[*Computers and the Humanities* 13 (1979), p.17-28]

# A Survey of Approaches and Issues in Machine-Aided Translation Systems

WAYNE W. ZACHARY

The much-proclaimed "information explosion" of recent years has produced an ever-increasing flow of scientific. technical, and scholarly literature, much of which is relevant to an audience wide enough to span several languages. This has, in turn, created a growing demand for more, faster, and better translation of printed material. The growth of international cooperative ventures, such as the Commission of European Communities and of multinational corporations, has provided further demand for quick and accurate translation capabilities. Translators and translation facilities, struggling to keep up with their burgeoning load. have turned more and more to the view that their problems can be solved only with help from computers. The earliest approach, first investigated three decades ago. was to get computers to do translation entirely by themselves. But as time progressed, this goal began to look unobtainable (at least in the immediate future), and interest shifted toward using computers in supporting, rather than starring roles, as aids to human translators. Progress on this front has been substantial and rapid, until today there are over twenty "computer-aided," or "machine-aided" translation systems in operation in the United States. Canada, and Western Europe,

*Wayne W. Zacharv is an associate in the Systems and Operations Analysis Division of Analytics, Willow Grove, PA.*

*This research was conducted under a subcontract to Analytics from Chase. Rosen and Wallace, Inc.. with funds provided by the Office of Research and Development. Central Intelligence Agency. Thanks are also due to the many individual researchers, too numerous to mention, who responded graciously to requests for information.*

This paper examines the current state-of-the-art of machine-aided translation, henceforth referred to as MAT. The emphasis will be on the organization of the various *operational* MAT systems focusing, in particular, on the different divisions of labor between man and machine that have been used, and on the philosophical and technical problems that have been considered in designing MAT systems. Less attention is paid to detailed descriptions of the various systems because the interested reader can find them in the references listed, and the Appendix contains summaries of the capabilities of the major systems discussed below.

In general, current MAT systems can be divided into three categories: (1) "pure MAT" systems, in which only vocabulary or terminological reference is mechanized. (2) "human-aided machine translation" (HAMT) systems, in which a greater portion of the translation process is carried out by the machine with the human providing only clarifying information to it, and (3) "pure machine translation" (pure MT) systems, in which the machine attempts to do the entire translation itself but still requires pre-input and/or post-output editorial assistance from human translators. This three-way division also corresponds to the three-stage characterization of language processing that is used in linguistics. Of course the correspondence is not accidental, for any MAT system must be based on some explicit model of language and there is wide agreement among linguists as to how language should be modeled, at least at a general level. As a result, it is possible to discuss MAT systems in the context of their practical organization *and* their underlying theoretical models simultaneously. This is the approach taken here. To preface the remaining discussion, a brief review of current linguistic theory is given first.

***Language Processing and Computational Linguistics.*** "Language" is defined in linguistics as the process by which people produce understandable sentences from underlying abstract thoughts and ideas, and understand the abstract ideas contained in the sentences written or spoken by others. This process occurs in several hierarchical stages and can be described or analyzed in two different ways: (1) emphasizing the structure and organization of language at each of these stages or levels (the view of "structural linguistics"), and (2) emphasizing the information processing which actually results in the production or recognition of a given sentence (the view of "computational linguistics"). The computational linguistic approach is not only much more compatible with the interest here in mechanizing part of the language process, but was originated and developed as a result of early interest in machining translation and the failure of the structural approach to develop workable computational models for MT. The remainder of the discussion will therefore consider language only from the computational viewpoint.

Modern computational linguistics originated with the work of Chomsky who in his 1959 critique of Skinner's *Verbal Behavior*, demonstrated the fundamental errors in the then-accepted behavioristic model of language and subsequently suggested (e.g. Chomsky 1957, 1965) a new, more mathematical approach. In this approach, it is not language *performance* (the sentences people actually say. also called "verbal behavior") that is to be modeled because people are notoriously careless in their use of speech, particularly everyday speech. Rather, it is their underlying *competence*, or ability as native speakers to speak, use. and think in a given language that is of interest. Linguistic models are, therefore, idealized models of the way sentences are produced from abstract ideas and the way abstract ideas are understood from spoken/written sentences. This production/recognition process is divided into four hierarchical components. Phonology, the first of these components (first from a recognition viewpoint[1]) refers to the identification of speech sounds and is not relevant to our interest here only in written language. The other three components are morphology, syntax, and semantics.[2]

***Morphology***. The second stage in recognition, morphology, refers to the process by which individual units of information are identified from the sentence or "input string" to be understood. Frequently, these units are words but many words contain several such units, through a concatenation of a root (or a compound) with one or more affixes, each of which independently adds different information. Not all of these information units or "morphemes" convey actual meaning in the idea underlying the sentence. Many merely convey information about the relationship among the other morphemes in the sentence, for example, prepositions or conjunctions. These "function" or "marker" morphemes are used in the next component in the recognition process, syntax.

***Syntax.*** In the syntactic component, all the relationships between the morphemes in the sentence are identified through analysis of either the word order, the marker morphemes, or both, depending on the language being considered. The syntactic analysis results in identification of information structures at a higher level than morphemes. The relationships among these structures and their morphemic constituents are usually represented computationally as a tree structure, which is input to the last stage of recognition, the semantic component.

***Semantics.*** In the last (semantic) component in the recognition process, the morphemes and syntactic relationships are analyzed to reconstruct the meaning of the sentence. This "meaning" is understood by the reader or listener in terms of the underlying conceptual structures with which he/she thinks and reasons. These structures are to a lesser degree different in each individual, and to a greater degree different in each culture, but are obviously similar enough to allow understanding of speech and language to take place.

