P-1409-C

VOLUME 1

NATURAL LANGUAGE AUTOMATED PROCESSING AND ARTIFICIAL INTELLIGENCE: STATE-OF-THE-ART

COGNOS

1984 FEBRUARY 29

DSS FILE 12ST.41414-2-03X2

DOC: David Waung

NOTE

This report, State-of-the-Art, is the first volume in a series of reports addressing Natural Language Automated Processing and Artificial Intelligence. The reports in this series are:

- The State-of-the-Art
- Implications of New Technology Thrusts
- Possible Social and Economic Impacts
- Opportunities for Canada
- A Program Plan, and
- Priorities for Canada

ACKNOWLEDGEMENTS

The material in this report has been assembled by the contractor from reports submitted by subcontractors (Isabelle, Kittredge, Pylyshyn, Skuce, Stein) and consultant (Mermelstein).

We are especially grateful to Dr. Paul Mermelstein and to Bell Northern Research for their help with the speech recognition and synthesis section of this document.

EXECUTIVE SUMMARY

This report surveys a variety of research endeavours and technological achievements in Artificial Intelligence (AI), Natural Language Processing (NLP) and Machine Translation (MT). Among the more successful of the AI products are the Expert Systems, which exhibit human-level expertise in narrow areas of specialization (examples include diagnostic, advisory, designer, or tutoring systems for medicine, science, engineering, or business applicaions -- many of which are now commercially successful). After a brief examination of these systems, the bulk of the report focuses on areas of AI that relate to NLP, with further special emphasis on MT research. We consider the survey of accomplishments in general AI to be germane, however, inasmuch as we take the position that development of MT must be viewed as part of the research on NLP, which in turn has to be viewed as part of the general effort in AI. We take this view because we believe that in the long term success in MT will draw on the development of techniques, research methodologies, and even particular systems designs emanating from the worldwide effort in AI in general, and in NLP in particular. Thus we devote some space to an analysis of successes in Expert Systems, Speech Recognition (and Synthesis), and Character Recognition. We attempt to show that there are lessons to be learned from these successes, as well as direct benefits to be had from successful products developed in these areas. Considerable space is devoted to the question of why certain commercially viable products have become possible at this time. This discussion introduces the important theme of suboptimization of designs as a compromise between technical feasibility and commercial needs.

In one of the main sections of the report we discuss a variety of areas of research concerned with NLP. Perhaps the most successful of these are the database query systems which allow users to ask questions of a database in English or some other natural language. Such systems, believed to be beyond the technical horizon just a few years ago, are now fast-selling commercial products -- and are even available on micro-computers. Other NLP developments (most of which are still in the prototype-building stage) include 'text critiquing' systems which help authors improve their grammar and style, text synthesis systems which generate coherent text (such as summary reports) from structured data (e.g., stock market data), and computer-aided language instruction systems.

A central part of our survey in this report concerns progress in machine translation, which we understand to include not only automatic translation of languages, but also machine-augmented human translation and human-aided or human-enhanced machine translation. In this section we consider various suboptimization techniques, such as those involving specialization to sub-languages, as well as pre and post-editing. We also review a variety of currently operational (or near-operational) commercial systems, promising developmental projects in Canada, the U.S., Japan, France, and the European Economic Community. Finally we briefly describe approaches being taken by one or two longer-term proposals and survey some of the areas of pure research we believe will feed into a successful design for MT. We observed in these discussions that there is a wide range of performance demonstrated by these ideas and that the most developed of the commercial products, while capable of relatively limited performance, nonetheless can still be economically viable in special circumstances. The critical issue in the short term appears to be the development of the right sorts of suboptimization techniques and settings. In the long term, however, it appears that it will be necessary to develop (or import from other areas of technology) a variety of specialized techniques such as the ones being explored in AI laboratories and NLP centres around the world.

An interesting facet of our investigation of products and research programs in machine translation is that it revealed a significant isolation between the development of commercial MT products, carried out largely in the private sector using relatively old technologies and motivated by strong market demand, and the main body of research in AI. We feel that this split is not due to the nature of the technical problems in MT so much as it is a legacy of an overcorrection to the unrealistic expectations of early (1960s) research in MT. For this reason it may be that the relative lack of longer term MT research (especially in large AI laboratories in North America) represents a special opportunity for Canada to enter the field with a serious program of research. However, we emphasize that in our view such a program must be (a) diverse in its approach (it is too early to put many eggs in one basket), (b) varied in its goals (long-term goals must be kept in mind as well as suboptimal solutions to satisfy critical immediate needs), and (c) cast as part of the more general goal of developing AI and NLP research programs, as well as basic research programs into knowledge representation, computational linguistics, computational reasoning, and the human-computer interface. Furthermore a program to promote the development of MT in Canada, while extremely worthwhile, must not be veiwed as a substitute for a more general plan to foster AI research. In the final report we argue that AI in general is too important to our national interest to be left at its current level of support, especially in view of the efforts being put into such research by our international technological competitors.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

CHAPTER 1
INTRODUCTION AND BACKGROUND

Although this report will be concerned primarily with describing the state-of-the-art in Natural Language Processing (NLP) and Machine Translation (MT), we will mention a number of related areas of Artifical Intelligence (AI). Part of the reason is that several of these areas, for example speech recognition and synthesis and optical character recognition, provide important potential products that are directly relevant to the problem of making NLP products more useful and more generally available. But there are some additional, and in some ways more important reasons why in the long term NLP and MT should be viewed as at least closely related to, if not an integral part of, the field of artificial intelligence. We list a few of these below.

- The more advanced techniques of NLP are to a considerable extent drawn from work in AI, although, in some cases they are unfortunately not drawn from the most current generation of AI technology. The same programming languages are typically used and new techniques, search algorithms and representational data structures developed in other areas of AI find their way into the more ambitious NLP projects. The most advanced NLP researchers find themselves to be part of the same research community as other practitioners of AI though later we will consider some exceptions to this general principle especially in recent commercially motivated MT developments.

Furthermore, we expect that as time goes on this will be more and more true; as a greater variety of AI techniques are developed and as more ambitious high-performance MT and NLP projects are attempted, there will be more and more need for such AI techniques as those associated with knowledge-based AI methods. Such systems might, for example, require the use of special AI techniques, and even specialized highly parallel hardware architectures such as those being developed in artificial intelligence laboratories.

- It is important to maintain the broadest possible approach to MT and NLP in order to maximize the chances of significant breakthroughs. Recently a panel of experts gathered to discuss MT at a meeting of the Association for Computational Linguistics emphasized repeatedly "...that it is important for work to proceed in parallel on a number of different fronts" (Kay, 1982). One front that many people believe holds the most promising key for MT in the long run involves combining NLP with knowledge-based approaches such as those used by 'expert systems' of the kind we will describe below. Plans for NLP and MT in other countries - particularly Japan - also take this approach. For example, the authors of the Japanese Fifth Generation Computer Plan (to be sketched below) believe that significant progress in NLP and MT requires the simultaneous development of more generally useful AI tools and techniques.

- Finally, one of the most important reasons for considering NLP, MT and AI together in this report is that there are important lessons to be learned from the experience in other areas of AI. For example, one could learn something about the factors that determine how and where technical feasibility and market demand come together, what sorts of factors make it possible for extremely advanced software enterprises to succeed, what the life cycle of a complex piece of AI software has turned out to be in other areas, and how much help we could expect from new innovations in hardware and software technologies or from basic research into such related questions as how human translators carry out their task. We shall return to this issue in the concluding section of this report.

In Chapter 2, we will point out the recent dramatic increase in interest in the field of AI among both governments and industry as shown in a number of national plans and industrial initiatives. In Chapter 3 and again in Chapter 7, we shall examine more carefully the reasons why AI has achieved successes. We will do this because we believe that such an approach gives us a rational basis for examining the likelihood of Canada making significant progress by using various different approaches to NLP, MT and perhaps other AI areas.

CHAPTER 2

NATIONAL PROGRAMS IN OTHER COUNTRIES

First we will briefly sketch the motivations and general outline of national plans in Japan, Britain, the USA and the EEC. These national plans are in most cases described clearly in the various reports cited in the bibliography (Appendix A). A brief consideration of the Canadian scene will also be given. An inventory of Canadian research and a consideration of special opportunities for Canada, in the light of other national plans along with our existing commitments and capabilities are discussed in a separate report called Opportunities for Canada.

2.1 The Japanese Fifth Generation Computer Plan

One the most ambitious and carefully thought-out research and development programs is the Japanese Fifth Generation Computing project launched in October 1981. In June 1982 the Institute for New Generation Computer Technology (Japanese acronym ICOT) was formed under the direction of Dr. K. Fuchi. This institute is currently in its second year of a three year assured government funding allocation totalling about Y11 billion (roughly $60 million Canadian) and is expected to be funded at over half a billion dollars over a ten year period and to eventually involve some 200-300 persons. At the Tokyo conference which marked the launching of the plan, the Japanese spoke of the goal of this project as the development of computer systems, using the most sophisticated AI techiques and new hardware designs, which would narrow the gap between computer technology and the average nontechnical user.

In initiating this project, according to Feigenbaum and McCorduck (1983), the Ministry of International Trade and Industry (MITI) intends Japan to capture at least 30% of the computer market, thus making it the world leader in computer technology by 1990. Another goal stressed by ICOT is to increase the productivity of all sectors of Japanese life, including those that are currently far removed from computer technology, and to improve the quality of life of Japanese citizens.

To this end the project is directed at increasing the ease of computer access by broadening the range of inputs to include voice, pictures, print, and handwriting. Thus, at the centre of this plan is the development of a natural language input capability. Feigenbaum & McCorduck describe the importance of NLP in the Japanese plan as follows:

> "Because nonexperts will be the largest group of users, natural language processing is one of the most important research goals of the Fifth Generation. Research here will cover speech wave analysis, phonetic and syntactic analysis, semantic analysis, and pragmatic analysis, which derives understanding by extracting themes or foci in a given sentence, detecting focus shifts, and so on. For speech output, sentence generation will also be studied. Text analysis is also considered a part of natural language processing by the Japanese, although they are quite aware that the techniques used for large-scale text analysis are quite different from the techniques needed to smooth the way for an individual user to talk to his machine.

> Natural language processing will...be put to use in the development of a highly ambitious machine translation program (initially between English and Japanese) with a vocabulary of 100,000 words. The goal is 90% accuracy (the remaining 10% to be processed by humans). The translation will be part of an integrated system that takes part in each of the processes from compilation of the text to printing the translated documents."

The Japanese plan also recognizes that making computers accessible to the widest range of potential users requires not only that input be in natural language, but also that the computer systems contain a great deal of knowledge of various subjects, including in particular the sort of knowledge that goes into 'common sense' reasoning. The idea is that in order for a user to be able to obtain access to computer systems with minimum knowledge of their operation it will be necessary for the computer systems to have knowledge of the users and to be able to make rapid inferences based on this knowledge. Thus, the development of knowledge-based inference systems, with all that this implies regarding the need for new software and hardware tools, will have to play a crucial role in the plan. The plan, and its implications for the computer and information based industries in the West, is dramatically discussed by Feigenbaum and McCorduck.

2.2 The British Initiative: The Alvey Report

The British have been quick to respond to the Japanese initiative. A report commissioned by the Minister for Information Technology recommended that a national programme for Advanced Information Technology be launched at a cost of 350 million pounds (about $650 million) over 5 years (Alvey, 1982). Our latest information (from our British AI colleagues who were involved in formulating proposals for the committee) is that the report was favourably received and that a substantial portion of the requested funds (two thirds was to come from government) will be approved (a large portion of the work has already been funded by the Science and Engineering Research Council). In fact, The Economist (September 16, 1983) reported funding of about 180 million pounds (about $333 million) was being organized "to subsidize joint research into microchips, advanced software, new ways of talking to computers and artificial intelligence." The report recommended that the existing information technology activities in Britain be doubled and the program be run as a UK effort, though it strongly urged that close co-operation with other countries be maintained because of the complexity of the enabling technologies.

The Alvey Committee identified four major areas of information technology as essential enabling technologies and therefore as requiring special consideration in the overall plan. These are: (i) software engineering in general, including the development of better software tools, (ii) the man-machine interface, including the development of input and output devices using visual, speech, and touch sensors, (iii) intelligent knowledge-based systems (IKBS), including deductive data-bases and various types of expert systems, and (iv) hardware technologies, particularly new types of processors suitable for IKBS, and technologies for the design and manufacture of very large scale integration (VLSI) chips. We believe that this list of priority technologies is exactly right and will return to this point in our analysis of requirements for a Canadian effort to be discussed in the Opportunities for Canada Report.

2.3 The American Program

The United States of America does not have a national program specifically directed at guiding the research and development effort in MT, NLP and AI. However, these fields are receiving massive funding through the Department of Defence (DOD) (DARPA, ONR, AFOSR, ARO), the National Institutes of Health, and the National Science Foundation (NSF). At the present time the U.S. remains the world leader in AI and NLP, though its efforts in MT are much more limited in comparison. There is great concern in the U.S. about the economic impact of the Japanese initiative and they are studying the Japanese plan carefully.

One DOD initiative which can be viewed as a response to the Fifth Generation Plan is the DARPA project known officially as 'Strategic Computing and Survivability' but generally called 'The Supercomputer Project'. It is estimated that it will involve between $350 million and $750 million over a five year period.

U.S. computer companies are particularly concerned about the possibility of losing their competitive edge to the Japanese. A number of these companies have formed the Microelectronics and Computer Corporation (MCC), under the direction of Admiral Bobby Inman (formerly of the National Security Agency). The corporation was spearheaded by Digital Equipment Corporation (DEC) and Control Data Corporation and involves over a dozen U.S. computer manufacturers. Its goal is to support advanced R & D work that will allow the U.S. to keep ahead of the Japanese initiative. According to The Economist (September 16, 1983) they currently have "an annual budget of about $50 million, rising to $100 million in three years or so."

A recent change in U.S. tax laws has resulted in a large influx of research funding to universities. For example, a group including IBM set up the Semiconductor Research Corporation (SRC) to finance research on VLSI at academic research laboratories. Stanford University also has set up a new Centre for Integrated Systems, financed by 19 companies at a cost of $19 million, as well as a Centre for the Study of Language and Information, financed by the System Development Foundation at a cost of about $14 million. Other academic sites, including Carnegie-Mellon, Massachusetts Institute of Technology (MIT), Brown and Yale also have major laboratories financed by computer companies or consortia of private industries. For example, Carnegie-Mellon has a Centre for Robotics Research and MIT has just established Project Athena with a budget of $70 million ($50 million of which is to come from DEC and IBM) to develop a campus-wide experiment in computer-aided education using distributed computing.

The recent announcement by Mr. William Baxter, the Assistant Attorney General in charge of antitrust policy, that the U.S. government would not challenge cooperative research has also opened up new joint ventures in the computer industry, paralleling the MCC and SRC corporations mentioned above, to help offset the Japanese initiative.

Although as of this date there is no equivalent of an Alvey or Fifth Generation Report within the U.S., various government agencies have produced reports on different aspects of AI research, and a Congressional Committee is examining the Japanese Fifth Generation 'threat'. There is a major report prepared by the National Bureau of Standards and National Aeronautics and Space Administration called 'An Overview of Artificial Intelligence and Robotics'. A recently published report of an NSF information technology workshop (Waltz et al., 1983) provides an assessment of the state-of-the-art in natural language processing and expert systems, and makes recommendations for future directions. Also the three volumes of the Handbook of Artificial Intelligence (Barr & Feigenbaum, 1981; Cohen & Feigenbaum, 1982), or the much briefer chapter on Artificial Intelligence in the NSF sponsored Computer Science and Engineering Research Study (Arden, 1981) provide summaries of some of the accomplishments and directions of research in artificial intelligence in the United States.

2.4 Continental Europe and the European Economic Community (EEC)

A number of European countries have active research programs in AI and NLP. These include Italy (with a modest investment and about 50 researchers), West Germany (with considerable activity in NLP, vision, knowledge representation and theorem-proving, involving several hundred professionals), France (which conducts AI research at the Centre National d'Etudes des Télécommunications, in Lannion, at the Centre Mondial in Paris, and in Marseille and Grenoble - these will be described later), in Sweden (which is known for its theoretical studies, as well as work on speech recognition, dialogue systems and expert systems), and in Norway (which is currently also conducting a study for a national plan). The only countries we are aware of that have national plans for AI and 'Fifth Generation' are France (CNRS, 1983) and the Netherlands. One component of the French plan is a large-scale MT project called ESOPE which has been budgeted at 56 million Francs (about $9 million) over three years. In preparation for their national effort the Dutch have produced an excellent report (Dutch Ministry of Education and Sciences, 1983) recommending urgent expansion of research into areas of speech, NLP, expert systems, cognitive modelling and human-machine interaction, and various support technologies. We will examine the Dutch proposal in connection with our proposals for support of research in Canada because of similarities of scale.

The EEC has a major interest in information technology in general, as exemplified by the ESPRIT information technology program on which $1.5 billion will be spent over the next 5 years (Commission of European Communities (ESPRIT), 1982) and the FAST forecasting and assessment program (Commission of European Communities (FAST), 1982, vols 1 & 2). Moreover, in December 1982 the EEC launched one of the largest efforts ever undertaken in MT - the EUROTRA project. Over the next five years 27 million ECUs (about $28 million) will be spent on development of a large-scale system that is intended to be capable of translating several types of texts among all 7 languages of the EEC (i.e. between 42 language pairs). This project will be described later.

2.5 AI in Canada

There has so far been no major study or recommendations made in Canada for a national plan to support the development of artificial intelligence, although there is considerable interest at the federal level in related fields such as microelectronics (for example the first phase of a 'Microelectronics Cooperative' has just been funded by NSERC at an initial level of $17.5 million). Indeed, Canada is one of the few industrialized countries without a national plan that includes AI and NLP as an explicit component. In spite of this, there is a significant amount of excellent AI and NLP research in Canada, although it is virtually all in universities and is funded through the normal science funding channels at a level far below that of comparable facilities in the U.S., Britain, or France. For example, there is some innovative work in machine vision at McGill University and the University of British Columbia, some work on schemes for knowledge representation at the Universities of Ottawa, Toronto, Saskatoon, British Columbia and Simon Fraser University.

Canadian researchers have built up certain areas of excellence during the past 10 years. Much of this work has been reported at meetings of the Canadian Society for Computational Studies of Intelligence/Société Canadienne pour Les Etudes d'Intelligence par Ordinateur. In fact this society was formed in 1973 eight years before the U.S. formed such a society. Our current areas of special expertise include image analysis (particularly satellite imagery), basic research and theorizing in machine perception of certain kinds of photographs (such as

X-Rays, pictures of metal castings), basic problems of knowledge representation, and basic research in the human-machine interface. We have a strong concentration of expertise in this country in what has now become a central part of the Japanese Fifth Generation Plan, namely Logic Programming with experts at the following Universities: British Columbia, Western Ontario, Waterloo, and Ottawa. We also have several centres of excellence dedicated to the study of human factors in computing including campuses in Calgary, Edmonton, Guelph, London, Montreal and Ottawa. See, for example, the inventory of some of this work carried out for the Department of Communications by Roger Kaye of Carleton University. One of our major areas of expertise relates to language, expecially machine translation, and we shall discuss this in Chapter 5.

Not only do we possess areas of special competence for which we have research programs already in place, but we also have certain problems of special national concern that lend themselves to AI methods. Some of the current AI research has been applied to particularly Canadian problems, such as aids for the management of the forestry industry by the use of computer analysis of Landsat imagery or aids for medical research by computer aided analysis of X-Ray and other kinds of pictorial data. In addition, of course, there is the well-known METEO system for translating meteorological reports. Developed by the former TAUM Project, this system is used daily by the Canadian Meteorological Centre (CMC) at Dorval International Airport.

There are also areas of application that are important to Canada which have not been pursued. For example, one of the most successful applications of expert systems AI work, carried out in the U.S.A. was on the mechanization of expertise in prospecting for minerals and petroleum.

The work was done in collaboration with a Canadian prospecting expert from British Columbia, yet it was carried out in California at SRI International. Other obvious areas of special concern to Canada include those associated with communications and with language. In the case of communications, AI techniques have had almost no application so far.

CHAPTER 3
STATE OF THE ART IN SELECTED AREAS OF ARTIFICIAL INTELLIGENCE (OTHER THAN NLP AND MT)

The above sketch of several national plans for AI, NLP and MT shows the extent of interest in these fields and also the extremely ambitious nature of the anticipated results. What evidence is there that such expectations are warranted?

Any discussion of such topics is bound to be controversial. There have been unrealistic expectations in the past and many people remain skeptical. Some of the problems of general reasoning, robotics, vision, automatic speech understanding and other related areas present difficult theoretical problems. Nonetheless, there have been some unquestionable successes in recent years. Below we will very briefly mention a few of them and indicate some of the hallmarks of the successful programs. Once again details will be limited because the focus in this report is primarily on the current state of progress in NLP and MT, rather than AI in general. The main objective of the present sketch is to illustrate some points concerning the potential for impressive accomplishments even where many of the more general theoretical problems remain unsolved. In the concluding section we shall again take up the question of the lessons that can be learned from these and related projects; lessons concerning sub-optimization, short vs long term goals, methods of financing and organizing large-scale projects, and so on. These lessons will later be used as the basis for some of the recommendations in our final report.

3.1 Expert Systems and Knowledge Engineering

3.1.1 Background

Expert Systems (ES) are one of the longest-established application areas of AI. By application area, we mean particular practical problems in which AI methods have been used to try to automate mental tasks previously done only by humans. The purpose of an ES is to perform a task in such an area in a manner which approximates that of a human expert, both in how it reasons, and how it interacts with the user. The reasoning depends on a body of knowledge encoded in a manner easy for humans to understand and verify, while the interaction should be as natural as possible and preferably in natural language.

It is not our purpose here to give a history of ES research (see, e.g. Barr and Feigenbaum, 1982, Gevarter, 1982, or Nau, 1983) but rather to categorize and summarize the current essential trends in ES types, designs, applications, and development.

3.1.2 Types of Expert Systems

We distinguish five types corresponding to major application areas of expert systems.

3.1.2.1 Diagnostic Systems

These are the most common ES types, at least so far. Their function is to decide on the basis of many pieces of evidence, some possibly vague, into which category a particular case of some phenomenon should be classified. The medical systems, typified by MYCIN (Shortliffe, 1976) are of this type.

3.1.2.2 Designer Systems

This category is somewhat broader, typically including systems whose task is to design or synthesize an instance of some object or system which must adhere to certain rules. Examples of such systems are the EL circuit design system (Stallman and Sussman, 1979), the VLSI design system of (Campbell et al. 1982), and the well known XCON system (McDermott, 1982) which configures VAX installations routinely at the main plant of Digital Equipment Corporation. The latter is one of the few designer systems in routine use.

3.1.2.3 Advisor Systems

Advisor systems work interactively with a human expert to amplify his/her level of informed decision-making in situations where there may be no clear right/wrong answer or design to be determined. The boundary between these and diagnostic systems is not always clear; for example the geophysical systems like the DIPMETER ADVISOR (Smith and Baker, 1983), play a role somewhat like that of a medical doctor trying to infer, on the basis of various clues, what possible things might be happening below the earth's surface.

3.1.2.4 Tutor Systems

A closely allied function is in teaching a budding expert, perhaps even a first-grader. The behaviour of the system is that of an experienced human colleague, providing hints, suggestions, facts, and reasons why the user's or its statements about the subject are good or bad.

An early such system, SOPHIE (Brown et al., 1974) was clearly a tutor, providing a student electronics repair person both a simulated laboratory and teacher. Some more recent systems of this type are reported in (Sleeman and Brown, 1982). A descendant of MYCIN, NEOMYCIN (Clancey and Lettinger, 1982), makes use of the architecture of MYCIN to tutor medical students who are learning the knowledge that MYCIN has.

3.1.2.5 Inventor Systems

These systems have as their goal the totally automatic discovery or creation of designs, concepts, even works of art, which are considered new and interesting by humans, and which would require inventiveness on the part of humans. Such a system might invent a new circuit, molecule, or melody. None of the systems described previously would be considered inventive, but the BACON system (Langley et al., 1983) is claimed to actually 'discover' laws of physics from experimental data, though these laws had already been discovered by physicists. The AM and EURISKO systems (Davis and Lenat, 1981) evidently discover new concepts in mathematics, though not yet any that have excited mathematicians. Recently, the latter program has been given the task of proposing new types of VLSI designs with some success (Lenat, Sutherland and Gibbons, 1982). In computer programming, analogous methods are being explored for program synthesis. In these, a program is written either automatically or semi-automatically by describing what it is supposed to do, leaving it to the synthesizing program to decide how it should actually be written.

3.1.3 Current Applications of Expert Systems

3.1.3.1 Scientific and Engineering

Some of the earliest ESs were tools for chemists. One of the earliest of these was a system called HEURISTIC DENDRAL (Lindsay et al., 1980). A product of a Nobel laureate in genetics (Joshua Lederberg) and an AI expert (Edward Feigenbaum) this system accepts mass spectrometer data and nuclear magnetic resonance spectrometer data from an unknown organic substance and works out possible molecular structures for it. It does this not by having a precise model of chemistry, for that is both too complex and too incomplete for the purpose, but by having built into its program all the rules-of-thumb and intuitive knowledge and experience that the researchers could uncover by patiently interviewing working chemists. Most of this knowledge is implicit and not the sort of thing that is explicitly taught in chemistry courses. This system and its successors METADENDRAL and MOLGEN (Stefik, 1981) have been so successful that they have a number of papers in chemical journals to their credit.

More recently systems for engineering applications have been developed; EL (Stallman and Sussman, 1979), an electrical engineer's assistant, and a nuclear reactor assistant (Nelson, 1982). The VLSI design system and XCON mentioned above are also notable electrical engineering design aids. The geophysical programs such as PROSPECTOR (Duda, Gaschnig and Hart, 1979) and the DIPMETER ADVISOR (Smith and Baker, 1983) have attracted considerable attention.

3.1.3.2 Medical

Medicine is the other classic ES area, since MYCIN has served as a parent to a number of other medical systems, and as a widely-described example of a simple yet effective ES methodology. INTERNIST (Pople, 81) is probably the most impressive ES running today, containing tens of thousands of facts, and competing very successfully with senior medical doctors. Systems developed at Rutgers (Kulikowski and Weiss, 1982) and MIT (Solovits, 1982) also perform impressively on more restricted domains. Many of these systems are now undergoing clinical trials to determine their strengths and weaknesses.

3.1.3.3 Educational

Computer-aided instruction (CAI) has been a long-sought goal of educators and industry. For example, Control Data Corporation has spent nearly a hundred million dollars on its PLATO system. However, to date it has not been financially successful. Traditionally, these systems have been large, expensive, difficult to program, and not remarkable in use.

Recent developments in intelligent CAI (ICAI) systems, which incorporate features in common with other AI systems, particularly other expert systems, have rekindled hope for widespread use of computer-based learning systems (Sleeman and Brown, 1982). For example, NEOMYCIN (Clancey and Letsinger, 1981) is built on MYCIN's rule base of medical rules, to which is added a second rule base encoding the tutoring strategy. Like many of these systems, it also has a model of the student's knowledge and behaviour. NEOMYCIN and its related system, GUIDON (Clancey, 1979) has been used in teaching medical students, but it is still experimental.

BUGGY (van Lehn and Brown, 1981) typifies the considerable group of ICAI systems devoted to teaching primary grade skills such as arithmetic. It attempts to infer a model of the student's incorrect performance from examples, borrowing from basic computer science research in inferring rules from examples. It then tries to debug this model, like correcting an incorrect program, borrowing again from AI research in automatic program debugging. If it can locate the bug, it trys to explain to the student why it is incorrect, as a good teacher should.

EXCHECK (Suppes, 1981) represents the other end of the academic spectrum. It attempts to teach university students basic principles of mathematical reasoning and proof techniques, a notoriously difficult task. It has been in use at Stanford University for several years.

