Japanese Technology Evaluation Center

JTEC

JTEC Panel Report on

Machine Translation In Japan

Jaime Carbonell, Chair
Elaine Rich, Co-Chair
David Johnson
Masaru Tomita
Muriel Vasconcellos
Yorick Wilks

January 1992

Coordinated by

Loyola College in Maryland
4501 North Charles Street
Baltimore, Maryland 21210-2699

JAPANESE TECHNOLOGY EVALUATION CENTER

SPONSOR The Japanese Technology Evaluation Center (JTEC) is operated for the Federal Government by Loyola College to provide assessments of Japanese research and development (R&D) in selected technologies. The National Science Foundation (NSF) is the lead support agency. Other sponsors include the National Aeronautics and Space Administration (NASA), the Department of Commerce (DOC), the Department of Energy (DOE), the Office of Naval Research (ONR), the Defense Advanced Research Projects Agency (DARPA), and the U.S. Air Force.

PURPOSE JTEC assessments contribute to more balanced technology transfer between Japan and the United States. The Japanese excel at acquisition and perfection of foreign technologies, but the U.S. has relatively little experience with this process. As the Japanese become leaders in research in targeted technologies, it is essential that the United States have access to the results. JTEC provides the important first step in this process by alerting U.S. researchers to Japanese accomplishments. JTEC findings can also be helpful in formulating governmental research and trade policies.

APPROACH The assessments are performed by panels of about six U.S technical experts in each area. Panel members are leading authorities in the field, technically active, and knowledgeable about Japanese and U.S. research programs. Each panelist spends about one month of effort reviewing literature, making assessments, and writing reports on a part-time basis over a twelve-month period. All recent panels have conducted extensive tours of Japanese laboratories. To provide a balanced perspective, panelists are selected from industry, academia, and government.

ASSESSMENTS The focus of the assessments is on the status and long-term direction of Japanese R&D efforts relative to those in the United States. Other important aspects include the evolution of the technology, key Japanese researchers and R&D organizations, and funding sources.

REPORTS The panel findings are presented to workshops where invited participants critique the preliminary results. Final reports are distributed by the National Technical Information Service (NTIS), 5285 Port Royal Road, Springfield, Virginia 22161 (703-487-4650). The panelists also present technical findings in papers and books. All results are unclassified and public.

STAFF The Loyola College JTEC staff members help select topics to be assessed, recruit experts as panelists, organize and coordinate panel activities, provide literature support, organize tours of Japanese labs, assist in the preparation of workshop presentations and in the preparation of reports, and provide general administrative support. Mr. Cecil Uyehara of Uyehara International Associates provided literature support and advance work for this panel.

Dr. Duane Shelton
Principal Investigator
Loyola College
Baltimore, MD 21210

Mr. Geoff Holdridge
Director
JTEC/Loyola College
Baltimore, MD 21210

Dr. George Gamota
Senior Advisor to JTEC
Mitre Corporation
Bedford, MA 01730

LOYOLA
COLLEGE
IN MARYLAND

February 13, 1992

Dear Colleague:

This is a letter of transmittal for the JTEC report on machine translation. Additional copies will be available from the National Technical Information Service (NTIS) of the U.S. Department of Commerce as PB92-100239, but let me know if there is someone who should get one right away.

The report has an acknowledgement section, but I would like to add my thanks to all of the sponsors of this report: Joseph Clark, Deputy Director of NTIS; Phyllis Genther of the Department of Commerce's Japanese Technical Literature Office; Paul Herer, Emily Rudin, and Y.T. Chien at NSF; Charles Wayne at DARPA; and Rob Billingsley of the Defense Technical Information Center. Special thanks are due to Dr. Clark for conceiving of the project and organizing its sponsorship. We also appreciate Jaime Carbonell's work in organizing the workshop presentations, and are indebted to Elaine Rich and her staff at MCC for their work in coordinating and putting the report together. And, of course, I want to express our appreciation to all of the members of the JTEC panel on machine translation for their efforts, particularly their patience and perseverance in responding to the many comments we received on the draft report from our Japanese hosts, from other panel members, and from our very capable editor, Arnett Holloway.

Most of all, I thank the Japanese who generously hosted the panel's site visits in Japan. Professor Makoto Nagao greatly assisted in organizing these visits. JTEC panels have always been received with great hospitality and openness by our Japanese colleagues. Without their spirit of open scientific exchange and their hospitality, this study would not have been possible.

Sincerely,

Geoffrey M. Holdridge
Director
Japanese Technology Evaluation Center

4501 North Charles Street
Baltimore, Maryland 21210-2699
301-323-1010

JTEC Panel on

Machine Translation in Japan

FINAL REPORT

January, 1992

Jaime Carbonell, Chair
Elaine Rich, Co-Chair
David Johnson
Masaru Tomita
Muriel Vasconcellos
Yorick Wilks

This document was sponsored by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency, and the United States Department of Commerce, under NSF Grant ECS-8922947, awarded to the Japanese Technology Evaluation Center at Loyola College in Maryland. The Government has certain rights in this material. The views expressed herein are solely those of the authors and do not necessarily reflect those of the United States Government, the authors' parent institutions, or Loyola College.

JTEC/WTEC STAFF

R. D. Shelton, Principal Investigator

Geoffrey M. Holdridge, JTEC Director

Michael DeHaemer, WTEC Director

Bobby A. Williams, Assistant Director

Aminah Batta, Administrative Assistant

Catrina Foley, Office Assistant

Arnett Holloway, Editor

Table of Contents

List of Figures

Foreword

This report is one in a series of reports prepared through the Japanese Technology Evaluation Center (JTEC), sponsored by the National Science Foundation (NSF) and administered by Loyola college in Maryland. The report describes ongoing research and development efforts in Japan in machine translation, the automated translation of text between different languages.

Over the past decade, the United States' competitive position in world markets for high technology products appears to have eroded substantially. As U.S. technological leadership is challenged, many government and private organizations seek to set policies that will help maintain U.S. competitive strengths. To do this effectively requires an understanding of the relative position of the United States and its competitors. Indeed, whether our goal is competition or cooperation, we must improve our access to the scientific and technical information in other countries.

Although many U.S. organizations support substantial data gathering and analysis directed at other nations, the government and privately sponsored studies that are in the public domain tend to be "input" studies. That is they provide measurement of expenditures, personnel data, and facilities, but do not provide an assessment of the quality or quantity of the outputs obtained. Studies of the outputs of the research and development process are more difficult to perform since they require a subjective analysis by individuals who are experts in the relevant technical fields.

The National Science Foundation staff includes professionals with expertise over a wide range of technologies. These individuals provide the technical expertise needed to assemble panels of experts who can perform competent, unbiased, scientific and technical reviews of research and development activities. Further, a principal activity of the Foundation is the review and selection for funding of research proposals. Thus the Foundation has both experience and credibility in this process. The JTEC activity builds on this capability.

Specific technologies, such as machine translation, or telecommunications, or biotechnology, are selected for study by individuals in government agencies who are able to contribute to the funding of the study. A typical assessment is sponsored by two or more agencies. In cooperation with the sponsoring agencies, NSF selects a panel of experts who will conduct the study. Administrative oversight of the panel is provided by Loyola College in Maryland, which operates JTEC under an NSF grant.

Panelists are selected for their expertise in specific areas of technology and their broad knowledge of research and development in both the United States and in the countries that are of interest. Of great importance is the ability of panelists to produce a comprehensive, informed and unbiased report. Most panelists have travelled previously to the host countries or had professional association with their expert counterparts. Nonetheless, as part of the assessment, the panel as a whole travels to host countries to spend one full week, as a minimum, visiting research and development sites and meeting with researchers. These trips have proven to be highly informative, and the panelists have been given broad access to both researchers and facilities. Upon completion of its trip, the panel conducts a one-day workshop to present its findings. Following the workshop, the panel completes a written report that is intended for widespread distribution.

Study results are widely distributed. Representatives of the host countries and members of the media

are invited to attend the workshops. Final reports are made available through the National Technical Information Service (NTIS). Further publication of results is encouraged in the professional society journals and magazines. Articles derived from earlier JTEC studies have appeared in *Science, IEEE Spectrum, Chemical and Engineering News*, and others. Additional distribution media, including video tapes, are being tested.

Over the years, the assessment reports have provided input into the policy-making process of many agencies and organizations. A sizable number of the reports are used by foreign governments and corporations. Indeed, the Japanese have used JTEC reports to their advantage, as the reports provide an independent assessment attesting to the quality of Japan's research.

The methodology developed and applied to the study of research and development in Japan is now proven to be equally relevant to Europe and other leading industrial nations. In general, the United States can benefit from a better understanding of cutting-edge research that is being conducted outside its borders. Improved awareness of international developments can significantly enhance the scope and effectiveness of international collaboration and thus benefit all our international partners in joint research and development efforts.

Paul J. Herer
National Science Foundation
Washington, D.C.

Preface

This report is based in large part on a visit to Japan by the JTEC panel on November 25 - 30, 1990. During that week, we were able to visit 25 sites: Advanced Telecommunications Research Institute International (ATR), Bravice International, Center of the International Cooperation for Computerization (CICC), Digital Equipment Corporation (DEC), Electronic Dictionary Research Institute (EDR), Fuji Electric, Fujitsu, Hitachi, IBM, International Business Service Inc. (IBS), Institution for New Generation Computer Technology (ICOT), Inter Group, Japan Electronics Industry Development Association (JEIDA), Japan Information Center of Science and Technology (JICST), Kyoto University, Matsushita Electric Industrial Co. (Matsushita), Ministry of International Trade and Industry (MITI), NEC, Nippon Hoso Kyokai (NHK), Nippon Telegraph and Telephone (NTT), Oki Electric Co. (Oki), Ricoh, Sanyo Electric (Sanyo), Sharp Corp., and Toshiba Corp. Following that, three additional sites (Catena-resource, CSK, and Systran) were visited by individual panel members. The panel gratefully acknowledges the hospitality of our Japanese hosts during all of these visits. Their willingness to interrupt their prepared demonstrations to run translation examples that we had brought with us was particularly appreciated. (See Chapter 7 for a discussion of the results of this process.)

The panel would also like to acknowledge several other people who contributed to this effort. Cecil Uyehara coordinated the entire trip to Japan. Joseph Clark, Charles Wayne, and Tamami Davidson accompanied the panel and helped to prepare this report. We'd like to thank Linda Mitchell for her work in pulling the report together.

We owe a special debt of thanks to Professor Makoto Nagao. Our trip and thus this report would not have been nearly as productive without his help. Professor Nagao helped to arrange a large number of our visits to individual sites. He met with us on Sunday, our first day in Japan, to go over the plan for the week. He met up with us several times during the week. He offered general advice as well as specific suggestions throughout the process of producing this report, including detailed comments on an earlier draft. We would like to thank him for all of his diligent efforts.