All four of these components are clearly hierarchical, with the output of one being the input to the next, but their separation is not as complete as this information processing model suggests. Syntax plays a role in morphology, for example, in the identification of word boundaries which, particularly in spoken language, are often identified only by syntactic context. Semantics likewise plays a role in syntax; ambiguities are resolved, for example, only through semantic information or the context of the sentence. The understanding and incorporation of such intercomponent relationships in the hierarchical model has been a major difficulty in constructing an accurate and complete computational model for any language.

Translation is a special type of language performance in which:

1. a sentence is heard in one language and understood through the morphology syntax and semantics of that language
2. the underlying idea is mapped into concepts relevant to another language, and then
3. expressed in that second language through the rules of its semantics, syntax, and morphology.

Translation must take place this way, at least for a human translator, because the true meaning of sentence is not retrieved until the last stage of processing, and each language has different morphology, syntax, and semantics. However, it is widely believed that the underlying concept structures. although different across languages, *can* be mapped onto one another by any human being who is competent in both. Thus, any MAT system must deal with both source and target language morphology, syntax, and semantics, automating part and leaving the rest to the human translator.

*Structure of MAT Systems.* The essentially hierarchical model described above has serious implications for the design of MAT systems. Since the components of the process are related in a sequential processing fashion, it is impossible to automate one component without also automating all lower components in the hierarchy. Morphology can be automated alone, for example, as it interfaces directly with natural written language, but mechanization of syntactic processing requires that morphology also be mechanized in order for the syntactic model to have an interface with written language. If semantic processing is done by the machine, then the entire process must be automated. Therefore, the division of effort in the translation process between man and computer can be thought of as partition of the component hierarchy, where lower components are automated and higher ones are left to the human to process.

*Morphology, Lexicon, and Pure MAT.* While morphology is a useful and psychologically valid construct in modeling language, people are more comfortable dealing with words than with morphemes. To conform to this inclination, morphology is frequently represented in MAT systems by lexicon, which is the set of words (rather than the set of morphemes) used in a given language. Dictionaries, printed or computerized, represent attempts to collect all the words or lexical items in a language along with their meanings. The lexicon appears to be the area of language translation that human translators have the most difficulty with, particularly in scientific and technical areas; it is frequently reported (e.g., Burge 1978, Dubuc 1972) that more than one-half of a translator's time is often spent in looking up the translations of technical terms. But the lexicon also appears to be the area of language processing that is easiest to automate. One approach to MAT system design, called here the pure MAT approach, is simply to computerize the lexicon of two or more languages in order to supply a human translator rapidly and automatically with target language equivalents of source-language lexical items. Usually, only specialized vocabulary, particularly scientific and technical, are included in the automated lexicon, leaving the common "core vocabulary" or everyday words to the human translator. In this type of MAT system, the machine simply stores lexical equivalents and supplies them to the human translator on request. The translator himself deals with all problems in source- and target-language syntax and semantics.

Pure MAT systems of this type are in operation at Carnegie-Mellon University (the TARGET system, see Burge 1978); the Chinese-English Translation Association (see Mathias 1975); IBM Corporation (see Lippman and Plath 1970, Lippman 1971, 1975); Smart Communications Incorporated; Canadian Government, Ottawa (TERMIUM system, see Dubuc 1972, Dubuc and Gregoire 19741; West German Federal Bureau of Languages (LEXIS system, see Krollman 1970, Daley and Vechino 1973. unknown 1976): Siemens Corporation. West Germany (TEAM system, see Schultz 1975. Schmidt and Vollnhals 1974. Brinkman 1974): and the Commission of European Communities in Luxembourg (EURODICAUTOM system, see Goetschalckx 1974).

Both the Siemens TEAM system and the Federal Bureau of Languages LEXIS system, emphasizing off-line operations, focus on the construction of text-related glossaries (Krollman 1970, p. 124). All unidentified terms in a specific text to be translated are submitted to the computer before written translation takes place. For the terms, the computer then produces a list of target-language equivalents, which is used in the translation of that specific document and can then be discarded.

Other systems operate in an interactive fashion, emphasizing on-line look-up of terms as they are requested by the translator. Some of these systems, the IBM system and the Carnegie-Mellon TARGET system, for example, also provide multi-window and text editing capabilities. By allowing relevant dictionary entries to be retained on one pan of a computer terminal screen, source-language text to be retained on another part, and target-language text to be composed and edited on still another part, these systems free the translator from the need to remember or write down the various translations of words or phrases from the current passage, or the source-language passage itself while he "polishes" the final translation. In this type of system, the terminal is not merely an automated part of the translator's work area, but it becomes the work area itself.

Beyond this on-line/off-line distinction, however, there are still other, more philosophical differences among these pure MAT systems. These differences can be described in terms of the design and implementation problems that the different systems are addressing.