3.1.3.4 Business

A new dimension is opening in ES: business applications. In May, 1983, the Graduate School of Business Administration of New York University held a three day seminar on ES for business; they received more requests for admission than they could accomodate. They have launched a project sponsored by IBM in conjunction with Metropolitan Life, to develop an 'insurance expert' coupled to existing databases. Two examples of developments in this area are a portfolio manager (Cohen and Lieberman, 1983), and a system which encodes laws (McCarty and Sridharan, 1982)

3.1.4 Commercial Development of Expert Systems

Commercial development of ESs is being vigourously pursued in probably thirty to fifty different organizations in the U.S. alone. To this must be added corresponding efforts in academic locations, which tend to be closely linked to surrounding companies, many of them started or peopled by academics or ex-academics. In the U.S., probably of the order of 500 persons are now involved in ES development, with perhaps 100 more working on basic research which feeds ES development directly. The areas being pursued include the traditional medical, engineering, scientific and educational fields, plus a newcomer, business. Several companies have been launched in the U.S. specifically to develop ESs for business applications.

Canada has no commercial activity in ES as of mid-1983. Otherwise, a few projects have received very minimal funding. For example, The Defence Research Establishment Atlantic is developing systems to identify sonar signals and Skuce at the University of Ottawa (Skuce, 1983) has been developing the LESK language and system for general-purpose ES applications. Tsotsos (Tsotsos, 1981) at the University of Toronto has developed a system for analysing cineradiographic images of the human left ventricle. Elio at the Alberta Research Council has begun work on geo-physical ES applications. Dahl and Cercone at Simon Fraser University are developing a natural language system for advising students.

3.1.5 Suboptimization and Performance

What is significant about most of these recent expert systems is that they achieve their powerful performance not because they represent important new computer science innovations, but because the designers discovered (a) a domain in which it was possible to extract from a human expert a great deal of relevant knowledge that contributed to his particular expertise, and (b) a way to encode this knowledge in a system that does not just store and retrieve it, but uses it in making inferences (both rigorously logical deductions and merely plausible or probabilistic inferences). This is the basic secret behind virtually all expert systems. They produce expert level performance by encoding the knowledge of a human expert and using it in a systematic rule-governed way.

Part of the methodology of expert system design is well developed and explicit (e.g. the use of techniques such as production systems, backward chaining, progressive refinement, problem reduction, constraint propagation, and other methods developed in the study of search and problem solving). Indeed, the development of expert systems has become easier by an order of magnitude in the last year or two because of the availability, for the first time, of certain specialized system building tools. Tools such as the production system language OPS5e, the expert system development packages KES (Knowledge Engineering System, marketed by Software Architecture and Engineering), ART (Advanced Reasoning Tool, marketed by Inference Corporation), and LOOPS (Marketed by Xerox and described in a paper by Stefik et al., 1983) make the process of developing expert systems very much simpler.

Nonetheless, part of the methodology of creating a new expert system remains an art. In particular it is the art of uncovering the right kind of limited domain that will yield to this sort of special treatment, together with the art of knowing how to extract the relevant knowledge from experts and encoding it in one of the expert-system building tools. The methodology for the latter is now relatively well developed and described by a panel of experts in a recent book edited by Hayes-Roth, Waterman & Lenat (1983).

According to Dr. John McDermott, the creator of the first commercially used system R1 (also known to its user, Digital Equipment Corporation, as XCON) it takes between 3 months and a year to develop the right intuitions for uncovering appropriate application areas and for becoming familiar with the whole methodology. Even then, of course, some expert systems may take years to produce, depending on the complexity of the application. First the designer has to become thoroughly familiar with the application, then he must try to uncover the implicit rules that the expert uses. The first rules and intuitions that an expert provides are typically completely inadequate. It is only by trying to run them and iterating back to the expert that a high-performance system can be honed. Take a system such as R1 which automatically configures computer components (drawn from a set of about 400 possibilities) to a customer's needs; a complex task that took a highly experienced expert days of work. The initial design of this system required roughly one man year of work and its subsequent tuning took more than another full year and involved dozens of iterations between designer and expert. It grew gradually to contain some 2500 distinct rules.

The most knowledge-intensive system is said to be INTERNIST, a medical diagnosis system which considers almost 500 diseases and contains over 100,000 pieces of knowledge (Feigenbaum, 1982).

It is important to ask what makes it possible for such systems to achieve high levels of performance despite the fact that the general problem of logical inference and of general problem solving remains elusive. The answer typically given is the same one that is usually given for why language translation systems like METEO are successful, namely that the domains involved are limited. While this is no doubt true it misses an important point. It fails to explain why limited domains were not successfully exploited by earlier systems. The limited domains that earlier AI systems tackled were not so much limited as trivial toy domains such as block worlds. It also fails to take account of the fact that not any limited domains will do, only domains with special features that practitioners can intuitively discern. Knowledge engineering companies such as Teknowledge of California will not touch some domains even though they may appear limited by some criteria. Similarly, natural language customization contracts (say for natural language interfaces to databases) have been turned down because the companies that provide this service felt that the application was not one in which high performance could be demonstrated. The exact form of the limitation or sub-optimization is critical and so far it remains part of the art of artificial intelligence.

3.2 Speech Recognition and Synthesis

3.2.1 Speech Recognition

3.2.1.1 Introduction

Although much of the current speech recognition work is not viewed by its developers as a subset of the larger field of artificial intelligence, it represents an area of application for techniques perfected by artificial intelligence researchers. The prospect of spoken natural language communication between people and computers has motivated speech recognition research for the last 20 years and promises to continue to do so. It is only recently that commercial products in this area have started to justify in economic terms the extensive research investments of the past.

Speech recognition encompasses tasks varying in complexity over many dimensions. The most important of these dimensions are in the form of input (isolated words or connected sentences), the size of the vocabulary, the class of utterances accepted (the extent to which it approaches the full complexity of natural language sentences), the linguistic background of the speaker population, and the acoustic environment through which the speech passes before reaching the recognition device. While simple speaker-trained devices capable of recognizing vocabularies of 20 to 50 words have been available for at least 10 years, it is only with the most recently attained capability of incorporating very significant computing power on individual integrated-circuit chips that speech recognizer costs have dropped to the point of costeffectiveness. A recent announcement from Texas Instruments of a voice processing board for the TI personal computer incorporating a singlechip word recognizer is but one example of this capability (Dusek, Schalk and McMahan, 1983).

3.2.1.2 Current Status

Simple word recognizers have already been incorporated into a variety of voice-controlled toys thus increasing the public's exposure to simple forms of voice recognition. More complex recognizers with improved performance are becoming available. The prospect for costeffective exploitation of simple forms of speech recognition has driven industrial research in this area for the last few years and the results have not been disappointing. Cost reductions of two orders of magnitude have been achieved over only 5 years.

Clearly, only the simplest forms of recognition systems function in an economically viable mode today. These are generally speaker-trained, isolated-word recognizers with vocabularies not exceeding 50 words and cost under $1,000. NEC of Japan has offered a limited-vocabulary connected-word recognizer on the market for the last two years at a cost of roughly $50,000. More complex systems accepting isolated words from a large vocabulary (Bahl et al., 1983), or connected words from sentences with highly restricted syntactic structures (Levinson and Rosenberg, 1978), are under study in various research laboratories. Suen and De Mori (1982) provide further information on current research activities. Commercial exploitations of these developments require matching the enhanced but still limited capabilities to applications that justify the significant additional expense of processing from the more natural spoken communication that the enhanced capabilities might provide. Significant human factors problems arise from the need to remind speakers that the input is restricted to a small subset of natural language. Thus, the more complex applications of recognition are likely to be closely integrated into the application systems they support.

Advanced research in speech recognition is conducted today at the highest scientific level in the United States and Japan. IBM, Bell Laboratories, Texas Instruments and International Telephone and Telegraph are the noteworthy industrial participants. The Massachusetts Institute of Technology and Carnegie Mellon University provide examples of well-focused university-based research in this area.

The link to significant advances to computing capabilities in both these countries is not accidental. Research accomplishements are difficult to exploit without the best implementation technology. Significant additional work is under way in France, England and West Germany. Of these, the research effort in France should be noted as having expanded the most rapidly in the last five years. Speech recognition research in Canada is mostly limited to scattered efforts at a number of universities, except for the cooperative effort pursued by Bell-Northern Research an INRS-Telecommunications in Montréal.

3.2.1.3 Prospects for the Future

The performance of speech recognizers today is limited by the power of the models that we can construct and train with the aid of natural speech data. All currently used models make assumptions concerning the nature of the speech process to allow the use of mathematically tractable techniques and computationally reasonable algorithms. One can expect gradual refinement and extension of these models that allow their training with the aid of small vocabularies and few speakers. Improved techniques to allow recognition systems to shift their focus so as to concentrate on those temporal and spectral features of the speech signal that are likely to disambiguate specific confusion promise to be of significant help.

The language models employed today for speech recognition are particularly restrictive. Some researchers employ word trigram probabilities to predict the most likely words that can follow each other (Bahl et al., 1983), others construct hierarchical transition networks (Hunt et al., 1980). None of these approach in complexity the best currently available natural-language parsing systems. However, the need to exploit these models in an environment of incomplete and possible ambiguous acoustic information, makes the use of more elaborate natural language models difficult. To a large extent more complex syntactic structures have failed to be explored because limitations in the acoustic recognition capability have made the tasks impractical. As our acoustic models improve, we can expect additional progress towards more challenging syntactic structures. Finally, semantic information is known to be of potentially significant help, but has also not been exploited greatly because of limitations at the level of acoustic-phonetic modelling.

Full natural language recognition capability, such as might be required to support fully automatic translation systems, is not likely to be achieved before the end of this century. However, structured natural language input, such as might be used for limited forms of information retrieval could likely be attained by 1990. Applications that allow the conversation to be under the control of the machine so that information input by speech can be carefully structured, somewhat along the lines of hierarchic menus of today's terminals, are most likely to be successfully attacked by speech recognition techniques in the near future. Significant applications are foreseen in the office environment, such as for dictating, where the opportunity to edit and revise allows one to accept error rates that exceed those tolerable in other applications. This, of course, is equally true for interactive computer-aided translation systems.

The role foreseen for artificial intelligence in this process is that of suggesting new ideas, tools and techniques to be tested in the spoken language environment. In particular, efficient, goal-oriented search techniques have played an important role in speech recognition research in the past and promise to do so in the future. The integration of multiple sources of knowledge, including speech acoustics, phonetics, lexical information, syntax and semantics into a multi-layered but compatible recognition structure represents another important objective.

3.2.2 Speech Synthesis

Techniques for speech synthesis can be grouped according to the requirements of the particular application, a predefined restricted ensemble of spoken messages or an unpredictable unrestricted capability to convert all English/French text to voice. These are considered as follows.

3.2.2.1 Restricted Speech Generation

With the reduction in cost of digital memories, quite large bodies of text can be digitized and regenerated on demand. Depending on the storage limitations, voice compression techniques of increasing complexity can be employed. Storage requirements can be reduced to 10 thousand bits per second of speech with no quality degradation and to 800 bits per second with increasing degradation in quality. Further research can be expected to provide quality improvements over the entire bit-rate range. The economic benefits of such quality enhancement at the lower bit-rates will be most important.

3.2.2.2 Generation of Unrestricted Text

Speech synthesis boards that attach to personal computers and can synthesize unrestricted English text, specified as a sequence of ASCII coded characters, are available today for $3000 - $5000. The speech quality is generally intelligible, but somewhat unnatural. Comparable French systems are functioning in laboratories and are expected on the market soon. Future research is expected to focus in improving the naturalness of the synthesized speech. The incompleteness of the available models for syntactic and semantic analysis of natural language can be expected to act as a limitation of the quality of speech that one can generate. However, current quality is still limited at the segmental level, the specification of appropriate parameters for the speech sounds in the variety of contexts where they may occur. Accordingly, we would expect significantly better but still somewhat unnatural speech synthesis to be achievable this decade. Simultaneously further significant cost reductions can be expected, as much as an order of magnitude within 5 years.

3.3 Optical Character Recognition (OCR)

3.3.1 History and Significance

Optical character recognition represents one of the earliest attempts to simulate a uniquely human ability; that is the ability to recognize and classify patterns. Although research in artificial intelligence has diverged from the techniques employed in classical pattern recognition in recent years, there is reason to believe that both disciplines could benefit from a reconciliation.

Patents in optical character recognition antedate the digital computer and were for the most part directed towards applications in the postal system. The first practical systems were developed by Jacob Rubinow, a research engineer employed by the U.S. Post Office who founded Rubinow Systems in 1953. The company was acquired in the late fifties by Control Data Corporation, who virtually monopolized post office OCR applications for the next decade.

Many companies entered the OCR field in the sixties including IBM, and the other major mainframe computer manufacturers in the U.S. and Europe. Almost all of them withdrew by the end of the decade with little to show for their efforts. Then in the early seventies there was some hope that OCR would prove competitive against pooled data entry systems and at least eight new OCR companies attempted to develop low cost systems to meet this need. But the trend towards direct machine input, lower cost terminals, and simultaneous preparation of typed and machine readable material on word processors, combined with the relatively high error and reject rates of OCR systems resulted in the defeat of these efforts.

Since the mid-seventies optical character recognition has concentrated on the high speed recognition of postal codes. Even the postal systems of countries where cheap labour is in abundant supply, such as India and Nigeria, have purchased optical character recognition systems for evaluation. These systems all have a multi-font recognition capability and accept smudged print. Although their accuracy and throughput rate is significantly better than that of manual sorters their performance still leaves much to be desired.

A large part of the mail addressing is still handwritten or handprinted, and the need to pre-sort these envelopes imposes a large manual overhead on the recognition systems. In recent years there has been a major effort in Japan under the auspices of the Pattern Information Processing System (PIPS) project to develop systems which can read machine and handprinted Katakana characters. The companies developing these systems have as an additional incentive the award of a contract for more than a thousand OCR systems for the employee insurance system of the Japan Labour Market Centre.

The recent flurry of activity related to office automation has resulted in a demand for very low cost and very versatile OCR systems for entering data into electronic filing cabinets. An opposing school of thought has argued that since most mail will be transmitted electronically from terminal to filing cabinet and from filing cabinet to filing cabinet, there will be little real need for OCR. In general, however, most authorities agree that there will exist a large market for OCR systems whose price performance ratio enables them to penetrate the data entry market.

Another very significant market for more expensive and very high speed hand print OCR systems lies in the recognition of handprinted numeric entries on cheques, credit card vouchers, and other forms. At present these amounts are manually entered although the credit card numbers are optically read. Each number is entered twice but the error rate remains sufficiently high to be a major source of customer complaint.

For the reader who would like a broader perspective of the OCR field there are relatively few recent references outside of the journal literature. However, two useful reviews can be found in (Auerbach Publishers 1971 and Datapro Research Corporation 1976). A more comprehensive survey of the state of the art in handprint recognition is contained in (Suen et al., 1980). Professor Suen is at Concordia University in Montréal.

3.3.2 A Summary of the Present Status of Commercial Character Recognition Systems

In this section we will briefly summarize the capabilities and limitations of existing systems.

Large systems for post office and printed form applications are manufactured by Recognition Machines Inc. of Dallas, Texas, Control Data Corporation of Minneapolis, International Computers Limited in the United Kingdom, and Siemens in Germany. In Japan; Hitachi, Mitsubishi, NEC, Oki and Toshiba are active in the field. ITT, Philips, Litton Industries, and IBM no longer have efforts in this area.

In general all these machines recognize a wide variety of fonts; in some cases as many as thirty, and a number of them can recognize handprinted numerics. All of the Japanese machines can recognize handprinted Katakana characters as well as carefully printed Roman characters which are constrained within boxes. With unconstrained characters their recognition rate is worse than the 97% correct recognition and the 1% error that is considered mandatory in large scale applications (92% rejection rate is acceptable). These systems have prices ranging from $300,000 to more than a million dollars. A large part of the price is accounted for by the high speed paper loading, transporting, sorting and stacking mechanisms.

The Japanese systems are generally acknowledged to be superior. Leigh Instruments of Kitchener, Ontario manufactures a post office character recognition system based upon a Toshiba design. Leigh improved upon the original 1972 Japanese design, but has not been able to sustain the pace of development of the Japanese companies. A relatively modest financial committment to pattern recognition research has given Japan a world lead in this field.

A variety of machines for office systems are produced by Japanese, American and European companies. Most of these machines recognize between three and seven fonts but they must be told which font they are reading by a human operator who initializes the systems. Some systems have automatic feeders while others permit documents to be read by a hand-held wand. The automatically fed systems all require the typed material to be double spaced and preferably typed with a carbon ribbon. The hand held readers can read single spaced pages but the need for a human operator makes these systems only marginally competitive with typed data entry systems. Smudged characters or pages which are printed on both sides cause a marked deterioration in the performance of all available devices. There are no low cost systems which can read pages printed on both sides containing mixed fonts of various sizes with an acceptable recognition rate and at a speed which would make them commercially viable. Almost all systems are associated with mini computers. A machine which is representative of this class of OCR is manufactured by Hitech Canada Ltd. of Ottawa. A more sophisticated system with handprint capability is currently under development by National Cash Register (NCR) in Waterloo, Ontario. It is worth noting that Leigh Instruments, NCR, Hitech and earlier Litton Industries, ETC Systems, and Ferranti Packard were supported by Canadian government funding programs in their OCR endeavours.

The limited font capability, double spacing requirement, slow speed, and high price of existing systems clearly indicate why OCR is not more widely employed. It is also apparent that a low cost device that would also accept handprinting would satisfy a need among those who do not type or whose typing speed is not significantly faster than their printing speed.

3.3.3 The Economic Significance of OCR in Natural Language Processing

In the area of machine translation many publications are not available in machine readable form. This is the case particularly with newspapers, periodicals, and journals from eastern Europe. There exists a steady demand for the rapid translation of these publications from a number of U.S. and civilian and military agencies. The process of entering these data by means of keyboards using operators unfamiliar with the language being entered is costly and error prone. In many instances the errors introduced may result in incorrect translation or in a quality of translation which requires an uneconomic level of human intervention and post-processing. OCR systems which could directly accept newspapers, books, and magazines, and eliminate the human operators even to the extent of turning pages, could mean the difference between profit and loss in large scale machine translation.

Widely distributed facsimile readers communicating with a central high speed OCR system would enable a large number of remote users to share a central machine translation facility. In this case again the remote user would not be required to key in the data. The central system would read the document into the computer and return the translation to the client at some later time. In order to avoid excessive storage at the central facility the OCR would be required to convert the characters on the document into binary form for subsequent processing. The need for OCR is evident from the fact that the conversion from analogue to binary would result in more than eight million bits for a page. Character recognition would permit a reduction of the required storage by more than three orders of magnitude. The practicality of this type of operation depends upon a high speed, reliable, and versatile optical character recognition system.

In the not too distant future the holdings of the national central libraries could be converted into machine readable form. The high cost of OCR at present is an obstacle to the conversion process. Should low cost, high speed, multi-font reading devices with automatic page turning capability become available there is little doubt that large numbers of them would be utilized in this area. It is not unreasonable to suggest that the existence of large quantities of machine readable literature, much of it untranslated into other languages, will result in a significant increase in demand for machine translation. It follows that OCR and automated language translation are synergistic technologies.

The design and development of 'intelligent' database systems is a major problem area in artificial intelligence and natural language processing. In their final form these database systems will permit untutored users the flexibility of using natural language to inquire about information contained within the database. Each database system will be capable of permitting a series of searches and deductions and produce a natural language response. As yet only small prototype database systems of this sort have been constructed. When real world intelligent database systems are implemented the problem of data entry will be massive. A historical biography database confined to a single century of one country could require the entry of hundreds of historical texts. Such a database is under development at the Centre National de la Recherche Scientifique under the direction of Gian Piero Zarri (1981).

Although the problems surrounding the organization of knowledge are formidable there is little doubt that intelligent databases will be a major area of AI research and that OCR devices will be essential to entering the vast bodies of data they will require.

For the present, textual database systems for scientific, medical and legal applications are in need of effective OCR systems for entering a wide variety of literature that is not available in machine readable form.

Optical character recognition is not limited to the recognition of alphabetic characters. Ideographically writing systems such as Chinese have in fact a major need for OCR equipment. In the area of automated language translation a system which could identify the more than twelve thousand ideograms in the Mandarin dialect would be of enormous assistance to the human translator. It is not unreasonable to suppose that given an OCR reader that could identify all the ideograms, Chinese might yield more easily to automatic translation than alphabet-based languages. In the case of data input the very obvious impracticality of devising a keyboard for the Chinese character set provides the best argument of OCR as a means of input.

Since almost one quarter of the world's population will speak Chinese by the turn of the century the need for systems to automatically translate from what may become the world's largest source of publications is obvious. An OCR system for Chinese characters is a significant point on the critical path to achieving a practical automated system.

Dynamic or real time character recognition will likely be the basis for the recognition of ideograms and for languages whose alphabet is essentially cursive such as Arabic and Urdu. As in Chinese, dynamic character recognition systems are the only reasonable substitute for the keyboard as a means for entering data. There already exists a co-operative program involving Chinese scholars working in Canada on this problem at Concordia University. This co-operation could be extended and provide opportunities for jointly funded research in OCR as well as machine translation.

3.3.4 Some Representative State of the Art OCR Systems

The Kurzweil Computer Products (Cambridge, Mass.)

The Kurzweil OCR system was developed in the mid 1970's by David Kurzweil and several of his colleagues. The basic aim of its inventors was to provide a reading machine for the blind so that the system provides a voice synthesizer output which can be readily understood. The lack of emotion and emphasis in the voice has been the source of considerable complaint from blind users. In 1980 control of the company was purchased from Kurzweil and his associates by Xerox Corporation.

The Kurzweil system's performance is significantly superior to that of any other system intended for office and data entry applications. The basic strength of the system lies in its ability to 'learn' to read different font styles very rapidly. A human operator is required to 'train' the system by providing it with a sample of the font to be recognized and then informing the system of any errors it has made in the course of a trial run. A large number of fonts can be rapidly added to the system's repertoire enabling it to deal with multi-font documents containing characters with a wide range of sizes. The system can read single spaced documents and its speed is between five and ten times that of conventional typists. It is however, not without weaknesses. In particular, it will not perform satisfactorily with documents that are even slightly skewed with respect to its reading head or whose lines are skewed relative to the page. Also, the larger the repertoire of font styles that it can recognize, the slower the recognition rate and the larger the error and substitution rates. This means that given a periodical with a variety of font styles the level of performance is not too much better than that of a human operator.

It must be manually adjusted to read multi-column publications such as magazines and newspapers. Furthermore, it is not equipped with an automatic page turning peripheral device and its scanning system has not been designed to accomodate books or magazines. Nor does the system yet have a handprint capability.

The Kurzweil system has a price of between $80,000 and $150,000 depending on the peripherals. If it is fully occupied it displaces four data entry operators. In general at least one person is still required to identify rejected characters, to enter stacks of documents and remove stacks that have been read.

<u>The Caere Corporation</u> (Los Gatos, CA.)

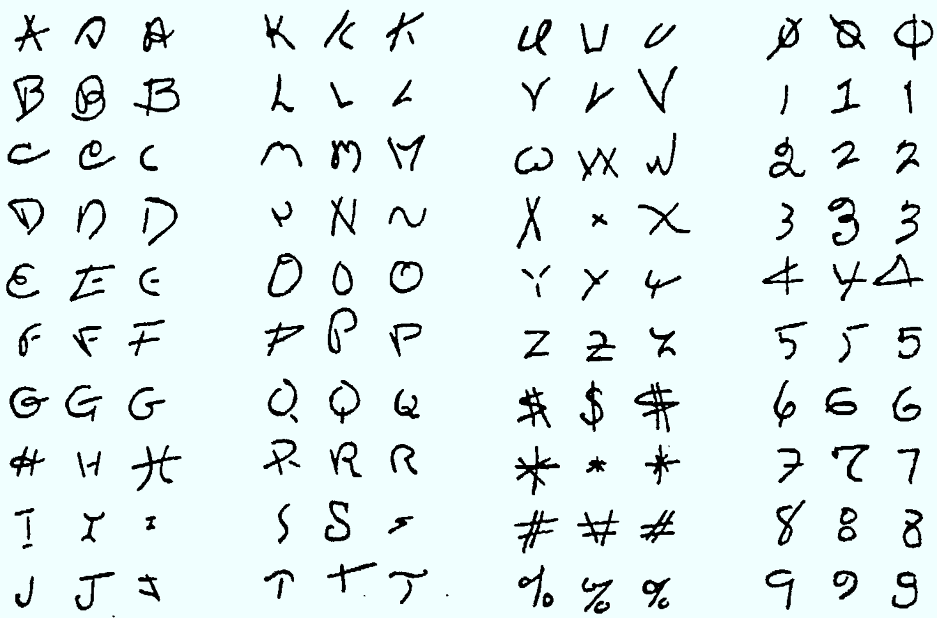
The Caere OCR system recognizes six fonts. The fonts cannot be mixed and the user must tell the system which font he intends to read. The user employs a hand held wand to read the document which can be either single or double spaced. The system is compatible with the IMB PC and has a retail price of $1,150.00. An automatic paper loader and transport is planned. The price for this system will be under $5,000.00.

Caere has built a slow and relatively primitive but effective recognition system around a microprocessor which relies on the manual dexterity of the user rather than electronics to correct for line skew. Since the user of the wand is present at all times, a relatively high rejection rate can be tolerated. The cost of off-line storage correction and display which constitutes a large part of the cost of a stand-alone OCR system disappears since these functions are performed by the personal computer.

With the availability of powerful micro-processors such as the Motorola 68000 and the National Semiconductor 16000; 256K-bit memories and low cost photo transistor arrays, it is likely that the font repertoire and performance of these low-cost systems will improve without an accompanying cost increase. Such devices provide still another means of enabling remotely located clients to transmit documents to a centrally located utility for translation and subsequent re-transmission to the client.

The Pencept Corporation (Waltham, Mass.)

The Pencept OCR reader employs a dynamic OCR recognition algorithm; that is to say the characters are recognized as they are written. In essence the computer recognizes the sequence and directions of the strokes that a writer makes in producing a character. The system can recognize characters that are sloppily and eccentrically printed such as those shown below.



The unit consists of a 15.5" x 15.5" writing tablet which can accomodate paper sizes of up to 11" x 11", an electronic pen, a control unit containing a Motorola 68000 processor, and a serial asynchronous TTY compatible port, enabling it to be associated with the IBM PC as well as with Apple, Wang, and DEC computers. Along the edge of the pad is a command area containing function boxes which can be programmed to accomodate the commands used in a particular application. When checked, these function boxes send a character string to the personal computer. Part of the edge of the pad is pre-programmed to emulate a number of the keyboard features of the IBM personal computer. The system is priced at $2,500.00 (U.S.).

The major advantage of such a system is that it permits nontypists to gain access to a number of the powerful features of a word processor. It may also be a psychologically more convivial means for some people to communicate with a natural language database and for translators who are not necessarily good typists to improve the quality of their translations with the assistance of the editing, spelling-correction, and other word processing features.

3.3.5 Paper Transport Mechanisms

The economic utility of a practical character recognition system is in many situations dependent on the availability of a convenient page-turning, paper-loading and transport mechanism. This is particulary true in the creation of databases where the documents are often in bound form. At present there do not exist reliable low cost page turning devices which can be used in conjunction with OCR systems. The fact that the print on one side of a page is often visible from the other side further complicates the task. Most users avoid this problem by first photocopying the text and then entering the separate pages into the reader. This pre-processing cost is usually ignored in the manufacturer's analysis of system economics.

3.3.6 Scanning Techniques

Early character recognition systems employed rotating mirrors and flying spot scanners. The mirror scanners were fragile and unreliable and the flying spot scanners were large, expensive, and difficult to interface with paper transport mechanisms. For the last ten years linear arrays or matrices of large scale integrated photodetectors have been employed in all practical systems. The demand for these arrays in facsimile systems has resulted in improved devices with steadily declining prices.

