[From: *Georgetown University Round Table on Languages and Linguistics 1989*, James E.Alatis, ed.]

Machine translation: Achievements, problems, promise

Winfred P. Lehmann
The University of Texas at Austin

1 Achievements. Among its many achievements, the demonstration by the School of Languages and Linguistics that human language can be manipulated by computer is unique. Not surprisingly, the interest in such a possibility, and the demonstration as well, were concerned with a practical benefit, translation. As in many developments, others had concerned themselves with the possibility earlier. Locke has given a clear and precise account of early history, including the patent application filed by Smirnov-Troyanski of Moscow in 1933 (1985:129-32; see also Hutchins 1986:21-24). And Warren Weaver's memorandum of 1949 entitled 'Translation' must be credited for its suggestion. But these efforts and the activities they encouraged remained matters of academic concern until 1954, when the founder of the School, Leon Dostert, in cooperation with Paul Garvin and others, demonstrated with a small set of Russian sentences that a computer could actually be programmed to carry out translation, indicating in this way its capability of controlling human language. By indicating the feasibility of computers to manage texts, the demonstration brought wide attention and financial support for research in the humanities that previously seemed to require little but sabbaticals for scholars and occasional subsidies to assure subsequent publication. For somewhat over a decade the area of language study that came to be called 'computational linguistics' thereupon enjoyed broad support.

In hindsight, the basis for the manipulation is obvious and straightforward. Language is a symbolic system. The computer is a symbol-manipulating device. Yet many accomplishments that subsequently seem obvious require someone with insight, energy and daring to bring them about initially. While the basis of the control of human language by computer is straightforward, Georgetown University in its School of Languages and Linguistics achieved the first demonstration of that activity.

However great the initial enthusiasm, computer manipulation of language would not have gained academic and funding support without continuously added evidence that the expectation of success was valid. It was fortunate that the School was established with a dual aim, as indicated in its designation: 'of languages and linguistics'. Its faculty included members who controlled both language and the ability to carry out the necessary analyses, notably the organizer of this session, Michael Zarechnak. These linguists carried out the necessary and often wearisome tasks of intricate analysis that provide the

means for a mechanical device to deal with the exceedingly complex symbolic system used by human beings to communicate.

The work encountered many obstacles. Among the chief was criticism and even derision from fellow humanists. It's scarcely necessary to recall examples; but it is important to state that such responses illustrate the general ignorance of language, even among scholars who concentrate on it, such as literary specialists, and also the unawareness of its uses in society, which may be even more widespread. With a bit of thought anyone who ever bought a car might have noticed that the manual in the glove compartment is longer than works like *Paradise Regained* and many other literary classics. Moreover, manuals are not written in the style of Milton's poems, nor of his essays, nor of law briefs, nor even of newspaper articles. Attention to literary works, including their translation, might readily achieve results with no further equipment than a pen, paper and a desk. But technical materials, such as manuals for complex equipment, and texts produced by administrative organizations like governments, are infinitely longer and might well merit treatment with the help of additional equipment, both in the interests of time and completion. I might note that I didn't cover the distance from my hotel to this room by means of a 747. Much as members of society apply different tools for differing objectives, they also apply differing forms of language in different contexts for differing purposes. The promise of computerized translation to manage one of these forms—the language of technical documents—brought funding for the computerized study of language for some time after the successful Georgetown demonstration.

That study during a brief period of about a decade yielded both practical and theoretical achievements. Among practical accomplishments were translation systems that long remained in use. The Russian-to-English system developed here was installed in sites in this country and abroad, with results that met the approval of its users (see Hutchins 1986:70-78). The School also served as training center for specialists, a contribution often overlooked when evaluating centers of computational linguistics. I mention only two such specialists. Peter Toma, Georgetown's primary systems specialist, went on to develop SYSTRAN, the translation system that has been most widely applied, as by the European Community with its massive language requirements. And Muriel Vasconcellos is leader of the group that developed Spanish-English and English-Spanish systems for the Pan American Health Organization.

In spite of such achievements, the activity was discredited and the funding stopped by the report of a committee of the National Academy of Sciences, supported by the National Science Foundation. It is difficult to overstate the damage caused by this report to the national welfare and to research in the humanities. Without using time and space to detail results, I merely mention that this last year another committee of the National Academy of Sciences, its Computer Science and Technology Board, listed among six preeminent challenges a translating telephone. The committee had learned that the Japanese had appropriated 120 million dollars for a seven-year project to 'build a telephone that will translate from Japanese to English' (Waldrop 1988: 1436). The committee also understood the difficulties adequately to indicate that a necessary step on the way to the aim was achievement of machine translation. Twenty-two years after the Academy killed research and funding,

not only here but abroad as well, it suddenly resuscitated the goal, asking federal funding agencies like the National Science Foundation to provide support. In the meantime the Foundation has invested funds for investigating frog calls and other animal communication systems. We may wonder when it will assign funds to meet the revived goal.