The first such problem results from the variability of written representation for many languages. The writing systems for most European languages (English is a notable exception) require diacritics and accents. In these languages and in many others as well, the written form may vary further according to context or usage, so that the same word may have several orthographies. It is impossible to include all the relevant diacritical information and alternate orthographies of each entry in an automated lexicon. Therefore, before words can be accessed, they must be transformed to a standard dictionary look-up form. This transformation process, called "lemmatization," poses a problem in MAT system operation, because the human user of the system must somehow know the procedure used to reduce the term to its lemmatized form. Learning the procedure, while not impossible, still places a burden on the translator by requiring him to know more than would otherwise be necessary. It also adds to the overall time required to access a term and adds to the likelihood that a term actually in the lexicon will be overlooked merely because of orthographic inconsistencies. While some work on computerized or automatic lemmatization is being done (e.g., Hann 1974, Weber 1976), all the MAT systems listed above require users to enter words in their lemmatized form, usually one where all diacritical marks and accents have been removed and only standard, prearranged spellings are used. While lemmatization continues to be a problem in all MAT systems, it is not so serious that any system has been designed solely to deal with it.

A second problem, similar to lemmatization, is how variants of a lexical item, which differ only slightly in meaning, are to be included in the lexicon. Common examples are the various conjugations of verbs which may be considered as separate items or as derivatives of the infinitive, and common root-affix combinations which may be considered as forms of the root or as separate entries. This problem of lexical extension is severe in synthetic languages.[3] where the number of potential compound word forms is infinite. If each form of each lexical item is considered a separate word, then the automated lexicons will grow cumbersomely large. On the other hand, if only the most generic forms are included, then the user of the system must know which form of a particular term will be considered the generic form. The most commonly used compromise is to maintain separate entries for roots and affixes. But in many languages compounding alters root orthography and the translator may have difficulty reconstructing the generic form for the root of an unknown term. This problem of morphological decomposition is also severe in ideographic languages, such as Chinese, where word boundaries do not coincide with ideograph boundaries, particularly in technical terminology. Automated lexicons which provide separate entries for roots and affixes, are *automated dictionaries,* like LEXIS and TEAM, while those which contain only whole words. like SMART and Agnew Tech-Tran, are *automated glossaries.* One solution to the lexical extension problem is to provide an additional means of access to the lexicon based on meaning. In this approach, the position that all lexical variations correspond to a single lexical entry is taken one step further so that all lexical representations of a single concept are considered to be a single lexical item. The access of concepts, accomplished by including an index of the concepts contained in the lexicon, is particularly useful in scientific and technical fields, where a single concept may be adequately represented by several different words with only subtle differences in meaning among them.[4] Thus, a translator may want to find any or all possible target-language translations of a given source-language concept. Alternately, if it can be deduced from the term's context and if he cannot reduce the term to a lemmatized or generic form, he may want to locate a translation of a term through its meaning.

Terminology, words, or phrases that represent technical . or scientific concepts are of greater interest in most pure MAT systems than is core vocabulary. Terminology is distinguished here from technical lexicon by including technical descriptive phrases (e.g.. "end-point sewage treatment plan") while technical lexicon includes single words and idiomatic phrases only. Technical terminology, therefore, includes technical lexicon, which is a subset of it. In an automated lexicon, the inclusion of terminology creates additional difficulties in the construction of concept indices. for many terminological phrases consist of a cover term and one or more qualifying terms: in engineering, for example, the phrases "control engineering," "control theory," "optimal control theory," and "non-linear optimal control theory" each have a distinct meaning, yet all can be subsumed under the concept of "control." This requires the use of multilevel or hierarchical indices that would allow the retrieval of such phrases through the entry of the entire phrase or a cover concept contained in the phrase. In fact, many systems, such as TERMIUM, allow full "permutational" access of terminological phrases (i.e., access through any word in the phrase).

Automated lexicons, which are concept-based but not terminological (that is, they contain only single-word entries), are *automated thesauri,* while ones that also contain terminology are called *terminology banks;* the TERMIUM. TARGET, and EURODICAUTOM systems are all terminology banks.

While the use of concept indices as additional means of access adds flexibility to the MAT system, it also creates both operational cost and some philosophical problems of its own. As great care must be taken that a single interpretation of the indexing scheme is used throughout, the required indices add considerable software costs to the system and make the entry of new terms more complex. Moreover, no indexing scheme will contain all the possible concepts needed by translators, but must instead impose on users of the system some static categorization of concepts. While it may alleviate the problems of lemmatization and root-affix morphology, the indexing system itself must be learned by the user.

A possible method of obtaining a dynamic indexing system would be to automate the classification process by having the computer analyze large amounts of source- and target-language text to identify terms that are used similarly and hence refer to a common concept. Research, particularly in the Soviet Union (e.g.. Ivanova 1969, Berzon 1971), is currently underway in this area but results are far in the future.

A third problem, referred to as "polysemy ," is generated **by** words with multiple meanings and is thus the reverse of the lexical extension problem: lexical extension occurs when a single concept is represented by many lexical items, while polysemy occurs when a single lexical item represents many concepts. Like lexical extension, polysemy is a greater problem in technical and scientific vocabulary. where the meanings of a given term may vary in different fields. To present all possible meanings or translation equivalents each time a polysemic term is looked up not only is a waste of effort but, more importantly, is potentially confusing to the translator using the system as he tries to locate the desired translation in a screen filled with irrelevant information.

One solution used by the TERMIUM, TARGET, and TEAM systems is to classify concepts or meanings according to the technical field of interest. This subject index, used in conjunction with the concept index, allows a concept or part of a terminological phrase to be accessed, and then one or more specific usages to be chosen for display on the basis of technical field of interest. This capability, although it adds even more flexibility to the terminology bank, shares with all classification schemes certain limitations: potential obsolescence, the need to be learned by the user, and additional cost to the system overhead. In addition, while there is currently an effort, particularly in western Europe, to standardize subject indices, no standard subject classification has superseded the several competing schemes.