Most modern scanners using solid state scanning array containing about 2048 photoconductors in a half inch length. An optical system extends the length of this array to the width of the page so that only uni-directional translation of the document is required. Such a system is inherently faster, less mechanically complicated and more reliable than the rotating drum class of device. It is still, however, sensitive to line skew and hence requires a sophisticated mechanical page feeding mechanism in order to function reliably with single spaced documents. There are two solutions to skew problems which have been suggested in the literature but which have yet to be implemented. The first solution employs a retina or two dimensional line scanner containing about 300,000 photo-conductors; devices of this sort are currently employed in the Sony solid state video camera. It is apparent that the ability to take in an entire line of characters at a time rather than a thin slice of a character expedites the task of the recognition system and makes it possible to tolerate a greater degree of skew. Another, even better solution is to employ a set of scanning arrays along the length of the page. In effect the image of the entire page is stored in the memory. The line skew can then be electronically corrected within the memory. In this case two or more facsimile scanning line arrays may be required and the memory for storing the document is approximately half a megabyte in size. With declining memory costs the tradeoff of memory against complex mechanical paper feeding and transport mechanisms appears justifiable. In addition, storing the entire page in memory permits a number of pre-processing operations to be performed which considerably improve the overall recognition capability of the system.

3.3.7 Pre-Processing

Before recognition takes place a number of operations must be performed on the characters. First, the characters must be separated so that each character can be separately examined. Following this the character is centered and if necessary rotated to bring into a standard orientation. Operations are then performed to enhance the contrast ratio of the character. Finally various algorithms are employed to remove extraneous noise, usually in the form of black dots in the blank areas of the character field, and to fill in various white gaps caused by the printing process in the black parts of the character.

Recent work in the theory of low level vision has resulted in the implementation of a pre-processor called the convolution box at M.I.T. The box enables the edges of characters to be emphasized against their background and binarizes the image. In effect this improves the contrast and would likely eliminate the necessity for noise removal operations. As yet the convolution box approach has not been applied to character recognition, although Lisp machine Inc., is preparing to market this device for such applications.

3.3.8 Recognition Techniques

Recognition techniques have received the most attention in the journal literature and have been the subject for thousands of theses over the past twenty years. This is in large part due to the fact that they often have an elegant mathematical basis and can be conveniently simulated and tested against standardized test data.

In the real world the problem of single font recognition is relatively simple once one solves the mundane problems of paper contrast ratio and the line skew introduced by the paper feeder. Once the characters have been separated, centred, and oriented almost any recognition technique will be effective for the recognition of the standard OCR fonts such as OCRA, OCRB, and Courier. for multi-font and handprint recognition the problem only becomes difficult when practical time and cost constraints are introduced.

In general low cost systems employ mini-computers or micro-processors for recognition while higher cost systems employ dedicated hardware. There is little doubt that special purpose VLSI devices would reduce the cost of OCR systems and improve their speed and accuracy but as yet manufacturers have been reluctant to commit themselves to the very large initial investment.

3.3.8.1 Template Matching

In template matching a reference pattern is stored for each character to be recognized and correlated bit for bit with the finalized input characters. If a special purpose system is employed the correlation is performed by a set of correlators, one for each character; thus a multi-font system might employ hundreds of correlators (or alternatively one set of fifty correlators and hundreds of stored reference patterns). If a digital computer is employed the correlation must be performed one character at a time. It is apparent that speed must be sacrificed or that a special purpose very high speed computer must be employed. It is also apparent that this approach necessitates that the characters be of the same size as the reference patterns and they are precisely centred and oriented in the scanning field.

3.3.8.2 Transformation and Feature Extraction

In this approach the characters to be analysed are first subjected to a transformation (Fourier, Walsh, Karhunen-Loeve) which ideally renders the resulting transformed pattern invariant with respect to rotation and translation within the area of the scanning field. Particularly small areas within the scanning field are called features. Each character is uniquely associated with such a set of features.

The techniques employed to select features and to design the interrogation tree are highly proprietary. Almost all low cost computer-based systems employ a feature extraction technique. The wide variation in the performance of these systems is a function of the scanning, pre-processing, and feature extraction procedures employed.

The feature extraction technique as described is not applicable to characters of different sizes unless some magnification-invariant transformation is applied. In addition characters derived from different fonts which appear very different to a human observer are frequently incorrectly identified by this technique; an élite G and a pica E are often confused. For this reason this technique has not been successful in multifont recognition systems.

3.3.8.3 The Topological or Linguistic Approach

The topological approach is based on the common sense observation that a character from virtually any character set, regardless of its size and orientation, can be recognized by a description of the character in terms of: lines, slopes, open and closed loops, ends, and corner, and the relative position of these features. Thus, an A could be described as two lines, one with a positive slope and the other with a negative slope which meet, with a line below the meeting point touching both sloping lines. The task of the character recognition system is to extract these features and the relationship between them from the pattern. The features and their relationships are then systematically combined into descriptions employing a set of rules or a 'grammar' which are then analysed to yield the identity of the character.

3.3.8.4 Dynamic Optical Character Recognition

Dynamic, on-line, or real-time character recognition utilizes the sequence and orientation of the strokes made in the course of constructing a character. The linguistic technique lends itself naturally to the recognition task. As the character is written a list of primitives and their relationships is constructed. The list is analysed within milliseconds to determine the character. The success of dynamic pattern recognition in recognizing poorly written characters is perhaps the best argument for the effectiveness of the topological handprint and eventually cursive character recognition applications.

3.4 Machine Vision, Graphic Communication and the Human Interface

In some respects the area of machine vision has had an even more checkered history than most other areas of AI. Recently, however significant breakthroughs have occurred. The most important of these have been primarily at the theoretical level, though some commercial vision products have been developed (e.g. the trainable visual recognition module for robot applications being marketed by Machine Intelligence Corp.). The latest round of progress in vision research has been spurred by the theoretical work of the late David Marr at MIT, and more recently by some important research at Fairchild, University of Massachusetts at Amherst, Universities of Rochester and Maryland, Carnegie Mellon University, Stanford University, and the Electrotechnical Laboratory in Japan. Here in Canada there is also important work being done by Steve Zucker at McGill, Alan Mackworth and Robert Woodham at the University of British Columbia, and John Tsotsos at the University of Toronto. Such research has led to a new understanding of the relation between the specialized low-level computations carried out by the visual system and the knowledge-based processes that go on in the rest of visual perception. It has turned out, for example, that a great deal more knowledge-independent processing goes on in vision (carried out by local parallel computations) than had heretofore been believed. As a result of this work people have developed ways of encoding visual scenes into a form of representation (called, for example, the 'primal sketch' or the '2.5-D sketch') which makes it possible to carry out relatively high-level shape, texture, and motion recognition (e.g. recognizing something as having a certain 3-D shape and as moving and rotating in certain recognizable ways) without having to carry out knowledge-based inferences. This is not what researchers believed 10 years ago. It is quite close in spirit to the finding in language processing that it is possible to do a great deal of analysis of sentences before bringing in knowledge of the domain of discourse (such analysis is what is referred to as "parsing").

This aspect of machine vision research represents yet another sort of general finding, namely that it sometimes turns out that difficult AI problems find partial solutions in unexpected places - such as in methods that lie in large part in electrical engineering, optics, acoustics, acoustical phonetics, or elsewhere outside the mainstream work in AI (which typically involves knowledge-based processing). Of course another very impotant part of the vision problem lies squarely within the AI knowledge-processing paradigm, as the elegant work of Alan Mackworth has shown. The distinction between knowledge-independent and knowledge-based processes is an important general dichotomy, and one that might well fruitfully inform the current research in speech and cursive handwriting recognition (e.g. a large part of the speech and handwriting recognition problem might be approachable within the signal-processing techniques of electrical engineering, rather than within the knowledge-based methods of AI). A serious study of the relative promise of different approaches ought to take account of such case histories and ought to keep an open mind about the widest possible range of options - especially where the commercial viability of the R & D enterprise is of concern, since in such cases it is even more likely that partial or suboptimal solutions may be deriveable by the judicious application of conventional engineering techniques.

Graphical communication covers a wide spectrum of ideas - from inputing data using a graphic tablet or 'mouse' to providing user-friendly interactions over pictorial displays. Computer graphics has been one of the most dramatic areas of progress in applied computer science, as witnessed by the spectacular computer-generated special effects in recent science-fiction films. In many respects this work is quite closely related to the machine vision problem and to the problem of interpreting images for such purposes as retrieving them by their meaningful content, transmitting them without loss of meaningful content, and answering questions about what is contained in the pictures.

Despite the impressive advances in vision research, we have little to report in this document on image communication and graphical input, partly because there has been very little work involving the application of AI techniques to image communication, or involving the use of NLP techniques in association with image transmission, say in two-way Telidon networks. A few small-scale research projects concerned with the interface between image and language communications are, however, underway in several institutions. For example, in the analysis of computer dialogue systems, special problems arise in the case of dialogues in which some of the information is presented in the form of pictures. Sidner and Bates, (1983) at BBN have begun to study these problems by collecting protocols on subjects who examine and modify a database using both natural language and a graphics display. Joshi et al. (1982) at the University of Pennsylvania are examing the problem of dialogues with databases of images used in various medical applications (e.g. radiographs and other laboratory test images). The Canadian Department of National Defence (DCIEM, Toronto, under the dirction of Dr. Martin Taylor) is also interested in this problem and is exploring some of the ramifications in a project called the Integrated Spatial Information System. This system stores a collection of information about a terrain in the form of both maps and discrete descriptive data. The goal is to eventually enable the system to act as an assistant for planning field operations.

The area of image storage, retrieval, and transmission involves both technical and human problems (the latter including ergonomic and social considerations). Such human interface problems are among the most important kinds of problems that will be faced in the design of 'fifth generation' systems. It will certainly be among the important considerations which determine the commercial viability of certain natural language systems and computational aids for professionals like lawyers, doctors, business executives, and especially translators. For example, even if a technology is developed enabling speech or typed natural language inputs it will almost certainly place some constraints on the user (i.e. the user will obviously not be as free in interacting with such systems as he would be in conversing with another person). The nature of these constraints will be instrumental in determining the ultimate success of the system. Thus, human interface research will have to play an important support role in all of these developments. For this reason it is important to pay close attention to the psychology of the user. Consequently, there is a need for R & D efforts to focus in this area.

An example of how it is possible to investigate the utility and the difficulties of using certain technical facilities, subject to certain constraints, is in the recent work by Gould, Conti & Hovanyecz (1983) at IBM. These researchers studied the problems people will run into in using a transcribing typewriter, which has certain limited capabilities, by having another person simulate this facility using various computer aids and displaying the simulated transcription on a screen. The researchers kept detailed records of the frequency and nature of certain problems encountered and compared performance and preferences for systems with different forms of limitation (i.e. for different suboptimizations). In this way they were able to better assess the future prospects of such a product even before the product becomes available. For example, they found that people inexperienced at dictation (e.g. administrators at IBM) rated an isolated-word speech recognition system capable of recognizing a vocabulary of 5,000 words as about as convenient to use for composing letters as their favorite methods of composing, although more experienced users of dictation equipment were dissatisfied with the speed and with the inability to recognize continuous speech. Users were happy to have the ability to edit their material on a screen with vocal commands. They were most disturbed when the occurrence of uninterpreted words was high (as in the smaller vocabulary simulations), and the experienced dictators even objected to the fact that 9% of the words were not identified in the 5000-word system. Since a 5000-word recognizer is only about a factor of 5 better than what is currently possible, this suggests that there may soon be systems that will have some special-purpose commercial application in the office environment. The sort of experimental technique explored by Gould and associates could very well be used in the case of proposed computational aids for human translators, and we shall raise this prospect again in discussing proposals for Canadian efforts.

CHAPTER 4

NATURAL LANGUAGE PROCESSING (NLP)

In this chapter, we summarize some of the major recent developments and research trends in the computer processing of natural languages. For the specific purposes of this report, we have omitted from this chapter any reference to machine translation systems, since this topic is given special treatment in Chapter 5.

In order to impose some order on the vast and quite diverse field of NLP, we consider first in section 4.1 work oriented towards applied NLP systems or system components whose primary function is language analysis (i.e., the passage from sentences, texts, or other linguistic forms into representations of the structure, content, purpose, etc. of those forms). In section 4.2 we treat separately the work oriented towards systems for text synthesis (i.e., creating linguistic forms from their representations). Discussion of some of the more theoretical work which bears on possible future break-throughs in applied NLP systems is treated in Chapter 6, since that work is relevant to both MT and NLP.

4.1 Applied Systems for Language Analysis

Work on natural language is proceeding on an extremely wide variety of fronts. In this section, we review a half-dozen identifiable application areas which characterize some of the important approaches. The application areas are discussed in order of increasing difficulty. We thus begin with areas in which systems can now be engineered for specific applications, and end with application areas where major innovations and even theoretical breakthroughs are required before specific implementations can be considered.

4.1.1 Text Critiquing

One of the NLP areas in which a good deal of knowledge is now available concerns the grammatical analysis of individual sentences. A wide variety of syntactic parsers and morphological analyzers have been implemented, particularly for English, over the past decade. Although no syntactic parser yet provides coverage of all grammatical constructions in the language, several are sufficiently complete that certain restricted applications are now in sight. In particular, it is now possible to design systems which verify not only the spelling, but also the grammar and superficial stylistic characteristics of standard prose (e.g., business letters, students' compositions, etc.).

The prime example of text critiquing is the EPISTLE system (Heidorn et al., 1982), currently under development at IBM's Watson Research Center. EPISTLE has been developed and tested specifically for the critiquing of business letters such as an office manager might expect his secretary to write. Its parser assigns a single syntactic structure to any input sentence, even if the sentence is ambiguous, ungrammatical, or represents only a fragment of a full sentence. When EPISTLE's 'core grammar' does not suffice to assign such a structure, the parser forces a 'fitted parse', selecting the largest string of sentence like material and adding the leftover words and phrases to it in some reasonable manner. Using the fitting procedure, EPISTLE always gets an analysis for each word string.

Even in the absence of a perfect parse (i.e., when the fitting procedure is applied), EPISTLE usually has enough syntactic and (shallow) semantic information in each parse tree to allow certain kinds of grammatical and stylistic verification. For example, several kinds of common grammatical errors can be recognized (and highlighted in colour on the monitor screen) and solutions offered to the author of the text. This is possible because most such errors involve a sufficiently local and unambiguous context that even a partial or approximate parse will reveal them. Or else, if the relaxation of a constraint is required to obtain a parse (e.g., improper number agreement between subject and verb), the system flags the violated constraint. EPISTLE includes routines to recognize a certain number of stylistic unfelicities (e.g., awkward phrases, jargon, heavy constructions, or overuse of particular words or constructions). When an unfelicitous word structure or word usage is detected, an improved form is usually proposed by the system. Spelling correction is also provided, based on an online dictionary of about 130,000 forms, each of which carries a certain amount of grammatical information for the parser as well.

Current limitations of EPISTLE seem to stem from the relative shallowness of its syntactic parse, which severely limits the kinds of stylistic critiques possible, as well as confusing the analysis of deeper grammatical problems. The parser seems incapable of recognizing constructions with complex scope (e.g.,'...each of whose...'). Nor does it seem able to provide global sentence restructuring as a critiquing service.

Although EPISTLE's grammatical capabilities are somewhat limited, the system would appear to be fairly close to commercial application in an office environment. With proper adaptation, it could certainly be used in the teaching of grammar and the critiquing of composition in schools. However, the size of the program, which requires a mainframe computer, would preclude economical implementation at present. IBM has not announced any plans for commercialization, but the number of people actively working on the project (about five, counting one recently added) might indicate such an intention. It would not be surprising to see at least some field tests in the next year or two.

4.1.2 Simple 'Level 1' Database Query Systems

Probably the most active application area for language processing concerns interfaces to database query systems. Over the last decade, steady progress has been made in the analysis of queries within a simple man-machine dialogue. Interfaces have been built which demonstrate competence in understanding user input of the following kinds (Hendrix et al., 1982):

- direct questions ("What is Smith's salary?")
- queries which require co-ordinating multiple database files
- simple uses of pronouns which refer to words in earlier queries
- input with mis-spelled words or minor grammatical errors
- elliptical inputs ("How many cruisers are in the Pacific Fleet?" (answer) "Submarines?")
- commands for display of graphic material (e.g., tabular) data ("By age and sex, list the salary and title of employees in New York")
- user definitions introduced at run time ("By 'underachiever' I mean a salesman under 35 who has not met his sales quota for the last three months. How many midwest underachievers do we have?")
- queries which reveal improper assumptions on the part of the user ("How many Japanese carriers have inoperative radar?"); instead of giving a literal response ("None"), some systems analyze the cause of the null answer and may reply: "There are no Japanese carriers.")

To date, most database query systems incorporate only a very limited theory of the domain of application. They do not in general translate queries into logical form or incorporate the degree of knowledge representation required for true dialogue with the user (which anticipates the goals and degree of knowledge of the user, for example). They allow only limited updates of the database, if at all. Hendrix et al. 1982 call such systems 'level 1' systems.

Commercial level 1 systems for English are beginning to appear on the market, and these should evolve significantly in sophistication over the next few years. The first example of a commercial system is INTELLECT, marketed since 1981 by the Artificial Intelligence Corporation of Waltham, Massachusetts (see Harris, 1978 and AI Corp., 1981). INTELLECT appears to have all the abovementioned features except the ability to analyze null answers. It has been designed to allow interfacing with the most common database management systems. Its particular compatibility with IBM database systems has led to IBM's decision, announced in June 1983, to market INTELLECT as one of its offerings for office-oriented software. As of June 1983, AI Corp. had sold 140 copies of the system at an average price of US $70,000 with the exact price depending on the size and complexity of the database. A database-specific lexicon must be constructed by hand for each user; the company has been known to refuse customers whose databases may be poorly served by the system's capabilities. Although AI Corp. has no plans to develop a French language version of INTELLECT, it appears open to proposals for a joint venture in this area.

Despite the fact that INTELLECT has the ability to analyze certain types of ungrammatical input and disambiguate ambiguous requests, this capacity is highly dependent on the direct mapping between the lexicon and the data fields. The system is weaker than many operational research systems in its coverage of English, particularly in its ability to handle quantified expressions ("Does every manager earn more than each of his secretaries?").

Since there is no model of the user or analysis of null answers, its responses may sometimes be too literal or unrevealing. It does not incorporate an interactive facility for transporting the system to a new domain, something that will be available for the IBM PC computer within a few months (at a price rumored to be in the $300 range). In this market, it is likely to have competition from Symantec which will likely draw on its experience with a number of interesting features from SRI's research systems.

Several other substantial level 1 systems have been developed, which are in the advanced prototype stage. We mention here only three of the better-known ones: the TQA system at IBM-Yorktown Heights, the USL system at IBM Heidelberg, and the ASK system at the California Institute of Technology.

The TQA system, under development at Yorktown Heights since the early 1970's, has undergone a constant evolution, but is still based on a transformational parser developed by Petrick and Plath. During 1978-79 the system was given an extensive test by the White Plains municipal office for querying their database on zoning and land use. Statistics collected during that trial (Damereau, 1981) showed that some 65% of the 800 queries to the system were correctly parsed and answered. Users sometimes had to reformulate a query to stay inside the artificial limits of the system's syntax and vocabulary (a typical problem for present query systems). Although IBM plans field trials at its own New Jersey sales office this year, it has not announced plans to market TQA. In fact, TQA does not seem to incorporate significant improvements over INTELLECT at the moment. And since AI Corporation's R & D staff is several times larger than the manpower commitment to TQA at IBM, it is unlikely that TQA will ever by marketed in a form similar to the present one.

The USL system at IBM-Heidelberg represents about the same degree of advancement as the TQA system, although it uses a different parser and semantic approach. Its market advantage lies in the fact that there exists a version for German as well as one for English. Both have been developed and tested on a database of geopolitical information. A number of interesting additions are planned for the system, but many of these concern long-range theoretical work, which is unlikely to pay off in the next five years. No plans have been announced for pre-market testing or any other moves towards commercialization.

The ASK system is being developed at the California Institute of Technology (Thompson & Thompson, 1983) for commercialization by Hewlett-Packard Corporation. ASK uses semantic networks to give a simple knowledge representation of the database domain. In addition to rapid parsing and analysis, its features include a facility for tailoring an existing database to a particular user's 'context' through an interactive dialogue. This includes the ability to add new definitions and extend the database structure through dialogues.

Level 1 systems can be improved in a number of ways without involving the detailed knowledge representation requisite for higher level systems (see 4.1.4 below). For example, one of the most promising techniques for allowing natural language interfaces to be transported to new database domains (with their associated differences in input vocabulary) is to have the system acquire this linguistic information during a dialogue with a database administrator who has no knowledge of computational linguistics. The TEAM system at SRI (Grosz, 1983) has an acquisition component which queries the database administrator about the data types to automatically set up a grammar and dictionary usable by the interface component. After that acquisition dialogue, which takes hours instead of the weeks to hand-code the same information in INTELLECT, the acclimated system is ready to answer the queries of end users.

A major problem arises in natural language 'updates' to databases. Even though natural language is not necessarily the most convenient medium for bulk data entry, some facility for limited changes is usually called for. At the very least, one wants to be able to add or modify individual facts. But unless very carefully controlled, natural language updates are potentially dangerous. The potential ambiguity of update commands may not be obvious to the user, and allow damage to data which is hard to undo.

Other improvements to level 1 systems include giving 'concise responses', which instead of answering a question like "Who drives a company car?" with a list of people (an extensional reply), would give a more meaningful (intensional) response such as: "The president and the vice-presidents". It would also be useful to allow systems to answer meta-questions such as: "What information is in the database?" or "Can you handle relative clauses?". Such facilities are still in the research stage.

4.1.3 Semantic Analysis and Information Retrieval from Extended Text

Several areas of research and development involve the semantic analysis of extended (multi-sentential) text, as opposed to the short queries or commands of database interfaces. Many systems which analyze extended text are not interactive, since the author of the text may not be on-line and the demands of high volume processing normally discourage any interactivity. Therefore extended text systems must usually be richer in linguistic detail, since there is no 'second chance' to rephrase the input. All the processing applications considered in this section are generally considered to involve the extraction of some representation of part or all of the content of the text, hence the term 'concept extraction' which is sometimes applied to this problem area.

One of the most significant advances in text analysis over the past decade has been the refinement of techniques for mapping texts from specialized subject areas into 'information formats', which are tabular representations of the data contained in the texts. These 'informatting' techniques have grown out of work done at New York

University (NYU) (e.g., Sager, 1978) which has concentrated on scientific and technical writing in medicine and related fields. This work has several applications for information science. One of the most important ones is in creating a database from full text. For example, Hirschman & Sager (1982) report on the conversion of hospital discharge summaries, written by an attending physician in telegraphic style, into a relational database. This access to information contained in the text opens up a new source of medical data for statistical analysis. Grishman (1978) also reports on the use of such techniques for query systems, where the query can be processed into semantic form using the same techniques. Central to this approach is a detailed linguistic study of the particular technical 'sub-language', which produces a kind of grammar in which sentence patterns are stated in terms of word classes (where words are put in the same class if and only if they function the same way grammatically in the sublanguage's texts).

Although a number of experiments have been carried out on converting sublanguage texts to information formats, this technique appears to be at least a few years from substantial commercial application, at least for complex medical texts. The reason for this is that while a large percentage of sentences in a typical report can be mapped into formatted data form, not all sentences can be formatted. In part, this is due to the size of the descriptive problem (languages are very large and complex objects!), so that linguistic knowledge of many less frequent phenomena is still incomplete. But in part, it is due to the fact that even technical reports will typically contain information which is extraneous to the subject matter of the sublanguage (e.g., remarks on the personal history of the patient and his family in a hospital record). This may involve the use of language which is hard to anticipate in a lexicon and grammar for the specialized domain. For such digressions one needs a much larger grammar and lexicon, essentially that of the language as a whole.

Despite the lack of perfect processing techniques for complex sublanguages, formatting techniques can easily be applied in simple sublanguages or where partial analyses are satisfactory. This is particularly the case where large amounts of text must be monitored and some analysis which can flag interesting sections is better than none.

Another major application area for the analysis of extended texts lies in the classification and dissemination of written documents. Much of the frontier work in this area has involved analysis of military messages and the monitoring of natural language reports for intelligence purposes. A number of ambitious goals have been outlined for a class of such systems (see Montgomery, 1981). Much of the work is classified or proprietary. Still, it appears that no large systems which use state-of-the-art semantic techniques to classify or disseminate documents are operational or close to being operational.

One of the more ambitious goals in the area of text analysis is that of automatic abstracting. We have not encountered any work which employs state-of-the-art techniques to actually produce abstracts from texts. However, there are two identifiable currents of work moving in this direction within two entirely different traditions. First, there is a relatively recent project in progress at the U.S. Naval Research Laboratories (NRL) on the automatic dissemination and summarization of telegraphic messages concerning malfunctioning electronic equipment onboard ships at sea. During the past year, a group headed by R. Grishman has constructed a system which uses the NYU string parser and sublanguage techniques to convert paragraph-length messages into information formats. Format entries are analyzed for revealing combinations of semantic classes, leading to the choice of one entry (the equivalent of a single proposition) which best summarizes the whole paragraph. After about 40 man-months of R&D, the NRL team has built a prototype system which successfully produces single-sentence summaries for many of the simpler paragraphs. But since parsing and formatting succeeds for fewer than half of the input sentences overall, typical paragraphs cannot yet be reliably summarized. It seems likely, however, that a useful system based on this approach could be operational within five years if sufficient attention is given to the linguistic problems of telegraphic sublanguages.

Another approach to abstracting, which is perhaps farther from commercial application but nonetheless relevant, is the work on summarizing news reports, carried out by R.

Schank and a number of his former students from Yale (e.g., DeJong, 1977). They have used 'sketchy scripts' to represent the structure of stereotypical events and their subevents. The hierarchical structure of scripts allows a summarization (on the topmost level) of a story which has been 'understood' (i.e., matched) according to the script representation. It does not appear that this approach has been applied with success to any truly difficult domains. It is therefore impossible to estimate the effort required to extend these techniques towards commercial applications. But since a script-based approach allows summarization even if a text is only partly analyzed, some limited applications may well appear within a relatively short time (2-5 years).

One recent project which appears to combine some ideas from script representation with some approaches to sub-language analysis is the NOMAD system at the University of California at Irvine. NOMAD is designed to analyze telegraphic ship-to-shore messages in 'command and control' situations. The system uses script-based expectations to interpret messages and paraphrase them into full standard English. Specific 'syntactic' patterns of the sublanguage are also used. This system appears to be in the early experimental stage.

4.1.4 Level 2 & 3 Interfaces to Databases and Expert Systems

Most existing natural language 'front ends' to database query systems and expert systems are limited to the 'level 1' capabilities described in section 4.1.2 above. In particular, current operational systems do not employ either an explicit, detailed representation of the knowledge associated with the application domain, or a model of the user's goals, state of knowledge, and limitations. Hendrix et al., (1982) have called systems with extensive explicit domain knowledge 'level 2' systems and systems with a detailed model of the user (in addition) 'level 3' systems. A good deal of oriented research is taking place on modelling such systems or on the underlying problems of representing the linguistic and extralinguistic knowledge which they require.

A number of experimental systems which incorporate level 2 capabilities are now under construction. Representative of these is the KNOBS system (Pazzani & Engelman, 1983) under development at MITRE Corporation. KNOBS makes use of several knowledge sources during the processing of a query, including scripts with stereotypical knowledge of the particular domain and inferencing rules for explicating information which is missing from the user's input. Within the context of the problem domain (an expert system providing consultant services to an Air Force tactical air mission planner), KNOBS illustrates the feasibility of integrating several different kinds of knowledge-based processing in a natural language interface.

Some preliminary attempts are being made to integrate a (partial) model of the user into natural language interface to query systems. A recent project at the University of California at Berkeley is aimed at building a consultant ('UC') for the UNIX operating system. In particular, UC provides an analysis of the user's goals during interaction with the system, employing rules ('frames') of considerable generality. For an overview of UC, see Wilensky (1982).