2 Problems. The chief problem faced by linguists seeking to achieve machine translation is the widespread ignorance of language, its structure as well as applications. One can hardly read any comment on language, even by scientists successful in other areas, without finding confusion of language with its written form. And because any infant can acquire language, it is assumed to be simple. Further thought might suggest the error of such an assumption concerning a communication system embodying means to express any concept, however complex or however novel. Examples may be taken from any science, for example, biology; specialists had no difficulty devising ways for discussing the immune system, which was largely unknown before work to combat AIDS. Documentation of the immune system requires as many as a million terms. To cite one example of a well-known scientist's basic lack of understanding of language, we may recall Jerne's address on receiving a Nobel Prize. In that address he alluded to transformational grammar and proposed that research would one day uncover the biological bases of human language.

As any linguist knows, the essence of language is relationship. There is no biological basis for the word *cat.* It is associated with a value, to use Saussure's term, through its difference from other words like *sat* and *mat.* Moreover, longer segments of language such as Strawson's widely repeated sentence *The cat sat on the mat* has its value through its difference from other sentences like *The cat sat on the mat?* or phrases like *The basset on the mat...* By identifying and mastering those values in any given language, speakers are able to communicate, to express and understand meanings. There are no biological chunks for Jerne or any future investigator to identify.

Understanding of a system based on relationships of massive complexity is elusive, as the five millennia or more of attention to language demonstrates. Much of this attention was external, such as providing lists of words with definitions, or counting letters, as in cabalistic or koranic study. In time general insights were achieved, as by Charles Sanders Peirce, identified by Sir Charles Snow as one of the two outstanding minds produced by this country in the nineteenth century.

To achieve an understanding of language Peirce posited three important relationships: those among meaningful elements of language that he called signs, those between speakers and signs, those between signs and the outside world. Such a position, as is well known, sorts out three areas of attention: the relationships among signs, which Peirce called *syntactics;* the relationships between signs and speakers, which he called *pragmatics;* the relationships between signs and the outside world, which he called *semantics.* Peirce's framework, now generally referred to as *semiotics,* permits us to distinguish and gain control of segments of human language without achieving the mastery of a normal speaker.

In this way the framework is of crucial importance for what we may provisionally call abnormal approaches to language. Among such approaches

in syntactics is construction of artificial languages like Esperanto or Fortran. Among them in pragmatics is manipulation of a speaker's relationship to an audience, as in the PR associated with television. Among them in semantics is restriction of the sphere allowed in a limited communication system, such as expert systems. But Peirce's insights were scarcely noticed, certainly not by linguists.

A second major problem concerns support for the humanities. The field is associated with gentility. Members of humanistic professions are supposed to concern themselves with history, with literature, and other topics of intellectual but not practical benefits. Equipment, even a typewriter, might be useful for enterprising humanists but nothing that was more complex. And practical applications, like translation, scarcely achieved the dignity for discussion, let alone subsidy, except for works purified of any but academic merit. As a result, computerized attention to language remained, and remains, outside the domain recognized as proper for respectable humanists to pursue and for funding organizations in the humanities to cultivate. And linguists, notably those enjoying their self-adopted label 'mainstream', disdained the activity, except when they were able to profit from available funding.

3 Promise. To examine the promise of computational attention to language we can hardly do better than note the situation in chemistry, one of the oldest and most successful of sciences. As we all know, the field has become increasingly complex. Thirty years ago an eminent chemist of my acquaintance used to quip that his students had never used a test-tube. Some time later the department at our university set out to solidify its eminent position even further: they sought out and appointed computational chemists. As is well known today, chemists may approach no nearer to the experiments many of us carried out in our elementary chemistry course than a computer screen. And the designation of such specialists: theoretical chemists!

By contrast, in linguistics the term 'theoretical' has been appropriated by linguists who do not deal with language, but, if at all, with 'an ideal language spoken by an ideal speaker-listener in an ideal society'; preferably, they confine their attention to grammar. One may attach too much importance to designations, but I would like to propose that instead of 'theoretical' such linguists should be referred to as speculative. And attention to an 'ideal' or any other kind of mythical language should be labeled 'speculative linguistics.'

A major problem of academic or scientific attention to the humanities until recent times has been the absence of any possibility of testing one's hypotheses. The physical scientists, as Galileo demonstrated, could devise convincing tests. Even when dealing with abstruse hypotheses, like Einstein's, convincing tests were devised, as the British astronomers demonstrated to general delight by observations during an eclipse of the sun in 1918. Linguists can't look to the heavens for such demonstrations, nor can they find convincing evidence by positing fanciful structures like the Language Acquisition Device in the brain, but they can use computers. It may not be inaccurate to claim that the most important contribution of the computer is and will continue to be its provision of a device to test hypotheses formed about the means of communication developed by human beings, and thereupon to lest other humanistic activities.