***Automated Syntax and Human-Aided Machine Translation (HAMT).*** In the syntactic component of language processing, the relationships among the morphemes in a sen**tence** are identified and marked so that a clear, unambiguous meaning can be derived from the sentence by interpreting individual semantic components. The process of identifying these relationships is referred to as parsing. While a great deal is known in general about the form of grammar (rules for parsing) for natural language, the involvement of the morphological and the semantic components in some parts of the parsing process has prevented the construction of a complete formal grammar for any natural language. The morphological component is involved in several ways, for example, through the existence of "gradients" of applicability of certain parsing rules to lexical items in particular grammatical classes.[5] More important, however, is the semantic component, involved particularly in the identification of metaphorical or non-literal usages and the resolution of ambiguity. Although humans sometimes are deliberately ambiguous, it also frequently happens that a given sentence or phrase, parsed in two or more equally correct ways, can give rise to two or more possible meanings.[6] The native speaker/reader uses the context of the conversation to resolve the ambiguity but in doing so makes recourse to the semantic component. The use of metaphor, even in scientific terminology (e.g.. "black holes" which are not holes, or "flags" in computer programs, which are not flags), can be detected only through analysis of the semantic content and context of the usage.

By combining relatively powerful partial grammars with exceptions to the rule heuristics, it is currently possible to construct algorithms that can parse the majority of sentences in any given natural language. Such algorithms are used in the second type of MAT system, the HAMT system, in which the human is used only to resolve semantic or syntactic ambiguities, or to provide counsel on non-literal usage while the computer does the morphological and remaining syntactic processing in the source and target languages. A primary operational difference between this type of system and the pure MAT is the necessary expertise of the human translator. In pure MAT systems, a relatively knowledgeable human translator is required, on whom the bulk of the actual translation process is placed and who must deal with the syntax and semantics of the source and target language without help from the computer. In fact, the pure MAT systems currently in operation have been designed primarily as time-saving aids to professional translators. When the text being translated is also scientific or technical, as is often the case, then besides being bi- or multi-lingual, the translator must have expertise in the subject area. The number of people who can fill such roles is at best limited. In the HAMT system, the human translator, since he acts simply to supply information, is relieved from having to know in detail the syntax and lexicon of the source or target language.

The earliest HAMT system, the experimental MIND system (Bisby and Kay 1970) which translated English into Korean, used monolingual experts who answered (in the source language) questions posed by the computer concerning ambiguities or uncertain usages that it encountered. Once all the ambiguities had been resolved, the parser was able to reproduce the text in Korean in a simplified Korean grammar (i.e., one which could produce correct Korean sentences but only in certain regular grammatical forms). In this interactive system, in which the human resolved problems as they were detected, the obvious benefit lies in not requiring truly bilingual technical experts but merely monolingual ones. The system currently in operation at Brigham Young University (see Lytle et al. 1977) is a HAMT-oriented system and. in many respects, a logical descendant of the MIND system. One of the newest HAMT systems is that developed by the Weidner Communication Corporation. Using a relatively simple parsing algorithm, it makes an initial translation of Spanish text to English and then provides on-line lexical and text-editing assistance to a human translator who "cleans up" the output to a final form.

The problems with HAMT systems, unlike the problems with MAT, which are primarily philosophical, are primarily those of computational cost-effectiveness, in HAMT, problems such as lemmatization can be dealt with either mechanically (through automated input like optical character recognition) or through the use of trained input operators who do nothing but enter text to the system. The human translator, who does not interact with the dictionary stored in the computer, has no need to learn any concept- or subject-indexing scheme. But for HAMT to be practical, the parsing algorithms have to be sufficiently powerful to require only occasional human clarification. If every sentence needs human disambiguation or clarification, or if the machine parses more slowly than a human translator, the system is less efficient and certainly more expensive than a human translator alone. For the relaxed requirements on the expertise of the human translator to dominate, the syntax must be fairly complete. Among others, Kay (1976) feels that such a goal is both easily reachable and the best allocation of labor between man and machine. If a HAMT system is coupled with a capability for on-line text editing of output, then the entire system seems quite attractive. At present, however, parsing algorithms sufficiently general for a single HAMT system to translate material cost effectively in a variety of subject areas are not available, and until they are. HAMT is not practical.[7]

***Semantics, Understanding, and Pure MT***. In pure MAT systems, the computer aids the human translator by providing rapid retrieval of unknown words, terminology, concepts, and phraseology. In HAMT systems, the human aids the computer translator by providing syntactic and semantic disambiguation. If the semantic component is automated along with the syntactic and morphological components, then (in theory) the result is fully automatic machine translation—pure MT. In practice, however, all attempts at pure MT still directly involve human assistance and are therefore, even though they are not intended as such, considered here as a form of MAT.