This report is based primarily on the panel's visits to the 28 sites listed above. Of these, 19 (ATR, Bravice, Catena, CICC, CSK, Fujitsu, Hitachi, IBM, JICST, Kyoto University, Matsushita, NEC, NTT, Oki, Ricoh, Sanyo, Sharp, Systran, and Toshiba) have been developing machine translation (MT) systems. Some of these institutions (for example, CSK, IBM, and JICST) are also significant MT users . Four of the sites (DEC, IBS, Inter Group, and NHK) were users but not developers of MT systems. EDR is building computerized dictionaries that can be used by MT systems. Fuji Electric is doing work on optical character recognition (OCR). The remaining three sites (ICOT, JEIDA, and MITI) are organizations that have had considerable involvement with MT. Other sources of information (such as [JEIDA 89]) have also been used as appropriate. This report is not structured as a set of site reports. Instead, it is organized around issues. As a result, it may be difficult to get a complete picture of an individual site. To help solve that problem, Appendix I lists each of the sites that are mentioned in the report. In addition, there is an index entry for each site, and all of the references to that site can be found in the report under that entry. In general, the abbreviations shown above will be used throughout the report. A full listing of all the abbreviations used in the report is given in Appendix III.

Throughout this report, there are several places where a set of MT systems is listed and some collection of properties of the systems is described. In each such case, we have included all the systems

for which the relevant information was available, either from the JTEC visit or from some other source. As a result, the set of systems considered, and even the number of systems mentioned, necessarily varies from one description to the next.

Machine translation in some respects is not one monolithic technology area but many related ones. Some MT systems have as their goal increasing the throughput of human translators; others aim at fully automated translation. Some systems are designed to work in restricted domains, while others must be far more general. Users of MT systems are concerned with technology that is cost-effective today. Researchers are often more concerned with the technology of the future. Even the history of MT work appears controversial. Not everyone has the same view. As a result, it is impossible to write a report such as this *una voce.* We have tried to present our results as consistently as possible. But even among the JTEC panelists, there is not complete agreement on several key issues. As a result, this report is organized not so much as an integrated whole but rather as a set of chapters, each with an author or pair of authors' names attached. It is our hope that, by taking this approach, we are presenting a fair impression of the diversity of views within the machine translation world, both here and in Japan.

Executive Summary

The goal of this report is to provide an overview of the state of the art of machine translation (MT) in Japan and to provide a comparison between Japanese and Western technology in this area. The term "machine translation", as used here, includes both the science and technology required for automating the translation of text from one human language to another (for example from Japanese to English or from French to Japanese).

Machine translation is viewed in Japan as an important strategic technology that is expected to play a key role in Japan's increasing participation in the world economy. MT is seen in Japan as important both for assimilating information into Japanese as well as for disseminating Japanese information throughout the world as part of the export process. As a result, several of Japan's largest industrial companies are developing MT systems. Many are already marketing their systems commercially. There is also an active MT and natural language processing research community at some of the major universities and government/industrial consortia.

Although MT products are already available, their incorporation into the everyday translation process is just beginning. The JTEC team visited two translation service bureaus that were using MT for about 20% of their volume, and there are others that are also starting to exploit MT. The volume of translation done using MT may grow quickly in the near term, since new networking services are making MT services more easily accessible. MT systems are also being used internally in many companies, including both the companies that have built the MT systems as well as their customers. The primary use for MT today is in translating technical documentation for products to be sold abroad. The volume is still relatively small but appears to be growing steadily. There is also an increasing use of MT systems embedded in other applications, such as database retrieval systems, electronic mail, and (in the prototype stage) speech-to-speech translation systems.

Most of the effort to develop MT in Japan has been devoted to systems that translate from English to Japanese (E/J) and from Japanese to English (J/E). But over the last several years, several of the systems have been extended to cover other languages, including the common European languages (such as German, French, and Spanish) and other Asian languages (such as Chinese and Korean).

Users reported varying degrees of success with MT. Although some users reported reduced productivity, many continue to rely on MT for such benefits as consistency of translated terms. Other users report productivity gains of up to 300%. Productivity gains of 30% appear average, with higher numbers for restricted application domains and lower ones for broader domains for which the system has not been finely tuned. Most uses of MT require some human pre- or postediting to produce acceptable quality translations.

Most of the MT systems now available in Japan are transfer-based systems. The majority of them exploit a case-frame representation of the source text as the basis of the transfer process. There is a gradual movement toward the use of deeper semantic representations, and some groups are beginning to look at interlingua-based systems. The Electronic Dictionary Research Project (EDR) is building a dictionary that will contain at least 400,000 concepts, as well as the associated Japanese and English words. When this dictionary is available, it will provide one important tool for building interlingua-based systems. All the commercial MT systems available today translate a single sentence at a time, although

some exploit a small amount of information about the larger context.

The currently available commercial systems clearly reflect the significant investments that have been made by the Japanese. They have dictionaries ranging in size from 50,000 to 800,000 entries (the latter including specialized technical terms). Many have over 300,000 entries, but none are as rich or as detailed as the EDR dictionary.

There is a clear consensus among MT vendors in Japan that the dominant factor that influences MT acceptance is the accuracy and fluency (collectively termed "quality") of the resulting translations. Other factors, such as purchase price and friendly user interfaces, although important, are much less significant. The vendors are therefore concentrating most of their efforts on improving quality, primarily by enlarging their dictionaries and grammars, and, secondarily, by gradually moving toward systems that perform deeper levels of semantic analysis.

A comparison between the U.S. and Japan in terms of MT and related technologies shows that Japan is ahead of the U.S. in several important ways, including the commercial use of MT, the acceptance of MT among users, the development of knowledge sources such as dictionaries, and the use of optical character recogniton (OCR) as an input modality, as well as in funding levels for R&D in MT. The U.S. has led in funding for basic research in natural language processing (the scientific underpinning of MT), and continues to lead in technological diversity (the number of different approaches that are being considered), linguistic diversity (because of a greater interest in the U.S. in the European languages), and level of effort devoted to R&D in speech processing. In both the U.S. and Japan, total funding for MT (including R&D, commercialization, deployment, and day-to-day use) appears to be on a gradual but steady rise.

The Japanese have made, and continue to make, a very significant commitment to MT. This commitment is visible in several ways, including Japanese industrial and government funding levels, the Japanese view of MT as an international prestige technology, and, most recently, in the increasing, steady, day-to-day acceptance and use of MT in the Japanese marketplace. Overall, the Japanese commitment to MT is greater than the U.S. one, though the latter is by no means insignificant.

1. Introduction: Machine Translation in Japan and the U.S.

Jaime Carbonell

1.1 The State of the Art in Machine Translation

The goal of machine translation (MT) is to automate the process of translating natural languages: English to Japanese (E/J), Japanese to English (J/E), Russian to French (R/F), Spanish to English (S/E), and so forth. In the ideal case, the translation would be fully automated, highly accurate, stylistically perfect and applicable to any topic and any style of texts. In present day reality, MT is only partially capable of achieving these objectives, with trade-offs between degree of automation vs. accuracy, breadth of coverage and text type vs. stylistic appropriateness, etc. Research and development proceed inexorably forward, gradually improving the MT state-of-the-art. But for certain classes of applications, MT is already a viable commercial reality, as we see below.

Currently, there is one recognized European commercial MT system (METAL by Siemens), two to four major American commercial MT systems (SYSTRAN, LOGOS, etc.) and at least three times that many Japanese MT systems. (See Figure 1-2.) In general, all of these systems produce rough translations containing errors of content and style that are typically corrected by a human translator (the "posteditor"), who is fluent in both the source and target languages. In addition, some of them require that the input text be "pre-edited" by a person fluent in the source language, before the MT system is used. The primary economic benefit results when less human effort is required to produce and correct machine translations than to produce them from scratch --- at least in situations in which stylistic perfection is not absolutely required. Secondary benefits accrue from other sources, such as consistency across translations and the embedding of MT into an already automated text production process.

1.2 The Role of Machine Translation

There are two primary roles for machine translation:

1. Assimilation of information in multiple foreign languages into the native language.
2. Dissemination of text in the native language into multiple foreign languages for a variety of reasons, chiefly as product literature to promote exports.

Each of these two roles makes different demands on an MT system. Dissemination requires very accurate and stylistically sound translation, as the translated texts will usually be printed and disseminated widely, and, in the case of technical documentation, acceptance of the product or service in the foreign market will be determined in part by the quality and timeliness of the translated text. On the other hand, assimilation places less stringent requirements on stylistic quality and may impose relaxed accuracy requirements as well. Often texts are translated only to determine what they are about, which requires only the roughest of translations, and only the small percentage of those found potentially relevant require more accurate or complete translation. The information analyst (e.g., a scientist, a technology watcher, an economic monitor, etc.) can tolerate stylistically imperfect texts even for those texts that are fully translated, though accuracy may be at a premium.

Unlike accuracy and style, where the requirements for dissemination are far more stringent than for assimilation, when we look at the ability to exploit topic and style constraints, we see that assimilation

presents the more difficult problem. Documents for multilingual dissemination, such as operating and assembly manuals for electronic equipment, share a common subject matter and a common writing style. In many cases, the same organization that is responsible for the translation has control over the production of the original documents and so can enforce stylistic rules that enhance the accuracy of MT. In contrast, the assimilation task must be flexible enough to accomodate documents on diverse topics and written in diverse styles. We describe the use of MT for both assimilation and dissemination in more detail in Section 6.1.

MT systems should be evaluated with respect to the user's primary objectives --- either assimilation or dissemination. Within each category, the requirements for such factors as diversity, accuracy, volume, speed, etc., should be analyzed. Although most MT systems, especially Japanese ones, have not been developed to specialize in either primary function, current trends indicate that specialization to accommodate user demands may be a way of obtaining greater performance on the desired dimensions.

1.3 An Historical Sketch of Machine Translation

This section provides a brief sketch of the history of the development of MT systems. A summary of this discussion is shown in Figure 1-1. For additional background and historical perspective, see [Hutchins 86] and [Nagao 89]. We also return to this subject in Chapter 8, when we discuss the status of MT in Europe.

Work on machine translation started in the United States in the 1950s. Warren Weaver, a vice president of the Rockefeller Foundation, had been impressed by early projects undertaken in England by Booth, Britten, and Richens between 1946 and 1949, and in 1949 he wrote the "Weaver Memorandum", in which he proposed that there are language universals, that the basis of language is logical, and that the use of techniques from cryptanalysis to encode and decode the meaning of natural language would be the key to translating by computer [Weaver 55]. The first MT work in the U.S. was begun by E. Reifler at the University of Washington in 1950 and initially concentrated on German to English (G/E). A second G/E effort began at the University of Texas during the late 1950s. Other MT work in the U.S. during the 1950s focused primarily on Russian to English (R/E) translation, largely fueled by the Cold War and Sputnik and the perceived need for tracking Soviet technology. R/E MT was first demonstrated by Georgetown University in 1954. In 1956, larger efforts on R/E translation began both at the University of Washington and Georgetown University. During this period, the earliest small-scale Japanese MT systems also started.

By the 1960s the first European MT research had begun, most notably the GETA [Vauquois 84] project in Grenoble, France, led by Professor Bernard Vauquois. The efforts in Japan continued, although they remained small, but work in the United States increased. For example, work was revived at the University of Texas on the system that would be named METAL [Bennett 85] in the early 1970s. Development of the SYSTRAN system had also started by the end of this decade [Toma 76]. All these efforts encountered a series of scientific and technological difficulties. In particular, it proved difficult to produce semantically accurate translations by purely lexical and syntactic means. In fact, the philosopher-mathematician Bar-Hillel stated that lexical ambiguity could not be resolved without recourse to world knowledge, and without resolving ambiguity it was impossible to translate accurately. The state of the art in the 1960s did not permit semantic analysis, and therefore Bar-Hillel concluded that MT was not then possible. Perhaps more damaging to U.S.-based MT was the ALPAC report [ALPAC 66], produced under

YEARS	U.S.	EUROPE	JAPAN
1950s	Start Sizable MT Projects		Early MT Research
1960s	METAL GEORGETOWN ALPAC: MT drastically cut	Start MT GETA	Basic NLP R & D
1970s	SYSTRAN Basic NLP R & D	EUROTRA	Basic MT R & D
1980s	"grass roots" Restart MT R & D SYSTRAN	EUROTRA METAL, SYSTRAN	MU System MT boom in industry labs
1990s	"Official" MT R & D SYSTRAN Multi-Lingual NLP	End EUROTRA Basic NLP R & D	MT Products R & D too: CICC, EDR, ATR, . . .