As above, I cannot take time to discuss the early problems. Among these were the pitifully primitive computers; it is about as pointless to try to inform a contemporary student in computational linguistics about the IBM 650 or even 709 and similar advanced computers of thirty years ago, or about punch cards for inputting data, as to portray life during the depression to a teenager today. And to mention that early computational linguists had to program in machine language puts one among the head hunters of Papua. In any event, things have changed. Anyone with any interest can learn about Rumelhart's experiments with language acquisition by computer or Sejnowski's NETtalk, a neural net 'model that can learn to pronounce English' (Roberts 1989:481). Without any inborn or implanted language acquisition device, computers have acquired elements of human language.

However admirable these experiments, we here are concerned at least in part with control of language materials for specific ends. One may ask: why translation? Why not question-answer systems? Or expert systems? Or speech signals to a robot on an assembly line? And so on. We respond to these and similar questions with the general statement that such applications make use of tricks rather than a thorough control of human language. In a sense they are clever adaptations like the computer programs that get us a seat reservation from an airline, or tell us our current bank balance. For advances in control of human language a computer has to handle it in much the breadth of its human speakers.

So-called knowledge-based systems may be even more dangerous, through specious attractiveness. They are based on the assumption that meaning exists somewhere out there, and that shrewd techniques, notably those utilizing various kinds of logic, can be used to codify and in this way control meaning. Codification is attempted through words, often with the assumption that these are the carriers of all meaning. The notion is dispelled in the first day of elementary classes in linguistics with utterances like *oh yea?* in contrast with *oh?—yea.* Devotees of knowledge-based systems might review linguistic study of the eighteenth century. To it we owe Roget's *Thesaurus,* and similar works for other languages, which are useful occasionally to writers. But the nineteenth century learned that much meaning is conveyed through phonology and morphology and syntax. Further, these meanings are fundamental to communication, as utterances like *Did she?* vs. *She did.* illustrate. In short, for advances in control of human language a computer has to handle it in much the breadth of human speakers.

To achieve that breadth, translation is the most readily realizable application, and accordingly the optimum initial goal of linguists who seek to establish a responsible base in theoretical linguistics.

I support this statement briefly: Computerized translation is concerned with technical materials, that is, with language that is most circumscribed pragmatically and at the same time most simple semantically. *Oxygen* has one meaning, in contrast with a word like *love;* and verbs like *oxidize* have a far more circumscribed meaning than do verbs like *get* or *hold.* Moreover, syntactically, technical materials are more closely regulated than is other language, even that of government documents. It is no accident that METEO, the Canadian system for translating weather reports, is one of the most rewarding that has been devised and put in use. As it and other examples

illustrate, by achieving machine translation theoretical linguistics gains control of one strand of human language. In the way of science, it will move from that mastery to broader mastery.

We might review briefly some of the aims of that mastery. In linguistics, as in other sciences, advances are most solid through a theoretical approach. Practical goals disclose problems; their solutions are most fruitful when general principles are sought, for these often solve an entire range of problems, not merely the ones detected. In view of the interruption of research, insights achieved by specialists in machine translation are forgotten or credited to other researchers. I cite only a few.

In the realm of general structure, Victor Yngve distinguished between left-branching and right-branching modifiers in languages; English restricts left-branching severely, in contrast with permitting numerous right-branching. Moreover, the two processes are favored in specific languages, as is right-branching in English—with relative clauses and object clauses placed to the right of heads—and left-branching in other languages, like Japanese—with such clauses preceding heads, so that *the man who came to dinner* is expressed in Japanese with sequences corresponding to 'dinner to came man'. Today such study is carried out in typology, following the lead of Joseph Greenberg; specific characteristics are related to clause patterns, notably by attention to V(erb) O(bject) and OV languages.

In the area of pragmatics Erwin Reifler proposed distinguishing the vocabulary of what are now referred to as sublanguages or different registers. The distinction today is widely observed in sociolinguistics.

In the field of syntax we may note Zarechnak's study of the *'-sja* verbs in Russian' (1971). Using twelve features, he arrives at a set of formally determined classes. It is interesting to compare his results with those of Geniušienė in her more general work 'The Typology of Reflexives' (1987), especially since both credit Xolodovic for ideas. Zarechnak's conclusions can be pursued in greater detail because of 'its suitability for computer programming', as he points out (1979). Unfortunately, the cessation of funding checked the possibility of further such studies.

As a final example, Hutchins attributes the introduction of formalization into linguistics to work in machine translation (1986:59). These selected examples may illustrate the contributions we may expect for accurate understanding of language by resumption of funding.