The semantic component in language processing relates the morphemic and syntactic structure of a sentence to concepts and ideas understood by the reader/listener. It mediates between the structure of speech (or writing) on one hand and structure of long-, intermediate-, and short-term memory on the other to produce inputs to the cognitive processing of the message or meaning of the sentence. Unfortunately, of all the areas of computational linguistics, the study of semantics is, despite years of concerted effort, the least advanced. A decade of work on computational semantics is summarized in the volumes by Minsky (1968), Schank and Colby (1973), Anderson and Bower (1973), and Charniak and Wilks (1976). The primary stumbling block to the construction of adequate formal models of semantics has been the lack of a good model of (or even a clear intuition about) the structure of human memory and/or the organization of concepts in it. What results from semantic processing is understanding, and however it is computationally modeled, this understanding must be framed in terms of some conceptual knowledge base or memory structure.[8]

The question of modeling memory (or knowledge base) for use in translation systems, however, opens computational linguistics to one of the longest-standing issues in linguistics. This argument can be traced back (in its modem form) to the work of Sapir (1929. 1933) and Whorf (1955). who questioned whether the categorization of the world implicit in language originates with the experience of human speakers or whether the manner in which people experience the world is determined by the way it is categorized by their language. They accepted the latter hypothesis and, as no firm evidence has ever been obtained to disprove it, the issue has remained alive since. For translation, the question has deep implications: if meaning has no universal experience-based origin, then translation is not really possible in a computational manner. Although all languages are clearly "anismorphic" (Zgusta 1971, p. 194) in that there is not a one-to-one correspondence of concepts from one language to another, if there are really no deep isomorphisms of thought underlying all languages, then attempts to base translation of semantic understanding are doomed to failure, for this understanding would be rooted in the source language and not convertible to understanding in the target language. The hypothesis also has implications for automatic concept-based thesauri or terminology banks, as well as for pure MT. since it indicated that a universal conceptual categorization scheme is impossible.[9]

This problem is least critical in scientific/technical communication, where great effort is placed on the explicit definition of all relevant concepts and no room is left for the reader/listener to redefine experientially what is being stated. On the other hand, it is most critical in highly symbolic communication, such as poetry, where the primary goal is for the reader/listener to reconstruct the meaning of the text in resonance with a personal experiential view of reality. Thus, mythology or folklore may lose much more through translation than mathematics or physics. But semantics, lacking workable computational models, must be included in most pure MT systems at a much lower level than pure understanding of the source text. The majority of these systems do not consider semantics any further than sentence-by-sentence, using a series of heuristic semantic rules to translate the morphology and syntax of the source language sentence into an equivalent target-language form which is then turned into a target-language sentence. Cn this principle operate the translation systems at Oak Ridge National Laboratories (see Jordan, Brown, and Hutton 1977); the University of Texas (METAL system, see Lehmann and Stachowitz 1975); the University of California, Berkeley (QUINCE system, see Wang et al. 1975); The University of Montreal (METEO system, see Chevalier, Dansereau, and Poulin 1978); LATSEC, Inc.; LOGOS Corporation; and at Wright-Patterson AFB. The result is almost uniformly less-than-correct text in the target language; human editors are used to correct or smooth the style and syntax of the output text and to select from alternative translations when the system is unable to resolve ambiguities in the source text. This human intervention, termed post-editing, is usually coupled with a pre-editing of the source text to indicate certain syntactic or lexical information. The University of Texas METAL system, which requires only minimal post-editing and pre-editing, is still under development and not yet fully operational. A system in operation at the Chinese University of Hong Kong (Loh 1976) requires no post-editing, but since parsing could not proceed without very extensive human pre-editing to resolve syntactic/semantic ambiguity, it is probably better considered as a HAMT system.

Wilks (1973) has advocated a more extreme knowledge-based method which he terms the "artificial intelligence approach." He suggests that the semantic component should map the morphology and syntax of sentences in a text, first into strings in an interlingua semantic representation, and ultimately into strings in the predicate calculus. These strings, being representations in pure logic, can by reversing the operation of the semantic component then be expressed in terms of the syntax and lexicon of a target language. He has actually constructed a small pilot program that performs admirably, but a truly operational system of this type is not even contemplated. Still the success of this approach may lead to its imitation in the future, particularly if future attempts do not succeed in substantially improving the quality of sentence-by-sentence models through the addition of more and more heuristics.

Although interest in MT has persisted from the earliest days of computers (e.g., Weaver 1955. originally written in 1949), the report of the National Academy of Sciences Automated Language Processing Advisory Committee (ALPAC 1965), which concluded that high quality, fully automative machine translation was not around the corner (or even far down the block), led to a great curtailment of interest in and funding for pure MT research. The report's conclusions were opposed, however, by several observers who felt that they were based on inadequate and obsolete data (see Josselson 1971, p. 44-9,. for a summary of this position), and some MT research has continued to the present.

All pure MT systems that are operational today require the intervention of a human as either a pre-editor, post-editor, or both, before totally correct target language text can be obtained, and although some developmental systems minimize the human involvement, it is likely to remain essential for some time to come. Several studies (e.g., Bar-Hillel 1964, Sinaiko and Klare 1971, 1973) have shown that obtaining high quality finished translations by pure MT with human post-editing) is both costlier and slower than obtaining them from a human translator alone. On the other hand. Bar-Hillel (1971) has suggested that where high quality translation is not required but haste is. the removal of the post-editing could yield a trade-off of quality for speed that would make pure MT attractive to potential users. The Oak Ridge and Wright-Patterson pure MT systems both work in this kind of environment, providing rapid turnaround to those needing "quick and dirty" translations.

***Conclusions***. Interest in using computers as aids to translation has grown roughly in proportion to the awareness of the difficulties involved in using them as translators themselves. The reasons for the failure to achieve high quality, fully automatic machine translation are now clear. Although many of the problems were laid at the feet of inadequate computer hardware (or software), the real problem lay with the simplistic and fundamentally incorrect models of language that were applied to MT in the 1950's and early 1960's. The primary impact of the linguistic revolution begun by Chomsky was to point out that language is more complex than had been thought and that a complete computational model was not at hand.