Figure 1-1: MT Historical Highlights

the auspices of the U.S. Academy of Sciences in 1966. ALPAC stated that MT was not economically feasible, among other reasons, because of the high cost of computers in the 1960s compared with the relatively low cost of human translators at that time — a situation now clearly reversed. In consequence, most but not all U.S. funding for MT research and development (R&D) evaporated in the 1960s. In this post-ALPAC climate it was left to the private sector to step into the breach. During this period, SYSTRAN, a private company, developed an R/E system that was installed for the U.S. Air Force [Toma 76].

The 1970s witnessed continued progress in France, the first widely used MT system in Canada (TAUM-METEO, developed at the University of Montreal for translating weather forecasts between English and French), and continued progress in the remaining U.S. MT efforts, most notably SYSTRAN and METAL. In spite of the hiatus in most U.S. MT R&D, research into the underlying science and technology of natural language processing (also called computational linguistics) continued unabated. In Japan, the pace of MT efforts quickened. Earlier efforts that had focused on basic work in natural language processing became the basis for substantial work on MT. See [Uemura 86] for a summary of some of these early efforts. Professor Makoto Nagao's laboratory at Kyoto University became a well-known center of MT work in Japan. The MU project [Nagao 86] got underway at Kyoto and elsewhere, with substantial government money and the goal of building a comprehensive MT system. (See Chapter

9 for details on this effort.) The research projects of the 1970s were to pave the way for the large Japanese MT R&D efforts of the 1980s and their commercialization in the 1990s. Fujitsu with its ATLAS-I project in 1978 represented one of the first industrial commitments to large-scale MT.

Company	System	MT Method
Bravice	MICROPAK	Syntactic Transfer
Catena	STAR	Syntactic Transfer
CSK	ARGO	Interlingual
EDR	EDR Electronic Dictionaries	Interlingual
Fujitsu	ATLAS-II	Interlingual
Hitachi	HICATS	Semantic Transfer
IBM, Japan	SHALT	Syntactic Transfer
JICST	JICST	Semantic Transfer
Matsushita	PAROLE	Syntactic Transfer
Mitsubishi	MELTRAN	Syntactic Transfer
NEC	PIVOT	Interlingual
Oki	PENSEE	Syntactic Transfer
Ricoh	RMT	Syntactic Transfer
Sanyo	SWP-7800	Syntactic Transfer
Sharp	DUET	Semantic Transfer
Toshiba	ASTRANSAC	Semantic Transfer

Figure 1-2: Japanese Industrial MT Systems

The 1980s witnessed an historically unparalleled set of initiatives in Japanese MT. In the early 1980s, the MU project was in full swing, supported by substantial government funding. MU focused on translating abstracts of scientific papers between Japanese and English. The MU system has subsequently served as the basis of the JICST translation system, which is in everyday large-scale government use for scientific abstract translation. In the same time period, virtually every large computer and electronics company in Japan endeavored to build its own MT system, most with substantial R&D teams (dozens of researchers) over many years, as summarized in Figure 1-2. (See Section 1.6 for an explanation of the entries in column 3 of this figure.) The majority of these efforts have produced working MT systems, mostly for translating Japanese to English and English to Japanese. Several new government-sponsored efforts were started in the late 1980s in Japan, most notably at the Center of the International Cooperation for Computerization (CICC), where the focus is on interlingua-based translation between Japanese and several Asian languages (see Section 9.9), and at EDR, which has undertaken a large effort aimed at building shared dictionaries and knowledge bases to support MT (see Section 9.6). Another new R&D effort is the work at ATR on a translating telephone. (See Section 9.10.)

Europe witnessed the EUROTRA phenomenon [Johnson 85] in the 1980s. This effort, spanning over a decade, was sponsored by the European Community (EC), and had the ambitious goal of translation between every EC language pair. Although not successful at this ambitious goal, EUROTRA energized European computational linguistics and MT activities. (See Section 8.3.) Also during this period, the

METAL system was acquired by Siemens/Nixdorf and moved to Europe.

In the late 1980s, MT research started again in earnest in the United States, emerging from the long shadow of ALPAC with the establishment of several substantial initiatives, including the Center for Machine Translation at Carnegie Mellon University, several MT efforts at IBM, the University of New Mexico, MCC, New York University, and the University of Southern California's Information Sciences Institute. Although representing greater technological and linguistic diversity, the U.S. and European MT efforts have not yet matched those of the Japanese in terms of sheer numbers, budget, longevity, and commercial maturity.

The 1990s have followed trends that were established in the late 1980s. These include increased MT R&D in Japan, and initial commercialization of MT systems in the Japanese market. A significant development in the Japanese MT sector is the emergence of MT in established translation services, such as Inter Group (which uses Fujitsu's ATLAS-II) and IBS (which uses Sharp's DUET system and NEC's PIVOT). MT services may be offered both with and without postediting, with even the former priced somewhat below the higher quality human translation. Some customers prefer MT, while other customers prefer human translation. (See Chapter 6 for a more detailed discussion of this phenomenon.) Other developments of the early 1990s were the cancellation of the EUROTRA project by the EC and the increasing pace of MT R&D in the United States.

1.4 The Japanese View of Machine Translation

The Japanese see MT as being very important. The following excerpt from a brochure handed out by Fujitsu at MT Summit III provides a good example of this attitude:

> Japan is well known as an exporter of manufactured goods, but perhaps less well known as an importer of knowledge and information. As far back as the 7th century AD, there was a steady flow of knowledge into Japan from China and the Korean peninsula. After a period of self-imposed isolation, the late 19th century--the so-called Meiji Restoration--saw the floodgates opened and knowledge and know-how poured in from Europe and the United States.
>
> Even now, news, information and data from overseas are beamed into Japan, translated into Japanese, and disseminated by the mass media and other communication channels. And conversely, as Japan's influence in the world becomes ever greater, there is more and more overseas demand for information originating in Japan. The latter means that, more than ever before, there is an enormous need for Japanese-to-English technical translation.
>
> Translation has always been demanding of manpower, money and time. Every year in Japan, about 100 million pages are translated at a cost of around 500 billion yen. A golden opportunity for translators? Maybe, but for a variety of reasons, including the boycotting of English during the Second World War, the number of Japanese with the necessary skills is severely limited.
>
> Fujitsu, itself demanding vast amounts of Japanese-to-English translation, has been trying to bridge the gap by developing the ATLAS machine translation system. Machine translation has been a goal of computer science since the late 1950s, but only recently, due to advances in artificial intelligence and computational linguistics, has it become even remotely practical.

Throughout this report, we provide evidence of the substantial investment that a large number of Japanese companies have been and are making in MT technology. We will also describe the results that these investments have already produced, and offer some hints of what the future may hold.

1.5 A Comparative Analysis of Japanese and U.S. Machine Translation

The state of the art in Japanese and U.S. MT can be analyzed comparatively from several perspectives, as illustrated in Figures 1-3 through 1-14. These figures represent rough composite estimates based on the knowledge the panel has about MT efforts, both here and in Japan.

Figure 1-3 shows that funding for MT R&D in Japan is substantially higher than in the U.S., although U.S. funding promises to increase gradually. New Japanese corporate funding is more focused on productivity and commercialization while maintaining an active and steady R&D effort level. Figure 1-4 indicates the expected increase in commercial MT in Japan in response to this trend. It also shows little growth in commercial MT in the U.S. during this period. (The gestation period for new MT systems is fairly long, so the favorable commercial effects of increased R&D may only be expected to be evident in the U.S. starting around the turn of the century.)

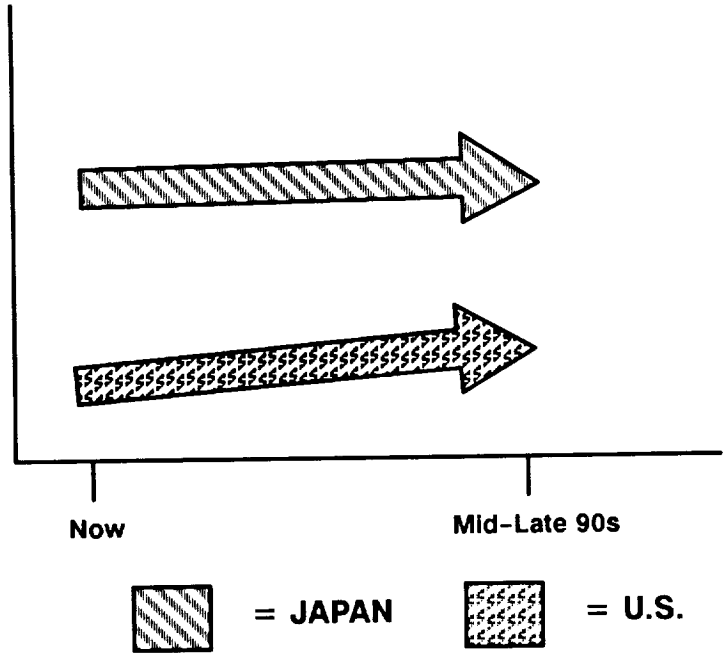

Figure 1-3: Funding for R&D in MT Technology

Improved accuracy (see Chapter 7) appears to be the single most important factor in determining how widely accepted MT will become. Both Japanese and U.S. efforts are expected to show steady improvement in accuracy, as shown in Figure 1-5. The largest differential will be in special-purpose (in terms of topic and style of texts) vs. general-purpose MT. Neither country enjoys a clear advantage in terms of accuracy. Special-topic MT gives greater accuracy now and promises more substantial improvements in the near-term future than general-purpose MT. The latter will also improve in accuracy, slowly but inexorably, extrapolating from present trends.

