MACHINE TRANSLATION:

see also Natural Language Analysis and Processing

AIM AND REQUIREMENTS

Aim

Machine translation (MT) has the aim of translating material from one human language into another by computer. It is designed primarily for the scientific and technical materials produced in great quantities by our advanced society, and in large part rapidly superseded. Since these lose much of their value unless translated virtually at once, MT seems an ideal means of making them more generally accessible, especially if it can be carried out reliably and cheaply.

Its most influential early proponent was Warren Weaver, secretary of the Rockefeller Foundation. In 1946, shortly after the computer had been developed, Weaver proposed that it be used for translation, arguing that languages are codes and could thus be translated by a machine designed for the management of symbols. Formulating his views in a paper labeled "Translation," he sent this on July 15, 1949, to several hundred colleagues. The paper led to the first serious work on MT, to conferences, and publications [18, 19, 26].

One of the most prominent early scholars in the field was the late Yehoshua Bar-Hillel who organized a conference at M.I.T. in 1952 and published a paper on the state of research [1]. He subsequently guided the direction of research by successive articles and papers [2, 3]. Work was inaugurated in other countries as

well, notably England, France, Germany, and the USSR. Comprehensive surveys of this work were provided by Delavenay [9] and by Josselson [13].

Support for research was stimulated by a demonstration in 1954 arranged by Leon Dostert and Paul Garvin of Georgetown University in conjunction with Peter Sheridan of IBM. Subsequently various centers of research, notably Georgetown, Harvard (Anthony Oettinger), and M.I.T. (Victor Yngve), received grants and contracts to implement MT. By 1964 these had amounted to 20 million dollars, yet no system had been devised which produced translation without considerable post-editing [6].

Sponsors of research then asked the National Academy of Sciences to set up a committee to review the situation. The report of the Automatic Language Processing Committee (ALPAC) under the chairmanship of John R. Pierce, issued in 1966, was highly critical of MT. It did not recommend further support for MT, but rather for computational linguistics in general and for mechanical aids to human translators. This report had a severe impact. Research was greatly reduced, so that progress virtually ceased. Some curtailed projects were maintained, especially where there was particular awareness that rapid translation of large quantities of material is essential, as in the Common Market and in countries with emphasis on bilingualism, like Canada. Of late, this realization has been strengthening, so that once again there are prospects for research support, as noted below.

Linguistic Requirements

A primary reason for the disillusionment with MT and the slow progress in MT research is the immense complexity of language and the inadequacy of the. grammars and dictionaries available for even the most extensively studied languages.

A language is an open-ended system which can yield an infinite number of sentences. When people learn their native language, they master this system in such a way that they can produce and recognize sentences which have never been produced before. They also can distinguish between grammatical and ungrammatical sentences. No grammar, however, has been produced which encompasses this capability. The grammars which are available for learning a second language contain rules which permit speakers to recognize the similarities and differences between the target language and their own. But details are often mastered by individual observations rather than from carefully formulated rules. Second language learning, like first language learning, is then in part intuitive.

To carry out MT of any variety, extensive grammars with precise rules[1] must be produced. Precise rules are necessary because of the ambiguity in languages, and because one-to-one lexical correspondence between languages is rare; moreover, languages differ considerably in their surface structures.

A brief example may illustrate the ambiguity of natural language. In one context the statement He declined may mean "He refused"; in another it may mean "His health worsened." Words typically have multiple meanings, such as table in its concrete and figurative uses, and in its meaning "chart." Humans disambiguate

[1] "Rule" in the sense of determination of normal usage—i.e., a "regularity" rather than a "regulation."

by noting cotext or context of an utterance. Simulating this capacity on a computer involves considerable ingenuity and study.

Cotextual and contextual information must also be included in MT dictionaries because words can rarely be translated with single equivalents, especially common words. Thus English know must be translated with German kennen if its object is a person, with wissen if a fact, and with koennen if a skill. Rules can indeed be provided to assure such translations, but they involve considerable computer storage. Neither the early computer technology nor the computer programs were adequate to provide huge storage with rapid access. .