It was in this light that the ALPAC report concluded machine translation research should be suspended and replaced by basic research into computational linguistics and by development of ways to use computers as aids to translation. Since the report's publication in 1966, great advances have been made in the latter endeavor and considerable but less substantial progress in the former. Machine translation is still in the experimental or developmental stages, but machine-aided translation is currently an operational reality throughout the world.

## TABLE A-1. PURE MAT SYSTEM

| Institutional Location | Languages | Data Base Size (# of Entries) | Area of Specialization | Comments |
|---|---|---|---|---|
| Carnegie-Mellon University, Translation Center (TARGET System) | English (*)<br>German (*)<br>French (*) | 3-5K | Iron/Steel Mfg. and Commercial/Financial Terms | Spanish (*) currently being added. System used on-line, in conjunction with window-editing systems. |
| SMART Communications, Inc. | French (*)<br>English (*)<br>Spanish (*) | 500K | Engineering/Military Hardware Terminology | German (*) and Arabic (*) currently being added. Used in conjunction with text simplification system to translate technical manuals. |
| Chinese-English Translation Assistance (CETA) Group | Chinese (S)<br>English (T) | 500K<br>1K | Scientific/Technical Terms<br>Common Usage Terms | Includes data from many other Chinese-English data bases in this country. |
| IBM Corporation Thomas Watson Research Center | Italian (*)<br>English (*)<br>French (*)<br>German (*) | 50K | Scientific, especially Computer and Electronic Terms | Developed and used for internal translation of IBM documents into European languages. |
| Siemens, A.G. West Germany (TEAM System) | 8 European Languages<br>All (*) | .5M | Scientific/Technical Terms | Batch mode operation intended primarily for use in dictionary and technical glossary publication. |
| Commission of European Communities, Luxembourg (EURODICAUTOM System) | English<br>French<br>German<br>+3 Others (Unspecified) | 10K | Scientific/Technical and Steel Manufacturing Terms | Used mainly in-house for translation of commission-produced documents. |
| Canadian Dept. of State, Ottawa, Canada (TERMIUM System) | French (*)<br>English (*) | 4M | Engineering/Scientific Terms | Largest and oldest terminology bank in the world, can be used on-line throughout Canada. |
| West German Federal Language Bureau (LEXIS System) | German (*)<br>English (*)<br>French (*)<br>Russian (*) | .9M | Engineering/Technical Terms | Mainly interrogated through batch processing; printed and computer composed microfiche output available limited on-line service (i.e., ≤ 4 simultaneous users). |

## TABLE A-2. HAMT SYSTEMS

| Institutional Location | Languages | Data Base Size (# of Entries) | Area of Specialization | Comments |
|---|---|---|---|---|
| Brigham Young University | English (S)<br>Spanish (T)<br>Portugese (T)<br>German (T)<br>French (T)<br>Chinese (T) | 10K "word senses" = 50K words (for each language) | General Usage and Ecclesiastical Terms | |
| Chinese University of Hong Kong (CULTSystem) | Chinese (S)<br>English (T) | Unavailable | Mathematics and Physics | Used to translate the journals of ACTA MATHEMATICA SINICA and ACTA PHYSICA SINICA into English. |
| Weidner Communication Corporation | English (T)<br>Spanish (S) | 20K | General Usage Terms | English (*), Spanish (*), and French (*) are under development, and existing data bases are under expansion. |

### *Appendix: Operational MAT Systems*

Reference was made in the main body of the paper to numerous MAT systems. This appendix provides some detailed information on the major operational MAT systems in North America and western Europe, Bruderer (1977, 1976) lists many additional systems in operation in eastern Europe and Asia. In Table A-1, which presents information on eight pure MAT systems, each entry is identified by its name and/or institutional location, the languages it supports, the size of its lexical data base, and its subject areas of specialization and emphasis. Under the languages supported, source languages are indicated by an S, target languages by a T, and languages that can be used interchangeably as source and target by an asterisk (*). The sizes of the data bases are indicated in thousands of entries (K) or millions of entries (M). In some cases, additional comments are provided. It should be mentioned that today almost all major published dictionaries, glossaries, etc., are constructed with the aid of a computer, and most exist in a magnetic tape form. However, the mere fact that a dictionary, etc., exists in a computer-accessible form does not mean that it can actually be used as a MAT data base. In fact, quite the contrary is true. Without a well-conceived internal record segmentation structure, a software scheme for accessing it, or a practical searchable index, a dictionary on a magnetic tape is no more automated than its hardbound form. Therefore, included here are only data bases that have been expressly constructed for the purpose of MAT and can be operationally used as such.

Table A-2 provides data on three HAMT systems identical to that provided in Table A-l for pure MAT systems, and Table A-3 gives the same information for six pure MT systems. To aid in the comparison of entries from the three tables, it should be noted that the MT and HAMT data bases tend to be larger than the MAT data bases because they include "core" or everyday vocabulary not usually included in MAT data bases, as well as specialized or technical terms.

Tables A-l, A-2, and A-3, which are not an exhaustive survey of *all* MAT programs, are intended rather as an extended sample of the major operational systems. Excluded are those systems on which development has stopped short of operational implementation, as are those which were never intended to be operational, but were constructed only as applications of new principles of computational linguistics in limited experimental frameworks.