Figure 1-6 shows that in both Japanese and U.S. markets MT is gaining gradually in acceptance, with Japan having and maintaining a lead. As illustrated in Figure 1-7, the same situation and trends are

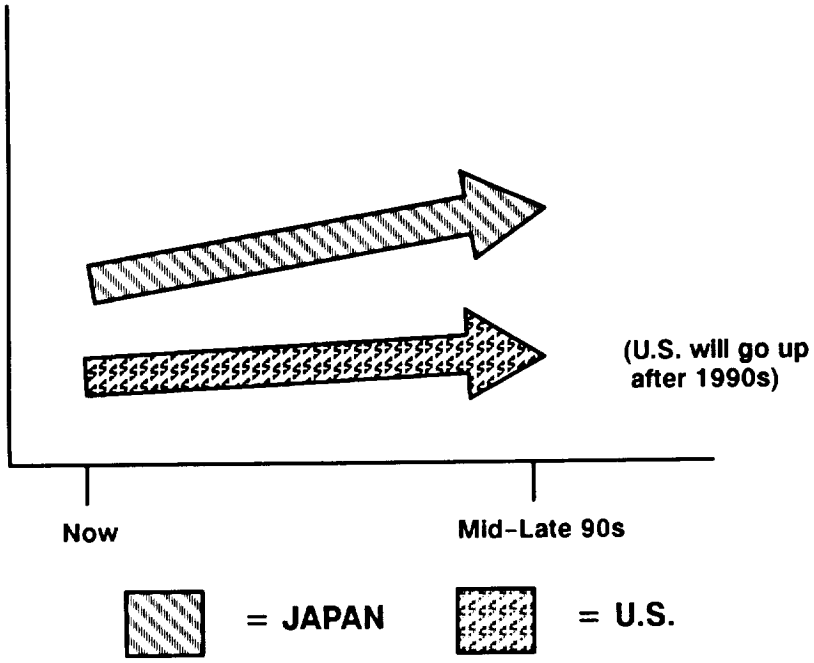

Figure 1-4: Commercial Use of MT

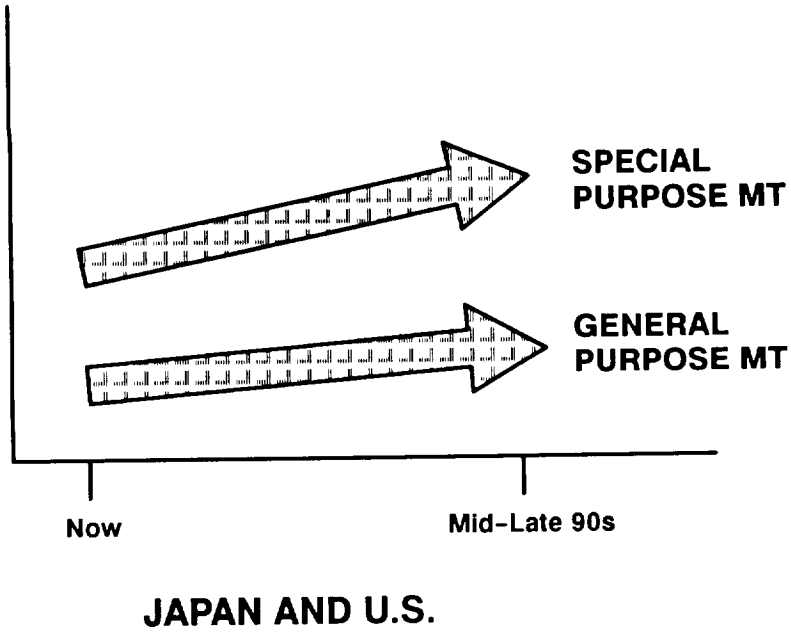

Figure 1-5: Accuracy of MT

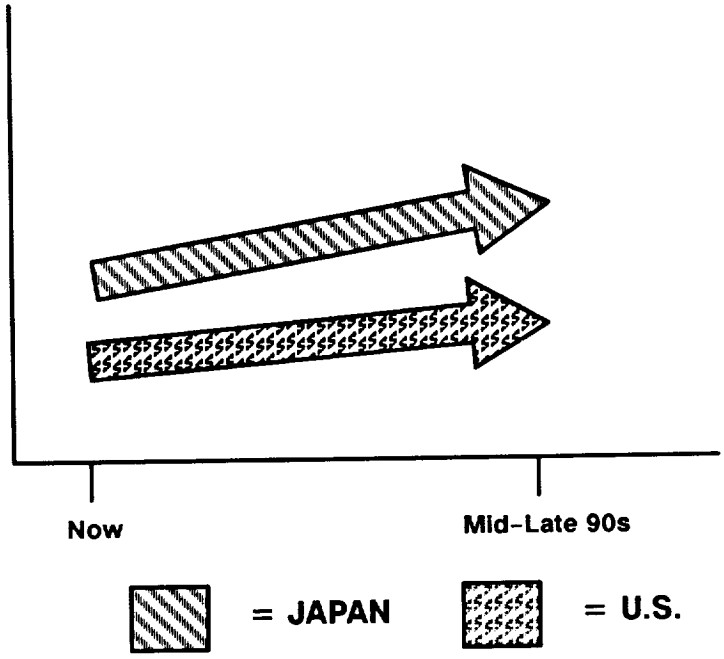

Figure 1-6: Acceptance of MT

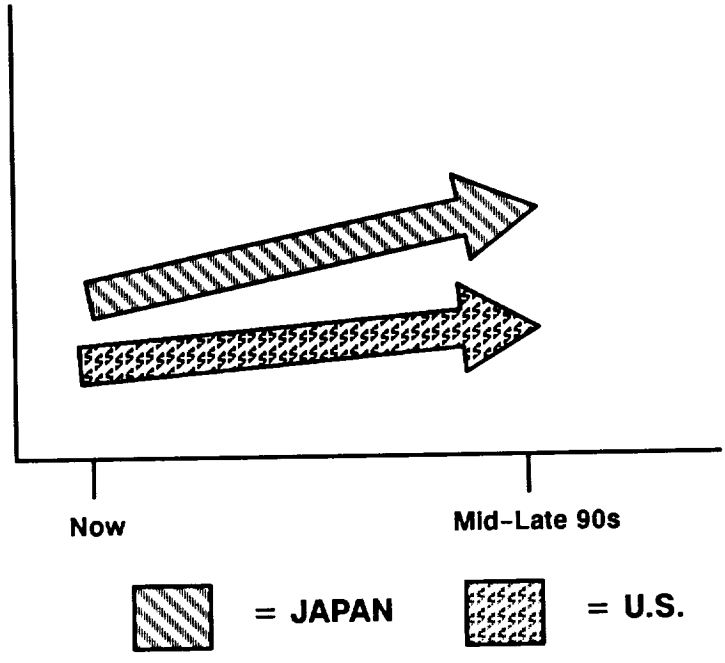

Figure 1-7: Integration of MT

present for the integration of MT systems into other text processing software: text editing, document production, printing, formatting, optical character reading (OCR), etc.

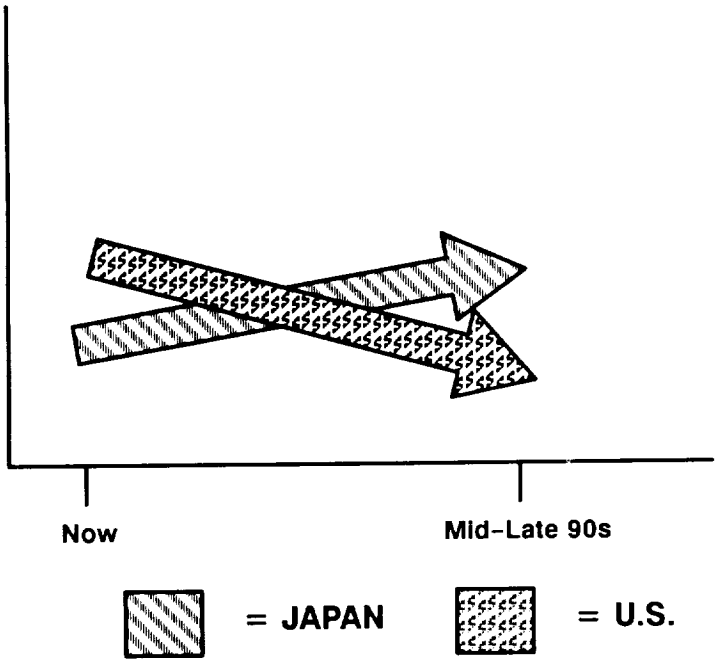

Figure 1-8: Funding for Basic Research in Natural Language Processing

The basic science and technology underlying MT is natural language processing (also called computational linguistics), which is the study of computer processing of language, including: parsing algorithms, language generation algorithms, grammar formalisms, knowledge representation, computational lexicography, and inference techniques. Traditionally, the U.S. has been a bastion of scientific research in this area, but research funds in the U.S. have been decreasing. (See Figure 1-8.) In contrast, Japanese and European funding for the basic research underlying MT is increasing and will surpass U.S. funding levels, if they have not already done so. The U.S. research infrastructure risks being surpassed in the one dimension where it has traditionally led: computational linguistics, both the basic theory and computational methods.

Figures 1-9 and 1-10 indicate that the U.S. leads Japan in technological diversity (for better or worse) and in linguistic diversity. The former refers to the variety of technological approaches to MT, as discussed in the following section, and the latter refers to the number of languages between which MT systems are being developed. Present trends indicate that the U.S. will maintain its lead in technological diversity, but the gap will narrow in linguistic diversity as new Japanese projects, such as CICC (translation among several Asian languages), get underway, and Japanese commercial systems continue their trend toward expansion into European languages, as exemplified by Fujitsu's efforts in multilingual MT (German, French, and Spanish).

MT requires multiple knowledge sources, including lexical, grammatical, and semantic ones. (See