A good deal of oriented research is taking place at several major American centres on knowledge representation and discourse pragmatics, with the specific intention of extending the performance of natural language interfaces. The University of Pennsylvania has recently been awarded some 4 million dollars by the National Science Foundation for a study of Flexible Communication with Knowledge Bases, with a strong emphasis on discourse pragmatics (see Joshi et al., 1982). One of the features of this research will be to acquire an integrated view of both linguistic and visual communication with databases. This requires a representation of certain types of knowledge which will interface with both linguistic structures and with two and three-dimensional images. Penn's recent research in language analysis has been particularly strong in recognizing various kinds of user misconceptions on the basis of rules for goal-oriented linguistic behavior.

4.2 Applied Systems for Language Synthesis

Compared to the research and development on language analysis, work on the synthesis of text has become popular only quite recently. To some extent, this may be due to the fact that linguistic knowledge of what makes text sound coherent is much more critical for text synthesis than for text analysis, at least in the context of current processing problems. These principles of text coherence and global textual organization have only recently come under intensive scrutiny. In general, knowledge of linguistic structure beyond the sentence level has lagged far behind our understanding of how to describe single sentences.

4.2.1 Report Generation (or Synthesis) for Stereotyped Domains

One of the simplest application areas for language synthesis is the construction of factual reports in natural language using as input only non-linguistic data. At the moment, this can be achieved only for very stereotyped kinds of reports, using shortcut techniques which do not yet reflect a deep understanding of the linguistic structures involved. In particular, a prototype system for generating English stock market reports from hourly stock quotations has been demonstrated by Kukich (1983), based on the idea that reports in certain factual sublanguages map easily onto relational data structures (Sager, 1978, Kittredge 1983). In the case of stock market reports it is relatively easy to extract from raw data the changes in value of stocks and group indices which investors consider important. Still, this 'interesting' data must be organized linearly into a text-like structure using general rhetorical principles, and each data segment must be encoded as a grammatical sentence which meets certain local constraints on avoiding word repetition, etc. But for stock market reports, it has proved feasible to forego a semantic representation of the text in favor of a simple mapping from data units to sentence fragments. This engineering approach offers some promise for text synthesis in simple sublanguage such as weather reports, commodity market reports, and reports for ad hoc data-bases. Such an approach will certainly prove inadequate for text synthesis in large or complex domains, particularly where there is no reasonable bound on the number of identifiable types of sentence content.

4.2.2 Synthesis of Bilingual Reports from a Common Database

The synthesis of natural language reports has much wider application if the texts can be generated by using linguistically well-founded grammars for sentence units and text units. Research into a linguistically general approach to sublanguage text synthesis is underway at the Université de Montréal where Kittredge & Mel'cuk (1983) are modelling the synthesis of stock market reports and two varieties of agricultural market report for both English and French. Although current work is primarily on describing the sublanguage semantics and an adequate computational model, a 'toy' implementation program should be ready within a year, and a useful operational implementation program could be achieved within two to three years. It should be stressed that the automatic synthesis of bilingual reports from one and the same non-linguistic database offers an alternative to machine translation in certain simple sublanguages. This technique does away entirely BOTH with the writing of the report AND with its translation. The very notion of bilingual report generation avoids giving either language priority, since conceptually the synthesis of the reports occurs in parallel (simultaneously).

4.2.3 Computer-Assisted Language Instruction

In the domain of computer-assisted language instruction (CALI), most systems that have been developed so far have made very little use of recent NLP techniques. Systems like Control Data's PLATO are mostly based on drills, fill-ins and multiple choice questions explicitly provided by the teacher. While such systems may be of some use in language teaching, they clearly lack most of the attributes expected of a language teacher. They know very little about: a) the language they have to teach, b) how language is used, in general, to convey meaning and intentions, c) how to detect and remedy the lack of such knowledge on the part of the student. Pre-canned exercises typically cannot anticipate the specific needs of each learner.

State-of-the-art techniques in NLP would permit the development of systems that can generate on-the-spot customized material for exercises; moreover, an NLP-based system could analyze the mistakes made by the learner (cf. text critiquing systems) and make helpful comments.

The experimental ILIAD (Interactive Language Assistance for the Deaf) developed at Bolt, Beranck and Newman (BBN) (Bates and Wilson, 1981) is, to date, the best illustration of these possibilities. The system has two components. The first one is a sentence generator which accesses grammatical, semantic and pragmatic knowledge to produce an unlimited number of sentences together with their formal description. The second component is an interactive monitor which communicates with a deaf student to select exercises, calls the sentence generator with a set of specifications, prompts the student to perform some action on the basis of the generated material (e.g. apply a syntactic transformation), and evaluates and comments the answer. A version of ILIAD which runs on a micro-computer is currently being field tested. With adequate funding, the system could become fully operational within a couple of years.

ILIAD has been designed to teach deaf children their native language, but it is obvious that similar techniques are applicable to foreign language instruction. In this regard, Soper (1982a) reports preliminary experiments on an APPLE microcomputer, and Soper (1982b) presents a detailed proposal for the development of a full-scale system. A related possibility (Sanders (undated)) is to develop computer games with a specially designed NL interface.

CALI systems seem well within the reach of present technology, because language teaching provides a context where language use is naturally bounded.

4.2.4 Extended Responses from Database and Expert Systems

One of the more active areas of research in text synthesis concerns the generation of multi-sentence responses in the context of database query systems or expert systems. In the TEXT system developed at the University of Pennsylvania (McKeown, 1982), three kinds of meta-level questions about databases can be answered with extended text: questions about information available in the database, requests for definitions, and questions about the differences between database entities. Some systems have used 'canned' responses for frequently asked questions of this type. But canned responses cannot adjust the response to the context of the particular request nor are they reliable in the case of constantly evolving databases. The TEXT system uses general principles of discourse structure, discourse coherence, and relevance in producing extended responses. It can pair rhetorical techniques (e.g., analogy) with discourse purposes (such as providing definitions). Currently, TEXT is implemented on a database of the US Office of Naval Research containing information about vehicles and weapons, but it must still be considered a research prototype.

There is now a particularly urgent need for a multi-sentential text synthesis capability in expert systems. Users of expert systems tend to doubt the reliability of a system's diagnosis or result unless the system's reasoning is justified in an extended explanation. The problem with traditional expert systems is that their rules do not explicitly reflect all the steps in the reasoning process. An expert system which simply converts into English a trace of the system's application of rules can give only an explanation of WHAT the system did, and not necessarily WHY it did it. The missing part of explanation requires a model of the user as well as a representation of several kinds of knowledge of the domain (e.g., textbook knowledge of domain taxonomy and heuristics for reasoning in the domain). The work of Swartout (1983) and others on XPLAIN is directed toward modelling some of the required knowledge and reasoning capabilities.

4.2.5 Synthesis of Multi-Paragraph Text Using Planning

Work in at least three research centres is oriented towards building systems which generate more than a single paragraph and which use a sophisticated linguistic model coupled with refined techniques for planning a discourse. McDonald (1982) at the University of Massachusetts has generated texts describing simple everyday scenes using as input the unordered, non-linguistic description of a physical configuration. Mann and others at the Information Sciences Institute of the University of Southern California have developed the NIGEL system (Mann & Matthiessen, 1983) which uses systemic grammar and three kinds of knowledge sources to 'plan' extended text. Work by Appelt and others at SRI International on TELEGRAM (Appelt, 1983) allows a large amount of feedback between the linguistic constraints encoded in a large grammar of English and a sophisticated planning mechanism for discourse which reflects a variety of communicative goals.

Techniques for generating lengthy texts (beyond a single paragraph) are being developed in a more theoretical framework than those for generating shorter texts for the purposes of query systems and expert systems. It may be too early to estimate the time required before such techniques will find their way into commercial systems. It would seem likely, however, that in 10-15 years it should prove possible to create texts (in one or more languages) which serve rather complex technical needs. For example, manuals which describe the operation and maintenance of equipment (e.g., automobiles and electronic devices) might be generated automatically from the kinds of nonlinguistic descriptions used by engineers during the process of design and manufacturing. Furthermore, such manuals could be tailormade (electronically) to the level of expertise exhibited by the potential user. Perhaps even sooner, it should be possible to synthesize texts which meet certain pedagogical needs (for computer-assisted instruction, etc.) in primary and secondary schools. Meeting these practical goals will require a considerable increase in research on representing knowledge, on goaldirected linguistic and non-linguistic behavior, and on common-sense reasoning, just to name a few critical areas.

We will now turn our considerations to machine translation before returning to areas of pure research required for breakthroughs in both natural language processing and machine translation.

CHAPTER 5
MACHINE TRANSLATION (MT)

5.1 Problem Definition: Translation and the Language Barrier

While the new technologies in the domain of communications facilitate the retrieval, processing and exchange of certain types of information, they are also partly responsible for feeding the 'information explosion'. At the same time, national economies and political systems become increasingly interdependant. As a result, more and more information has to be disseminated all over the world. Since natural languages are (and will continue to be) the most powerful vehicle for communication, most of the exchanges have to be made through them.

Some 3,000 different languages are spoken in the world, though of course their relative importance varies a great deal. According to figures quoted in Van Slype et al. (1981), English is the mother tongue of 8% of the world's population, used to some extent by 30% of the same population, and the language of more than 50% of all papers published in scientific and technical journals. Thus languages are unequal in their importance, whether it is in certain political institutions (e.g. the UN has only 6 official languages), or in the quantity of information to which they provide access. In addition to the problems created by the increase in international communications, countries like Canada have their own internal language problems, because of their multi-ethnic population.

New languages are difficult for people to master, and it is unrealistic to expect everyone to become multilingual. Moreover, there is very little reason to believe that these communication problems will eventually be bypassed by the takeover of some 'interlingua', whether natural (e.g. English) or constructed (e.g. Esperanto). In fact, there is an international trend for nations, large and small, to assert their language rights. Thus, in a growing number of political organizations (e.g. EEC or Canada), it becomes vital that the languages of all participants be placed on an equal footing. In the sphere of international trade, serving one's customers in their own language is important for securing sales.

The only solution appears to be in large scale translation and, therefore, translation has now become a large industry. Van Slype et al. (1981) estimate the volume of professional translations in the world to be somewhere between 20 and 60 billion words annually, which would represent a sales figure of 1.6 to 4.8 billion ECU's (or approximately 1.7 to 5.0 billion dollars). High as they may appear, these figures are probably overly conservative. According to a study group commissioned by the Japanese government (JEIDA, 1982), in Japan alone, 40 billion words are translated annually, at a cost of nearly $2 billion. Moreover, the world translation market is expanding at a rate of 9 to 10% per year. It is clear that any significant reduction in translation cost would result in a sharply increased demand.

In Canada, since the adoption of the Official Languages Act in 1968-1969, the workload of the Federal Translation Bureau has jumped from 78.1 to 276 million words in

1981-1982, and the demand is still increasing. Three quarters of a billion words per year would seem to be a very conservative estimate for the whole of the Canadian market. But in fact, the material actually translated represents only a very small portion of what would have to be translated for both languages to be equal in all relevant respects.

In industry, most of the larger exporting companies have been forced to develop sizable translation services to meet their customers' language requirements. For high technology companies, this is not an easy task. For example, the documentation of a single piece of sophisticated equipment (e.g. a computer system or an aircraft) can represent millions or even tens of millions of words. In fact, documentation already accounts for a large portion of the cost of such products: according to Brinkmann (1979), up to 50% in some cases.

Translation costs depend on a number of factors. Some of them are the type of text to be translated, the level of quality needed, the turnover rate and the geographic location. In Canada, the average is about 20 cents per word. So the Canadian market would represent no less than 150 million dollars annually, and this is very likely a conservative estimate.

If the technological evolution of our societies is partly responsible for dramatic increases in the bulk of material in need of translation, one can also ask to what extent information processing technology can help relieve translators, whose burden seems to be getting more and more out of hand.

Until very recently, little has been accomplished in the way of successfully applying computer technology to the solution of translation problems. However, the last 30 years have seen many attempts to use the computer in translation, occasionally with some moderate degree of success. Until recently, Canada was undoubtedly a leader in that respect. At the present time, several foreign countries are pursuing major efforts in the field of machine (aided) translation. In the following sections, we will review in some detail the history of these efforts and their current status.

5.2 Historical Background

Accounts found in the literature on the history of machine translation often differ in the way they characterize the successive stages of development. We follow Hutchins (1982) in distinguishing the following four periods: 1946-1954, 1954-1966, 1966-1975, 1975-present.

5.2.1 Early Experiments: 1946-1954

Although proposals for MT can be found much earlier, it was only in the late forties, with the advent of the first generation of digital computers, that serious experimentation began, simultaneously in Europe, U.S. and Soviet Union. One notable example is the work started in 1946 by A.D. Booth, in England, on automatic dictionary lookup. A memorandum written by Weaver (1955) gave some credibility to MT as a scientific enterprise, relating it to previous successful efforts in code breaking: the difference was that, in this case the code to crack was natural language and the message to be extracted had to be expressed in terms of a universal interlingua common to all languages. In this early period several projects started in universities, first exploring efficient dictionary lookup methods, and soon coming to the realization that other levels of linguistic analysis would have to be added to produce acceptable outputs.

5.2.2 Euphoria and Disillusionment: 1954-1966

Several factors had contributed to the early belief that practical MT systems were just around the corner. Because of this, and the critical need for translations, especially for Russian to English translation, the U.S. government was persuaded to provide massive support for MT: 20 million dollars over ten years.

But the lack of adequate theories in linguistics and computation forced MT system designers to adopt a 'brute force' empirical approach. This resulted in the development of huge, ad hoc and non-modular translation programs, in which unsystematic linguistic descriptions of two languages and algorithmic specifications were inextricably intertwined.

This is now known as the 'direct' or 'first generation' approach. Typically, these systems reached a complexity ceiling before producing an output worth revising. Some of them nevertheless did find useful applications in contexts where quality did not matter very much (e.g. intelligence purposes). That was the case with the Georgetown system used by the Atomic Energy Commission.

But that was far less than what had been promised. The realization gradually dawned (e.g. Bar-Hillel, 1960) that fully automatic high quality translation would require the machine to possess extensive knowledge about the external world. The early enthusiasm gave way to a deep pessimism. The National Academy of Sciences then set up an investigation committee, whose report (ALPAC, 1966) stated that the results obtained so far in MT projects did not warrant further funding. It recommended that available funds should be spent on a longer term perspective, solving basic theoretical problems in linguistics and computation.

5.2.3 Quiet Years: 1966 - 1975

5.2.3.1 MT in Disgrace

Between 1965 and 1975, there has been very little academic research directly focused on machine translation. After the publication of the ALPAC report, nearly all of the American laboratories turned their attention to broader NLP/AI issues, and this move spread abroad (even to Japan).

Although ALPAC had criticized only shortsighted development efforts, the very notion of MT fell into disgrace in a large section of the academic community. Other applications (e.g. question-answering systems) would rather be chosen as testbeds for new theories.

On the other hand, the market pull for MT systems continued to increase with the demand for translation. Thus, in spite of a widespread scepticism, the work on MT would slowly continue, both in the private sector and in a few universities.

5.2.3.2 Commercialization of First Generation Systems

There were still some people who believed that the type of approach criticized by ALPAC could succeed through incremental improvements. A few efforts got underway, mainly in the American private sector, to develop commercially acceptable products. The best known example is SYSTRAN, which was developed by Latsec Corp. as a modified version of the Georgetown system. It was acquired by the U.S. Air Force and by NASA, among others. But its use remained largely confined to situations where high quality translation was not required.

Another example is the LOGOS system which was built by Logos Corp. for the U.S. Air Force to translate English technical manuals into Vietnamese. The project was abandoned with the end of the Vietnam war.

5.2.3.3 A New Generation of Systems

In contrast with the earlier systems, the prototypes developed in Austin, Grenoble and Montréal were based on an approach to the problem of translation that has been called 'indirect'. Instead of stating translation operations directly in terms of strings of words, an analysis process first builds a complete structural description of the material to be translated. Only after that will synthesis of the target language material begin. Thus the descriptions of both languages are clearly separated.

Ideally, in an indirect approach, the analysis process would produce a language-independent representation from a text that can be synthesized in any target language. However, as attempts to build such an interlingua never succeeded for real-life texts, indirect systems were generally led to adopt a 'transfer' model. That is, a transfer component, specific to each language pair, is used to establish the mapping between language-dependent structural representations.

Indirect systems make it much easier to state translation rules that take into account a larger context (usually the complete sentence), and because of that they go beyond a simple word-forword scheme. Moreover they are modular. In principle, the same analysis and synthesis components can be used regardless of, respectively, the target and source language.

The analysis component produced a full structural description of successive portions of the source language text. These descriptions rarely extended beyond sentence boundaries, and were based more on syntax than on semantics. Still they made it possible to formulate translation rules that go beyond a simple word-for-word scheme.

Another important step in the direction of modularity was also taken: the linguistic data were separated from the algorithmic components. The best example of that may be the Q-SYSTEMS approach, (Colmerauer, 1970) developed at the TAUM group. This formalism enables the linguist to write linguistic rules in a perspicuous way, without reference to the particular algorithms that apply these rules. With the advent of such tools as Q-SYSTEMS a core portion of MT systems was thus becoming language independent.

Systems based on the principles of the indirect approach and the separation between linguistic data and algorithms came to be called 'second generation' systems.

The second generation systems built before 1975 were all research prototypes. They permitted several new concepts to be empirically evaluated.

5.2.3.4 <u>The Emergence of Artificial Intelligence</u>

The research done in U.S. laboratories in the 1966-1975 period did not focus directly on MT but rather was concerned with broader theoretical issues, some of which are crucial to the solution of the MT problem in the long range.

As soon as more powerful syntactic models became available, several American laboratories turned their attention to even more difficult issues: how, in general, could a machine be made capable of carrying out tasks that require understanding natural language statements? It was soon discovered (in fact this was already the essence of the objections raised by Bar-Hillel (1960) against early approaches to MT), that not only was it necessary to give the machine access to some meaning representation of the text, but it also had to be able to reason on the basis of a general ('extralinguistic') knowledge of the world.

Several research groups received adequate funding to attack these problems head-on. Some remarkable progress was made in the early 1970's though not in the context of MT. Two of the best examples are the LUNAR system, which can answer queries in natural language about the content of a database (Woods, Kaplan and Nash-Webber, 1972); and the SHRDLU system, which interacts with a user, carrying out commands and answering queries in English (Winograd, 1972).

The techniques developed in artificial intelligence research were followed closely by the MT research community. However, these techniques were not ripe for integration into systems aimed at dealing with texts of a realistic degree of complexity. The few agencies that sponsored MT research at that time apparently expected the work to deal with real-life texts.

5.2.4 New Enthusiasm and Diversification of Strategies: 1975-present

The fourth period is the current one. Hence, it will be examined in the next two sections, together with an analysis of the suboptimization techniques used in MT. An inventory of current MT systems and R&D projects are listed in Appendix I and II respectively.

5.3 Current Trends in Machine Translation

5.3.1 A Better Understanding of the Limitations and Possibilities of Current Technologies

It is now widely accepted that fully automatic high-quality translation of unrestricted text (FAHQTUT) would require systems capable of reaching a deep level of understanding of natural language. Such a capability could only be based on a 'general AI', which itself presupposes an ability to store, retrieve and use appropriately vast amounts of linguistic and extra- linguistic knowledge.

Our understanding of the linguistic forms (morphology, syntax) has improved considerably in the last twenty years, but is still wanting in several areas (e.g. conjunctions). When we turn to linguistic meanings, the situation is much worse: current semantic theories cannot characterize correctly the meaning of quantifiers, tenses, adverbial modification, and a host of other phenomena.

At the interface between linguistic and extralinguistic knowledge lie the little explored domains of pragmatics and discourse structures. To attain FAHQTUT, a system would have to know how, in general, discourse (a linguistic object) is organized and used to convey the intentions (non-linguistic objects) of the speaker.

Finally, FAHQTUT would require access to an arbitrarily large amount of knowledge about the external world: common sense knowledge, encyclopedic knowledge, contextual knowledge, etc.

Although significant progress has been made on the formalization of extralinguistic knowledge in AI research (as documented in Chapter 3), we are obviously still very far from a general solution.

Thus, it is clear that FAHQTUT, if feasible at all, can only be a very long term objective. It then becomes crucial for the MT community to show that, short of FAHQTUT, MT systems can nevertheless become useful. We have seen in section 3.1.4 that the success of some AI systems could be explained in terms of an 'art of suboptimizing', that is, an art of uncovering limited problems for which a practical solution can be engineered.

In the last few years, evidence has begun to gather that suboptimal MT technology can help solve real-life translation problems. In the next section, we will review the variety of suboptimization techniques that have been tried often with some degree of success.

5.3.2 Approaches to the Problem of Suboptimization

There are several ways in which MT might be helpful even if it falls short of FAHQTUT. One can compromise on one or several of the following attributes: 'fully automatic', 'high quality' or 'applicable to unrestricted text'. Below we present a classification of suboptimization methods grouped into the following categories: machine-aided human translation, restricted input systems and human-aided machine translation.

5.3.2.1 Machine-Aided Human Translation (MAHT)

It has frequently been claimed (e.g. Kay (1980) and others), that given the current state of NLP technology, the best way to help solve translation problems is not to develop MT systems, but rather to provide the translator with a set of tools that would help him do his work more efficiently while leaving the basic initiative in the translation process to the translator. A specialized translator's workstation could integrate in a coherent environment several tools, including:

- a powerful split-screen text processor that can align source language (SL) and target language (TL) texts;
- text-critiquing facilities: spelling checkers, and possibly more (see section 4.1.1);
- facilities for producing concordances (keyword in context) to speed up terminological research;
- on-line dictionaries, and facilities for maintaining them;
- access to multilingual thesauri;
- access to remote terminology banks;
- morphological analysis facilities working in tandem with dictionaries and concordance facilities;

- access to public databases of previously translated material.

Other elements can be added at will to this list. Since more than 90% of the texts are likely to be out of reach of fully automatic MT in the next ten years, it is obvious that the development of such work stations is highly desirable. As NLP research progresses, more sophisticated components can be added to the system. According to Melby (1983), MT systems should in fact be embedded within such a workstation; the translator should decide for himself whether or not to use MT, which would in any case do no more than provide 'suggestions'.

The technology to develop a basic workstation is certainly available. In fact, a few American companies (e.g. Weidner, Alps) are currently introducing products of this type on the market. These products run on relatively cheap micro-computers, and might soon be affordable for individual translators.

5.3.2.2 Restricted-Input Systems

While MAHT systems omit one of the attractive potential features of MT, namely that the translation process itself is not 'automatic', restricted input systems retain automation (that is, the machine has the initiative in the translation process). But this is done at the expense of specializing the system to certain well-defined types of texts. There are two sub-categories of such systems: those applying to artificially constrained subsets of language and those applying to natural sublanguages.

5.3.2.2.1 Artificially Restricted Input

From a technological point of view, artificially restricting the input is an easy approach since it amounts to tailoring the problems to the solutions that are at hand, instead of the reverse. To some extent, of course, this simply passes the hard problems to someone else. Some extra burden is clearly being transferred to the writing process.

This, however may be a reasonable alternative provided several conditions can be met:

- First, of course, the user of such a system must be in control of both the writing and the translation processes.

  This is by no means always the case: in Canada, probably more than 75% of all translations are produced by translation services which have no access to the document producers.

- Second, the subsets of both source and target languages accepted by the system must be broad enough to permit natural and acceptable expression of the content. This is obviously not easy to assess. Subjective quality judgments made by qualified writers, translators and users of the texts remain the prime criterion.

- Third, the restrictions on style enforced by the system must be simple and natural enough for the writer to be at ease with them (at least after some training period). In evaluating such a system one can use ratings by authors and also measures of productivity.

- Finally, the MT output must meet the same standards of quality, readability, etc. as any other translation.

It may very well be the case that these conditions can be met in certain situations. After all, the subsets of natural languages that state-of-the-art MT systems can handle are fairly large.

In the U.S., Xerox has been using SYSTRAN, and more recently, ALPS, on artificially restricted input, apparently with success. The SMART system provides an integrated environment to write and translate documents in controlled subsets of natural languages. Being custom-built to handle these subsets, SMART can provide on-line control of the writing process, and should be able to produce better translations than unrestricted systems.

5.3.2.3.2 Natural Sublanguages

In many cases, texts to be translated are written within a very narrow subject matter. This is particularly the case with technical manuals, medical reports, scientific articles, financial market reports, etc. Each such subset of the language (called a 'sub-language') uses only a part of the whole language's syntax and vocabulary. Even more important, the way words can be combined to make sentences is relatively limited. These limitations can be summed up in a specialized lexicon and grammar for each sublanguage. The recognition that natural sublanguages exist is not new, but the techniques for describing and exploiting the special properties of sub-languages for NLP in general and MT in particular have only recently begun to pay off.

The TAUM-METEO system (Chevalier, Dansereau and Poulin, 1978) is an excellent illustration of the practical relevance of the notion of sublanguage to MT. This system takes advantage of the very tight sublanguage restrictions found in weather forecasts to translate accurately and automatically (e.g. without any post-editing), more than 80% of the bulletins that are submitted each day. To our knowledge, this is the only reported case of fully automatic high quality translation.

The obvious disadvantage of this approach is that, since at least some parts of sublanguage-based systems are developed specifically for a given type of text, a high volume is required in order for such a system to become cost-effective. Before starting the development of a system, the level of complexity of the sublanguage must be assessed; some criteria can be found in Kittredge (1983). Simpler sublanguages with higher volumes of text offer the best prospects for cost effective MT systems.

Thanks to modern modular system design, it is not necessary to start an MT system from scratch for every sublanguage: several modules will tend to be sublanguage independent. For example, morphology rules tend to be constant across sublanguages. Even in the case of syntax, it is sometimes possible to 'tune' a general parser (cf. Sager, 1981); in other cases it will be preferable to develop (as in the case of METEO) a specialized parser for the sub-language. The most sublanguage-specific components will remain dictionaries and semantic analysis.

5.3.2.3 Human-Aided Machine Translation (HAMT)

Still another approach to sub-optimization is to drop the requirement that MT be fully automatic. While automation is basically retained, because the machine has the initiative in the translation process, the human will be requested to assist the machine with problems that are beyond its capabilities. Human intervention can take place either before, during or after machine processing.

5.3.2.3.1 Systems with Pre-editing

When the weaknesses of an MT system are known, it is possible for a human operator to manipulate the text in such a way that potential problems will be eliminated before reaching the machine. The most obvious way is to reformulate the text in a manner that does not affect its content, but expresses it in a form that the machine can handle. This approach is very similar to artificially restricted input systems (cf. 5.3.2.2.1), but can be used where there is no control over the writing process.

A different approach is to have a human annotate the text with extraneous symbols that the machine will interpret as clues to the structure of the text. For example, symbols may be inserted to help segment the text into appropriate translation units; or brackets may be added to make explicit the scope of conjunctions, the internal structure of compounds, etc.

An obvious disadvantage of this approach is that it consumes user time; if post-editing is still required, in the end, to make the output acceptable, there is certainly a risk that the system will not be cost-effective. The usual arguments in favour of pre-editing are: it requires less qualified personnel (because there is no need for bilingualism); and in a multilingual MT system, it is preferable to invest in preventing mistakes (because otherwise they will have to be corrected in each one of the target languages).

Several well-known systems (e.g. Weidner, METAL) require an operator to help segment the text into translation units, but very few make extensive use of pre-editing. One example is the CULT system (Loh, 1978), which is used to translate mathematical texts from Chinese into English.

5.3.2.3.2 Interactive Systems

Kay (1973) presented the first proposal, with his MIND system, to use interactivity during parsing to help the machine solve difficult lexical and structural ambiguities. When an unsolvable ambiguity was detected, parsing was interrupted until the operator answered an appropriate request. A few years later, a large-scale MT system based on this principle, the ITS system, was developed at Brigham Young University. Experiments with ITS have shown that too much intervention tended to take place for the sytem to be usable, and it was abandoned. The ALPS system was developed as a further attempt to turn the same idea into a useful product.