Other problems, as in the syntactic and semantic analysis of the languages concerned, had not been adequately identified, so that much linguistic research as well as improved computer technology was required before a partial MT system could be developed. Such problems may be illustrated by examining briefly the expression of definiteness in selected languages. In English this is conveyed largely by use of the definite article. Russian, on the other hand, one of the languages in greatest demand for MT, contains no articles; definiteness is expressed by other devices which are not described in available grammars. Before acceptable MT from Russian to English is possible, the Russian devices for indicating definiteness must be described.

Another language of concern for MT, Chinese, also lacks articles. Definiteness may be indicated in Chinese by placing a noun before the verb; thus the Chinese sequence: man dog see (to use English lexical equivalents) would correspond to English The man sees the dog, while man see dog would correspond to English The man sees a dog. Before MT research, grammars did not attempt to describe such details. English speakers who learned Russian and Chinese mastered them by practice based on observation and on trial and error.

Similarly, no adequate rules for use of the definite article in English had been provided. One rule was that nouns used a second time in an account require the definite article, as in fairy tales: Once upon a time there was a beautiful princess. The princess . . . . Yet this rule is not specific enough to program a computer to handle sequences like the following, which requires a definite article in the first use of a noun: She was reciting Hiawatha. The poem excited her students.

This example of the problems involved in achieving accurate translations of definite articles may illustrate that MT could not be instituted simply by programming the grammars that were available before the advent of computers. Rather, translations must be based on determination of deeper structures arrived at through syntactic and semantic analysis. This requirement, as noted below, led to the development of transformational grammars, which deal with underlying as well as surface structure in their linguistic descriptions.

Linguistic descriptions involve dictionaries as well as grammars. Languages have traditionally been described in two processes: production of grammars, which deal with the arrangement and classes of elements, and production of dictionaries, or lexicons, which define those elements. Just as grammars of the past are inadequate, so are the dictionaries.

Large, so-called unabridged dictionaries contain approximately half a million words. Even in this large stock, many technical terms are omitted—the technical terms in one science alone, chemistry, far exceed a half million. Often, too, the definitions in available dictionaries are insufficiently precise to meet the requirements of MT. For example, a large German dictionary defines beugen as "bend, bow down" and the like, but does not indicate its widespread meaning in technical

use: "diffract." Moreover, MT systems must include capabilities for dealing with new technical terms since MT is directed at translating journals, monographs, and other avenues of information while this is fresh and of vital concern. Dictionaries as well as grammars must therefore be amplified, and continuous lexical research must be maintained to account for new words as scientific disciplines develop. Even so, no MT system attempts to account for the entire language. Accommodation of technical and scientific lexica is the primary demand. In any event, the linguistic requirements for MT are formidable.

Computational Requirements

Computer requirements, both hardware and software, are similarly demanding. Early computers were particularly designed for purposes such as numerical analysis; MT, on the other hand, is carried out by data processing techniques. Early MT research therefore had to deal with steps necessary for adapting computational processes to data manipulation. Initially consideration was given to the development of special-purpose computers. Some were actually built, notably the Mark II used by the Foreign Technology Division (FTD) of the U.S. Air Force Base at Dayton, Ohio, at the time of the ALPAC report. Yet research in MT makes up such a comparatively small proportion of computer applications that the undertaking was soon abandoned and available general-purpose computers were used.

Likewise, the early computer languages such as FORTRAN were developed for the purpose of numerical manipulation. MT projects had the possibility of modifying these or of working directly with machine language. Either option required massive programming activity. This led to the development of the first computer language achieved for data-processing, COMIT, designed by Yngve. Currently, projects make use of the subsequently developed systems such as SNOBOL and LISP.

Whatever the choice of computer and programs, the storage requirements for even a carefully selected lexicon and the accompanying grammar vastly exceeded the storage capabilities available in computer. (Even today a dictionary as small as the Merriam-Webster Collegiate taxes the storage of the typical large computer facility characteristic of a university.) For an MT system to function at all, the lexicon had to be stored on tapes; access time was slow, as was translation.

The vision of MT proposed by Weaver thus required a huge amount of preparatory research before even simple MT systems could be tested. This research fell to two fields, one of which—computer sciences—was nonexistent in 1947, and the other—linguistics—was so new that there was only one small linguistics department in the universities of the United States. It is scarcely surprising that development was slow.

THE DESIGN OF MT SYSTEMS

MT is a problem in applied linguistics. To achieve the aim of automatic translation by computer, an algorithmic system for language analysis, including specialized grammars, lexicons, and programs, must be designed and implemented.