## TABLE A-3. PURE MT SYSTEMS

| Institutional Location | Languages | Data Base Size (# of Entries) | Area of Specialization | Comments |
|---|---|---|---|---|
| University of Texas Language Research Center (METAL System) | German (S)<br>English (T) | >100K | Scientific/Technical | Still under development. |
| Logos Corporation | English (S)<br>Farsi (T)<br>Spanish (T)<br>French (T)<br>Russian (T)<br>Vietnamese (T) | 5-150K (varies with language used) | Science, Electronics, Electrical and Civil Engineering | |
| LATSEC, Inc. (SYSTRAN System) | English (*)<br>French (*)<br>Spanish (*)<br>Russian (*) | 20-40K | Biology/Physics | Essentially the same translation program as that used at Wright-Patterson AFB, with a smaller data base. |
| Oak Ridge National Laboratory | Russian (S)<br>English (T) | 31K=300K words[1] | Scientific, especially Physics | Direct descendant of the original Georgetown MT system, first implemented in the early 1960's. |
| University of Montreal | French (T)<br>English (S) | 1.2K | Meteorology | Provides translation of Canadian government weather forecasts. |
| Wright-Patterson Air Force Base | English (T)<br>Russian (S) | 300K | Scientific/Technical | |

[1] The system is capable of recognizing alternate syntactic forms of the same terms (e.g., plurals or gerands) and, therefore, a given number of data base entries can actually provide responses for a larger number of words.

### *NOTES*

The discussion will be framed in terms of language recognition, but everything applies equally, in reversed order, to language production. Therefore the term "recognition" should be interpreted as "recognition/production."

2. Still another component, pragmatics, is often included in the hierarchy above semantics, but is not discussed here. Dealing with the situational or interactional considerations in language use, it is primarily relevant to verbal discourse rather than expositor or written language. Thus, its relevance to computational linguistics lies in the area of interactive natural language processing, not translation.

3. Languages which allow new words to be formed freely fan existing roots, stems, and whole words are termed "synthetic," and ones which allow instead a great deal of meaning qualification through syntactic context manipulation are termed "analytic."

4. Lippman (1975, p. 309). for example, lists twelve terms in English which each represent the concept of a "linking loader" as it is used in computer science.

5. This is a complex issue and cannot be treated here in any detail. See Ross (1973) for a further discussion.

6. The classic example (from Chomsky) is "flying planes can be dangerous" which has two obviously different meanings.

7. The new Weidner system purports to meet the above criteria for practicality, but only time and the marketplace can tell.

8. This position has become axiomatic for researchers in monolingual natural language processing (those attempting to construct natural language query systems for data bases. for example), but is still far from accepted by computational linguistics working in translation.

9. In particular, Kuhn (1972) suggested that all scientific discourse is framed in terms of scientific "paradigms," conceptual models of the interrelations among the ideas used in a particular field of science at a particular point in time. Political, social, or cultural barriers can result in quite different paradigms being used by scientists in different cultures; for example, as the American concept of pathology-based medicine contrasts with the mainland Chinese paradigm of medicine as homeopathic. It would be difficult to subsume such opposing paradigms in terms of a single underlying semantic model or concept index. Changes of Paradigms, such as the shift from classical to quantum mechanics, are accompanied by total restructuring of the underlying conceptualization.

## *REFERENCES*

1. Anderson. John and Bower, Gordon, eds., *Human Associative Memory* (New York: John Wiley and Sons), 1973.

2. Automatic Language Processing Advisory Committee (ALPAC), *Languages and Machines--Computers in Translation and Linguistics* (Washington: National Academy of Sciences. National Research Council). 1966.

3. Bar-Hillel, Y., "A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation," *Language and Information* (Reading, Massachusetts: Addison-Wesley), 1964, pp. 174-149.

4. Bar-Hillel, Y., "Some Reflections on the Present Outlook for High Quality Machine Translation," NTIS Document AD-737573. 1971.

5. Berzon, V.E.. "Some Techniques for Formalizing the Process of Establishing U-Relations Between Sentences in a Correlative Text." *Nauchno-Teknicheskaya Informatsia,* Series *2,* No. 8. 1971.

6. Bisby. R. and Kay, M., *The Mind Translation System: A Study in Man-Machine Collaboration,* Report P-4786 (Santa Monica, California: Rand Corporation). 1972.

7. Bruderer. Herbert. *Handbook of Machine Translation and Machine-Aided Translation* (New York: North-Holland). 1976. in German.

8. Bruderer. Herbert. "The Present State of Machine and Machine-Assisted Translation." in Commission *of* European Communities. *Overcoming the Language Barrier* (Berlin: Verlag Dokumentation), 1977.

9. Burge. John. "The Target Project's Interactive Multilingual Dictionary." Project Technical Report No. 13 (Department of Modern Languages and Computer Science, Carnegie-Mellon University), 1978.

10. Charniak, E. and Wilks, Y., eds., *Computational Semantics* (New York: North-Holland), 1976.

11. Chevalier. J., Dansereau, J., and Paulin. G.. *TAUM-METEO: Description du Systeme* (Montreal: University of Montreal), 1978.

12. Chomsky, Noam. *Syntactic Structures* (The Hague: Mouton). 1957.

13. Chomsky, Noam. *Aspects of the Theory of Syntax* (Cambridge: MIT Press). 1965.

14. Chomsky, Noam. "Review of Skinner's *Verbal Behavior," Language*. 35(1959). 26-58.

15. Daley. D.H. and Vechino, R.F.. USAF, "The West German Federal Bureau of Languages and Machine Aided Translation in Germany." *Federal Linguist, 5.* 3-4(1973), 14-18.