The idea of an interactive system in which man and machine can cooperate is very appealing. Unfortunately, it seems that interactive systems have not yet achieved such a fruitful symbiosis. Rather, the machine tends to force the translator/user to think in its own terms, asking countless, irrelevant questions. And finally, because efficiency is an important consideration for an interactive system, one may be tempted to cut on the linguistic processing: the burden left on the operator is then increased further.

5.3.2.3.3 Systems with Post-editing

When low-quality output is sufficient, MT works very well. But such situations are in fact exceptional. When normal or high quality is needed, all MT systems have to resort to post-editing. The only known exception is the TAUM-METEO system, whose output is accurate enough to make post-editing unnecessary. Still, even in that case, some material is rejected by the system and has to be turned over to human translators.

It has frequently been claimed (e.g. Carbonell, Cullingford and Gershman, 1981) that post-editing an unreliable output can be at least as time-consuming as conventional translation of the whole text, since it implies mentally retranslating the text, and then making awkward adjustments. Indeed this claim has been substantiated by some experiments, the best known being those made by ALPAC.

On the other hand, several experiments have reported opposite conclusions: for example, Van Slype (1979) for the use of SYSTRAN at EEC and Hundt (1982) for the use of Weidner at Mitel. We conclude that MT output can profitably be post-edited, with the following qualifications:

- The machine output must reach a certain threshold of quality (but there are no known methods of assessing that threshold in advance)

- It is essential that the post-editors have a positive attitude toward such work as repeatedly correcting the same mistakes, many of which are atypical of human mistakes (e.g. wrong articles). Such a positive attitude will be more likely to develop if the post-editors can be involved in improving the system (e.g. updating the dictionary).

- Finally, and most importantly, it seems clear that in the successful cases, system users are ready to accept a product which is different from conventional translation: essentially, a more 'literal' translation.

5.3.3 New Enthusiasm for MT

Since 1975, there has been a remarkable increase in the level of R&D activity in MT, in the U.S., Europe and Japan. Several factors may explain this rise, some of which are the constant increase in the demand for translation, the decline in the price of computer hardware and the optimism generated by some successes in AI.

But recent successes (or claims of success) with MT systems have also played an important role. The success story of the TAUM-METEO system has received worldwide publicity since 1977. The European Economic Community claims that after having invested four million dollars on the development of dictionaries, its use of SYSTRAN has become profitable. General Motors of Canada reports similar results. Xerox (Rochester) has apparently had some success with SYSTRAN, and more recently, with ALPS. Several multinational corporations are reported to be happy about their use of the SMART system. The Weidner system has received some publicity from satisfied customers, one of which is Mitel Corporation. Omniplex, a translation house in Palo Alto, has recently announced plans to translate the Encyclopedia Britannica into Arabic, using a modified version of the Weidner system. In Japan, the Electrotechnical Laboratory (a public R&D facility) reports success in its use of an MT system designed at Kyoto University.

Appendix I enumerates the best-known current MT systems, together with some information on their use and technical features.

In the U.S., the increase in MT R&D is most visible in the private sector. At least three new R&D companies have been formed since 1978, and have introduced new systems on the market: Smart Communications, Weidner Communications, and Automated Language Processing Systems. Also Logos Corporation has moved from the development of custom MT systems for government agencies to selling MT systems to the marketplace. Latsec Corporation, which introduced the well-known SYSTRAN, is still active. All of these companies distribute their products in a number of countries.

In Europe, the EEC has recently launched a large scale MT project which will generate activity in universities in each of the seven participating countries. The French have just started their own national project in MT: the ESOPE project. Siemens, the German multinational, has undertaken the development of a multilingual MT system, both for internal use and for commercialization.

But nowhere is this new enthusiasm for MT more visible than in Japan. Before 1975, MT activity was nearly non-existent there. Now, there are more than 15 different R&D projects in the country. The Japanese government is funding three relevant public projects: the Fifth Generation Project (see section 2.1) has as one of its long-term goals the development of an 'intelligent' multi-lingual MT system, the same organization which is responsible for the Fifth Generation Project, MITI, is also funding a shorter-term effort to develop an MT system for technical translation between English and Japanese and the Agency for Science and Technology is funding a large scale project (short term), also for technical translation. Moreover, at least two public laboratories, the Electrotechnical Laboratory of MITI and the Electrical Communications Laboratory (NTT) have MT projects of their own.

In Japanese universities, at least five different groups are developing MT prototypes. Finally, in the private sector, Fujitsu, Hitachi, NEC, Toshiba and others have recently started R&D projects in MT, some with commercia-lization goals for the shorter mid-term.

In Appendix II, we enumerate the main centres of R&D activity in MT (in the U.S., Europe and Japan), together with some information on their work and funding level.

5.3.4 Trends and Opportunities

5.3.4.1 Commercial Systems

Generally speaking, there is an important gap between products that state-of-the-art MT/NLP/AI technology would make possible and the systems that are currently on the market. Most, if not all, commercial MT systems are based on a very old, first generation technology. Table 5.3.4.1 illustrates a number of current operational MT systems. More details can be found in Appendix I.

In such systems, the translation process is based on very crude syntactic components, and this is reflected in the fact that the output text is very often grossly ungrammatical. It is hardly a surprise, then, that these systems will never go beyond a very superficial and unsystematic semantic analysis. In fact some of the commercial systems have no semantic component whatsoever, besides the very limited technique of 'micro-glossaries' (selecting the translation of a word on the basis of domain codes manually assigned to the text). When semantic analysis is so weak, the systems are very prone to unfaithful translations (e.g. 'contresens'), resulting from an arbitrary selection among alternatives in cases of ambiguity.

A further consequence of the weakness of syntactic and semantic processing in commercial systems, is that translation rules have to be stated in terms of correspondences between words or strings of

Inventory of Operational MT Systems

System	Language Pairs	Hardware Environment
SYSTRAN	English to French, Spanish, Portugese, Italian, German, Japanese, Arabic; Russian, French, Japanese to English	IBM Mainframes
LOGOS	English to German, German to English, former pairs; Russian to English, English to Vietnamese, English to Farsi	WANG OIS
ALPS	English to French, Italian, Spanish, German, Arabic; French to English	DG MV-4000
WEIDNER	English to French, German, Spanish, Arabic, Portugese, Japanese; French, German, Japanese, Spanish to English	DEC PDP-11 series IBM PC
SMART	English to French, Italian, Spanish, German, Japanese	IBM 370 DEC VAX 780
METEO	English to French	CDC CYBER series
Nagao's System	English to Japanese	Fujitsu FACOM M200
ATLAS/I	English to Japanese, Japanese to English	Fujitsu FACOM M-180 II
SPANAM	Spanish to English, English to Spanish under development	IBM Mainframe
TITUS II	between English, French, German and Spanish in all directions	?
CULT	Chinese to English	ICL 1904A

Table 5.3.4.1

words, rather than between higher level patterns (as a human translator normally does). This means that these systems will at best produce what can be considered a very literal translation.

The most recently commercialized systems do not appear to fare any better on these counts, when they are not simply worse; broadly speaking, the improvements that they bring are in the global working environment (hardware, word processing facilities, etc.) rather than in basic MT technology.

The fact that, in spite of all their shortcomings, several commercial MT systems find numerous users means that the market is very eager for MT products. Better systems would no doubt meet a warm reception. Meanwhile, those who want to turn to commercial systems should be ready to invest in customizing dictionaries, have large enough translation volumes to warrant such an undertaking, and be ready to accept a relatively literal translation as end product (e.g. lacking some idiomaticity and good style).

5.3.4.2 Toward Large Scale Applications of Second Generation Technology

The second generation MT technology, as developed in university groups (e.g. GETA, LRC, TAUM) before 1975, has not yet really been transferred to industry. A number of commercial systems have adopted more or less systematically some features of this technology. But besides TAUM-METEO, very few (if any) MT systems truly based on second generation technology have so far reached the production stage. Table 5.3.4.2 identifies a number of second generation and two third generation prototype systems. Additional information on these systems can be gleaned from Appendix I.

This fact is not purely accidental. Second-generation systems are much more expensive to develop than their predecessors. One of the essential features of these systems is that they attempt to base the translation process on complete structural descriptions of the source language text. In principle, these descriptions should be semantic representations. Since too little is known about semantics, second generation systems are in actual practice based on syntactic representations enriched with some semantic information. In itself, building correct syntactic representations turns out to be quite difficult. Large grammars may become computationally intractable. In addition, these grammars cannot operate without very detailed lexical information and this makes dictionary development extremely costly.

Moreover, even if the goal is somewhat less than FAHQTUT, second generation systems have to go beyond syntactic processing. In real-life texts, syntax leaves countless ambiguities both at the

Inventory of Prototype MT Systems

Systems	Language Pairs	Generation
EUROTRA	EEC language in all directions	2nd
SUZY	German to English, French, Russian; French, English to German	sophisticated 1st
Projet Nationale (ESOPE)	English to French, French to English, possibly Spanish to French	2nd
TAUM - AVIATION	English to French	2nd
Siemens' Project (METAL)	German to English	2nd
ARIANE -78	Russian to French	2nd
Nagao's Manuals	Japanese to English	2nd
Japanese Agency for Science & Technology	English to Japanese, Japanese to English	2nd
Wilks' System	English to French	3rd
SALAT	EEC languages	3rd

Table 5.3.4.2

lexical level (polysemy) and at the structural level (e.g. prepositional phrase attachment, scope of conjunctions, bracketing of nominal compounds). The usual approach is to use semantic features that permit a state of 'selection restrictions' between classes of words (e.g. verbs and their objects). That type of technique is certainly useful. But it is insufficient for the solution of several problems (e.g. nominal compounds). And here we reach the limits of what has been called the second generation approach. Beyond that, we enter a sphere where more research is needed before systems of a 'third generation' can become a practical reality.

Several major development projects based on second generation technology have been launched recently: EUROTRA, Projet Nationale (ESOPE), Siemens' project (METAL), the project at the Japanese Agency for Science and Technology, etc. These projects may succeed in creating operational systems that will provide better translation than current commercial systems: if not always accurate and elegant, their output is likely to have at least a higher level of grammaticality. The success will depend crucially on a correct choice of suboptimization techniques. While there may be other possibili- ties, TAUM-METEO has shown that second generation systems could produce spectacular results with simple sublanguages. But if because of market necessities second generation systems are extended beyond precisely defined sublanguages, it is not yet clear exactly how well they will fare.

5.3.4.3 Longer-Term Research

When we look for longer term MT research, it turns out that very little is being done currently, with the possible exception of a few Japanese projects like the Fifth Generation. There is clearly a pressing need for longer-term efforts in which MT research should be brought in line with leading edge research in other areas of NLP/AI.

There is no reason why the most recent techniques in semantics, pragmatics, knowledge representation, etc. could not be applied to MT problems. Of course, any system that attempts to reach a deeper level of understanding of natural language will have, in the foreseeable future, to be somewhat limited in scope. But natural sub-languages provide a context where more powerful techniques can be tested, and, hopefully, permit the realization of better MT systems in a not too distant future.

CHAPTER 6

PROMISING AREAS OF RESEARCH IN MACHINE TRANSLATION AND NATURAL LANGUAGE PROCESSING

The preceding two chapters have described both research and development efforts aimed at producing specific systems for machine translation and other types of natural language processing, or for specific processing capabilities. In this chapter we summarize several research areas which are at the same time crucial for long-term breakthroughs in NLP and MT yet not easily categorized according to processing function. Since it is impossible to foresee exactly which areas of pure research will prove to be decisive, our purpose here is only to present a sampling of current thinking on the matter and identify how some of this work is being addressed.

During the past several months, considerable excitement has been generated by the establishment at Stanford University of the Center for the Study of Language and Information (CLSI). This center has now received a $15 million, 4-year grant from the System Development Foundation to promote interdisciplinary research on a broad range of topics. Directed by Jon Barwise of the Stanford Philosophy Department, the CSLI will provide a focus for work on 'situated language' (language as used by human or mechanical agents to exchange, store and process information in particular worlds of discourse). In this view both human and computer languages are part of a single subject matter, which is badly in need of a unifying theoretical foundation. The CSLI will bring together more than 15 well-known researchers from Stanford, SRI International, Xerox PARC and other institutions in the Palo Alto area, and support a worldwide network of affiliated institutions and researchers. The scale of the CSLI effort, which involves many of the top scholars from computer science, artificial intelligence, philosophy, logic linguistics and other disciplines, will give new meaning to the term 'critical mass'. With this kind of concentration and investment, an acceleration in our understanding of many critical theoretical areas is virtually assured.

The CSLI's initial program is divided into five projects:

- integrate theories of human utterances at several levels of description, including the higher levels of text and dialogue;

- develop testable and computationally tractable theories of the mental processing of utterances;

- reconsider the theories and practice of computer science under a general linguistic interpretation, viewing computation as a linguistic activity;

- develop and extend existing theories to provide a comprehensive theory of situated language use; this will include a more unified theory of reasoning for use in modelling both human languages and the more structured computer languages;

- develop unified foundations for the above mentioned theories.

The full benefits of a pure research program such as that envisaged by the CSLI will probably not be felt for at least ten years (although specific breakthroughs might occur in half that time). In the meantime, a number of specific areas of pure research are proving quite active, and progress in those areas is fairly certain, regardless of whether or not they figure in the scope of CSLI activities:

- viewing the human speech activity as a special case of planned, goal-directed behavior (work of Cohen, Perreault and Allen);

- developing a unified view of discourse phenomena such as anaphora, ellipsis and focus (Grosz, Sidner);

- understanding how the structure and semantics of specialized sublanguages relate to those features of the language as a whole (Sager, Kittredge, Grishman);

- developing more refined techniques for knowledge representation such as the KL-ONE language and other formalisms for networks, frames and logic;

- developing richer logics, including models of common-sense reasoning (McCarthy, McDermott, Doyle, Reiter);

In addition to the research areas mentioned above, there is an important movement to develop new tools for the processing of language and for implementing specific systems. Again, the number of areas is very large, and our selection must be somewhat subjective. We might mention here the development of new programming languages, including modifications and extensions of PROLOG (e.g., concurrent PROLOG developed by the Japanese ICOT center), as well as special machines for the efficient processing of widespread AI languages (e.g., LISP machines). Artificial languages for knowledge representation might also fit in here, although they involve problems of directly modelling cognitive functions or capacities.

In the following chapter we will consider broader issues of AI and how AI is evolving.

CHAPTER 7
BACKGROUND AND STATE-OF-THE-ART: CONCLUSIONS

In this concluding chapter we present a number of observations based on our interpretation of the information gathered for this report, as well as on discussions held over the past several months with various experts. The observations, it should be noted, are not uncontroversial. They are largely interpretive, impressionistic, and subject to revision. Nonetheless, they are professional evaluations based on our experience and we feel that they may be useful in helping to extract some general lessons from the more objective collection of observations contained in the earlier part of this report.

7.1 Summary: Why Are Some AI Systems Now Becoming Successful?

It might be worthwhile at this point to briefly summarize some of the factors, discussed in connection with various developments in AI and NLP, that have entered into the recent success of such enterprises. The question we ask is, what has changed in recent years to allow such successes as we have reported, after a long period of only mediocre level performance? Part of the reason undoubtedly is the high level of funding that research in AI has received in the U.S. over the last decade through DARPA. Another is the greater availability of computing resources for use in what had previously been prohibitively costly AI research, which resulted from the dramatic drop in the cost of computing with the advent of VLSI technology. In addition, there are a number of specific factors that contributed to the timing of a number of the successful achievements. The following are some general categories of such factors. Each of them applies in some, though not all, cases. Beside each are one or two illustrative examples.

1) Progress in the basic science of artificial intelligence, which led to the discovery of a number of new principles and techniques for making machines behave intelligently.

   - e.g. principles of search, constraint propagation and planning. Such principles have turned out to be among the main reasons why some systems - e.g. computer chess and the Harpy speech recognition systems - have attained relatively high levels of performance. For the first time there is a large enough set of well defined purely AI techniques of this sort that courses and textbooks now exist on these topics (e.g. Nilsson, 1981; Winston, 1983).

2) The discovery of new basic principles in fields outside of AI which make successful mechanization of problems in these application areas possible.

   - e.g. formal linguistics, mathematics (the Risch method for symbolic integration which forms the basis of the highly successful MACSYMA system), new proof methods for mathematical logic which have formed the basis of a new family of knowledge-based programming techniques, as in the PROLOG language.

3) The discovery of a set of problems, which seem to be within the field of AI, but which succumb to various conventional analyses and methodologies. Such analyses, in other words, need not be specifically ones that are characteristic of AI in general.

- Performance on certain tasks, like high level chess play, improves dramatically with the addition of new and faster search algorithms (as in Chess 4.5) and computer architectures (as in the Chess Machine being developed at MIT), certain robot manipulation problems were solved by the development of new types of transducers and manipulators by electrical engineers, speech analysis and synthesis receives a lot of help from new signal processing techniques, certain problems in low-level vision turn out to be 'non-combinatorial' problems that do not require knowledge-based processes and can be solved by relaxation methods, retinally local parallel computations, or linear programming methods.

4) The availability of certain computational tools for aiding the development of more complex computer systems. These include very general tools such as high level languages (like LISP and PROLOG) as well as more specifically AI development systems such as KRL (Knowledge Representation Language), OPS5, KES, ART, LOOPS (all of which were mentioned earlier), and more specific tools for building prototype systems, such as widely available standard parsing algorithms like the ATN.

5) The discovery of certain appropriately narrow domains, with applied interest, where limitations of existing approaches do not interfere. For example, in the domain of language, it was realized that there are a number of situations in which effective communication takes place through relatively simple and restricted forms of language:

- Cases in which people communicate with a machine in a way that involves a limited range of possible meanings and with immediate feedback,

- Cases in which a community of specialists have evolved a distinctive jargon or sublanguage in which the range of meanings and forms is naturally restricted.

In both cases existing NLP technology turned out to be powerful enough to yield useful products such as natural language front ends for database query (e.g. INTELLECT) and various systems for processing sub-languages (e.g. METEO).

6) The evolution of a technologically more sophisticated culture which:

- needs (or believes it needs) certain products or tools that fall naturally within AI.

- has a willingness to accept a suboptimal or not very general product or tool so long as it reaches a certain minimal level of practicality or usefulness (e.g. certain limited natural language or speech recognition systems).

7.2 Factors Shaping the Field: Technology Push Versus Market Pull

In recent years a number of areas of AI development (including ones relating to language) exhibited a split between market-driven forces (or market pull) and science-driven forces (or technology push). Some of the areas for which there appeared to be a strong commercial potential have led to splinter developments outside the mainstream academic AI research and scholarly reporting traditions. Most of this work has occurred in the private sector. Perhaps a very early (and not very important) example of this was the development of an inexpensive chess playing product (Chess Challenger) whose peformance, while not up to that of the best systems developed in AI laboratories as test beds for studying problem solving techniques, was nonetheless amazingly good and was achieved with very low computing cost using methods specifically adapted for the games market.

Other early examples of such a split include certain kinds of pattern recognition applications (such as character recognition) and in robot manipulator technology. Both had been important areas of AI research in the 1960's and early 1970's and had shown some slow theoretical progress. But in the late 1970's the development of these areas became almost entirely market driven. Immediate application needs dominated and designs were based mostly on existing engineering principles. Optical character recognition was divorced from the general pattern recognition problem and robotics became separated from the much harder (and in the long term more important) problem of perceptual-motor coordination or the design of integrated mobile vision-reasoning-manipulation devices.

These commercial robots concentrated on programable movements, sometimes with modest touch-sensing capabilities, that required highly controlled environments (e.g. fixed jigs). Only in very recent R & D work has the possibility of introducing visual feedback been seriously considered once again.

More modern examples of this split can be seen in speech perception, natural language processing and machine translation. Since these were described in some detail in this report, our purpose in raising them again here is simply to draw attention to a recurring pattern of research: accumulation of technological expertise (sometimes of a very general sort, rather than technology specific to a particular application), transfer of the technology to the marketplace under perceived market demand, and then a return of the problem to the basic science laboratory, often with new insights acquired from the attempt to use and market the product.

Perhaps one of the first clearly AI technologies to become a commercial product is a natural language analysis system based on the augmented recursive transition network (ATN). As early as 1978, the Artifical Intelligence Corporation prepared to market a system (initially called Robot, but now renamed INTELLECT) which translates relatively unrestricted English questions into a formal query notation which can be customized for interrogating a user's relational database. Almost no new scientific or basic computer science or AI techniques had to be developed for this system: the techniques were standard text-book methods. The primary 'discovery' made by this development was that the domain of database queries is sufficiently restricted (i.e. the range of interpretations of sentences that are meaningful in this domain is narrow enough) that the serious unsolved problems of language analysis can be overcome in this special context, in simple ad hoc ways. The problems of lexical, syntactic and semantic ambiguities, of anaphoric reference, of discourse presuppositions, and so on, which plague more general natural language systems, can be bypassed using methods that would not work in general, but do work in this special domain (e.g. when the system cannot correctly parse a sentence it can usually guess at the intended meaning based only on knowing which words were used and what kinds of queries concerning topics related to those words are legitimate for that particular data base). As a result of this discovery and the resulting success of the INTELLECT system, a large number of companies have been formed with the primary objective of developing similar or related systems based on other, perhaps better, established language analysis approaches.

As we saw in the survey, a high proportion of the commercial natural language systems continues to take the form of interfaces to databases, or in some cases to expert systems (so-called natural language front ends). But other domains have also been discovered in which either the set of possible interpretations is constrained in advance, or else (as in the case of the 'writer's workbench' style-correcting system EPISTLE, and perhaps in language-teaching systems) even an incomplete analysis system can still be very useful. Of course, as applications of various kinds proliferate, there develops a need to return to the laboratory for new basic knowledge (see, for example, the application of natural language interactions with a pictorial database mentioned earlier, where new data are required concerning the sorts of queries that people make, and new techniques have to be developed for dealing with locative indexical expressions such as "the one above that one"). In the meantime, however, new useful products continue to be developed based on well known previous-generation techniques.

Another dramatic split that can be seen is in speech recognition. This is an area that has been heavily financed in the U.S. over the last 10 years by the defence department (DARPA) and produced some interesting large systems, as well as some new approaches. The large-scale DARPA funding has now been terminated and a number of R & D projects have recently been mounted with the goal of producing less ambitious, but more commercially viable products (as we saw in our review). The approach and the techniques being adopted in this latter work is quite different from that adopted in the more ambitious DARPA projects. What has happened is that there has appeared a very clear need for a robust speech recognizer. The perceived need is so great, in fact, that even a limited system might become a very profitable commercial success, particularly if the capability is well matched to some particular application, as it was in the case of natural language front ends to databases. The new directions represent a clear case of a technology responding to market pull. As Peter Hart (1982), one of the researchers working on such systems at Fairchild put it, "Applications of systems that understand English text - as opposed to spoken English - face an additional obstacle that was unappreciated by many people only a few years ago: viz., the keyboard. A putative user of a text-based system needs to satisfy two conditions: the population must be comfortable using a keyboard, yet not be daily system users who will quickly want to express themselves in the most succinct way." The discovery of this 'obstacle' to the marketability of language understanding systems has spurred the recent growth of speech recognition R & D efforts in the private sector.

To close this topic we should also point out that nowhere is the split between mainline academic research and commercially motivated product development more visible than in the case of machine translation. This could be a legacy of the discouraging experience of the mid 1960's when MT was still a central problem of AI (e.g. it occupies a prominent position in the historic 1961 conference on 'Mechanization of Thought Processes' held at the National Physical Laboratory in Teddington, England). After the negative report of the Automatic Language Processing Advisory Committee (ALPAC) - which, incidentally, only criticized short-range efforts directed at development of fully automated MT systems - most granting agencies and research laboratories simply gave up working on the problem (see the discussion in the section on the state-of-the-art in MT). Partly because of the psychological effects of the negative evaluations given the field by the ALPAC report and partly because of the fact that almost the only sources of funding available during this period were from private companies, the few projects that continued and the newer ones that arose in more recent years have tended to work outside the AI community - mostly in the private sector where the economic potentials of MT still held enough attraction to keep a few commercially-oriented projects alive.

The commercial products that have arisen have helped to break through the barrier of discouragement by showing that MT does have commercial possibilities. These were important ventures, but they were carried out in isolation from most other AI work and rarely published in AI or even NLP journals. There are serious disadvantages of working in relative isolation from mainline NLP and AI research. One of these is that such MT systems were typically based on older AI and NLP techniques and did not benefit from the sort of accumulation of more general tools and techniques that has occurred, say, in the case of expert systems research, which did stay in touch with the core of AI even while it developed robust commercial systems. Another is that, conversely, the market-sensitive MT research did not feed back and inform the longer-range academic research. Such feedback of operational experience has in the past frequently produced beneficial effects in shaping basic research by, for example, showing researchers what kind of problems need priority attention. This isolationist trend now appears to be slowly changing. Many of the larger MT projects are paying close attention to the newest work in linguistics, computational linguistics, and other areas of AI, and conversely, commercial research is showing signs of slowly feeding back to the laboratories their operational experience with various styles of suboptimization. Once again people are beginning to realize both that the trend towards short-term commercial solutions at the expense of longer term research is shortsighted, and that research and development that is insensitive to operational experience can lead to the blind alley of pursuing unrealistic goals (for example, it could provide information about the dimensions on which systems have to be robust to work in the real world). Vision systems that only work with highly controlled coherent lighting and perfectly matt surfaces, or speech systems that only work in anechoic chambers are probably not based on the right principles for providing long-term solutions to the basic problems under study.

7.3 The Complexity Barrier

Useful artificial intelligence systems, whether they be language understanding systems, expert systems, vision systems, or speech recognition systems, are invariably very large programs. Sometimes, when their function is better understood, they can be pared down to their bare essentials (as in the chess playing products) and made much smaller. Alternatively, as systems become more mature, they can be modularized so that certain parts can be made relatively small and efficient while other parts remain large and extendable and can even be stored in different media. For example, as we pointed out in the section on natural language processing, more modern systems separate such components as the parser, the grammar, the semantics, and the 'inference engine,' while in vision systems we try to separate the data-driven 'low level vision' processes from the model-driven processes which recognize and categorize objects in the scene, from perhaps the inferencial component that takes more global expectations of what might be there into account.

This sort of modularization is an important strategy for dealing with the extreme complexity that is often required. Nonetheless modularization is only possible when the systems we are dealing with are well understood. Furthermore, modularization by itself does not solve the problem of the large size of AI systems. Large systems, or at least large databases, raise a whole new set of problems that small prototype or 'toy' systems do not have to face.

It has been a common strategy in AI to investigate problems by first setting up small but representative 'microworlds' on which to test ideas (a good example is Winograd's "Blocks World": Winograd, 1972). The idea, as it was expressed by Winston (1983), is that these micro-worlds are for AI what e. Coli are for biology - a small tractable model of the real thing. Recently, however, people have begun to take the view that microworlds are deceptive in their simplicity - that certain kinds of intelligent performances are only attained if we concentrate on the right kinds of limited interactions with real problems. Marr (1977) was one of the people who insisted that vision systems missed the important problems when they confined themselves to microworlds of blocks and other polyhedra. Similarly, natural language tasks raise very different problems if they try to deal with language as it occurs naturally in certain contexts, even when the contexts are restricted. Problem solving systems which try to deal with the sorts of real life problems that experts have to deal with also run up against quite different problems from those faced by certain 'games'. In such realistic contexts one frequently has to deal with unanticipated inputs ('noisy data') and with the problem of collecting, organizing, encoding, and manipulating very large databases. Furthermore, the extremely costly problem of maintaining and extending such databases, as new information is brought in or as the application requires that the scope of the system's performance be extended, add to the complexity. Large scale natural language and machine translation projects and real-life expert systems are good examples of this. They both face an enormous 'complexity barrier' in terms of the acquisition and continual extension of their databases, be they dictionaries, sets of rules, or additional relevant semantic or pragmatic information.