Linguistic Strategies in the Designing of an MT System

Traditional linguistics dealt with language as though it consisted of several strata. The recurrent patterns were treated in the grammar; the less frequent in the lexicon. The elements listed in dictionaries are generally words, but some are idioms. Thus expressions like go without saying are listed in the lexicon on the grounds that this is an idiom, but not sequences like go without saying goodbye. By appropriate entries for idioms one can distinguish between them and nonidiomatic patterns, such as go without one's overcoat. Each word of the latter type of sequence is listed separately in a dictionary, and the relationship between them is described in the grammar. Current linguistics also recognizes a distinction between the grammar and the lexicon, although the terminology is not always the same as that of earlier linguistics.

MT systems are designed to manage language by means of these two components. The lexicon is listed in stores; the grammar is handled by means of algorithms. In implementing these activities, various practical problems arise.

One must decide, for example, whether to analyze compounds, e.g., light-ray, and list each element separately, or to include the compound in the lexicon. The decision may vary, depending on the languages involved and on computer capabilities. German, for example, is notorious for its lengthy compounds, especially in technical writing: fire insurance company is one word in German: Feuerversicherungsgesellschaft. Moreover, such compounds may be created freely, like English nominal phrases. Accordingly, it seemed necessary in early MT projects to develop procedures for automatically analyzing compounds into their constituents rather than to list all possible compounds as separate entries. The decision was influenced by restrictions on storage capacity and access time to the data. Today, with the relatively recent development of huge storage capacities, a different decision might be warranted. Yet, however such problems are handled, the lexical difficulties are far smaller than are the grammatical.

One of the long-studied problems of grammars is the treatment of synonymous sentences. The formerly popular Advanced English Grammar of Kittredge and Farley [15] recognizes the equivalence of sentences like the following (p. 135): It was a pleasure to see him and To see him was a pleasure. Another example showing a similar equivalence relationship is: It is easy to understand you; compare You are easy to understand. Relating such semantically equivalent sentences (the relationships have been referred to as transformational) has occupied much of the attention of recent linguistics. Besides studying specific patterns like those with easy, linguists have explored the extent of all such relationships in languages, especially in English.

Another pattern which has received extensive attention is the passive. This also had been studied by traditional grammarians, who generally treated it as did Kittredge and Farley (p. 110):

> Any sentence of which the predicate is a transitive verb followed by an object, may be changed from the active to the passive without affecting the sense.
>
> ACTIVE Richard shot the bear.
> PASSIVE The bear was shot by Richard.

In recent grammars the passive has been considered a transform of the active, yet there have been no analyses which are generally accepted by all transformationalists, for some passive-type constructions are not simple transforms of actives, e.g., They were taken by surprise.

Whatever the terms used, or the specific rules, a grammar is clearly much simpler if such related constructions can be considered equivalent, and the simpler or more natural variant can be treated as more fundamental. An MT grammar with such an approach would obviously be more economical than one which set out to analyze each sentence into a subject and a predicate, regardless of its relationships with other sentences.

Such a transformational approach may be applied by analyzing surface structures of sentences and determining the simpler varieties, e.g., Richard shot the bear. This is the procedure pursued by Zellig Harris who originated transformational grammar.

The approach can also be applied by assuming an underlying "simple variety" comparable to a concept which then is variously expressed and modified to become the actual surface sentence. This is the transformational approach pursued by Harris's student, Noam Chomsky. In the 1960s this approach aroused tremendous enthusiasm among young linguists and among nonlinguists such as psychologists [10], for it seemed that the underlying patterns posited for the surface structures permitted insights into the functioning of the mind. Yet today, many of the more active syntacticians are returning to research which agrees with the approach of Harris, combining it with simultaneous "semantic" analysis, in procedures inspired by the logician Richard Montague. (See Partee [22], for example.)

Linguists concerned with MT also tried the two approaches, but much of the publication on which they attempted to draw was directed at providing support for transformational generative theory rather than at describing languages more precisely. Accordingly, much research carried out in transformational generative grammar, some of it funded by MT projects, had little pertinence for MT.

The transformational approach, however, brought attention to formal structures into linguistics. These are clearly well-suited for computer manipulation of language. Moreover, they can be examined for their adequacy by procedures known as grammar testers; these analyze the rules proposed in a grammar, whether it is designed for MT or for theoretical purposes. Computational techniques thus came to be a powerful research tool for linguists [12].