That funding must also be adequate to support research into further areas of language and broader spheres than technical language. Some semantic studies are being carried out, but only on selected problems. The report of the Academy's Computer Science and Technology Board indicates that the way to achievement of a translating telephone requires a massive effort, including 'a machine translation system capable of dealing with all the vagaries of human language, including ambiguity, nongrammatical phrases, and incomplete phrases' (Waldrop 1988:1436). This achievement then demands computer control of the pragmatic and semantic as well as the syntactic spheres of language. The Japanese, as the Board also reports, have allotted more than a hundred million for the telephone that 'will translate from Japanese to English'; this sum is in addition to the massive amount made available over the past decade, in comparison with no funding in this country.

And as the Board recognizes, control over coded texts is a far cry from control over speech. We have a bit of work ahead of us. It is curious that the imaginative computational work on language that enjoys considerable funding is carried on not by linguists but by psychologists like the Rumelhart group and by biologists at the Salk Institute, while the primary specialists dealing with language remain without support.

And even when the translating telephone is achieved—as things are going, by the Japanese—we will be only at the beginnings of computer control over language. Translation is important, as for the European community, for scholars dealing with special problems like translation of the Bible, for forward-looking leaders in the Third World who see computerized translation as the means to bring their countries up to date in education, technology and government.

We also need to yoke the computer to identify the material in publications that is pertinent to specialists. Even in linguistics, publication is so massive that no one presumes any longer to control all of it; and chemists gave up long ago even covering the abstracts of publications in their field. We might also allude to the massive intelligence communities. For all such groups translation is only a first and partial step, far less important than control over data. Computers must be harnessed to cull out desired data, to perform data retrieval.

After data retrieval, computerized procedures will be extended to fact retrieval, that is, not simply to determine the information in texts but the knowledge controlled by human speakers. We may then be able to use the power of computers for managing the data assembled when we deal with the social and political problems of the prospective ten billion inhabitants of the earth, the data of biological complexities like those of the immune system or harmful agents like viruses, and of problems in the physical world like aerodynamic systems, as well as those resulting from experiments with supercolliders, and so on. Even with funding comparable to that provided by international research groups, we can scarcely expect such achievements until well into the next century. Yet it is scarcely surprising that the foremost countries economically today are those supporting the research and technological applications involving computational linguistic activities.

In the meantime, the control over broader types of communication, among them oral language, will yield many applications. Possibly the most welcome among these would be assistance to the handicapped. We have seen how current computer capabilities permit the noted scientist Hawking to communicate still. Control over speech, combined with advances in psychological and biological research, like the several alluded to above, could greatly assist those similarly handicapped. Other advantages of computer mastery of speech have been projected by imaginative technological specialists, including such commonplace activities as directing a vehicle. Technologists and specialists in other sciences must never forget that all of these depend on understanding language, that is, results derived from research carried out by linguists with an accurate understanding of human language.

To return to the current scene and hopes of funding, we look forward to resumption of broader activities of this university in computational studies. Thanks to Professor Zarechnak's continued work, teaching and research in this

field has never ceased here, as documented by reports like that given by him at the Kentucky conference in 1988. And as this celebration indicates, the School has maintained its energy and productiveness. Few commemorations of its important demonstration thirty-five years ago would be more significant than reestablishment of a large-scale project.

Note

Linguistics Research Center, The University of Texas, Austin, TX 78713-7247. The Center is supported under a contract with Siemens AG, Munich, and by a grant from the Division of Research Programs, Texas Higher Education Coordinating Board.

References

Dym, Eleanor D., ed. 1985. Subject and information analysis. New York: Marcel Dekker.

Geniušienė, Emma. 1987. The typology of reflexives. Berlin: Mouton de Gruyter.

Hutchins, W. J. 1986. Machine translation: Past, present, future. Chichester: Horwood; New York: Wiley.

Lehmann, Winfred P. Machine translation, in Dym 1985:110-23.

Locke, William N. Machine translation, in Dym 1985:124-53.

Roberts, Leslie. 1989. Are neural nets like the human brain? Science 242:481-82.

Rumelhart, David E., James L. McClelland, and the PDP Research Group. 1986. Parallel distributed processing. Explorations in the microstructure of cognition. Cambridge, Mass.: MIT Press.

Waldrop, M. Mitchell. 1988. National academy looks at computing's future. Science 241:1436.

Zarechnak, Michael. 1979. The history of machine translation. In: Machine translation, by Bozena Henisz-Dostert, R. Ross Macdonald, et idem. Trends in Linguistics 11. The Hague: Mouton. 4-73.

Zarechnak, Michael. 1986. The intermediary language for machine language translation. In: Computers and Translation 1:83-91.

Zarechnak, Michael. 1988. Machine translation: Present status and future outlook in the global milieu. In: Global demands on language and the mission of the language academies, ed. John Lihani. Lexington: The University of Kentucky. 151-71.