This 'complexity barrier' is a major concern in certain branches of computer science and AI. The complexity is manifest in two places, in the problem of designing the system and in the problem of the cost and speed of a running system which ultimately determines whether it can be a viable commercial option. Below we summarize some of the techniques that may be relevant to dealing with these complexities. It is largely because of such considerations that it is essential that the project of developing systems for NLP, MT or ES needs to keep in close touch with developments in AI generally, and in such areas of computer science as software engineering, computer architectures, distributed computing and computer networks - any of which might provide a key for dealing with the complexity barrier.

1. The first strategy is to attempt to capitalize on any relevant techniques or tools developed in AI and software engineering in general. Computer Science has come a long way in the last decade in terms of the development of programming methodologies and general programming tools - many of which may not be known to specialists in, say, NLP or MT. It is the availability of such general tools and methodologies that has, more than anything else, made it possible for software development to be a much more cumulative discipline

than it has ever been before, and it is the cumulative effect that has made it possible to cascade the work of many people into extremely complex systems. The obvious candidates for attention in NLP and MT projects are:

- Software engineering methodologies - for example the design and use of 'specification languages'. (Software engineering is now a discipline. Although anyone can learn to program, there is such a thing as better and worse styles of programming which show up in the ease of debugging and extending programs - expecially complex ones.)

- High level languages such as LISP, PROLOG, MODULA, and derivatives such as LOGLISP, concurrent PROLOG, INTERLISP, GLISP (Novak, 1983), object-oriented languages like SMALLTALK (Goldberg & Robson, 1983), as well as more general tools such as syntax-directed compilers which enable researchers to design their own special-purpose language relatively easily.

- Software workbenches which help designers keep track of modifications to software and to debug it more easily, and workstations which allow multi-tasking to be used with multiple window screens, and so on.

- Special purpose software development systems, sometimes designed for purposes that initially appear different from the particular application at hand. For example, as we indicated earlier, there are now a large number of widely available development systems for creating knowledge representations (i.e. knowledge-representation languages such as KL-ONE - Schmolze & Brachman, 1982), for building expert systems (OPS5, KES, LOOPS, ART, etc, to mention only a few of the commercially available ones), for carrying out inferences (e.g. the LMA system, Lusk, McCume and Overbeek, 1982), for building prototype parsers to test out grammars and marker semantics (e.g., ATNs, Metamorphosis grammar parsers, Q-systems, Lingol, etc).

2. The second strategy also involves capitalizing on the 'division of labour' principle that pervades science and technology. Almost all AI and NLP projects involve knowledge that resides in other disciplines. For example, we have already mentioned the importance of modularity of NLP systems. Well, the workings of the various modules are typically studied by different disciplines, even though many workers have become proficient in several of the contributing disciplines. Nonetheless one frequently hears linguists lament that many of the people building NL systems do not appreciate some of the subtleties of language, of semantics, pragmatics, discourse structure, and so on, that linguists have studied, and conversely linguists are not aware of the developments and the techniques studied in computer science. AI and NLP are essentially interdisciplinary pursuits. Even knowing where the module boundaries might be can gain from multidisciplinary participation in the research

project. (This does not mean, by the way, that different specialists - each one ignorant of the other's field - should participate in the research team. It just means that the content and methodology of different fields should be attended to. Ideally all participants would have some competence in the other fields).

3. The third strategy is to adopt (or create) various computational aids for the task of acquiring the relevant data for building large databases. When you remember that expert systems like XCON contain 2500 rules (and others, like INTERNIST, even more), and natural language systems have dictionaries and grammar rules in the same order of complexity, you realize that the sheer task of assembling this database, not to mention extending it to new domains, is formidable. Several expert systems now contain facilities for incrementally adding new rules by carrying out a dialogue with the expert (e.g. the TEIRESIAS extension of the MYCIN medical diagnosis system, Shortliffe, 1976). Adding new grammatical and lexical knowledge to a natural language system is a task that has not been automated to any extent, although the TEAM system at SRI mentioned earlier is a particularly notable counterexample: it actually acquires linguistic knowledge in the course of a dialogue with a person who is not an expert in linguistics. The ultimate goal, of course, is to automate the process of rule acquisition - in other words to endow the system with a learning capability. While fulfillment of this goal is a long way off, it does remain one of the major

research areas for ES and NLP. For example, in their report on promising research areas for NLP, Waltz et al. (1983) single this out as one of the most important areas for future development, saying that "...it would be desirable to build a system that could learn both vocabulary and world knowledge by dialog in NL with a user, or by reading text (e.g. dictionaries, encyclopedias, texts, stories, etc.)". Allen Newell also singles out the problem of knowledge acquisition as one of the main bottlenecks to highly intelligent systems.

4. The fourth strategy, discussed by Hart (1983), is to attempt to build 'deep' as opposed to 'shallow' systems. A deep expert system is one which does not produce some level of performance solely by the incorporation of various ad hoc heuristic rules, but rather is based on a representation of a deeper causal model of the process for which it is an expert. Most of the expert systems one hears about these days, and most of the nearly-commercial NLP systems are 'surface' systems. 'Deep' systems tend to be fragmentary research vehicles, often with poorer performance but with greater extendability in principle. The idea that Hart puts forward is that as the task becomes more and more difficult the complexity of surface systems will go up exponentially, while that of deep systems will increase more gradually. Presumably the relative performance of these two kinds of systems will do the opposite - that is, the performance of deep systems will not degenerate as rapidly as they are applied to

more difficult problems. This is an instance of the well-known power-generality tradeoff. This sort of trade-off appears especially dramatic in MT where it has long been observed that one can get, say 70% of the job done with something like 10% of the effort. Shallow systems can often do things with techniques that cannot be extended to more general domains. It is now widely acknowledged that generality in MT can only be attained with 'deep' methods. Of course, as we have been pointing out all along, general systems are not the only kind of system worth pursuing.

Even if the conjecture about the relative growth of complexity in deep and surface systems is true, however, Hart agrees that research on surface systems needs to be pursued - not only because such systems may be immediately useful, but also because the performance limits of surface systems is far from being understood. For example, the HARPY speech recognition system is a surface system: it was designed primarily to meet the DARPA 5 year performance milestones. It does not fit our current view as to how a really good speech recognizer (such as a human) must work, as does, say, the HEARSAY system. Yet it continues to outperform the 'deeper' HEARSAY II system (also developed at Carnegie-Mellon University) and nobody is really sure that it won't continue to do so indefinitely. Newell, for one, believes that HARPY embodies important implicit and not well understood principles, despite its more pragmatic and engineering pedigree. For these and other reasons there is no doubt that research must continue on both kinds of systems.

5. The fifth strategy is concerned with the complexity of the realization phase, rather than the design phase. Many AI systems must solve combinatorial problems of great complexity. Others require vast amounts of storage. For these to be realized in an economical way it may be necessary to develop special-purpose high-power hardware. It's easy of course, to simply put the problem in someone else's field - to blame any lack of success on the slowness or other limitation of available hardware. But when we are concerned with economic viability, and especially when the application must be carried out in real time - as in the case of speech recognition and question answering - hardware can play a crucial role. In the section on speech recognition we mentioned some of the new inexpensive speech recognition and synthesis VLSI chips. These are examples where VLSI technology will make current computational ideas economically attractive. In this category there are also some very promising technologies for very large mass memories (e.g. the writeable laser disk is one important current product which may be an important component of NLP and MT systems). There are also some exciting possibilities for very fast processors using optical computing elements. But perhaps even more important in the long term is the recent crop of ideas for new computer architectures. These include various massively parallel computer designs (e.g. the 'connection machine' consisting of thousands of active message-passing processors instead of random access memory registers). There are also hardware architecture designs being developed which are especially

suited for different classes of software. There are crossbar networks like NETL (Fahlman, 1981) that are useful for certain kinds of inferences in semantic nets (especially 'property inheritance' inferences), Data Flow machines that are especially suited for object-oriented message passing languages (like SCHEME, Hewitt, 1981), associative memory and pattern-matching machines suited for certain kinds of pattern invoked processes - including logic-based languages like PROLOG, and hardware especially designed for running concurrent logical processes for programming systems like concurrent PROLOG. In addition there is considerable interest in special hardware for sensory processing (e.g. in vision) because of the large degree of parallel cooperative processing that appears to be taking place there (for example, in the human visual system's computation of shape, movement, stereopsis, and so on). There are already special hardware designs being studied for their potential relevance to OCR (i.e. the 'convolution box' which came out of the MIT vision research). Although there are many radical ideas for new architectures it is still too early to tell which, if any, of them lies on the critical path to new commercially viable AI or NLP products. In any case it seems very likely that some of these developments will indeed make a fundamental difference to some problems which now appear to be too resource-demanding to be commercially viable - vision and speech recognition being especially good candidates. Thus future AI, NLP, and MT work needs to keep in close touch with computer architecture developments - especially those aspects that relate to large fast memories and fast inferences.

7.4 Project Planning and Organization of Funding

The question of funding major projects of national concern - such as MT - will be considered in detail in a separate report. At this point we present a few remarks based on some observations of research planning in the U.S. As mentioned earlier, the U.S. does not have a national plan as do a number of other nations. Nonetheless, there have been cases of large scale planning and funding that deserve special consideration because there may be lessons to be learned from it that apply to the Canadian needs in MT. In particular, we will look at the case of the DARPA speech recognition project.

DARPA has been funding speech recognition research for a number of years. In the early 1970's it was decided that speech recognition was of special national concern and a plan for funding a number of promising projects on a large scale was devised. Planning for the DARPA initiative was done in a very systematic way. A blue-ribbon committee chaired by Allen Newell laid out a rational 5 year plan in 1973 (Newell et al., 1973), including detailed modest performance criteria to be met by the end of the period (e.g. it should be able to recognize in nearly real-time, and after a minimal period of training, a vocabulary of several hundred words spoken by a single speaker using a restricted syntax and one restricted subject matter). Some 10 or more laboratories had initially been funded and the number was systematically reduced on the basis of their promise and performance, until only 3 remained (BBN, CMU, and SRI). These were given large-scale funding to continue their research. Most of the projects took a serious long-term approach aimed not simply at meeting the performance criteria, but also at designing a system on the basis of sound principles that would enable the further development of the highest possible quality continuous speech recognition systems. Towards the latter part of the 5 year plan Carnegie-Mellon University added a second project, called HARPY, to their research effort. HARPY took a 'shallower', or more engineering approach, aimed directly at meeting the performance criteria. It was based largely on existing technology (e.g. it used a technique known as 'beam search') and on stochastic mechanisms (probabilistic transition nets) not favoured by the AI community because of their ultimate limitations. HARPY ended up with the highest performance of all the DARPA sponsored projects, comfortably exceeding the 5 year performance milestones.

Despite this success DARPA ceased funding the speech recognition research because the results did not show promise of producing a useable system in the forseeable future. Many people consider the experiment in large scale central planning and funding to have been a mistake. For example, Prof. John McCarthy, a pioneer in artificial intelligence and critic of the project, feels that what was missing in the experiment was a healthy diversity of approaches. Though no one doubts that the projects funded represented the most promising ideas available, the problem in this case may not have been at a stage where a Manhattan Project style all-out attack was the right approach.

There are arguments for both sides. On the one hand central planning does curtail diversity by "putting all one's eggs in one basket." On the other hand it is not clear that the same amount of research would have been done by equally able people if the funds had been more liberally distributed. The individual projects did require considerable resources as well as contributions from a variety of disciplines. The question of centralized versus decentralized mission-oriented research planning has to be considered in relation to such factors as the state of the basic underlying science and technology, the availability of appropriately trained professional personnel, the estimated size of a working system (e.g. the amount of sheer data that have to be gathered and assimilated), and the need for special facilities, such as computer power, nontechnical and technical support staff. It is not obvious, for example, that it would have been possible to focus as much total manpower on the projects if the individual centres of research had not been up to some critical size and had the intellectual and material resources to attract able professionals, who are quite rare and much in demand, to work on that problem rather than one with higher and more immediate payoffs. The issue is worth raising, even if there is no clear answer in this case, because it is one that will have to be faced in Canada, given the need to solve a problem of the importance and magnitude of Machine Translation.

It should also be noted, now that the major DARPA speech projects are no longer being funded, a number of less ambitious projects have appeared. Many of these are based on novel ideas and many are being pursued in private industry. It's hard to assess how much these have been influenced by the DARPA experiment. Certainly the fact that so much attention was being focussed on speech recognition in the 1970's through DARPA support had an impact on the interest that others took in the problem (for example, just as DARPA began funding the major speech research program, Bell Laboratories decided to undertake their own project). In addition it increased the basic knowledge in the field and provided important training for many people who later continued speech recognition research in commercial laboratories (e.g. many of the people working on speech at Fairchild, as well as those at Dragon Systems, came from the SRI speech project). The training factor alone might justify mounting large-scale centralized research projects.

The new projects of the 1980's represent a considerable diversity of approaches and goals. Many of them, in fact, are intended to produce commercial products in the near future (e.g. the work at Fairchild, Bell Labs, and Dragon Systems in New York). It is still too early to tell whether or not these new projects will produce a more useable product than came out of the large centrally planned DARPA projects, although the review of current research presented earlier does suggest that market forces will probably dictate that some limited products will become commercially significant in the near future.

7.4.1 A Note on National Funding Styles

The traditional Japanese style of funding new technological thrusts involves making a large national commitment, with central planning and close cooperation among government, industry, universities, and other groups (e.g. labour). A number of European countries have also pursued large-scale projects of national interest by creating national plans (e.g. Britain, France, Scandinavian countries and the Netherlands). In the U.S. there is a tradition of funding major programs of national interest through the Defence Department. In the case of AI and NLP, as we have already remarked, most of the funding has come through DARPA, with additional significant support through other defence agencies (Air Force Office of Scientific Research, Army Research Office, etc). In addition, where national and corporate interests coincide there has been industry-sponsored research, such as the current massive support of computer science by computer companies mentioned earlier (e.g. the Microelectronics and Computer Corporation, and such university-based laboratories as Stanford's Semiconductor Research Corporation, Carnegie-Mellon's Robotics Institute, MIT's Project Athena - all funded at the $10 to $70 million level).

There is also a significant infusion of research funding for high technology through smaller technology companies with venture capital backing. Indeed, the latter has become a major factor in recent commercialization of AI. Our observation at the recent conference of the American Association for Artificial Intelligence (AAAI) is that most academic AI researchers who have developed significant AI techniques or systems have formed private companies, usually with some venture capital backing, to further develop and market these ideas. The total amount of research support in this form is not known, but judging by the number of people involved it may be significant.

The point of these remarks is that such traditional sources of support for R & D (especially in high technology) are not generally available in Canada. Research in Canada is supported either by the few Canadian companies large enough to have R & D laboratories here, or by traditional science-funding agencies. The level of support that such channels can offer is extremely limited. Even in the U.S. funding by traditional science granting agencies would not be nearly adequate to support R & D into AI and NLP. We have been told that it is the capacity of agencies like DARPA to supply large contracts and superstructure grants (e.g. the ARPANET discussed below) that made it possible for the U.S. to reach the level of accomplishment in AI that it now enjoys. What we need in Canada is to find a new avenue of support for such large scale projects of national interest as those in NLP or MT or AI, to compensate for the lack of a tradition like that of Japan or the U.S.

7.5 Computer Networks: Developing a 'Research Community'

Another DARPA experiment that provides an important lesson for Canada was the establishment of the ARPANET, a packet-switching computer-to-computer network which linked together laboratories receiving funding from DARPA. Arpanet was the first such experiment in inter-laboratory communications. It was initially established to rationalize the cost of computer resources. The idea was that computer resources could be shared across projects. Although such sharing does take place, especially with centres having special facilities (such as the Illiac IV computer, or special medical computing facilities such as Stanford's SUMEX-AIM project), by far the most important result of the Arpanet was to bring together into an intellectual community researchers working on AI and related advanced projects in computer science. It enabled the main AI laboratories to keep in close touch by computer mail and allowed the rapid dissemination of documents (e.g. technical reports and theses) and even software. It would be difficult to overestimate the importance of this facility. Indeed it contributed to a distortion in the usual channels of scientific communications. A very small proportion of AI work (at least in the 1970's) was ever published. The research was reported in technical reports that could be accessed by computer file-transfer systems and printed. Network mail kept people informed about developments in different areas and demonstrations of operational systems could easily be arranged at different sites. In the last few years it has become common to submit reports for publication in magazines such as the AI Magazine and the SIGART Newsletter using the ARPANET (editors of these magazines publish their 'net address' along with their postal address for this purpose).

The ARPANET helped enormously in getting a major effort in AI started in the U.S. It provided a community of workers, a way of sharing not only information, but also computational tools and computational resources. It is no wonder that virtually every major effort in computationally-based science strived to provide network facilities for its workers (e.g. the medical applications of AI set up the SUMEX-AIM facility, cognitive scientists have proposed the establishment of COGNET, many major educational institutions are linked by EDUNET, and the computer science community in the U.S. as a whole is currently developing CS-NET). The establishment of a network is also one of the major proposals in the Dutch AI report and the Canadian Microelectronics Cooperative recently funded by NSERC.

7.5.1 Potentials for International Cooperation

The issue of research community also raises the question of international cooperation. The Dutch report points to the importance of computer networks, which cross national border at 'gateways' to link freely with the networks of other nations. They clearly have in mind international cooperation. Canada has a strong tradition of research cooperation with the U.S. and Great Britain and (especially in the case of NLP and MT) with France. The British plan, as laid out in the Alvey report, explicitly calls for international cooperation on the grounds that the problems posed by Fifth Generation systems are too large to be solved by Britain alone. The Japanese plan is also emphatic in placing importance on international cooperation for the same reason. Japanese industry have already made joint venture arrangements with U.S. researchers. Canadian researchers have also been approached by both Japanese industries (see the report by Manning and Wright on the Waterloo delegation visit to Japan, 1983) and by European industries (e.g. we have been approached informally by Siemens). Canadian planning should be sensitive to the possibility of joint ventures in high technology (especially MT and NLP) involving foreign countries. This would not only help to finance large projects of mutual interest, but would also have the enormous benefit of helping provide training for Canadian researchers in the highly specialized areas of AI/NLP/MT, for which there is presently an acute shortage of trained personnel not only in Canada, but throughout the world.

APPENDIX I

INVENTORY OF MT SYSTEMS

APPENDIX I: INVENTORY OF MT SYTEMS

In this appendix, we enumerate the most important MT systems. We distinguish between operational systems and prototype systems.

A - Operational Systems

Grouped under this heading are products which are currently being used for production purposes whether or not these systems are commercially available.

1) SYSTRAN

- Owner: Latsec Corp.; several licensees, including Kawasaki for the Japanese version.
- Development: Latsec Corp.; extended and improved by several users and licensees.
- Users: US Air Force, NASA, EEC (which has invested 4 million ECUs (about $4.2 million), in improving the system), General Motors, and several others.
- Language pairs: from English to French, Spanish, Portuguese, Italian, German, Japanese, and Arabic; Russian, French and Japanese to English.
- Hardware: IBM mainframes.
- Price: approximately $5000 per month for software for the first language pair.

- Suboptimization techniques: specialized dictionaries, post-editing.

- Technical features: direct approach; use of some syntactic and semantic information in dictionaries, interfaced with OCR and word processing.

- Evaluation: SYSTRAN is still one of the best systems on the market. Over the years, very comprehensive dictionaries have been developed. However, because SYSTRAN's linguistic processing is very weak, users are forced to continue expanding the dictionaries with large numbers of ad hoc 'idiom' rules. This system cannot achieve more than very little raw translation, with frequent mistranslations and ungramaticabilities.

2) LOGOS

- Owner: Logos Corp., Bedford, Mass.

- Development: Logos Corp.

- Users: in the past US Air Force; currently, at least one translation house (Omniplex) and several German companies including SAP, a large software house.

- Languages: former versions for Russian to English, English to Vietnamese and English to Farsi; currently marketed version for German to English; English to German to be released soon; longer term plans for several other pairs including English to French.

- Hardware: Wang OIS System.

- Price: 120,000 Deutsch Marks (about $56,000) for hardware (Wang OIS with 256 K memory and one terminal); software charges according to use, approximately 50 pfennigs (about 25¢) per line translated.

- Suboptimization techniques: some interactive pre-editing, specialized dictionaries, post-editing.

- Technical features: dictionary maintenance facilities; relatively rich lexical information including hierarchy of semantic classes; large dictionary for German-English (65,000 words) provided with the system; integrated word processing environment.

- Evaluation: It is too early to give a global judgment on this new version of LOGOS: it is not clear how much improvement there is over previous versions. Apparently, the German-English system is a scaled down version of a system which had been built in co-operation with Siemens, and which Siemens had finally judged unsatisfactory. The lack of any treatment of compounds is clearly a major problem (especially for German). The company claims that the system is based on an interlingua rather than transfer approach: examination of the dictionaries shows that this claim is based on a non-standard understanding of the notion of interlingua.

3) <u>ALPS</u>

- Owner: Automated Language Processing Systems, Provo, Utah.

- Development: same.

- Users: Agnew Techtran (a translation house) in Los Angeles, Xerox translation services in Rochester and Control Data Service Bureau in France among a few others.

- Languages: English to French, Italian, Spanish, German and Arabic; French to English.

- Hardware: Data General MV-4000 minicomputer.

- Price: Between $120,000 and $400,000 depending on the specific configuration.

- Suboptimization techniques: some pre-editing, specialized dictionaries, interactivity and post-editing; system can also operate in MAHT mode (two different levels of assistance).

- Technical features: interfaced with OCR systems, direct translation approach, integrated word processing environment.

- Evaluation: At the Canadian Translation Bureau, Alps has been found to be unsuitable for HAMT on the type of text for which it has been tried. With incomplete morphological descriptions, rudimentary lexical information, a weak syntactic component and total reliance on its user for semantic information, the system may be considered linguistically primitive. Raw translation (after interaction but before word processing) is frequently ungrammatical or incorrect, and never more than very literal. Interactivity is being used in place of powerful linguistic processing.

4) <u>ALPS Writer's and Author's Workstation</u>

- Owner and development: Alps, Provo, Utah.

- Users: the system has just been introduced on the market.

- Languages: applicable to almost any language.

- Hardware: Convergent Technologies microcomputer.

- Price: It varies according to the number of users, for example, the price is $45,000 for one user and $85,000 for 5 users.

- Suboptimization technique: machine-aided human translation.

- Technical features: a workstation comprising word-processing, on-line dictionary, multi-alphabet printing, etc.

5) <u>WEIDNER System</u>

- Owner: Weidner Communications, Provo, Utah a company controlled by Bravice (Japan). Several licensed distributors, including Cosmopole Communications (Montréal) and Omnitrans (for Arabic countries).

- Development: Weidner Communications.

- Users: reportedly, about 25 systems have been sold so far. In Canada, the best-known user is Mitel Corporation. Cosmopole Communications, a translation house, also uses the system on a daily basis.

- Hardware: the system runs on several models of the PDP-11 series; more recently, it has been implemented on an IBM Personal Computer.

- Prices: wide range of options from buying a PDP 11/44 at $150,000 US (plus options for more pairs of languages, dictionaries, terminals, etc.) to renting an IBM PC at $435 US per month.

- Suboptimization techniques: some pre-editing, specialized dictionaries, post-editing. Although the system is advertised as a machine aid to human translation, the translation process itself is fully automatic.

- Technical features: direct approach (no separate dictionaries for SL and TL), on-line dictionary maintenance, integrated word processing.

- Evaluation: Linguistic processing is clearly below state-of-the-art. Lexical descriptions are sketchy, syntactic analysis is local and there is very little semantic processing. The ran translation is frequently ungrammatical or incorrect, and never more than very literal.

6) <u>Weidner's workstation</u>

- Owner and development: Weidner Communications, Provo.

- Users: The system has just been introduced on the market.

- Languages: Applicable to most languages.

- Hardware: IBM Personal Computer.

- Price: Approximately $10,000 US.

- Suboptimization techniques: machine-aided human translation.

- Technical features: Word-processing, on-line dictionary.

7) <u>SMART</u>

- Owner: Smart Communications Inc., New-York.

- Development: same

- Users: several (undisclosed) multinational corporations; Employment and Immigration Canada.

- Languages: English to French, Italian, Spanish, German and Japanese.

- Hardware: IBM 370, VAX 780.

- Price: $85,000 to $125,000 US for the software (one pair of languages).

- Suboptimization techniques: artificially restricted input, post-editing.

- Technical features: on-line text critiquing facilities, integrated word processing facilities.

- Evaluation: No global judgment can be made at this point.

8) <u>METEO</u>

- Owner: Canadian Secretary of State Department.

- Development: TAUM group, University of Montréal.

- User: Secretary of State on behalf of Environment Canada, at CMC in Dorval. System is in daily operation since 1977. Approximately 4 million words translated per year.

- Languages: English to French.

- Hardware: CYBER

- Price: not a commercially available system.

- Suboptimization techniques: sublanguage approach; restricted to weather forecasts.

- Technical features: Second-generation approach; complete syntactic/semantic analysis; entirely written in a high level meta-language called Q-SYSTEMS; automatic detection of text that cannot be translated correctly; translation rules operate on sentence structures (not word-for-word).

- Evaluation: METEO is quite limited in scope, but it constitues so far the only reported case of a fully automatic high quality translation system.

9) <u>Nagao's system at ETL</u>

- Owner: Electrotechnical Laboratory, a public R&D lab, under MITI in Japan.

- Development: Nagao's group at University of Kyoto.

- User: The Research in Information Processing (RIPS) department of ETL, Ibaragi.

- Language pair: English to Japanese.

- Hardware: Fujitsu FACOM M200.

- Price: Not commercially available.

- Suboptimization techniques: input restricted to titles and indexes of technical papers.

10) ATLAS/I

- Owner, development and user: Fujitsu laboratories, Kawasaki City.

- Language pairs: English-Japanese, Japanese-English.

- Hardware: Fujitsu FACOM M-180II.

- Price: not commercially available.

- Suboptimization techniques: pre-editing, interactivity, restricted to computer manuals.

- Technical features: transfer approach, case frame analysis.

11) SPANAM

- Owner: Pan American Health Organization (PAHO), Washington, D.C.

- Development: PAHO developed the system between 1976 and 1980 as a modified version of the old Georgetown University system.

- User: PAHO has been using SPANAM in a production environment since 1980, and by May 1983, more than 1,100,000 words had been translated.

- Languages: Spanish to English; an English to Spanish version is currently under development.

- Hardware: IBM mainframe.

- Price: not commercially available.

- Suboptimization techniques: post-editing.

- Technical features: direct approach, integrated word processing facilities.

12) <u>TITUS II</u>

- Owner: Institut Textile de France.

- Development: same, 1972-73.

- User: same, since 1973 for translation of abstracts of technical papers on textiles.

- Languages: between English, French, German and Spanish in all directions.

- Suboptimization techniques: artificially controlled input.

- Technical Features: closed vocabulary; 16 predefined simple sentence patterns.

13) <u>CULT</u>

- Owner: Chinese University of Hong Kong.

- Development: same, began in 1968.

- User: same, for translation of 'Acta Physica Sinica' ad 'Acta Mathematica Sinica'.

- Languages: Chinese to English.

- Suboptimization techniques: pre-editing, interactivity and some post-editing.

- Technical features: not known.

B - Prototype Systems

1) SUZY

- Organization: Sonderforschungbereich 100, Elektronische Sprachforschung, University of Saarbrucken.

- Sponsors: Deutsche Forschung Gesellschaft has funded the development of SUZY between 1972 and 1982.

- Goals: General purpose MT system for several language pairs, including German to French and English, and Russian, French and English to German. Coverage varies with language pairs; Russian and German analysis components are fairly comprehensive (the German dictionary contains approximately 75,000 words). No real MT production is planned for the moment. The analysis component of SUZY, which is called SATAN, is being experimentally tested for information retrieval of German texts. For the sake of a few experimental uses like that, SUZY is now being stabilized, and most of the efforts in the next three years will be going to the development of an improved version of the system: SUZY-2.

- Approach: SUZY can be considered a sophisticated first-generation system. There is no clear separation between algorithms and linguistic rules. The parser is similar to the one used by SYSTRAN: homograph resolution is done separately from structural analysis, using matrices of forbidden category strings. Semantic analysis is rather primitive. There is no clear stand on sub-optimization strategies. SUZY-2 is going to be much closer to a typical second-generation system.