Grammars designed for MT must devise procedures for relating surface structures with units of meaning. For this purpose "deep" structures are posited and related to the surface structures. Grammars must be produced for these various levels. The use of these grammars may be illustrated by reference to problems resulting from the presence of discontinuous elements in language, such as took . . . down in She took the unusual words down. After their identification in a surface grammar they are brought together in a further grammar, so that the verb and the particle can be treated as a unit. In such a grammar of the deep structure they are regarded as a semantic unit, equivalent to "note." An MT system in this way analyzes sentences of a source language with a surface grammar and various additional grammars to identify the underlying structures. These are then transposed into the target language using a comparable set of grammars for that language.

The grammars of an MT system are thus designed in accordance with the understanding of language achieved in linguistics. The rules thy incorporate are highly precise and inclusive. In the same way, the dictionaries start from tradi-

tional dictionaries but proceed to more specific analyses of use and meaning, indicating these through subscripted features accompanying the lexical entries. Other components of an MT system, such as the recognition procedures, are determined at least in part by computer capabilities.

Computational Strategies: Search and Parsing

Search Procedures

However the linguistic analysis is produced, arrangements must also be made to recognize accurately the words of the initial input string and to relate them to other words in the sentence, that is, to simulate reading. The computerized process is known as search. A search may be done in various ways, one by identifying each word from left to right and matching it with an entry in the lexicon. After identification, parsing is produced. This procedure leads to complexities, especially in lengthy sentences, but also in a simple sentence like: She took the unusual words down. Here the analysis of took is determined in part by the last word. Procedures must then be included to move backwards as well as to read from left to right.

An alternate procedure, devised by Paul Garvin, focuses on the words of a sentence in accordance with their grammatical significance rather than by their linear order of occurrence. This procedure, called the fulcrum approach, first identifies the major syntactic elements, like the subject and its verb, and only later their modifiers. These, such as the adjective unusual in the example above, are then processed in their relation to the central structure.

At present the "reading" of input texts is readily accomplished as a result of extensive work in the early MT research.

Parsing Procedures

In designing an MT system, the rules for parsing sentences can be incorporated in the computer algorithm, or the algorithm and the grammar can be kept separate. By the first kind of design, the system would consist of two components: a lexicon and a translating algorithm. Such a system is known as bipartite. By the second, the system would be tripartite, consisting of a lexicon, a grammar, and a table of computational rules which could be used with any grammar of any language.

The initial MT systems, notably those developed at Georgetown University, were bipartite. These systems have been extensively used in production by the U.S. Air Force Foreign Technology Division (FTD), the AEC Oak Ridge National Laboratory, and the Scientific Information Processing Center of EURATOM, and commercially by LATSEC, Inc.

As late as 1971 Garvin argued in favor of a bipartite system, on the grounds that "the fundamental problem in the automatic recognition of grammatical structure of text is the correct sequencing of the application of the rules of the grammar which are supposed to effect the recognition. In this author's opinion, such a sequencing of the application of different grammatical rules can be effected only by making the rules of the grammar an organic part of the algorithm" [11, p. 109].

Yet bipartite systems have been criticized because they grow to be cumbersome. Such a grammar can indeed be readily programmed, but when it must be updated the added rules result in a minor program supplementing the original program. In time

the translating algorithm of a bipartite system may come to be a patchwork which is very difficult to improve further.

In a tripartite system, on the other hand, the grammars may be extensively revised, with no need to modify the computational rules. Moreover, the same set of computational rules can be used for many languages, while in a bipartite system a new translating algorithm is necessary for each additional language. The tripartite design then would seem vastly superior. These differing approaches to parsing may illustrate that the design of a system is determined by strategies which are inherent in the new application of linguistics made possible by the development of computers.

Besides drawing on linguistics and computer sciences, MT also incorporates findings of rhetoric and stylistics. Sequences longer than sentences, often referred to as cotext in the domain of rhetoric, must be taken into consideration when translating.