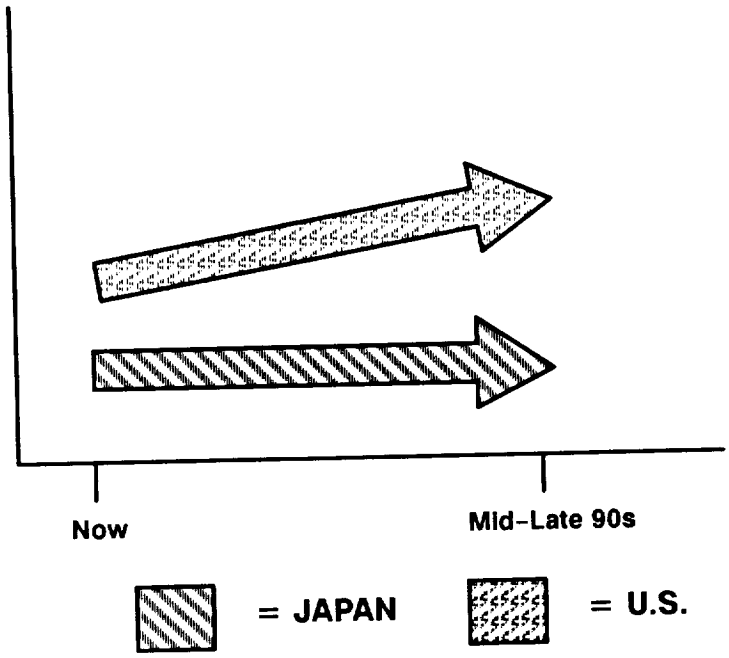

Figure 1-9: Technological Diversity

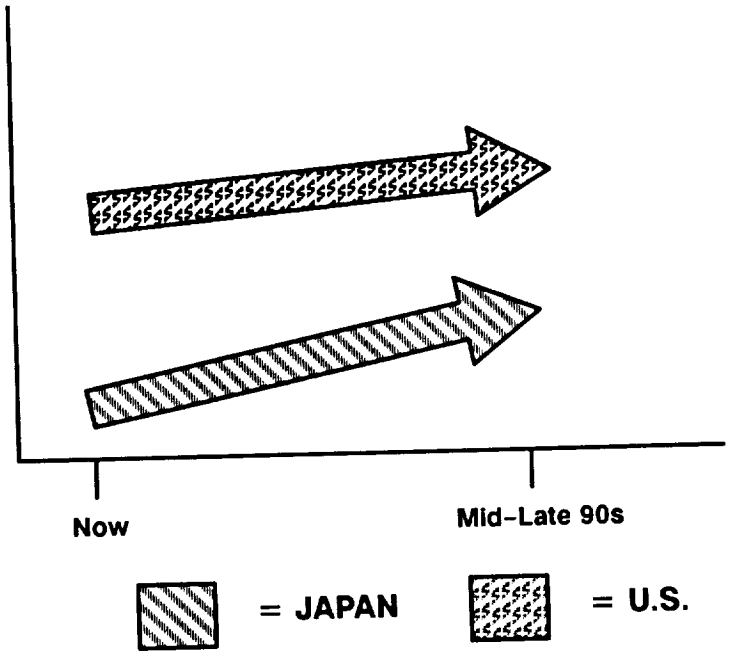

Figure 1-10: Linguistic Diversity

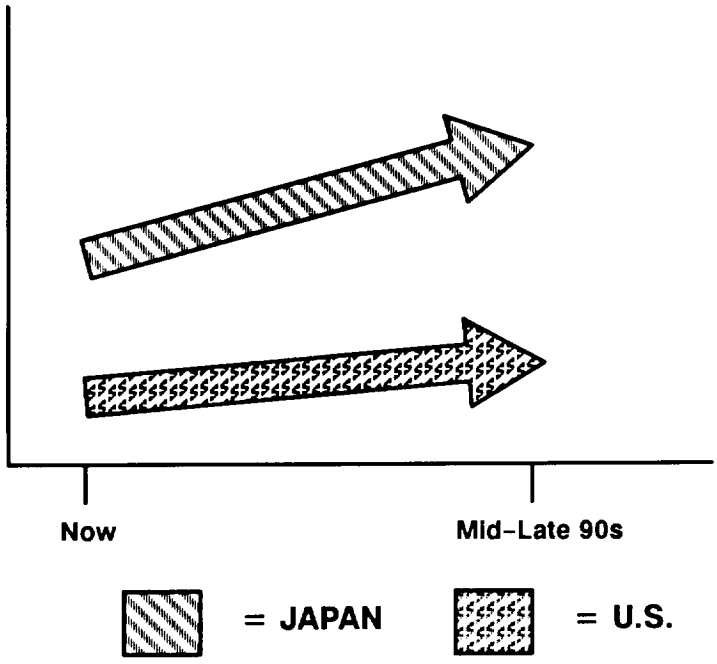

Figure 1-11: Private Knowledge Sources

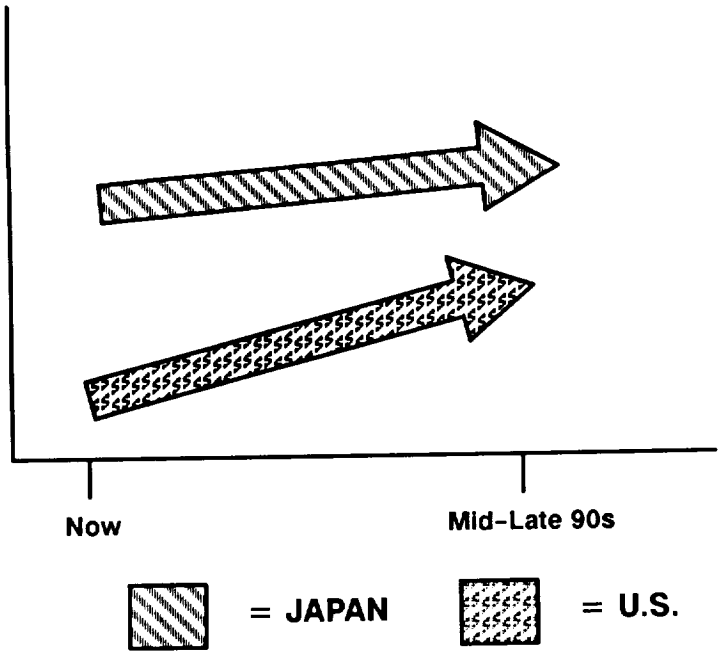

Figure 1-12: Shared Knowledge Sources

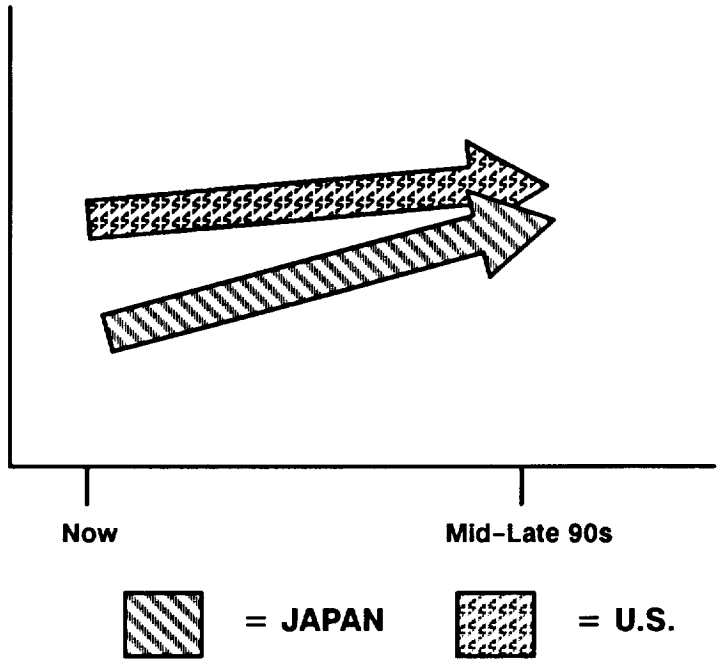

Figure 1-13: R & D in Speech Recognition and Speech-to-Speech MT

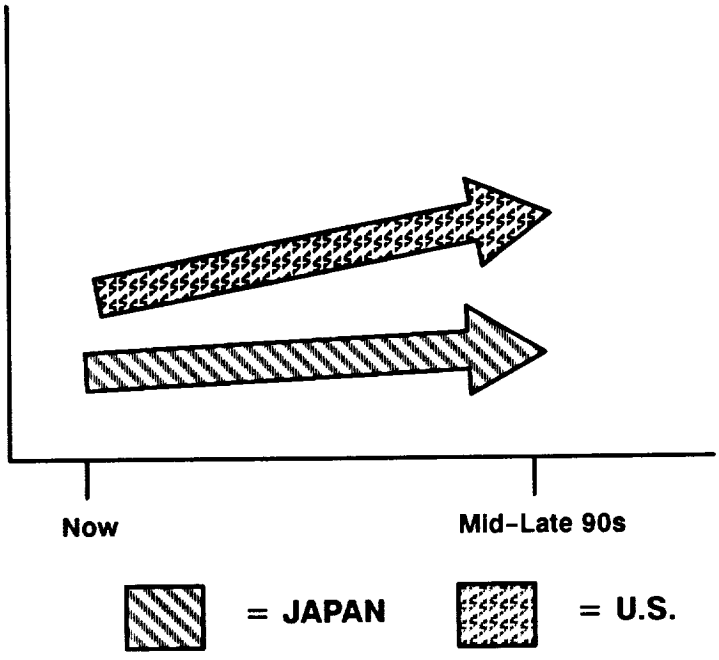

Figure 1-14: R & D in Other Natural Language Processing Technologies

Chapter 4 for a detailed discussion.) These knowledge sources are large and expensive to build and maintain. Therefore, they are valued resources in MT research and are even more important in successful MT system deployment. Figure 1-11 indicates a Japanese lead (and possibly widening gap) in private knowledge sources, i.e., those built and owned by the large computer and electronics firms that build and market MT systems. Although Japan also leads in shared knowledge bases (most notably EDR), the gap may narrow assuming continued funding from DARPA and other U.S. government agencies that are targeting some funds specifically at building shareable knowledge sources. (See Figure 1-12.)

The U.S. also maintains a lead in other related research areas. For example, Figure 1-13 shows that the U.S. leads in speech recognition technology, but both the U.S. and Japan are working on the early integration of speech technology into speech-to-speech MT. Specifically, in Japan this is going on at ATR and at some companies such as NEC. Figure 1-14 shows the status of related natural language processing (NLP) technologies such as automatic extraction of knowledge from text (e.g., to populate databases), NLP-based human-computer interfaces, routing and classification of texts for assimilation, etc. The U.S. has a narrow lead these areas that may widen if current trends continue.

1.6 Paradigms for Machine Translation

Historically, many different approaches to MT have been tried with varying degrees of intensity. Figure 1-15 captures all of the major paradigms for machine translation. The arrows represent transformations that may occur during the translation process. It should be clear from the figure that there are many paths from source to target texts.

Early MT efforts focused primarily on the bottom levels of the figure. Over the years there has been a steady trend upward toward systems that exploit deeper analyses of the source text as the basis for translation. More specifically, the first methods developed for MT were the direct methods; they attempted to map words, phrases, and entire clauses from the source to the target language, without performing any analysis of the source text beyond straightforward morphological processing. A common approach was to build huge numbers of specific direct transfer rules by hand. Because of the enormity of the task and the poor results that came from the necessarily incomplete rule sets, this approach was abandoned early. However, three modern variants of the direct approach are currently being investigated:

- Statistically-based approaches at IBM in the U.S.,
- Example-based (also called case-based or analogical) translation at Kyoto University and ATR in Japan. (See Section 9.2.), and,
- Direct-transfer translation, also at ATR. (See Section 9.3.)

All require large, bilingual, sentence-aligned corpora, but none requires millions of hand-built direct transfer rules. These projects are still in the early research stages, but show promise.

Moving up one step in the figure, we get to syntactic transfer systems. In these, the source text is "parsed" or analyzed into syntactic structures. These structures (known as "parse trees") are transformed into corresponding syntactic structures in the target language by applying a set of hand-coded transfer rules. The lexical items (at the leaves of the syntactic tree) are also transferred by a bilingual transfer dictionary. After transfer, a syntactic generator maps the lexically-bound target-language syntactic structures into the target language text. This paradigm is sometimes called "transfer-based MT" or

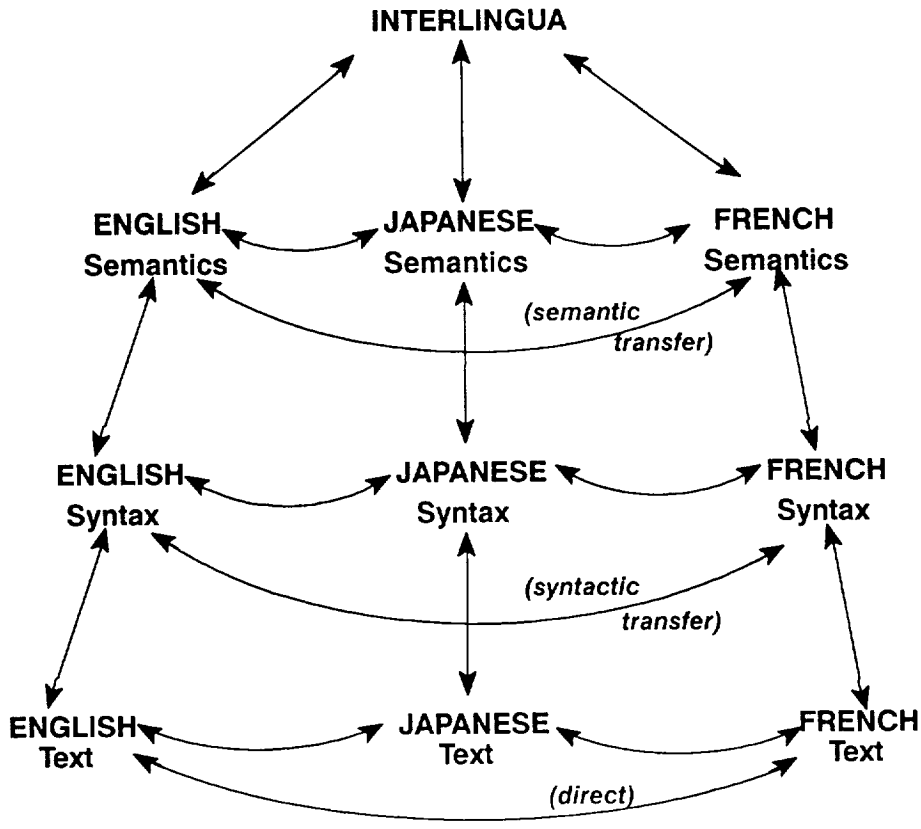

Figure 1-15: Interlingual vs. Transfer MT

"traditional transfer," and until recently has been the most popular paradigm for MT. Substantially fewer transfer rules are needed than were necessary in direct systems because the rules capture linguistic generalizations. However, there are problems with syntactic transfer:

- The lack of semantic analysis results in poor disambiguation and hence errors in the translation, and
- If a multilingual system is required among N languages, N^2 sets of transfer rules must be developed (one for each uni-directional language pair).