- Evaluation: SUZY is clearly below the state-of-the-art and is not likely to be turned into a useful product; it is too early to judge whether or not SUZY-2 will fare any better.

2) <u>TAUM-AVIATION</u>

- Organization: TAUM group, University of Montréal, between 1976 and 1981.

- Sponsors: The Canadian Secretary of State invested approximately 2 million dollars in the project.

- Coverage: developed for the translation of aircraft maintenance manuals. The project has been abandoned in 1981 after an evaluation showing that the system was not likely to become profitable for its sponsors in the next few years. At the end of the project a complete new second generation software had been developed, and fairly extensive grammars for analysis, transfer and generation for English to French had been built. The dictionaries were relatively small, being limited to core and hydraulics vocabulary.

- Approach: TAUM-AVIATION is a typical second-generation system, with a transfer approach. Its scope is restricted to the sublanguage of technical maintenance manuals. The analysis component is syntax-based but makes extreme use of semantic filtering. The parser is written in REZO, a compiled language based on augmented transition networks. The processing unit is largely restricted to the sentence. Translation rules are written in LEXTRA, a meta-language with full transformational power.

- Evaluation: Although the system did not incorporate the most recent developments in semantics, it approximated the state-of-the-art for texts of that level of complexity. Although post-editing was judged necessary, the quality of the output was high by MT standards: generally grammatical and frequently better than literal. Due to the haste with which the software modules were developed, they are somewhat heterogeneous. Plans to advance the system for commercialization where laid out by Isabelle (1981).

3) METAL

- Organization: Linguistic Research Center, University of Texas, Austin. Although LRC has been involved in MT since the early sixties, the current project really started in 1977.

- Sponsors: U.S. Air Force until 1979; since then, Siemens Corp. is the sole sponsor. Approximately 30 person/ years have been invested up to now.

- State of the work: as the first step in a larger development project (see Siemens project, Appendix II, section II-B-I), METAL has been developed for German-to-English translation of telecommunications equipment manuals. Currently, the German parser has a fairly extensive coverage. The dictionaries comprise approximately 10,000 words, mainly in the telecommunications domain.

- Approach: METAL is a second-generation system. It is based on a phrase structure grammar augmented with arbitrary procedures and transformation rules. The system is essentially syntax-based and makes little use of semantics. Some limited form of pre-editing is used and post-editing is judged necessary. Presently, the system could be said to be based on a sub-language approach, but it is not clear in what direction it will be extended by Siemens. The system includes dictionary maintenance facilities, spelling correction, fail-safe mechanisms (the system always produces some output), and word processing facilities. It is written entirely in LISP and runs on a Symbolics Lisp-machine.

- Evaluation: METAL certainly has a state-of-the-art syntactic parser. Lexical information does not appear to be extremely rich though. The transfer and synthesis components are weaker. In fact much of what is usually called synthesis is intertwined with transfer, which would appear detrimental given the goal of multilingualism. Finally, the semantic component is extremely weak: the system relies heavily on 'likelihood scorings' and idiom rules (e.g. for almost all German compounds).

4) ARIANE-78

- Organization: Developed between 1972 and 1980 by the GETA at the University of Grenoble.

- Sponsors: Centre national de la recherche scientifique and French Defence Department (DRET)

- Coverage: The software is language-independent, but the system has been tested on a large-scale application in Russian to French translation of scientific texts. Smaller experiments have been conducted for a large number of language pairs.

- Approach: ARIANE-78 is a nice example of second-generation MT technology. Several different specialized metalanguages are used for the various components of an MT system. Morphological analysis and morphological synthesis are effected through string-to-tree and tree-to-string transducers; syntactic/semantic components are realized by means of a tree-to-tree transducer. A transfer model is used, the intermediate representation being a multi-level interpretation of the text: surface syntax, deep syntax and semantics. As with other second-generation systems, the translation process is syntax-based, and semantic analysis is rather superficial. ARIANE-78 includes a complete translation environment, with post-editing facilities.

- Evaluation: ARIANE-78 is certainly one of the best second-generation systems currently available. The system is unfortunately not machine-independant.

(This will be corrected in th ESOPE Project). Some design principles are questionable. For example, using only trees (instead of 'charts' as in Q-SYSTEMS) for syntactic/semantic analysis makes grammar writing much more difficult for the linguist. Ongoing to an extensive use of fail-safe mechanisms the output sentences are occasionally ungrammatical.

5) <u>Nagao's system for computer manuals</u>

- Organization: The system was developed by Prof. Nagao's group at University of Kyoto between 1978 and 1981.

- Sponser: The Japanese government.

- Coverage: Developed for the translation of computer manuals from Japanese into English. The system has remained a relatively small-scale prototype.

- Approach: Second-generation system, transfer model, rich system of lexical rules.

- Evaluation: This system may be the best one realized so far in Japan. It will serve as a model for the current projects sponsored by MITI and by the Science and Technology Agency (see Appendix II, sections I-B-2 and I-B-3).

The systems listed below were designed as small-scale prototypes of a more advanced generation of systems.

6) Wilks' system

- Organization: Realized by Yorick Wilks and his students while he was at Stanford University.

- Coverage: very small scale system for the translation of simple texts from English into French.

- Main ideas: Wilks was one of the first to experiment with an AI approach to MT. His prototype MT system made very little use of syntax. Instead, his system would first assign semantic formulas to the words of the SL text, and then it would attempt to match 'semantic templates' onto sequences of formulas, in order to select intended word senses. The matched templates would then be linked together (eventually across sentence boundaries), using inference rules; these techniques permitted to solve certain anaphora problems. All semantic rules were based on the idea of preferential rather than categorical filtering, in order to deal with such phenomena as metaphors.

- Evaluation: Wilk's system had the merit to illustrate how AI techniques could be applied to MT problems, showing at the same time that MT was a worthwhile topic for AI research. Some claims that Wilks made in building his system were highly controversial and went much beyond what the prototype could really show. For example, the claim that syntax was largely irrelevant to understanding and MT had no strong basis: it can be argued that Wilks' system actually makes use of syntax (in disguise under a different terminology), and can at any rate handle only a very small subset of English.

7) <u>SALAT</u>

- Organization: Sonderforschungbereich 99, Heidelberg University

- Funding: Deutsche Forshung Gesellshaft funded a group of 3 or 4 between 1976 and 1980

- Main ideas: The research starts from the recognition that fully automatic high quality translation of unrestricted text requires a system that can access a data base containing general knowledge of the world, situational knowledge and inferencing mechanisms. The aim is then to explore possible ways of encoding and using such knowledge in an MT system, thus bringing a contribution to translation theory, linguistic theory and artificial intelligence. A prototype system, SALAT, has been developed (Hauenschild et al. (1979)). The system is based on a transfer approach and algorithms are separated from linguistic data. The system comprises a full-fledged syntactic component. The further use of a database containing knowledge of the world and of a deductive component makes SALAT a 'third-generation' system. These features are used during both transfer and analysis to resolve ambiguities.

- Evaluation: Although SALAT is not a very large system, it has been tested with real text in several languages. This project thus embodies some of the most interesting MT research that has been done in the last few years. It is unfortunate that the project has been abandoned. One of the former members of the group, C. Hauenschild, is apparently now pursuing some related work of her own at the University of Konstanz.

APPENDIX II

CENTRES OF ACTIVITY IN MT

APPENDIX II - CENTRES OF ACTIVITY IN MT

In this appendix, we try to identify the main actors and their funding levels, on the current scene of MT. Each activity is classified according to its institutional framework source (public, private, university), and its locus (Europe, U.S., Japan).

I - Public Projects

A) Europe

1) EUROTRA

- Organization: EUROTRA is a very large scale project with a complex organizational structure. The project was officially launched at the end of 1982 and is scheduled to finish in June 1988. The structure includes a central team located in Luxembourg which is responsible for project management, and for system specifications. In principle, each one of the EEC countries will set up an institute that will channel available money to one or more university groups. These groups will undertake the development of dictionaries and grammars, using the software defined in Luxembourg. A liaison committee (ISSCO in Geneva) will coordinate the work of the national groups.

- Goals: To produce operational prototypes that can translate restricted types of text between the seven current languages of the EEC. It is hoped that the new system will replace SYSTRAN at EEC and produce much better results.

- Sponsors: Over the 5 1/2 years of the project, the GD-13 Administrative unit of the EEC will provide 16 million ECUs (about $16.6 million) while the participating countries will provide an additional 11 million ECUs (about $11.5 million).

- State of the work: Preliminary software specifications have already been worked out, and are currently being implemented in an 'experimental assembly'. Linguistic specifications are much less advanced.

- Approach: From the software point of view, EUROTRA will borrow several ideas from Q-SYSTEMS, and develop a production rule system with richer control and data structures. Very little material has been made public on that, and even less on the linguistic framework. They will use a transfer approach, which means that 7 analyzers, 7 generators and 42 transfers will have to be developed. As far as we can tell, the system will be largely syntax-based.

- Evaluation: This project is obviously very ambitious, even if the immediate aim is only 'operational prototypes'. It is no doubt one of the largest MT projects ever undertaken. The construction of 42 transfer components is not an easy task, even with massive funding. Some of the participating countries have very few qualified specialists. However, one hopes that even if all the goals are not reached, significant progress will be made.

2) Projet National (formerly ESOPE)

- Organization: French national project; participants include the GETA group (University of Grenoble), Institut de Formation et de Conseil en Informatique (IFCI - a non-profit organization), and SG2 (a high-technology French company).

- Sponsors: The Agence pour le Development de l'informatique (French government) and SG2 would apparently each provide half of the budget of 56 million French francs (about $9 million) over the next three years.

- Goals: The project is considered to be a 'pilot' project, that is, a project that is oriented towards technology transfer between university and industry. The end product should be a 'pilot' multilingual MT system, meaning a system which, after some packaging and further development, should become an industrial product. A whole new software would be developed, and prototype systems would be created for two or three pairs of languages: English to French (and possibly Spanish), and French to English. Each prototype should have a broad grammatical coverage, a core dictionary (9000 entries) and at least one large specialized dictionary.

- State of the work: The project is just starting. Some preliminary work has been accomplished in the last 18 months with limited funding from l'Agence de l'Informatique (ADI): technical specifications of the software, linguistic specifications and methodology as well as some training.

- Approach: The systems developed by the Grenoble group in the past (e.g. ARIANE-78) were based on principles very similar to those used in TAUM's systems. The system planned in the ESOPE project, ARIANE-X, will essentially be an improved portable version of ARIANE-78 (cf. GETA, 1982). It is aimed at providing a complete environment for MAHT, HAMT and MT. Facilities for on-line dictionary consultation and maintenance, and for word processing will be integrated. High-level software for writing dictionaries and grammars will be developed. In addition to the set of tools very similar to those used in ARIANE-78, other tools will be developed as well, including ones that are close in spirit to Q-SYSTEMS and to transition networks. There are no definite plans to go beyond the usual semantic capabilities of second-generation systems.

- Evaluation: Although the GETA group has considerable experience in MT, it seems that the project could have difficulty in meeting its goals. First, the time schedule (approximately three years) is very short, given all that needs to be done. Since GETA plans to pursue its own activities in parallel, a massive injection of new personnel is necessary. Cooperation with the private sector is certainly desirable, but some adjustment period is likely to be necessary. The software envisioned looks very complex. Moreover, it will probably not be available before the final phases of the project, which makes upward compatibility with the previous system imperative.

B) Japan

1) Fifth Generation Computer Project

- Organization: R&D is carried out by the Institute for New Generation Computer Technology (ICOT), a public organization in which the main Japanese computer manufacturers have a participation (see section 2.1).

- Goals: The general goals of the project have been presented in section 2.1. Five major R&D areas have been identified: inference systems, database systems, intelligent interface systems, development support systems and basic application systems. Machine translation is seen as one of the main components in the latter area. The aim is to develop a multilingual MT system which, with a vocabulary of 100,000 words, will translate with an accuracy of 90% at one third of the cost of human translation.

- Funding: The current budget for the entire Fifth Generation project is estimated at more than half a billion dollars. But it is not known yet what fraction of this amount will be spent on MT.

- Schedule: The program of the Fifth Generation project runs from 1982 to 1992. MT R&D will take place during the middle and final phases of the project, that is, between 1985 and 1992. Results of basic research in NLP carried out during the initial phase will be used.

- Approach: It is too early to have a clear idea of what exactly will be done. But the general direction of the project is to develop sophisticated AI-based products, with the help of new computer architectures and logic programming.

- Evaluation: The general feeling about the whole Fifth Generation project is that while the explicit goals may be overly ambitious, important breakthroughs are likely to be made. This applies to the MT part of the project as well.

2) MITI's system for internal use

- Organization: Independently from the Fifth Generation project, in 1982 MITI decided to fund the development, in the short term, of an MT system for its own internal needs.

- Funding: $240,000 to be spent over three years.

- Goals: Produce a 'simplified' English to Japanese MT system, for the translation of articles from leading American and European newspapers and journals.

- Approach: Use results from research done at University of Kyoto and at the Electrotechnical laboratory, to produce a simplified system suitable for a production environment.

3) The Science and Technology Agency Project

- Organization: The Science and Technology Agency of the Japanese government has launched a project in which Prof. Nagao's group (University of Kyoto), the Electrotechnical Laboratory (ETL -- a public R&D lab) and JICST (a national database of science and technology) will jointly develop an MT system. Nagao's group (see below) will provide the basic software, ETL the hardware, and JICST large scale dictionaries.

- Funding: Between 1982 and 1984, the Agency will provide $3.2 million.

- Goals: Develop in the short term a system which will translate automatically scientific and technical information in both directions between English and Japanese.

- Approach: The system will be a simplified and improved version of Nagao's system for computer manuals (see below).

4) Electrotechnical Laboratory, Ibaragi

- Organization: ETL is a major national laboratory under the responsibility of MITI; within its Pattern Processing Department, a small group directed by H. Tanaka has been working on MT for several years.

- Funding: For 1982, approximately $48,000.

- Goals: Develop an English to Japanese system for the translation of general documents.

- State of the work: In 1982, the prototype had a very small vocabulary: 170 words.

- Approach: Using what they call a 'fusion method', analysis, transfer and synthesis are done simultaneously.

5) NTT, Musashino Electrical Communication Laboratory

- Organization: ECL is the largest research lab within NTT; within the First Laboratory of ECL's Research Divison, a group of researchers headed by H. Nomura is working on a natural language understanding and translation system called LUTE.

- Funding: The project started in 1980; in 1982, nine researchers were involved and the budget was of $140,000 'excluding salaries'.

- Goals: Develop an intelligent natural language understanding system which can perform several tasks, one of them being to translate scientific and technical papers between English and Japanese (both directions).

- State of the work: In 1982, the system had a vocabulary of only 300 words.

- Approach: The analysis component of LUTE performs a deep semantic analysis, using AI techniques and an extralinguistic knowledge base. The system includes an interactive knoweledge acquisition component. Translation is performed using a transfer component.

II - Private R&D

A) United States

1) Latsec Corp., La Jolla, California
   - Start of MT R&D: 1964
   - Activities: Development and commercialization of the SYSTRAN system
   - Staff: average 20-25 people

2) Logos Corp., Bedford, Mass.
   - Start of MT R&D: mid-sixties
   - Activities: Development and commercialization of the LOGOS system
   - Staff: Approximately 90 people

3) Smart Communications, New-York
   - Start of MT R&D: 1977
   - Activities: Edition and translation services for technical documentation; development and commercialization of the SMART system
   - Staff: Approximately 35 people.

4) Weidner Corp., Provo, Utah
   - Start of MT R&D
   - Activities: Development and commercialization of: a) the Weidner translation system, and b) Weidner translator's workstation; translation services.
   - Staff: approximately 60 people

- Approach: See the description of METAL below. The system will be coupled to TEAM, Siemens' terminological database, which with its 1.5 million concepts in up to 8 different languages is one of the largest in the world.

- Evaluation: See evaluation of METAL (Appendix I-15). Siemens' project appears quite ambitious. Although METAL features one of the best syntactic parsers currently available, massive infusion of vocabulary is likely to create semantic problems.

2) <u>Philips' Research Project, Philips Laboratories, Eindhoven</u>

- Project leader: Jan Landsbergen

- Funding: only one person is involved full time

- Research themes: The use of Montague grammar in MT

- Main ideas: Landsbergen is developing a small MT prototype called ROSETTA. This system is based on Montague grammar, mainly in the sense that it incorporates a type of compositional semantics, the semantic rules working in tandem with the syntactic rules.

- Evaluation: The coverage of ROSETTA is still extremely limited; it is not clear how much the system can be extended to handle real-life texts,

- Approach: See the description of METAL below. The system will be coupled to TEAM, Siemens' terminological database, which with its 1.5 million concepts in up to 8 different languages is one of the largest in the world.

- Evaluation: See evaluation of METAL (Appendix I-15). Siemens' project appears quite ambitious. Although METAL features one of the best syntactic parsers currently available, massive infusion of vocabulary is likely to create semantic problems.

2) <u>Philips' Research Project, Philips Laboratories, Eindhoven</u>

- Project leader: Jan Landsbergen

- Funding: only one person is involved full time

- Research themes: The use of Montague grammar in MT

- Main ideas: Landsbergen is developing a small MT prototype called ROSETTA. This system is based on Montague grammar, mainly in the sense that it incorporates a type of compositional semantics, the semantic rules working in tandem with the syntactic rules.

- Evaluation: The coverage of ROSETTA is still extremely limited; it is not clear how much the system can be extended to handle real-life texts, and it is not clear what contribution its design principles can bring to solve the main linguistic problems of MT.

c) Japan

1) Fujitsu Corp. (ATLAS/U)

Fujitsu has already developed an operational MT system, ATLAS/I (see Appendix I), which is currently being tested. In addition, they are currently developing a more ambitious system: ATLAS/U, and we describe this project here.

- Laboratory: Software Research Division, Kawasaki City.

- Funding: The project started in 1979, and should be completed by 1987. In 1982, Fujitsu spent $600,000 on the development of ATLAS/U and the testing of ATLAS/I.

- Goals: Develop a multilingual MT system for computer manuals.

- State of the work: In 1982, a large-scale prototype (40,000 dictionary entries) for Japanese to English had been developed.

- Approach: Ideally, the input text would be analyzed into a language-independant conceptual structure, similar to that proposed by Schank (see below), so as to eliminate the need for transfer; but it seems that the current prototype still makes use of a transfer component.

2) Hitachi Corp. (ATHENE)

- Laboratory: Hitachi's Systems Development Laboratory, in Kawasaki City.

- Funding: The project started in 1980, and for 1982 the budget was approximately $880,000.

- Goals: Develop a system called ATHENE, for the translation of various types of text between English and Japanese.

- State of the work: In 1982, a prototype was running for English to Japanese translation of technical texts, with a vocabulary of 10,000 words.

- Approach: Transfer model; some semantic analysis; post-editing facilities.

3) NEC (TRAP)

- Laboratory: Computer System Research Department, Kawasaki City.

- Funding: R&D in MT began in 1980, in 1982 five researchers were involved, with a budget of approximately $250,000.

- Goals: Develop an English to Japanese translation system, called TRAP, for computer manuals.

- State of the work: the project is still in its initial stage and TRAP has not been demonstrated yet.

- Approach: TRAP will be based on a transfer model and written in PROLOG.

4) <u>Other companies</u>

Several other Japanese companies are reported to have R&D work going on in MT, about which we have so far very little information. Among these are the following:

- Toshiba: apparently, a group of 5 people is working on a system that will translate scientific and technical documents from English, French and Italian into Japanese.

- Mitsubishi: reportedly working on an English to Japanese system.

- KDD Laboratories: working on a English to Japanese system for technical papers.

- Joint System Development Corp.: Working on an English to Japanese system for computer manuals.

III - University Research Groups

A) United States

In the U.S., we know of only two university goups that are curently involved in MT. One of them is the Linguistic Research Center (LRC) of University of Texas in Austin. Since the present work of this group is oriented more toward development than reserch, we will not describe it here; refer to section on Siemens' development project (I-B-I) and to the description of the METAL system in Appendix II. The other group is Schank's AI group at Yale University:

The Yale AI Group

- Project Leader: Roger Schank.

- Research Themes: Applying knowledge representation theories (scripts, plans and, more recently, memory organization packets (MOPS)) to the solution of some translation problems.

- Funding: No specific source identified; the MT work is done on a very small scale, with one or two students.

- Main Ideas: This group claims that the transfer approach cannot work and that to solve translation problems correctly it is necessary to map natural language texts onto a deep and universal semantic representation. Moreover, they claim that syntax-based systems cannot work because syntax is largely irrelevant to meaning. Finally, mapping the text onto an adequate meaning representation requires extensive use of various types of knowledge of the world which can be organized in structures called MOPS (memory organization packets).

- Evaluation: Schank's work on MOPS (like the work of several others on representing world knowledge) certainly has long-term relevance to the solution of the MT problem. However, some of the theoretical claims that are made are very controversial. For example, very few people believe that syntax is such a minor part of NLP. The possiblity of building a universal interlingua has also been repeatedly challenged on theoretical grounds. Finally, it should be noted that the Yale approach has not been demonstrated on more than a few well-chosen examples.

B) <u>Europe</u>

1) Groupe d'etudes pour la traduction automatique (GETA), Université de Grenoble.

   - Project Leaders: Prof. Bernard Vauquois and Prof. Christian Boitêt.

   - Research themes: Software and linguistic models for MT.

   - Funding: Since its creation, around 1960, the group has been supported by the Centre national de la recherche scientifique with an average of four or five full-time researchers. In addition, over the years, the group has received funding from various agencies for the development of MT prototypes. For example the DRET (Department of Defence) has been supporting the development of a Russian to French system (four or five researchers) for several years. A significant amount of research work is accomplished by graduate students. Finally, the GETA is now involved in the ESOPE development project (see I-A-2).

- Main ideas: GETA has been a pioneer in the development of the basic principles underlying second-generation systems. More recently, they have shown a concern for making such systems usable in a production environment, resistance to errors and user-friendliness. The GETA has been very influential. Over the years they have had cooperations with groups in several countries, including Canada, Brasil, China, Japan and Malaysia.

- Evaluation: The GETA is currently one of the strongest centers of expertise in MT in the world, thanks to some innovative work and the consistency of their funding over the years. One problem with their approach may be that they have had a tendency to concentrate much effort on the mathematical properties of their programming tools and less effort on the search for more powerful NL understanding models.

2) <u>Sonderforschungbereich 100, Saarbrücken</u>

- Project Leader: There are several subgroups in the Saarbrucken project, and there are research aspects in the work of most of these groups. However, the subgroup headed by Barbara Sandig (subproject A3) seems to be the most committed to longer term research.

- Research Themes: Text grammars, text representation, transfer of noun phrases.

- Funding: Sandig's group has been funded for several years, at a level of approximately 250,000 Deutsch marks (about $116,000) per year.

- Main Ideas: The research is done in the context of a German to English and French MT system. The parser from the SUZY system, (cf. 5.4.2) is used, with additional components added to it. A. Rothkegel is developing a prototype component called TEXAN (text analysis) which carries out a textual analysis based on Van Dijk's text grammars. H. Weber is developing COAT (coherence analysis of text), a module which analyses its input in terms of a knowledge representation very similar to Schank's MOPS. Finally, M. Weissberger is working on NOTRA, a component that can transfer German noun phrases into English and French, on the basis of a semantic analysis.

- Other Subprojects: Subproject A2, directed by Prof. D. Maaf, is planning the development of a new MT system called SUZY-II.

  Subproject C, directed by Prof. H. Scheel, investigates problems in the analysis of French for MT purposes.

  Subproject E, directed by Prof. G. Hotz, concentrates on the development of linguistically oriented programming tools.

  Subproject K, directed by Prof. W. Wilss, works on the analysis of English in an imply context.

4) Other universities in Europe

   a) As mentioned in section I-A-1, several universities are involved in the EUROTRA project, and this work will not be described further here. Several of these groups have had no other activities in MT so far (Utrecht, Leuwen, Colchester, Torino and Athens). Three groups that are active both within EUROTRA and independtly have just been mentioned: Grenoble, Saarbrücken and Manchester. Besides these, the group in Nancy is not a newcomer in MT. Under the direction of Prof. Bourquin, three or four people have been involved (part time) in the development of prototye systems for several years. Most notably, they have developed, in cooperation with the Grenoble group, analysis and transfer modules for an English-to-French system.

b) At the University of Konstanz, Christa Hauenschild, who has been involved in the development of the SALAT system (see Appendix I), is pursuing research of a more theoretical nature on MT problems.

C) Japan

1) University of Kyoto, Nagao's group

- Project leaders: Prof. M. Nagao and his assistant, Prof. Tsujii are generally held to be the leading experts in MT in Japan.

- Funding: The group has been funded since 1962 by the Japanese government; but it really took a new start in 1977, when Prof. Nagao became the leader. In addition to money received for its involvement in external projects (e.g. the project of the Science and Technology Agency, see section I-B-2), Nagao's group had an internal budget of $56,000 in 1982.

- Main ideas: The group has so far developed at least two systems based on second-generation technology. More recently, they have been investigating a new prototype system for the translation (Japanese to English) of the titles and indexes of technical papers; this newer system is based on deterministic parsing.

2) Kyoto University, Doshita's group

- Project leader: Prof. S. Doshita

- Funding: In 1982, the project involved four part-time researchers and had a budget of $16,000.

- Research themes: The use of Montague Grammar in MT.

- Main ideas: A formalism derived from Montague grammar can become a practical and powerful tool for machine translation. In 1982, a small prototype, with a vocabulary of 700 words, had been developed.

3) University of Kyushu

- Project leaders: There are two small groups at Kyushu, one headed by Prof. T. Tamachi, the other by Prof. M. Yoshida.

- Funding: MT research at Kyushu started in 1955; it was interrupted after the ALPAC report, and started again after 1975. In 1982, the two small groups received together approximately $75,000.

- Research themes: MT systems for translation of technical papers between Japanese and English.

- Main ideas: Prof. Tamachi's group is developing a system called IMAGES-1, based on a transfer model. Prof. Yoshida and his group are concentrating their work on the lexical components of an MT system.

4) University of Osaka

- Project leader: Prof. F. Nishida.

- Funding: Development of MT systems started in 1980; no figures on funding level are available, but in 1982, six researchers were involved.

- Research themes: Translation of general texts from English into Japanese.

- Main ideas: a case structure can be constructed on Hornby's dictionary verb patterns; this case structure helps remove ambiguities that arise during parsing of the English input and during transfer into Japanese structures.

APPENDIX III

IMPLEMENTED SYSTEMS FOR NATURAL
LANGUAGE PROCESSING

APPENDIX III

IMPLEMENTED SYSTEMS FOR NATURAL LANGUAGE PROCESSING
(not including machine translation systems)

TEXT-CRITIQUING SYSTEM:

EPISTLE

Development: IBM - Watson Research Center
Yorktown Heights, NY

Status: advanced prototype for an application system; no plans announced for commercialization.

Capabilities: provides critical analysis of English text for spelling, grammar and style, highlighting errors or doubtful passages with suggestions for improvement.

Possible applications: development of business correspondence, teaching of composition.

Advantages: relatively broad coverage of English syntax and vocabulary (130,000 dictionary entries); ranking of ambiguities: 'fitted parse' procedure for guaranteeing analysis of fragments, rare structures, etc.; tested on 2254 sentences from actual business letters.

Limitations: relatively shallow syntactic analysis precludes critiquing clarity or logical organization in a general way.

Software: LISP/370 for IBM machines; bottom-up parser.

SYSTEMS FOR SEMANTIC PROCESSING OF EXTENDED TEXT (WITHOUT DIALOG):

NOMAD

Development: AI Project, Department of Information and Computer Science, University of California at Irvine

Status: experimental prototype.

Capabilities: provides analysis and paraphrase of telegraphic, errorful naval messages into normal disambiguated English; relies on knowledge of domain to correct spelling and resolve ambiguities, ellipsis, missing sentence boundaries, etc.