One example was cited above: use of the definite article in English. Moreover, certain characteristics of language long studied in stylistics are especially prevalent among specific groups of language users or language used for specific purposes. Thus technical writing makes use of different patterns of expression from those of materials designed for a general audience. Technical German, for example, contains frequent preposed participial modifiers of nouns, which correspond to relative clauses; rarely found in the everyday form of the language, these modifiers constitute a special problem for any translation of scientific and technical German.

Support for this conclusion has been provided, especially in the computer-aided work of the New York University Linguistic String Project [24]. Concentrating on pharmaceutical texts, this has determined that technical writing makes use of special syntactic patterns as well as special vocabulary. "Intellectual verbs, like find, hypothesize, discover, which characteristically take human subjects" are found at the bottom of an analysis tree and "concrete words referring to objects in the science at the top," as in the example: "Carvalho and Leo found that the Ca^{++} uptake of skeletal SR involves the exchange of Ca^{++} with other cations in SR." Sager's analysis not only distinguishes the "words referring to the objects and relations in the science proper from statements the human investigator makes about those objects and relations" but also determines "definite classes, like quantity operators and causality operators" in characteristic positions [24, pp. 37-39]. Mechanical parsing has in this way led to the determination of characteristic semantic and syntactic patterning in scientific writing. Such results illustrate contributions of MT and related procedures to a deeper understanding of language. They also support the conclusion that MT is a distinct activity, requiring procedures that are best determined by drawing on a number of disciplines in relation to one another [16].

Summary of Strategies of Design

For more detailed information, interested readers might wish to refer to the specialized reports which are issued periodically by active research groups. Only a selection will be reviewed here, to illustrate the current situation and requirements for further improvements.

Any system must take the complex structure of language into consideration. Accordingly it must deal not only with the lexical elements, but also with syntax in its surface and deeper levels. In the brief history of MT there has been a progression from systems which basically translated word-for-word, to those which

included attention to surface syntactic rules, and finally to those which deal also with deeper syntactic and semantic levels of language. These three progressively more sophisticated types of systems are referred to as first, second, and third generation systems. Current research groups generally aim at third generation systems [6, 17, 28].

THE CURRENT SITUATION

In view of its independent status, the field has suffered from not receiving steady support. Researchers, whether drawn from linguistics or computer sciences, may well develop competence in MT. But computational possibilities change so rapidly that continued experience is helpful for original research. Virtually all support for such research has been governmental, as is generally true for developing fields—and this has been interrupted, in Europe and the USSR as well as the United States, in-great part as a result of the ALPAC report which led to virtually complete cessation of funding. Research groups were gradually disbanded, so that currently these are limited to relatively small projects, of which those at Grenoble, Saarbrücken, and Montreal will be noted briefly here. As evidence of the decline of research after the publication of the report, it may be noted that the March 18, 1977 issue of Science devoted to electronics does not even mention MT.

On the other hand, The Commission of European Communities recently arranged the Third European Congress on Information Systems and Networks, with the theme: Overcoming the Language Barrier [8]. The proceedings of this meeting contain reports on some of the ongoing research projects.

The Centre d'Études de la Traduction Automatique (CETA) at Grenoble, under the direction of Bernard Vauquois, has the longest record of continuous research [4, 25]. Moreover, its system has been tested with the largest quantity of texts. CETA draws on various linguistic theories, making use of dependency as well as transformational and phrase structure analysis. It deals with translation in three stages: analysis of the source language; interlingual mapping; and synthesis of the target language sentences, making use of a "language pivot." A parser has been developed for labeled tree structures in which the nodes of the tree include the morphosyntactic and semantic/logical data necessary for transfer to the target language. The primary aim of the project is translation from Russian to French, although grammars have been studied and tested for other languages as well, such as Japanese and English.

The Saarbrücken Automatic Translation System (SUSY) is being developed to carry out translation from Russian to German, with some work also being done on other languages such as French and English. Considerable attention is being given to automatic identification of lemmata (words reduced to their basic forms) in German texts. This is done by means of a mechanized lexicon containing 100,000 entries [20]. The grammars under development classify elements not simply by their syntactic use, but also by their function; thus pareil "similar" is treated as a "deictor" or article word rather than an adjective [27]. SUSY thus illustrates how MT may develop dictionaries and grammars designed specifically for the automatic analysis of languages rather than the work of traditional or contemporary linguistics.