16. Dubuc, Robert. "TERMIUM: System Description," *META*, 17, 4(1972). 203-19.

17. Dubuc, Robert and Gregoire, Jean-Francois, "Banque de Terminologie et Traduction," *META,* 20, 4(1974). 180-84.

18. Hann, Michael. "Principles of Automatic Lemmatization." *ITL: Review of Applied Linguistics,* 23, (1974). 3-22.

19. Ivanova, I.S., "Problems of Automatic Thesaurus

Construction." *Nauchno-Tekhnicheskaya Informatsiya,* Series *2.* No. 1, (1969) 17-20,

20. Jordan, S.R., Brown, A., and Hutton, F.C., "Computerized Russian Translation at ORNL," *Journal of the American Society for Information Science,* pp. 26-33, 1977.

21. Josselson, Harry, "Automatic Translation of Languages Since 1960: A Linguist's View," *Advances in Computers.* Franz Alt and Morris Rubinoff, eds. (New York: Academic Press, 1971), pp. 1-59.

22. Kay, Martin, "The Proper Place of Men and Machines in Translation," *American Journal of Computational Linguistics,* Microfiche 46, 1976.

23. Kuhn, Thomas, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press), 1962.

24. Lehmann, W.P. and R.A. Stachowitz, *Development of German-English Translation System,* Final Technical Report. NTIS Document AD-A008525 (University of Texas Linguistic Research Center), 1975.

25. "The Lexicography Information System (LEXIS) of the Bundeswehr Language Service." *Machine-Assisted Translation in West German.* JPRS Document 68726. 1974.

26. Lippman, Erhard. "An Approach to Computer-Aided Translation." *IEEE Transactions in Engineering Writing and Speech.* 14. 1.1971), 10-33.

*27* Lippman, Erhard. "On-Line Generation of Terminological Digests in Language Translation," *IEEE Transactions on Professional Communications,* PC-18, 4(1975). 3-9-18.

28. Loh, Shiu-Chang. "Translation of Chinese Language Journals by Computer," *Association of Literary and Linguistic Computing Bulletin.*

29. Loh, Shiu-Chang and Kong. L., "Computer Translation of Chinese Scientific Journals," in Commission of European Communities. *Overcoming the Language Barrier* (Berlin: Verlag Dokumentation, 1977).

30. Lytle, E.G. et al., "Junction Grammar as a Base for Natural Language Processing," *American Journal of Computational Linguistics,* Microfiche 77, No. 3, 1977.

31. Mathias, James, "The Chinese-English Translation Assistance Group and in Computerized Glossary Project." *Federal Linguist, 5,* 3-4(1973), 7-13.

32. Minsky, Marvin, *Semantic Information Processing* (Cambridge: MIT Press). 1968.

33. Ross, John Robert, "A Fake NP Squish," *New Ways of Analyzing Variation in English,* C.J. Bailey and R. Shuy, eds. (Washington: Georgetown University Press, 1973), pp. 96-140.

34. Sapir, Edward, "Conceptual Categories in Primitive Language," *Science,* 74(1931), 578.

35. Sapir, Edward, "The Status of Linguistics as a Science," *Language.* 5. 207-14.

36. Schank, Roger and Colby, Kenneth, *Computer Models of Thought and Language* (San Francisco: W.H. Freeman), 1973.

37. Schmidt, R., and Vollnhals, O., "The Use of the Lexicographical Branch of a Data Bank System to Produce a Phraseological Technical Glossary," *Machine-Assisted Translation in West Germany,* JPRS Document 68726, 1974.

38. Schulz, Joachim, "Lexicography with TEAM—Automatic Dictionary Composition," *Machine Assisted Translation in West Germany,* JPRS Document 68726, pp. 23-24, 1974.

39. Sinaiko, H.W., "Translation by Computer," *Science,* 174(1971), 1182-84.

40. Sinaiko, H.W.. and Klare, George R., "Further Experiment in Language Translations: Readability of Computer Translations." *ITL: Review of Applied Linguistics* 15(1972). 29

41. Sinaiko, H.W.. and Klare, George R., "Further Experiments in Language Translation: A Second Evaluation of the Readability of Computer Translations." *ITL: Review of Applied Linguistics.* 19(1973), 29-52

42. Skinner, B.F., *Verbal Behavior* (New York: Appleton-Century-Crofts, 1957).

43. Smith, Raoul N.. "Computational Bilingual Lexicography: A la Recherche du mot juste." Paper read at Foreign Broadcast Information Service Seminar on Computer Support to Translation. 1978.

44. Wang, William et al., *Chinese-English Machine Translation System,* Final Technical Report (Berkeley: University of California at Berkeley, 1975).

45. Weaver, W.. "Translation," *Machine Translation of Languages,* W.N. Locke and A.D. Booth eds. (Cambridge: Technology Press, MIT; New York: John Wiley and Sons, 1955).

46. Weber, Heintz Josef, "Automatische Lemmatisierung—Zielsetzung und Arbeitsweise eines Linguistischen Identifikationsverfahrens," *Linguistische Berichte,* 44(1976), 3-47.

47. Whorf, B.L., *Language. Thought, and Reality: Selected Writings of Benjamin Lee Whorf,* John B. Carroll ed. (New York: John Wiley and Sons, 1956).

48. Wilks, Yorick, "An Artificial Intelligence Approach to Machine Translation," *Computer Models of Thought and Language,* Roger Schank and Kenneth Colby, eds. (San Francisco: W.H. Freeman and Co., 1973). 114-15.

49. Zgusta. Ladislav. *Manual of Lexicography* (The Hague: Mouton, 1971).