
<tr>
<th>Name</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austin, William M.</td>
<td>39,120,121,176</td>
</tr>
<tr>
<td>Belmore, Dan A.</td>
<td>163,167</td>
</tr>
<tr>
<td>Buckland</td>
<td>168</td>
</tr>
<tr>
<td>Crossland, R. A.</td>
<td>71,175,176</td>
</tr>
<tr>
<td>Fadner, Reverend Frank L., S. J.</td>
<td>1</td>
</tr>
<tr>
<td>Garvin, Paul L.</td>
<td>19,46,47,113,118,119,121,167,169,173,176,177</td>
</tr>
<tr>
<td>Halliday, M. A. K.</td>
<td>(see Crossland, R. A.)</td>
</tr>
<tr>
<td>Heasly, Clyde C.</td>
<td>53,114,115,116</td>
</tr>
<tr>
<td>Hodge Carlton</td>
<td>48,117</td>
</tr>
<tr>
<td>Joos, Martin</td>
<td>13,45,46,47,48,49,169,170,171,172</td>
</tr>
<tr>
<td>King, Gilbert W.</td>
<td>79,112,113,114,116,117,176</td>
</tr>
<tr>
<td>Kirsch, N. A.</td>
<td>168</td>
</tr>
<tr>
<td>Lehmann, W. P.</td>
<td>118,119,125,167,174</td>
</tr>
<tr>
<td>Mann, Jesse</td>
<td>120,121,122</td>
</tr>
<tr>
<td>Mueller, Hugo J.</td>
<td>89,167,168</td>
</tr>
<tr>
<td>Newman, Simon M.</td>
<td>103,113,114,117,118</td>
</tr>
<tr>
<td>Oswald, Victor A., Jr.</td>
<td>63,112,113,114,115,167,172,175,176</td>
</tr>
<tr>
<td>Paper, Herbert H.</td>
<td>48,49,168,169</td>
</tr>
<tr>
<td>Pyne, Jane A.</td>
<td>143,172</td>
</tr>
<tr>
<td>Rabinow, Jacob</td>
<td>102,112,113,115,116,117</td>
</tr>
<tr>
<td>Raun, Alo</td>
<td>47</td>
</tr>
<tr>
<td>Richens, R. H.</td>
<td>(see Crossland, R. A.)</td>
</tr>
<tr>
<td>Thomas, Richard B.</td>
<td>151,168</td>
</tr>
<tr>
<td>Yngve, Victor H.</td>
<td>170,171</td>
</tr>
<tr>
<td>Zarechnak, Michael</td>
<td>48,135,169,174</td>
</tr>
</tbody>
</table>