MT research in Canada is prompted by the government's policy of biculturalism. In keeping with it the Canadian State Department's Translation Bureau "has under-

taken to develop a second-generation system for machine translation of technical manuals" [7]. The system, being developed under a contract with the TAUM project of the University of Montreal, is to be put into operation in November 1978. Time requirements necessitate the immediate development of a system rather than further research, and hence the restriction to a second-generation level, involving only lexical and morphosyntactic analysis, on which English-French transfer will be based. Confidence in useful results is derived from the success of the METEO system "which is used for the translation of weather forecasts intended for the general public" [7].

In the United States, research at Austin and Berkeley is being done to incorporate advances in computational activity into systems devised earlier. The aim is production of a third-generation system. There is hope of close cooperation among these various research groups and also of utilization of previous accomplishments.

EVALUATION OF MT

As noted above, work on MT has been severely criticized. Taking no consideration of the obsolete system in use at the time, the ALPAC report concluded that translations performed by translators supported by the government were cheaper and better than those produced by FTD [23, p. 28]. It cannot be denied that the new field of MT attracted some inadequate workers, nor that the translations produced with TDS's Mark II required considerable postediting and accordingly were expensive. Computational costs have, however, been reduced as systems have been improved, while the costs of human translators are constantly rising. It is also acknowledged that there have been considerable advances in MT research [14, pp. 225-226]. Whether these advances could have been made without waste is a question that may be viewed in the larger context of all technological developments. The history of MT research might be compared with that of research to meet future energy needs; all the funds used for MT would scarcely maintain some energy projects for a fraction of a year. But as Kay states [14, p. 219], after the ALPAC report "any application … for financial support for a project involving language and computers . . . could expect a swift and categorical refusal." Human frailty seems to involve error and waste in any technological advance, whether it concerns transportation, energy, or communication.

MT gains its appeal and support from the problems arising from the huge amount of publication currently produced, and from the short-lived usefulness of that publication. Technological and scientific materials are made available in vast quantities, whether in English, Chinese, French, German, Italian, Japanese, Russian, Spanish, or other languages, and much of it is soon out of date. (Some scholars assert half-seriously that any scientific article is out of date by the time it is printed.) A scientific article written in a foreign language may then be of little use if it is not translated at once.

Moreover, bilingual countries like Canada mandate by law the translation of huge quantities of government documents virtually overnight. There are also projects such as the one currently under way in Canada for the translation of technical manuals associated with the employment of the Aurora missile [7]. And the European Common Market countries arrange translation of many materials from any of the official languages—English, Danish, Dutch, French, German, Italian—into any other, making a total of 30 combinations. These requirements gain urgency

as educational systems have greatly reduced the study of foreign languages, even in Europe where the tradition of foreign language teaching was strong.

Yet the adverse criticisms have discouraged investigators as well as suppliers of funding. Linguists who consider themselves devoted to theory give MT little attention. In the 50 year index to Language, the journal of the Linguistic Society of America, there is only one reference to MT (Vol. 32, 1956) and this is a review of a collection of essays. Although MT does not enjoy a favorable evaluation today, its future will no doubt be determined by social needs, which when sufficiently pressing will change funding and attitudes.

Future research will be greatly assisted by the remarkable improvements in computer hardware and software. These will also affect the method of output. The huge data stores required for manipulation of human language are no longer such a problem for the computers of today, when rapidly accessible disk storage facilities are available. But systems will have to be carefully designed for economical storage of these data, with rapid access in spite of their bulk. Moreover, computer languages, like LISP have now been designed for handling large data bases, so that programmers no longer have to modify the technically suited computer languages. Further, input and output facilities have been greatly improved. While scanning devices have not been developed as rapidly as was promised, language material is now widely available on magnetic tape; it can be transmitted to the computer in more flexible ways than through the cumbersome keypunch machine. Moreover, output can be quickly displayed on CRT terminals, diminishing the need for printout. These advances in computer technology will no doubt lead to MT which is displayed on a CRT, with the possibility of simultaneous running display of the original. Users will then have the option of rapidly scanning a desired text in much the same way they would scan texts available in their native languages, and the cumbersome process of printing huge amounts of text of uncertain value can be dispensed with.