The first of these problems can be at least partially solved by moving upward again, toward a greater degree of semantic analysis. Almost all syntactic transfer systems in use today extract some semantic features and exploit them during translation. A further step has been the development of the "semantic transfer" paradigm, which is now widely used in Japan. Here the analysis is deeper, including syntactic and at least partial semantic analysis of the source text prior to starting the transfer phase. (This approach is shown in the upper horizontal arc in Figure 1-15.) The transfer occurs between corresponding semantic representations (typically case frames) from source to target, and then the generator maps these tranferred case frames into the target text. In principle, translation accuracy improves because some disambiguation occurs prior to transfer, and since semantic representations in different languages are more similar to each other than syntactic ones, the transfer component is smaller (although still N^2 for N languages).

If complete syntactic and semantic analysis is performed, then it is possible to produce a meaning representation, called an interlingua, independent of source or target language. The interlingua may contain just the meaning of the source text, or it may also contain a language-independent description of

the linguistic form that was used in the source text so that effects such as focus can be recreated properly in the target. Once an interlingual representation is obtained, the text is generated into one or more target languages. No explicit transfer phase is necessary. In essence, the interlingua approach trades off more effort at analysis and generation for no effort at all in transfer.

There is considerable debate in the scientific community as to which is the best approach to MT. It is generally recognized that transfer (either syntactic or semantic) may be the easiest to build for a single language pair, whereas the interlingual approach may provide translations of better quality and/or provide the most extensible paradigm in a multilingual environment. Most everyone agrees that there are cases in which the ability to extract meaning is critical if high quality translations are going to be produced. People differ, however, on how hard it will be to do this in a practical way, and so there is disagreement on when (if ever) a true general-purpose interlingual system will be able to be built. Another reason for disagreement about the role of interlingual systems is that, in some researcher's eyes, using an interlingua necessarily means discarding surface linguistic facts from the source text, thereby reducing the fluency of the translation in some cases. Many proponents of the interlingual approach reply by saying that linguistic facts, as well as semantics, can be captured (in a linguistically neutral way) in the interlingua.

Despite the intensity of the "interlingua debate", it turns out that existing Japanese commercial MT systems can mostly be accounted for by the middle range of Figure 1-15. There are no direct systems and there are no "pure" interlingual systems, although some, such as NEC's PIVOT and Fujitsu's ATLAS-II, are closer to being purely interlingual than most others. This can be seen in Figure 1-2, with the following caveat: It is important to keep in mind that the transfer-interlingua dimension is a continuum rather than a discrete choice, so there are some borderline cases that are difficult to characterize precisely. For example, Ricoh's system uses slightly more semantics than most syntactic transfer systems but less than most semantic transfer systems. Similarly, commercial interlingua-based systems also exploit some transfer rules, even though that is not part of the "pure" interlingual paradigm. Throughout this report, we have attempted to describe the approach of each system in the terms that a particular system's designers use, even though we recognize that these terms are not always applied consistently.

Although there is debate about which point along this spectrum shows the most promise as a basis for MT systems (and, as we just mentioned, there is not even full agreement on exact terminology), there does seem to have been a gradual movement upwards over the years towards deeper analysis and interlingual systems. One example of this is the long-term, large-scale CICC project in which several of the major Japanese MT companies participate. (See Section 9.9 for a discussion of this effort.) Chapter 9 will elaborate on this trend and describe in more detail the research that is being done on interlingua-based systems.

In addition to the transfer-interlingua dimension, systems differ with respect to where the human intervention occurs. Postediting is the most typical situation: The MT system attempts its best translation and a human posteditor corrects the mistakes after careful comparison of source and target sentences to fix errors of both content and style. Pre-editing the source text to make it easier to translate is also sometimes used. Typical pre-editing operations include breaking up long sentences into shorter, more easily analyzed ones and replacing ambiguous passages and words with less ambiguous ones. Many Japanese systems (and some American ones such as Xerox's specialized use of SYSTRAN) combine

both pre- and postediting in an attempt to do less of the latter.

Another possibility is "just-in-time editing" or "user query", where the MT system queries the author (or translator) interactively during the analysis to resolve ambiguity as needed. Finally, research also addresses machine-aided translation (MAT) where the human translator is in control and the MAT system provides productivity tools such as on-line terminology banks and grammar checkers. The latter approaches are being investigated more in the U.S. than in Japan, which partly accounts for the greater U.S. technological diversity discussed in the previous section.

1.7 Structure of the Report

The rest of this report is organized around a set of topics that together cover the most important aspects of the state of the art of MT in Japan. Chapter 2 describes in more detail the technical ideas that underlie MT systems. Chapter 3 describes the languages and the application domains that are receiving the most attention in Japanese MT work today. Chapter 4 discusses the knowledge sources (primarily dictionaries) that are used in Japanese MT systems. Chapter 5 outlines the life cycle of a typical MT system in Japan. Chapter 6 surveys the uses of MT in Japan and briefly discusses both a set of user sites and a set of vendor sites that were visited. Chapter 7 talks about the major factors (quality and productivity) that influence the acceptance of MT in Japan. Chapter 8 puts this analysis of MT in Japan in perspective by describing the status of MT efforts in the United States and in Europe. Chapter 9 moves away from a focus on deployed systems and describes the main thrusts that current MT research in Japan is taking. Finally, Chapter 10 offers a very brief analysis of the future of MT in Japan.

Following the main body of the report is a list of the references that were cited. Although we had access to literature in both English and Japanese, we have made a conscious effort to cite works in English (because of their much greater accessibility) whenever possible. (Of course, when fully accurate MT comes of age in the next century, such linguistic biases in citations should no longer be necessary.) Three appendices follow the list of references. The first summarizes the sites that are mentioned in the report. The second contains short biographies of the panel members. And the last is a list of abbreviations used throughout the report.

2. Technical Infrastructure

David Johnson

This chapter is concerned with providing a high-level overview of the basic technology underlying the linguistic processing characteristic of typical state-of-the-art, Japanese MT systems on the market today.

Figures 2-1, 2-2, and 2-3 contain schematic charts, taken almost verbatim from [Hutchins 86], of each of the three main kinds of MT systems (interlingual, transfer, and direct) shown in Figure 1-15. Note that "SL" means "source language" and "TL" means target language.

Since most of the MT systems on the market today are either syntactic or semantic transfer-based systems, this chapter describes the transfer-based approach. In chapter 9, we describe other approaches, including interlingual and example-based systems. Knowledge sources (lexicons, thesauri, etc.) are explored in Chapter 4.

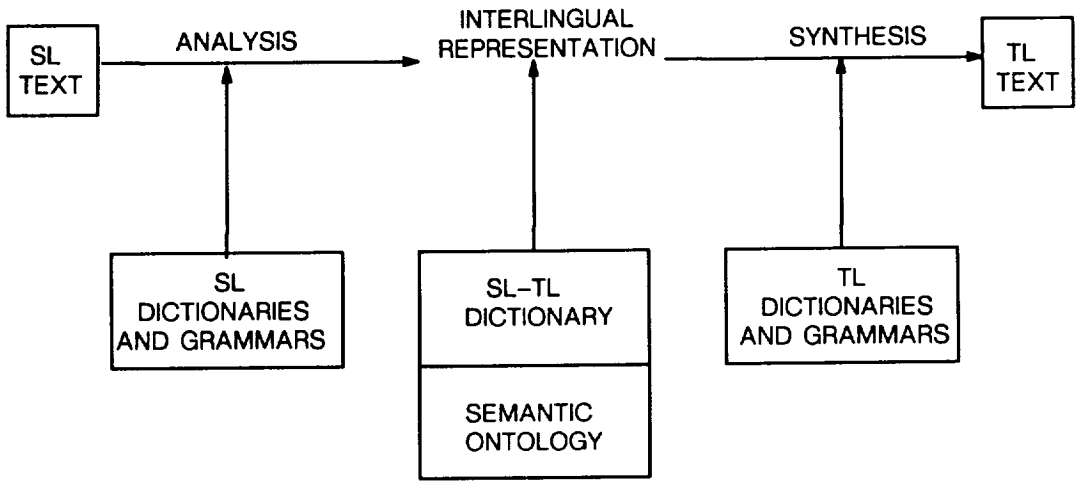

Figure 2-1: Interlingual MT System Architecture

We will first sketch the major stages involved in machine translation, and then examine the three key functions of a typical MT linguistic processor — analysis (parsing); transfer; and generation (synthesis) — effecting the overall translation from source sentence to target sentence. Detailed comparisons or evaluations of specific systems will not be made, although specific systems will be used as examples of the approaches we describe.