Possible applications: general approach applicable to a variety of telegraphic messages (including weather reports) for which domain semantics can be described.

Advantages: extensive use of domain knowledge allows recovery from errors or ellipses which normally require human expertise.

Limitations: appears to have very partial coverage of the relevant vocabulary and domain semantics; it remains to be seen whether complications will arise when system is scaled up to operational size.

NRL - CASREP

Development: Naval Research Laboratories
Bolling AF Base
Washington, DC.

Status: experimental prototype.

Capabilities: produces single-sentence summaries of telegraphic reports regarding equipment failures aboard U.S. Navy ships at sea.

Possible applications: approach may be useful for summarizing reports in other highly restricted sublanguages.

Advantages: methodology extendible to other domains; based on large corpus of actual ship-to-shore messages.

Limitations: does not include inferencing mechanism for solving deep ambiguities.

FRUMP

Development: Yale University
AI Project
New Haven, CT

Status: experimental prototype (no longer being developed).

Capabilities: produces summaries of news stories for about 10% of stories on UPI wire service, particularly those dealing with plane crashes, wars, forest fires and diplomatic relations.

Possible applications: scanning certain reports for interesting content or for summarizing when key elements are relatively predictable; intelligence monitoring.

Advantages: relatively robust (can extract useful information for summaries even when cannot 'understand' all details; relatively fast (20 sec CPU time per story on DEC 20).

Limitations: method so far appears difficult to scale up to large domains or complex subject matters; may stumble on texts because of limited syntactic information; may misinterpret some texts.

INTELLECT

Development: Artificial Intelligence Corporation
Waltham, MA

Status: commercial; marketed since 1981 under several names (e.g., as OnLineEnglish by Cullinet Software, as GRS Executive by Informaton Science Inc., as part of Express system by Management Decision Systems and as part of Honeywell's Multics environment).

Capabilities: converts English queries regarding a particular database into commands to database management system; converts database replies into English when necessary; handles many elliptica queries and some usages of pronouns; constructs tables and graphs as reply to certain varieties of questions (e.g., "Show me..."); compatible with several database management systems.

Advantages: compatible with wide variety of database systems; distributed by IBM under sales agreement; first commercial query system.

Limitation: must be hand-tailored to each new domain and its corresponding vocabulary and word usages; high price ($70,000) makes it accessible only to larger corporations; gives literal replies; limited ability to handle quantified expressions.

TQA (transformational question-answering)

Development: IBM Watson Research Center
Yorktown Heights, NY

Status: pre-commercial prototype; no announced plans for commercialization; planned in-house (sales office) use in fall 1983.

Capabilities: parses replies to English queries to a database on zoning and land use.

Advantages: underwent a year of testing by naive users (in White Plains municipal office) for which performance statistics have been published (AJCL, v&,no.1); relatively large grammatical coverage.

Limitations: apparently has no performance capabilities beyond those new included in INTELLECT; poorly transportable to new domains.

USL

Development: IBM Scientific Center
Heidelberg, West Germany

Status: advanced prototype; no plans announced for commercialization.

Capabilities: provides answers to queries in English or German (using 2 separate versions of the system) to a database of geopolitical information; resolves pronominal reference within same sentence; allows adding individual facts to database using natural language.

Advantages: only large prototype which exists for more than one language; uses relational model of language; semantics organized on solid logical foundations, allowing for interesting extensions.

Limitations: limited treatment of pronouns and sentence fragments; tested in only one domain; no facility of rapidly transporting to a new domain; some grammatical limitations on handling quantified expressions.

ASK (A Simple Knowledgable system)

Development: California Institute of Technology under contract with Hewlett Packard Corporation.

Status: pre-commercial prototype; planned for commercialization, but no release date set.

Capabilities: provides relatively fast analysis and response to English queries to databases; designed for user who provides own knowledge base; fairly transportable to new domains; allows user definitions of words; uses semantic network with inheritance of attributes, etc.; allows creation of 'context' for individual user, based on larger database which is customized for individual through a system-supported dialog.

Advantages: relatively independent of domain; one of fastest pre-commercial systems; runs on desktop computer.

Limitations: moderate coverage of English grammatical constructions; less transportable to entirely new domain than to new users within same general data environment.

Implementation: runs on HP9836 desktop computer; programs written in PASCAL; needs hard disk for larger contexts.

TEAM (Transportable English Access interface Manager)

Development: SRI International
Artificial Intelligence Center
Menlo Park, CA

Status: advanced experimental prototype.

Capabilities: produces a lexicon and pertinent grammatical information required for querying a new database by means of a system supported dialog during which a database manager not familiar with computational linguistics answers general questions about database categories and word usage in the domain; effectively creates the domain-dependent component of a natural language interface during a single dialog, replacing the hand-tailoring process which usually requires several man-weeks.

Advantages: first general (transportable) interface to databases for a natural language; tested on a number of new domains.

Limitations: some limitations in range of database types handled; verb types limited to transitive verbs.

SYSTEMS FOR GENERATION (SYNTHESIS) OF TEXT FROM DATA

ANA

Development: University of Pittsburgh
Department of Information Science
Pittsburgh, PA

Status: experimental prototype (individual Ph.D. thesis project).

Capabilities: produces short reports on activity of the New York Stock Exchange in stereotyped style based on semi-hourly quotations of individual stocks and NYSE composite (Dow Jones).

Possible applications: production of stereotyped reports on stocks and other commercial exchanges.

Advantages: tailoring of linguistic component to domain improves efficiency.

Limitations: linguistic approach is highly dependent on word usage of domain (must be completely redone for new domain); does not incorporate knowledge structures for the domain in explicit form.

MUMBLE

Development: MIT AI Laboratory (begun as individual thesis project) & University of Massachusetts at Amherst.

Status: experimental prototype.

Capabilities: produces short coherent texts in English (written or spoken) which refer to two-dimensional scenes.

Advantages: input to system has been tested in a variety of knowledge representation systems including predicate calculus, KL-ONE, FRL, OWL and other; only system to include spoken component with large grammar and sophisticated facilities for producing pronouns, ellipses, etc.

Limitations: some problems with thematic constraints are evident.

TEXT

Development: University of Pennsylvania
Department of Computer and Information Science
Phildelphia, PA

Status: experimental prototype.

Capabilities: provides paragraph-length text as responses to three types of queries about database structure: (1) questions about information available in the database, (2) requests for definitions, (3) questions about the differences between database entities.

Advantages: incorporates principles of rhetoric and considerable linguistic information into strategies for planning and executing utterances; most advanced system for textual responses to database queries.

Limitations: depends on relational database structure.

APPENDIX IV

BIBLIOGRAPHY

BIBLIOGRAPHY

Advisory Committee on Artificial Intelligence for the Ministry of Education and Sciences, Artificial Intelligence in the Netherlands, March 1983

ALPAC, Languages and Machines: Computers in Translation and Linguistics, Report by the Automatic Language Processing Advisory Committee, National Academy of Sciences, Washington, 1966

Alvey, J. et al., A Programme for Advanced Information Technology, The Report of the Alvey Committee, HMSO, 1982

Appelt, D., TELEGRAM, A Grammar Formalism for Language Planning. Proc. of the 21st Annual Meeting of the Association for Computational Linguistics. (June, 1983)

Arden, Bruce W. (ed.), What Can Be Automated, Cambridge: MIT Press, 1980

Artificial Intelligence Corporation, INTELLECT User's Manual. Waltham, Mass., (1981) Bahl, L.R. et al., "Recognition of Isolated-Word Sentences from a 5,000-word Vocabulary Office Correspondence Task", Proc. IEEE Int. Conf. Acoustics, Speech and Sig. Proc. ICASSP' 83, 1065-1067, 1983

Bar-Hillel, Y., "The Present Status of Automatic Translation of Languages", Advances in Computers, 11 (1960), 91-163

Barr, A. and Feigenbaum, E. (Eds.), The Handbook of Artificial Intelligence, Vols. 1&2, William Kaufman: Los Altos, 1981

Bates, M. and Wilson, K., ILIAD: Interactive Language Instruction Assitance for the Deaf, Bolt, Beranek and Newman, TR 4771, 1981, 74p.

Bornat, R. and Brady, J.M., "Using knowledge in the Interpretation of Handwritten FORTRAN coding sheets", Int. J. Man-Machine Studies, Vol. 8, pp. 13-27, 1976

Brinkmann, K.-H., Perspectives d'avenir de la traduction automatique, in META, 24:3, 1979.

Campbell, A. N., Hollister, V. F., Duda, R. O. and Hart, P. E., "Recognition of a hidden mineral deposit by an artificial intelligence program", Science, 217 (Sept. 1982) 927-929

Carbonell, J., Cullingford, R. and Gershman, A., "Steps Toward Knowledge-Based Machine Translation", IEEE Transactions, PAMI-3:4 (July 1981), 376-392.

Chalmers, J. and Stein, A., "Iterative array for the preprocessing of handprinted characters", Electron Lett., Vol. 5., pp. 431-433, Sept., 1969

Chevalier, M., Dansereau, J., Poulin, G., TAUM-METEO: description du système, Groupe TAUM, Université de Montréal, 1978, 47p.

Chiu, L. Y. and Chen, Z., "On an invariant feature selection for numeric characters", Proc. Int. Computer. Symp., Vol. 1, pp. 644-662, Dec. 1978

Clancey, W. "Tutoring rules for guiding a case method dialog." International Journal of Man-machine Studies, 11 (1979), 25-49.

Clancey, W. and Letsinger, R. NEOMYCIN: reconfiguring a rule-based expert system for application to teaching. In: Proc. IJCAI-81, 829-836

CNRS, Version Préliminaire Rapport Sico CNRS, April 8, 1983.

Cohen, P., The Handbook of Artificial Intelligence, Vol. 3, William Kaufman: Los Altos, 1982.

Cohen, P. and Liebermann, A report on FOLIO: an expert system for portfolio managers. In Proc. IJCAI-83. (Aug. 1983) 212-214.

Commission of the European Communities, Communication from the Commission to the Council on Laying the Foundations for a European Strategic Programme of Research and Development in Formation Technology "ESPRIT": The Pilot Phase, Brussels, August 1982.

Commission of the European Communities, FAST, The Fast Programme, Volume 1, Results and Recommendations, Brussels, December 1982.

Commission of the European Communities, FAST, The Fast Programme, Volume 2, Summaries of the 36 Research Projects, Brussels, December 1982.

Damereau, F. "Operating Statistics for the Transformational Question Answering System." American Journal of Computational Linguistics 7:1, (1981) 30-42.

Davis, R. and Lenat, D. Knowledge-based Systems in Artificial Intelligence. McGraw-Hill, New York, 1982.

DeJong, G. Skimming Stories in Real Time: an Experiment in Integrated Understanding. Res. Rep. No 158, Yale Computer Science Department, (1979).

Duda, R., Gaschnig, J. and Hart, P. Model design in the PROSPECTOR consultant system for mineral exploration. In Expert Systems in the Microelectronic Age, D. Michie (ed), Edinborough University Press, Edinborough, 1979.

Dusek, L., Schalk, T.B. and McMahan, M. "Voice Recognition Joins Speech on Programmable Board", Electronics, v. 56, 128-132, April 21, 1983.

Economist, 16 September 1983, p. 77.

Fahlman, S. E., NETL: A System for Representing and Using real World Knowledge, Cambridge, Massachusetts: M.I.T. Press 1979.

Feigenbaum, E. and McCorduck, P., The Fifth Generation, Addison-Wesley, Reading, 1983.

Gevarter, W., An Overview of Expert Systems, U.S. Department of Commerce, Report NBSIR 82-2505, 1982.

Goldberg, A. and Robson, D., Smalltalk-80: The Language and its Implementation, Addison-Wesley, Reading, 1983.

Gould, J. D., Conti, J. and Hovanyecz, T., "Composing letters with a simulted listening typewriter." Comm. A.C.M., 26, 1983, 295-308.

Granger, R. et al. Scruffy Text Understanding: Design and Implementation of the NOMAD System. Proc. of the Conf. on Applied Language Processing, Santa Monica, (1983).

Grishman, R., and Hirschman, L. "Question Answering from Natural Language Data Bases". Artificial Intelligence 11:25-43, (1978).

Grosz, B. TEAM: Transportable Natural Language Interface System. Proc. of the Conf. on Applied Natural Language Processing, Santa Monica, (1983).

Harris, L. "The ROBOT Natural Language Interface." Comm. of the ACM, (1978).

Hart, Peter E., "Directions for AI in the Eighties" SIGART Newsletter - ACM #79, 11-16, January 1983.

Hauenschild, C., Huckert, E., Maier, R., SALAT: Machine Translation via Semantic Representation, paper presented at the Konstanz Collogium on Semantics, (Sept. 1978), 29 pp.

Hayes-Roth, R., Waterman, D. and Lenat, D., Building Expert Systems, Addison-Wesley, 1983.

Heidorn, G. et al. "The EPISTLE text-critiquing system", IBM Systems Journal, 21:3 (1982) 305-326.

Hendrix, G. et al. "Natural Language Interface." American Journal of Computational Linguistics, 8:2 (1982) 56-61.

Hewitt, C., "Viewing Control Structures as Patterns of Passing Messages", AI Journal, Vol. 8, 1981, 323-364.

Himmel, S., "Some real-world experience with handprinted optical character recognition". IEEE Trans. Supt. Man. Cyber., Vol. 8, pp. 288-292, Apr. 1978.

Hirschman, K, and Sager, N., Automatic Informatting of a Medical Sublanguage. In Kittredge, R. an Lehrberger, J. (eds.) Sublanguage: Studies of Language in Restricted Semantic Domains, de Gruyter, (1982).

Hundt, M., Working with the Weidner Machine Translation System, in V. Lawson (ed) Practical Experience of Machine Translation, North-Holland, 1982, 45-52.

Hunt, M. J., Lennig, M., and Mermelstein, P., "Experiments in Syllable-based Recognition of Continuous Speech", IEEE Int. Conf. Acoustics, Speech and Sig. Proc., ICASSP80, Denver, Col., 880-883.

Hutchins, W., The Evolution of Machine Translation Systems in V. Lawson (ed) Practical Experience of Machine Tranlation, North-Holland, 1982, 21- 37.

Isabelle, P., Bourbeau, L., Chevalier, M., Lepage, S., TAUM-AVIATION: description d'un système de traduction automatisée de maneuls d'entretien en aéronautique, COLING-78, Bergen, 1978, 29 pp.

Isabelle, P., Some Views on the Future of the TAUM Group and the TAUM-AVIATION System, Groupe TAUM, Université de Montréal, 1981, 35 pp.

JEIDA, The Present Status of Document Translation in Japan, Japan Electronic Industry Development Association, 1982.

Joshi, A. et al. Modelling Interactive Processes: Flexible Interaction with Knowledge Bases in Three Dimensions (a proposal to the NSF), University of Pennsylvania, (1982).

Kay, Martin, (Chairperson), "Machine Translation - Report of a Panel", American Journal of Computational Linguistics, Vol. 8. No. 2, April-June 1982, 74-78.

Kay, M., The MIND System, in R. Rustin (ed.) Natural Language Processing, New-York, Algorithmic Press, 1973, 155-188.

Kay, M., The Proper Place of Men and Machines in Language Translation, ms., Xerox/PARC, 1980, 20 pp.

Kittredge, R., Sublanguage-Specific Computer Aids to Translation: A Survey of the Most Promising Application Areas. Report to Secretary of State Department, Government of Canada, 1983, 95 pp.

Kittredge, R. "Semantic Processing of Texts in Restricted Sublanguages." Computers and Mathematics with Applications, 9:1 (1983).

Kittredge, R. and Mel'cuk, I. Towards a Computable Model of Meaning-Text Relations within a Natural Sublanguage, Proc. of the 8th IJCAI, (1983).

Kukich, K., Design of a Knowledge-Based Report Generator. Proc. of the 21st Annual Meeting of the Association for Computational Linguistics, (1983).

Kulikowski, C. and Weiss, S., Representation of expert knowledge for consultation. In: P. Solovits (ed) Artificial Intelligence in Medicine, Westview Press, 1982.

Langley, P., Zytkow, J., Bradshaw, G., and Simon, H., Three facits of scientific discovery. In: Proc IJCAI-83, (Aug. 1983) 465-468.

Lenat, D., Sutherland, W. and Gibbons, J. Heuristic search for new microcircuit structures: an application of artificial intelligence. The AI Magazine, 32 (Summer 1982) 17-33.

Leun, C. Y., "Advance in optical character recognition" Proc. Canadian Computer Conf., pp. 263-268, May 1978.

Levinson, S.E., and Roseberg, A.E., "Some Experiments with a Syntax Directed Speech Recognition System", Proc. Int. Conf. Acoust. Speech and Sig. Proc., Tulsa, OK, 700-703, 1978.

Lindsay, R., Buchanan, B., Feigenbaum, E. and Lederberg, J., Applications of AI for Organic Chemistry: The DENDRAL Project. McGraw-Hill, New York, 1981.

Loh, S. et al., Machine Translation of Chinese Mathematical Articles, ALC Bulletin, 6 (1978), 111-121.

Lusk, E.,McCume, W., and Overbeek, R., "Logic Machine Architecture: Kernel Functions", In: Proceedings of the Sixth Conference on Automated Deduction, Lecture Notes in Computer Science, 138, Springer-Verlag, 1982.

Mann, W. and Matthiessen, C., NIGEL: A Systemic Grammar for Text Generation, Information Sciences Institute, University of Southern California, (1983).

Manning, E. and Wright, A Trip Report to Japan, University of Waterloo, Summer 1983.

Marr, D., "Artificial Intelligence - A Personal View", Artificial Intelligence, 9 (1977), 37-48.

McCarty, L. and Sriharan, N. A computational theory of legal argument. Technicl Report LRP-TR-13, Laboratory for Computer Science Research, Rutgers University, (1982).

McDermott, J., "R1: A rule-based configurator of computer systems." Artificial Intelligence 19 1 (Sept. 1982).

McDonald, D. Salience as a Simplifying Metaphor for Natural Language Generation. Proc. of the 2nd AAAI Conference, (1982).

McKeown, K. "Generating Natural Language Text in Response to Questions about Database Structure" (dissertation), Univ. of Pa., (1982).

Melby, A., "The Translation Profession and the Computer", Calico Journal, 1:1 (1983), 55-59.

Mermelstein, P. and Eden, M., "A System for automatic recognition of handwritten words." Proc. Fall Joint Computer Conf., pp. 333-342, 1964.

Montgomery, C., "An Active Information System for Intelligence Analysis". Signal Magazine, (1981).

National Physical Laboratory, Symposium No. 10, Mechanisation of Thought Processes, Vol. 1&2, Her Majesty's Stationery Office: London, England, 1959.

National Physical Laboratory, Symposium No. 13, 1961 International Conference On Machine Translation of Languages and Applied Language Analysis, Vol. 1&2, Her Majesty's Stationery Office: London, England, 1962.

Nelson, W., REACTOR: An expert system for diagnosis and treatment of nuclear reactor accidents. Proceedings of the Second Conference of The American Association for Artificial Intelligence, (Aug. 1982) 296-301.

Newell, A. et al., Speech Understanding Systems, North Holland: Amsterdam, 1973.

Nilsson, Nils J., Principles of Artificial Intelligence, Tioga Publishing Co: Palo Alto, California, 1980.

Novak, Gordon S., Jr., "GLISP", The AI Magazine, Vol. IV, No. 3, Fall, 1983, 37-47.

Pazzani, M. and Engelman, C. Knowledge-Based Question Answering. Proc. of the Conf. on Applied Natural Language Processing, Santa Monica, (1983).

Pople, H., "Heuristic methods for imposing structure on ill-structured problems: the structuring of medical diagnostics." In: Artificial Intelligence in Medicine, Westview Press, 1981.

Sager, N., Natural Language Information Processing, Addison-Wesley, Reading, Mass., 1981, 399 pp.

Schmolze, J. and Brachman, R., Proceedings of the 1981 KL-ONE Workshop. Fairchild Technical Report 618, Laboratory for Artificial Intelligence Research, Fairchild Corp.

Shortliffe, E., Computer-Based Medical Consultations: MYCIN, Elsevier, New York, 1976.

Sidner, C. L. and Bates, M., Requirements for Natural Language Understanding in a System with Graphic Displays. BBN Technical Report No. 5242, March 1983.

Skuce, D., The LESK Tutorial. Technical Report TR-83-1, Department of Computer Science, University of Ottawa (1983).

Sleeman, D. and Brown, J., (eds.), Intelligent Tutoring Systems. Academic Press, New York, 1982.

Smith, R. and Baker, J., The dipmeter advisor: A case study in expert system commercial development. In Proc. IJCAI-83 (Aug. 1983) 122-129.

Solovits, P., (ed.), Artificial Intelligence in Medicine, Westview Press, 1982.

Soper, J., ANGE: a French Tutor, ms., Monterey Institute of International Studies, 5 pp.

Soper, J., Personal Computer-Assisted Foreign Language Instruction: an Innovative Linguistic Intelligence Approach, ms., Monterey Institute of International Studies, 17 pp.

Stallman, R.M. and Sussman, G.J., Problem solving about electrical circuits. In: Artificial Intelligence: an MIT Perspective, Winston, P. and Brown, R. (eds.) The MIT Press, New York, 1979.

Stefik, M. and Conway, L., "Towards the principled engineering of knowledge." The AI Magazine 2 3 (Summer 1982) 4-16.

Stefik, Mark, Borrow, Daniel G., Mital, Sanjay and Conway, Lynn, "Knowledge Programming in Loops", The AI Magazine, Vol. IV, No. 3, Fall, 1983, 3-13.

Suen, C. Y. and De Mori, R., "Computer Analysis and Perception, Vol, II, Auditory Signals", CRC Press, Boca Raton, FLA., 1982.

Suppes, P. (ed) University-level computer assisted instruction at Stanford: 1968-1980. Stanford University Institute for Mathematical Studies in the Social Sciences. (1981).

Swartout, W. Explaining and Justifying Expert Consulting Programs. Proc. of the 7th IJCAI. (1981).

Thompson, B. and Thompson, F. Introducing ASK, a Simple Knowledgable System. Proc. of the Conf. on Applied Natural Language Processing, Santa Monica, (1983).

Tsotsos, J. Temporal event recognition: an application to left ventricular performance. In: Proc. IJCAI-81, 900-905.

Van Slype, G., Evaluation de la qualite de la traduction automatique, Bureau Marcel Van Dijk, Bruxelles, 1979, 11 pp.

Van Slype, G., Guinet, J., Seitz, F., Benejam, E., Mieux traduire pour mieux communiquer, Infotecture pour la Commission des Communautés européennes, Luxembourg, 1981, 174 pp.

Waltz, David et al, "Artificial Intelligence: Assessment and Recommendations, The AI Magazine, Vol. IV, No. 3, Fall, 1983, 55-67.

Weaver, W., Translation, in W., Locke and A. Booth (eds.), Machine Translation of Languages, New-York, Wiley, 1955, 15-23.

Wilensky, R. Talking to UNIX in English: an Overview of UC. Proc. of the 2nd AAAI Conf., (1982).

Winograd, T., Understanding Natural Language, Academic Press, New York, 1972.

Winston, Patrick Henry, Artificial Intelligence, Second Edition, Addison-Wesley Publishing Company: Reading, Massachusetts, 1983.

Woods, W., Transition Network Grammars for Natural Language Analysis, Comm. ACM 13:10 (1970), 591-606.

Woods, W., Kaplan, R., Nash-Webber, B., The Lunar Sciences Natural Information System: Final Report, Bolt, Beranek and Newman, TR 2378, Cambridge, Mass., 1972.

APPENDIX V

ACRONYM LIST

AAAI	American Association for Artificial Intelligence
ACM	Association for Computing Machinery
ADI	Agence pour le Development de l'Informatique (France)
AFOSR	Air Force Office of Scientific Research (US)
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
ALPS	A machine translation system
AM	An expert system (physics)
ARIANE-78	A machine translation system
ARO	Army Research Office (USA)
ART	Advanced Reasoning Tool
ASCII	American Standard Code for Information Interchange
ASK	'A Simple Knowledgeable System'
ATN	Augmented Transition Network
BACON	An expert system (physics)
BBN	Bolt, Beranek and Newman, Inc. (USA)
BUGGY	An educational expert system (basic arithmetics)
CAI	Computer Aided Instruction
CALI	Computer Aided Language Instruction
CMU	Carnegie-Mellon University
CNRS	Centre National de Recherche Scientifique (France)
COAT	COherence Analysis of Text
COGNET	A computer network (cognitive sciences)
CSLI	Centre for Study of Language and Information

CULT	A machine translation system (math, Chinese to English)
DARPA	Defense Advanced Research Projects Agency (USA)
DCIEM	Canadian Department of National Defense (p. 60)
DEC	Digital Equipment Corporation (USA)
DENDRAL	An expert system (chemistry)
DOD	Department of Defense (USA)
ECU	European Currency Unit
EDUNET	A computer network (educational institutions)
EEC	European Economic Community
EPISTLE	a text-critiquing system at IBM - Yorktown Heights
ES	Expert System
ESOPE	French Machine Translation Project
ESPRIT	European Project for Collaboration in Information Technology
EURISKO	An expert system (mathematics)
EUROTRA	EUROpean TRAnslation (multinational MT Project of EEC)
EXCHECK	EXpert CHECKing (an educational expert system)
FAHQTUT	Fully Automatic High Quality Translation of Unrestricted Text
FAST	Forecasting and Assessment of Science and Technology
GETA	Groupe d'etude pour les Traduction Automatique (France)

GLISP	A high level programming language
GUIDON	An educational expert system (medicine)
HAMT	Human Aided Machine Translation
HARPY	A speach recognition system
HEARSAY	A speach recognition system
IBM	International Business Machines, Corporation
ICAI	Intelligent Computer Aided Instruction
ICOT	Institute for New Generation Computer Technology (Japanese acronym)
IEEE	Institute of Electrical and Electronics Engineers
IJCAI	International Joint Conference on Artificial Intelligence
IKBS	Intelligent Knowledge Based Systems
ILIAD	Interactive Language Instruction Assistance for the Deaf
INTERLISP	A high level programming language
INTERNIST	A medical expert system
ISSCO	Liaison committee of European national groups
ITS	An interactive machine translation system
KES	Knowledge Engineering System (expert system development package)
KL-ONE	A knowledge representation language
KNOBS	KNOwledge Based System
KRL	Knowledge Representation Language

LISP	LISt Processing language
LOOPS	An expert system development tool
LUNAR	A natural language processing system
MAHT	Machine Aided Human Translation
MCC	Microelectronics and Computer Corporation
METADENDRAL	An expert system (chemistry)
METAL	A machine translation system
METEO	system for English to French translation of weather bulletins
MIND	An interactive machine translation system
MIT	Massachusetts Institute of Technology
MITI	Ministry of International Trade and Industry (Japan)
MODULA	A high level programming language
MOLGEN	An expert system (chemistry)
MOPS	Memory Organization Packets
MT	Machine Translation
MYCIN	A medical expert system
NCR	National Cash Register
NEC	Nippon Electric Company (Japan)
NEOMYCIN	A medical expert system
NETL	A crossbar network
NL	Natural Language
NLP	Natural Language Processing
NOMAD	A natural language processing system

NRL	Naval Research Laboratories
NSERC	National Sciences and Engineering Research Council (of Canada)
NSF	National Science Foundation (USA)
NYU	New York University
OCR	Optical Character Recognition
OCRA	an OCR font
OCRB	an OCR font
ONR	Office of Naval Research (USA)
OPS5	Production System Language
PARC	(XEROX) Palo Alto Research Center
PC	Personal Computer
PIPS	Pattern Information Processing System (Japan)
PLATO	Computer-assisted learning language (Control Data Corporation)
PROLOG	PROgramming in LOGic (language)
PROSPECTOR	An expert system (geophysics)
R&D	Research and Development
SALAT	A machine translation system
SCHEME	An object-oriented message passing language
SHRDLU	A natural language processing system
SIGART	Special Interest Group on ARTificial intelligence
SL	Source Language
SMART	A machine translation system