On the other hand the linguistic contributions to MT have made fewer advances than has the computational area. The necessary large, precisely defined lexical stores were not being assembled in the absence of funding. Nor were the necessary syntactic and morphological analyses being produced. To be sure, great attention is being given by linguists to syntactic study; but much of this focuses on specific problems in English which have come to be focal points in constructing theories. For example, though the relative clause and its relation to its antecedent has been examined copiously, the aim of the studies is to test theory rather than to provide improved understanding of the construction, which previous grammarians understood and described thoroughly. But for MT, complete grammars are equally required for both source and target languages. Hence it is necessary to continue research in compilation of large linguistic data bases. The development of the requisite systems is accordingly a long-range undertaking.

IMPLICATIONS AND SPIN-OFFS OF MT RESEARCH

As mentioned earlier, when MT research was undertaken languages were not precisely described nor were their lexical elements defined in detail. Language learning is so intuitive a process that even punctilious scholars failed to note all the properties of lexical items or grammatical constructions until computers produced the nonsensical output which incomplete descriptions of language yield. While computational linguistics on its own might have led to the detection of

inadequacies, and then to improved grammars and dictionaries, work toward MT has made clear the accuracy and thoroughness required for the analysis and description of language.

Besides MT, computerized analysis is involved in automatic indexing, in data retrieval, and in fact retrieval. Indexing has been carried out from time immemorial, especially for texts of highest prestige in our society. Thus the Bible and the Homeric poems have been provided with indices and concordances which in the past might involve the life effort of one or even several scholars. Today such tools are produced readily by computer programs which are routinely used in MT research. Indexing, however, has remained at the lexical or first-generation level. The widely used KWIC (keyword-in-context) indexing has indeed developed skillful techniques . But these techniques yield only that information which can be derived from identification of word-forms. When indexing systems incorporate the parsers developed for mechanical translation, their usefulness will be greatly enhanced.

Research in MT will make contributions to other facilities as well, such as question-answering systems, and will profit from insights in such research. These systems are now used with relatively homogeneous data bases as in airline ticketing, in banking, and in accessing medical data. When combined with large lexical stores such as the syntactic and semantic algorithms developed in MT research, they may be extended to provide access to a wide variety of information.

The syntactic and morphological analysis carried out in second-generation MT research will permit the construction of data processing systems. When semantic analysis is added, systems can be devised which process facts. Queries then will not be restricted in their scope to the individual items included in data stores, but they will encompass related "facts," regardless of their linguistic expression. -

Projection of such computerized capabilities has led to proposals in artificial intelligence whose essential aim is to simulate the activities of the human brain. This involves not only the retrieval of facts and data but also the understanding of a situation. Such understanding requires concern for the setting in which a linguistic utterance is produced as well as the analysis of that utterance. In aiming at such achievement, artificial intelligence research has led to the proposal of restricted settings, known by various names, among them "scenarios," which are important in determining the meanings of utterances. For example, speakers of English will interpret differently the following utterance depending on their knowledge of the actors involved in a given setting of the scenario: She ought to be here by now. If the referent of she is in some way under the authority of the speaker, the utterance will be interpreted: "There is an obligation which normally would require her to be here at this time." If the referent is someone not under such authority, the utterance will be interpreted: "She generally would have arrived by this time. (Some difficulty must have interfered with her expected behavior.)" The computational study of language has in this way led to fuller recognition of the complexities involved in human communication, which extends to the study of perception as well as of language.

When such complexities are revealed they are by no means studied only by computational linguists or specialists in MT. Miller and Johnson-Laird have produced an intriguing study relating perceptual knowledge about the world to language use [21]. This has led to their proposal of a procedural approach to semantics, and to the study of memory systems. In their introduction, Miller and Johnson-Laird state that "the wide range of theoretical possibilities that modern computers suggest provides much of our present motivation to search for psychological, linguistic, anthropological, or other substantive boundary conditions on hypotheses about lexical memory" [21, p. 6].

The manipulation of language by computer which began in the effort to achieve machine translation has accordingly disclosed many problems and opened many areas of research. As our knowledge of language and its use is increased, these will be illuminated, though possibly not before the accomplishment of improved mechanical translation. MT of technical materials is one of the simplest linguistic applications involving computers. Its progressive advances will provide data which will then be available for more complex research areas and applications, such as the study of perception, of memory, and of storage of information in the brain. MT thus has numerous implications for other areas of research, as well as its central purpose, the rapid and automatic translations of texts.