2.1 Overview of the Translation Process

The flow of control from input to output in NEC's PIVOT system, a typical, sophisticated, production-level MT system, is shown in Figure 2-4. In this system, there are eight stages: (1) text input, via keyboard, optical character recognition (OCR) or other means; (2) pre-translation, where words not registered in any source-language dictionary or other knowledge source are located and placed in a file; (3) and (4) registration by a human pre-editor of "unknown" words (note that, in fact, many such words will

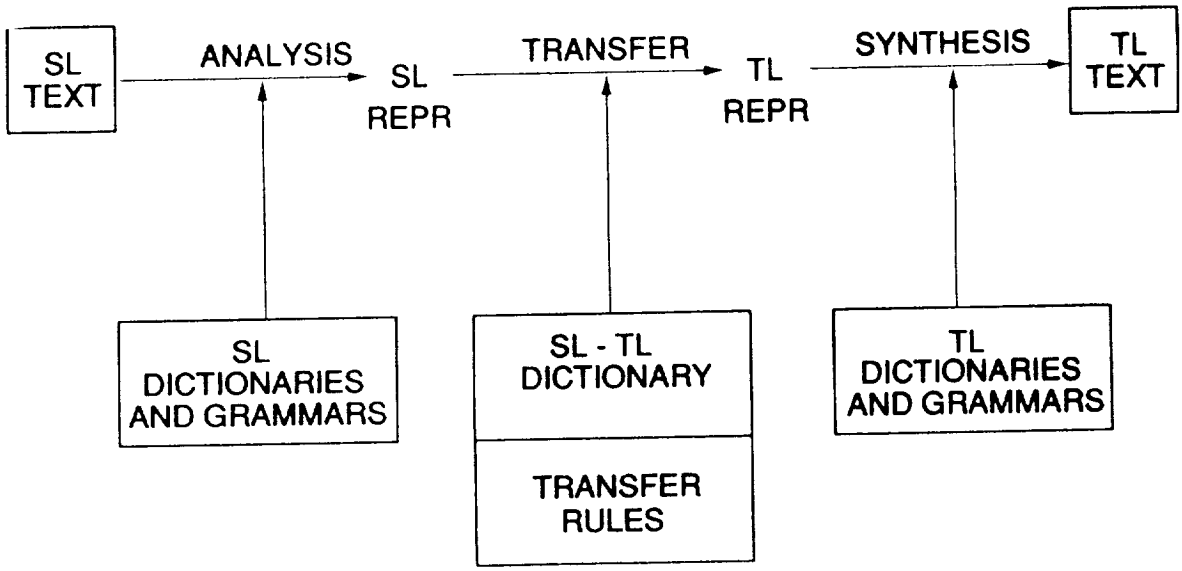

Figure 2-2: Transfer MT System Architecture

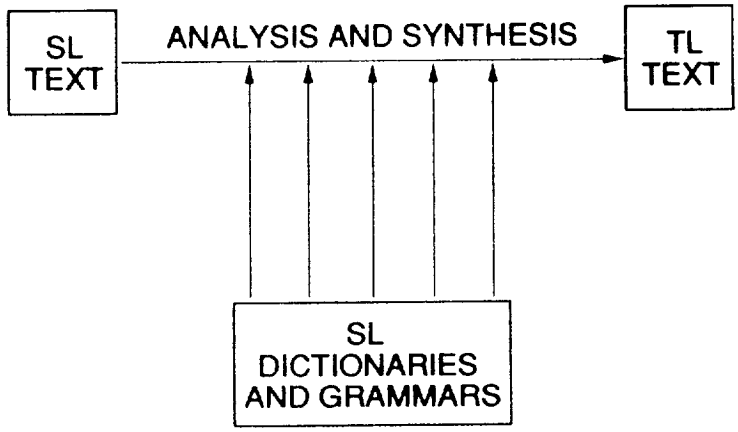

Figure 2-3: Direct MT System Architecture

not have translations, e.g., names, addresses, acronyms); (5) translation (by machine). This corresponds to the box "Machine Translation Process Software" in Figure 2-5 (taken from [Nagao 89]), which shows in more detail how the translation process typically works. Many systems omit (6) correction of analysis. It is followed by (7) translation, a repeat of (5), and (8) post-editing. As the stages listed in Figure 2-4 suggest, the time taken by the computerized translation process will generally comprise only a very small part of total translation time. Thus, speed of translation in (5) is not the crucial factor in the cost-effectiveness and user acceptability of a particular system. Instead, the key issue is the quality of this stage of translation, and it is in this arena that all MT systems fall down.

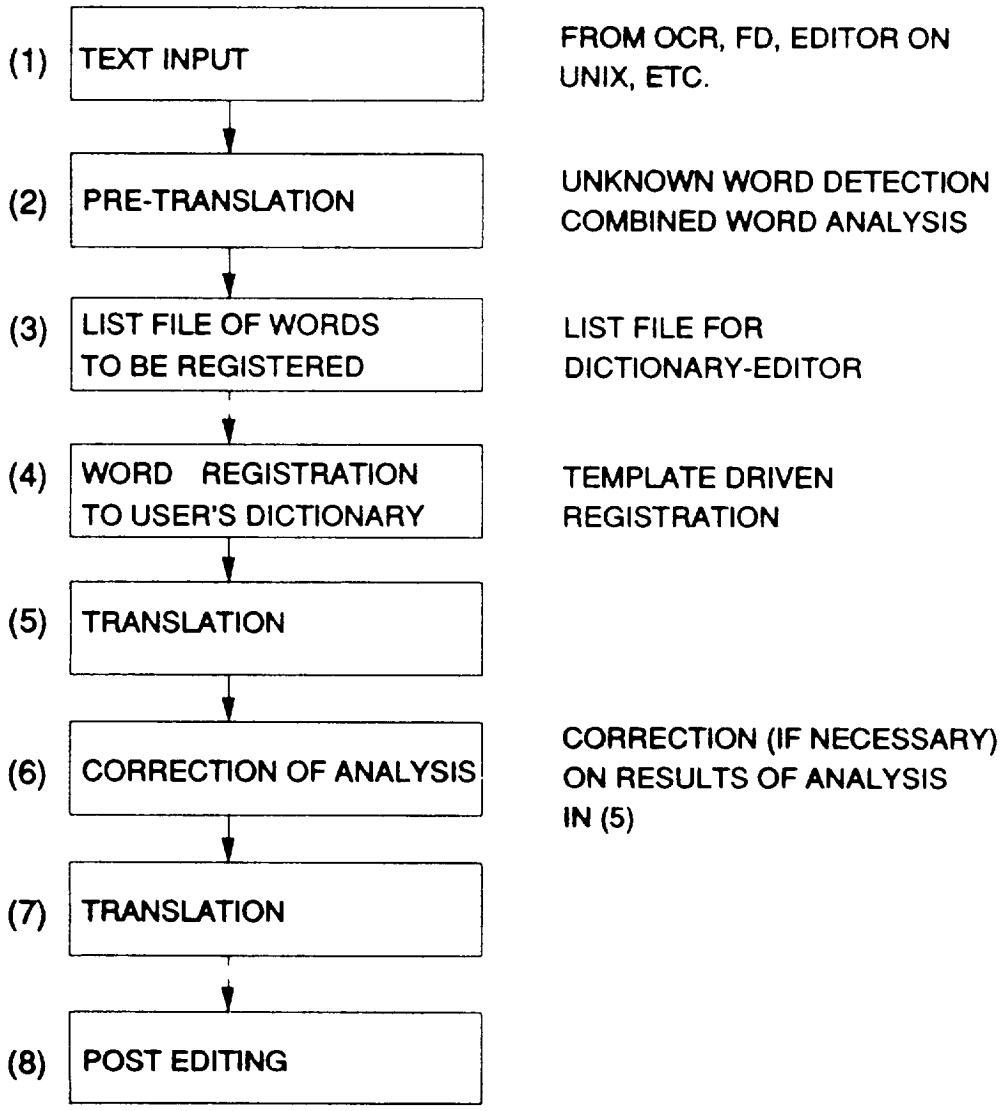

Figure 2-4: Flow of Control in the NEC PIVOT System

2.2 Translation Stages of the Linguistic Processor

Linguistic processing (Step 5 of Figure 2-4) generally proceeds through six basic stages: (1) morphological analysis of the source language; (2) syntactic analysis of the source language (parsing); (3) semantic analysis of the source language (semantic feature analysis); (4) transfer (mapping the internal representation of the source-language sentence into the internal representation of the target-language sentence); (5) syntactic generation of the target language sentence; and (6) morphological generation of the target language sentence.

Figure 2-6, which depicts the translation flow in Ricoh's MT system, is an example of a syntactic-transfer system that exploits some semantic feature information. This should be compared to the flow chart shown in Figure 2-7, for Hitachi's HICATS/J-E semantic-transfer system. Notice that both systems go through the same basic stages, i.e., they are not very different in terms of overall system organization.

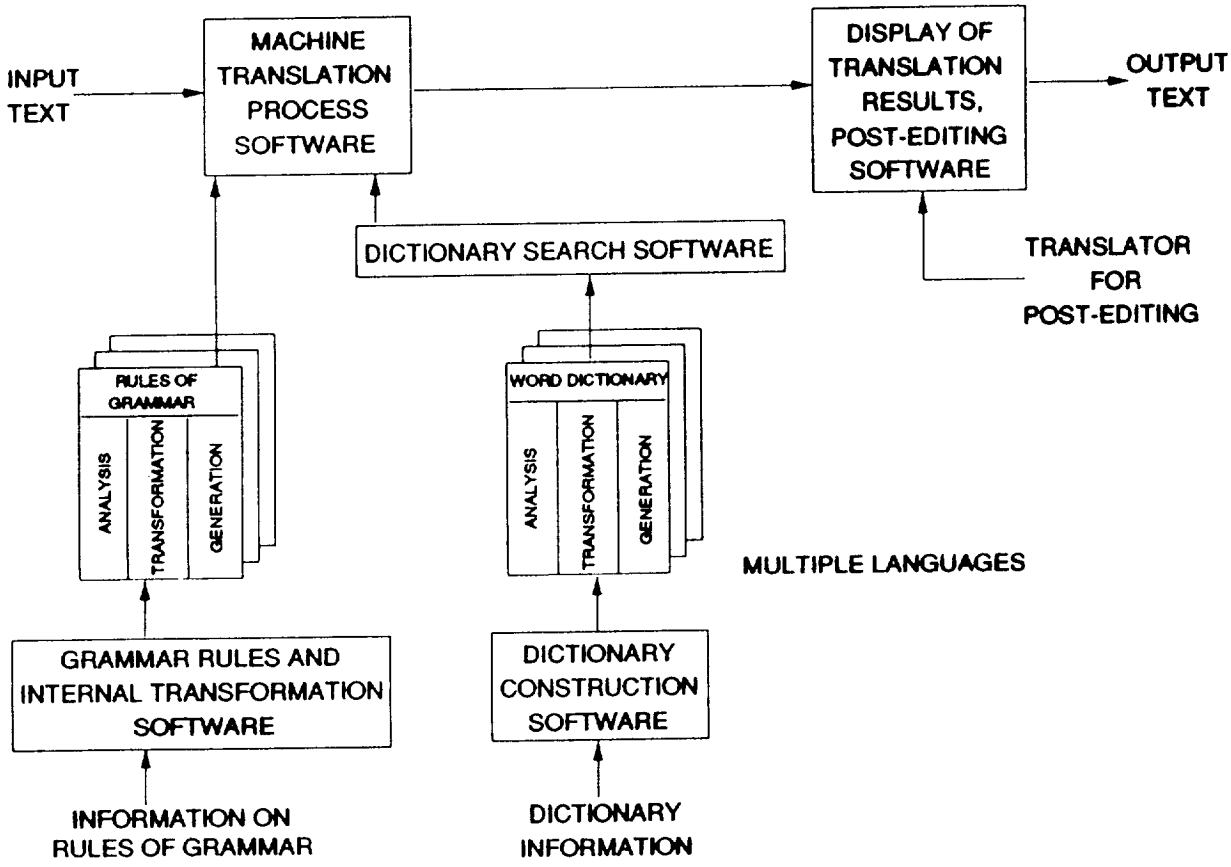

Figure 2-5: The Basic Software of a Machine Translation System

All of the systems described here are transfer-based, i.e., they are not direct nor are they true interlingual systems. But, as we noted in Section 1.6, the transfer-interlingua dimension is really a continuum, consequently, it is often difficult to characterize systems precisely according to one of the three types. In some transfer systems the level of transfer is syntactic; in others it is asserted to be semantic. It is not clear, however, what is at issue here, since some so-called semantic representations are often language-specific and not very detailed, while some so-called syntactic transfer systems can be highly abstract and fairly language neutral.