ACKNOWLEDGMENTS

I am grateful to Helen-Jo Hewitt, Zbigniew L. Pankowicz, and Solveig M. V. Pflueger for assistance in preparing this article.

REFERENCES

1. Bar-Hillel, Y., The present state of research on mechanical translation, Am. Doc.2, 229-237(1951). (Reprinted in Ref. 2.)
2. Bar-Hillel, Y., Language and Information, Addison-Wesley, Reading, Massachusetts, 1964.
3. Bar-Hillel, Y., Some reflections on the present outlook for high-quality machine translation, pp. 73-76 in Vol. I of Ref. 17.
4. Boitet, Ch., Un essai de réponse à quelques questions théoriques et practiques liées à la traduction automatique. Définition d'un système prototype. Thèse d'État, Grenoble, 1976.
5. Booth, A. D. (ed.), Machine Translation, North-Holland, Amsterdam; Wiley, New York; 1967.
6. Bruderer, H., Handbuch der maschinellen und maschinenunterstützten Sprachübersetzung,. Fink, Munich, 1977.
7. Chandioux, J., Creation of a second-generation system for machine translation of technical manuals, pp. 613-621 in Ref. 8.
8. Commission of the European Communities, Overcoming the Language Barrier (Third European Congress on Information Systems and Networks, Luxembourg, May 3-6, 1977), Vol. 1, Verlag Dokumentation, Munich, 1977.
9. Delavenay, E., La Machine à Traduire (Que Sais-Je No. 834), Presses Universitaires de France, Paris, 1959.
10. Fodor, J. A., J. G. Bever, and M. F. Garrett, The Psychology of Language: An Introduction to Psycholingutstics and Generative Grammar, McGraw-Hill, New York, 1974.
11. Garvin, P. L., Operational problems of machine translation: A position paper, pp. 95-118 in Vol. I of Ref. 17.
12. Hays, D. G., Introduction to Computational Linguistics, Elsevier, New York, 1967.
13. Josselson, H. H., MT in the sixties: A linguist's view, Adv. Comput. 11, 1-58 (1971).
14. Kay, M., Automatic translation of natural languages, in Language as a Human Problem (E. Haugen and M. Bloomfield, eds.), Norton, New York, 1974, pp. 219-232.

15. Kittredge, G. L., and F. E. Farley, An Advanced English Grammar, Ginn, Boston, 1913.
16. Lehmann, W. P., Towards experimentation with language, Found. Language 1, 237-248 (1965).
17. Lehmann, W. P., and R. A. Stachowitz, Feasibility Study on Fully Automatic High Quality Translation, Vols. I and II (RADC-TR-71-295), Rome Air Development Center, Griffiss Air Force Base, Rome, New York, 1971.
18. Locke, W. N., Machine translation, in Encyclopedia of Library and Information Science, Vol. 16 (A. Kent, H. Lancour, and J. E. Daily, eds.), Dekker, New York, 1975, pp. 414-444
19. Locke, W. N., and A. D. Booth (eds.), Machine Translation of Languages, Wiley, New York, 1955.
20. Maas, H. D., The Saarbrücken automatic translation system (SUSY), pp. 585-592 in Ref. 8.
21. Miller, G., and P. N. Johnson-Laird, Language and Perception, Harvard University Press, Cambridge, Massachusetts, 1976.
22. Partee, B. H. (ed.), Montague Grammar, Academic, New York, 1976.
23. Pierce, J. R. (chairman), et al., Language and Machines: Computers in Translation and Linguistics (A Report of the Automatic Language Processing Advisory Committee (ALPAC), Publication No. 1416, National Academy of Sciences-National Research Council, Washington, D.C., 1966.
24. Sager, N., Evaluation of Automated Natural Language Processing in the Further Development of Science. Information Retrieval (String Program Reports No. 10), NYU Linguistic String Project, New York, 1976.
25. Vauquols, B., La traduction automatique à Grenoble, Dunod, Paris, 1975.
26. Weaver, W., Translation, pp. 15-23 in Ref, 19.
27. Weissenborn, J., The role and form of analysis in machine translation: The automatic analysis of French at Saarbrücken, pp. 593-611 in Ref. 8.
28. Yngve, V. H., Syntax and the problem of multiple meaning, pp. 210-213 in Ref. 19.

Winfred Lehmann