However, as a matter of design methodology, in the case of so-called semantic transfer, there is an explicit effort, on the one hand, to purge the output of analysis of obviously language-specific material such as specfic case markings and word order. On the other hand there is an effort to make use of a limited set of putatively universal semantically oriented labels such as "agent", "theme", recipient", and so on. This use of a common representation language --- unordered trees with explicit labeling of relations

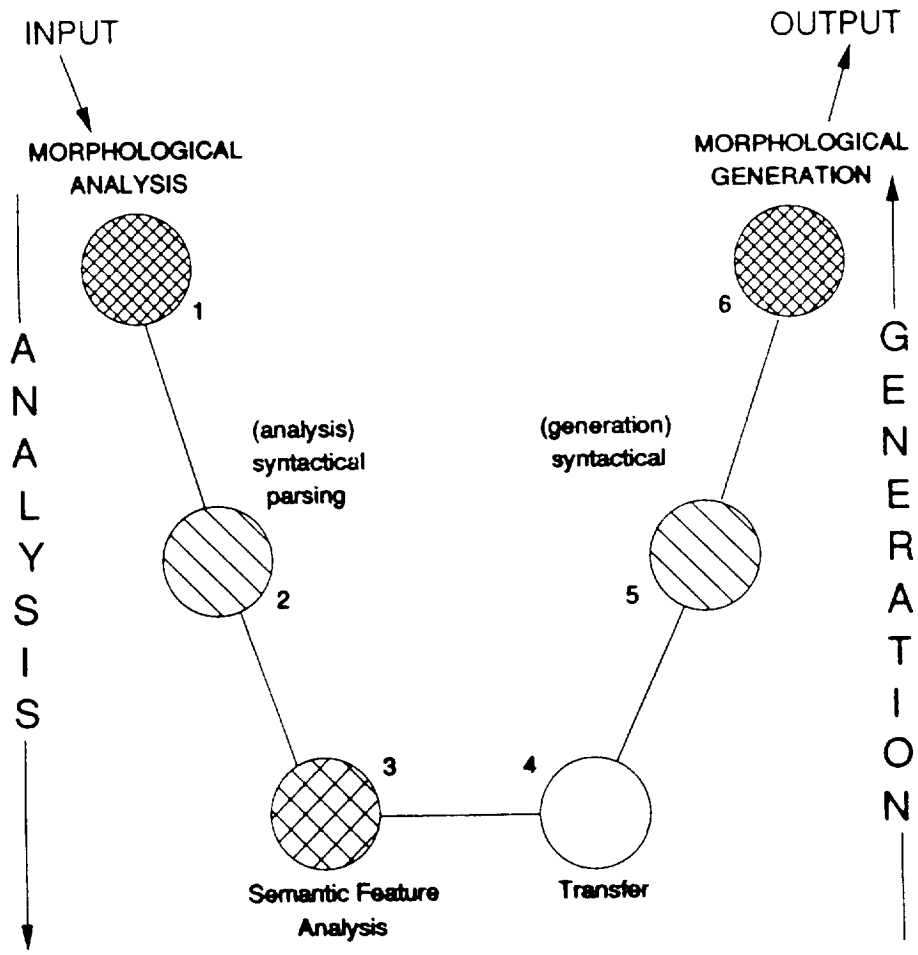

Figure 2-6: Flow of Control in the Ricoh MT System

like "agent" (dependency structures) — for source and target languages would seem to facilitate the minimization of superficial differences, and thus be a practical compromise between overly language-bound syntactic representations and a true interlingual representation.

2.3 Analysis

Two steps are essential to a thorough analysis of the source text — morphological analysis and lexical look up. Then the input string is parsed with respect to a formal grammar of the source language. The analysis grammar is often an augmented context-free phrase-structure grammar, but augmented transition networks (ATNs) are also commonly used.

The result of parsing is one or more phrase-structure trees that represent the syntactic constituent structure(s) of the source sentence. Constituent structure encodes the part-whole relations of words and phrases, as well as word order.

For example, the constituent structure of the English sentence, "He ate cake," is shown in Figure 2-8. The constituent structure for this sentence indicates that "he" is a Noun(N), which is, in turn, a Noun Phrase (NP). Together, the verb (V) "ate" and the noun phrase (NP) "cake" constitute the verb phrase (VP) "ate cake". The NP and VP together make a sentence (S).

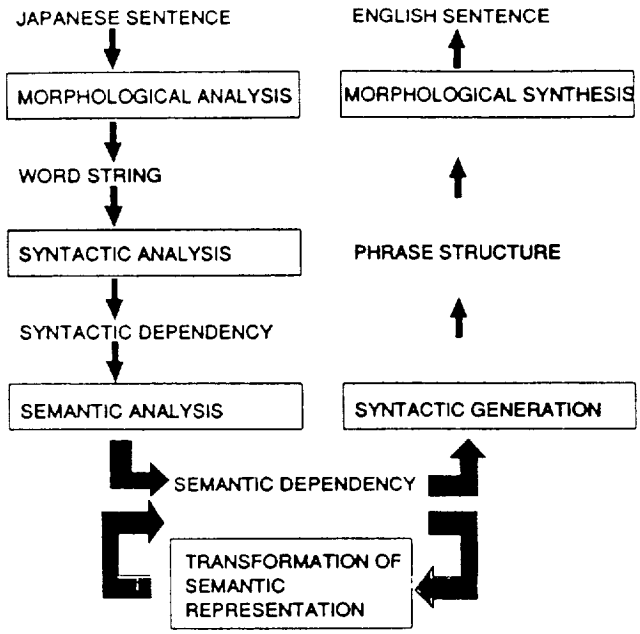

Figure 2-7: HICATS/J-E Translation Process

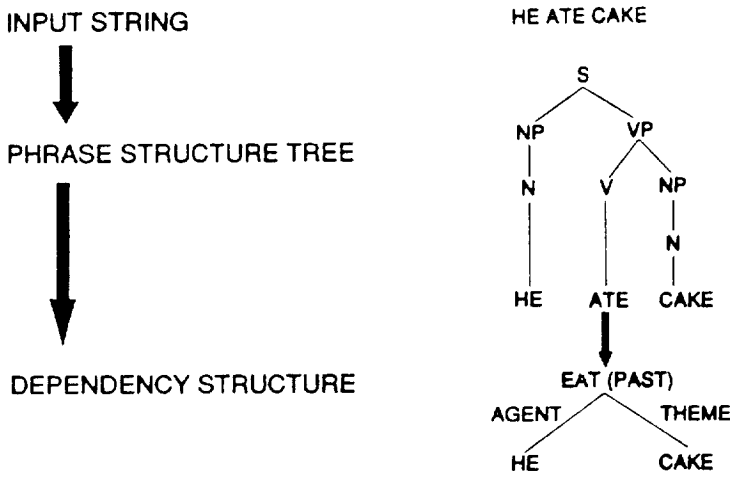

Figure 2-8: The Result of Parsing

Japanese sentences are organized as sets of phrases called *bunsetsu.* A *bunsetsu* is typically composed of a content expression and a function word that indicates the role of the content expression in the overall sentence. In some cases these function words behave the way prepositions do in English (except that function words come after their corresponding content expression). In other cases, the function words mark roles (such as subject and object) that are indicated in English by word order. A complete syntactic analysis of a Japanese sentence shows how the *bunsetsu* it contains relate to each other to form the overall analysis of the sentence.

The next stage of analysis is the conversion of phrase-structure trees to dependency structures. Dependency structures contain fewer nodes, in general, than corresponding phrase structure trees, since

each governing node in a dependency structure is a lexical item --- the head of the phrase (i.e., phrasal nodes) are thrown away. The explicit representation of the linguistically crucial notion of phrasal head is one of the key advantages of dependency representations. This is an important advantage for natural language processing systems, because the head carries most of the linguistic information governing the properties of the phrase. This information can be associated with specific lexical entries, and, after lexical look up, will be uniformly located at the roots of dependency phrases and hence readily accessible to rules during processing. To illustrate, in Figure 2-8, the head of the clause is the verb "eat." The lexical entry for "eat" would, no doubt, specify that any phrase headed by "eat" takes an agent that is an animal and a patient that is edible.

The second advantage of dependency representations is the explicit representation of grammatical relations, often called *cases*. Explicit representation of grammatical cases such as agent and theme facilitates the matching of linguistic information contained in dictionary entries with requirements specified in rules. In contrast, constituent structure trees only implicitly represent grammatical relations in terms of the part-whole relation and linear order of constituents. Since grammatical relations are crucial for carrying out the transfer process, the explicit representation of such relations is a significant advantage. In the case of the simple example in Figure 2-8, the dependency structure of the English sentence is structurally the same as its correspondent in Japanese. Therefore, in this very simple case, transfer would only involve lexical substitutions.

```
(5700) NP POSTP
           --> PP(%NP,CASE=CASE(POSTP),
           TOPIC=TOPIC(POSTP),KOOU+KOOU(POSTP),
           PSMODS=PSMODS...POSTP)

(5840) VERB --> VP(sVERB,HINSISEI='YOU')

(5851) PP(~DAI,~NO,SF,CASE)
       VP(~NO,CASESO,CASE(PP),ISIN.CASESO,
         CASE(PP).NOTIN.CASES,
         <@CASE(PP).EQ.'DIV"|
         @CASE(PP).ISIN.SF(PP)>)
         -->VP(-NAI,CASES=LISTIFY<CASE(PP)>...CASES,
              PRMODS=PP...PRMODS,N=N(PP)+N+4,
              <TOPIC(PP),+DAI>)
```

Figure 2-9: Example Analysis Rules from JETS

As mentioned above, analysis grammars are typically collections of context free phrase structure rules augmented with various features that impose conditions on the applicability of the rules. Three sample rules for Japanese, taken from IBM's JETS system are shown in Figure 2-9. The first rule (5700), for example, states that a noun phrase (NP) followed by a postposition (POSTP) makes a postpositional phrase (PP), provided that the conditions specified as features on the PP node are met.

Analysis grammars can become quite large --- ranging anywhere from several hundred to several thousand rules. Note, however, that it is difficult to judge the coverage of a grammar based on its rule size, since the coverage of an individual rule can vary dramatically. In many cases, rules can call program subroutines, and so are able to perform complicated manipulations.

The crucial unresolved problem in this stage of the translation process is ambiguity --- both lexical and

syntactic. For instance, given the sentence, "He saw a dog with a rhinestone collar," a purely syntactic analysis cannot determine that the phrase "with a rhinestone collar" modifies "dog" rather than "saw." Compare, "He saw a dog with his telescope," where the opposite bracketing is appropriate. Quite often, an input string will result in a set of parse trees. Hopefully, the correct one will be among this set (although it might not be!). At this stage, the semantic features attached to words in the source language dictionaries will be used to block inappropriate structures. For instance, in the English examples above, the English dictionary might indicate the semantic information that "saw" is typically modified by a prepositional phrase "with ..." whose object is some sort of instrument for viewing. On the assumption that "telescope" but not "collar" is marked in the dictionary as "+ viewing instrument," the unwanted parse can be filtered out.

One can readily see from this simple example that the set of required semantic features could, for general English at least, be open ended. This is the main reason why attempts at filtering out all and only the bad parses have failed for the general (domain independent) case, but work somewhat better for specific domains. If one knows that an MT system is to be used for computer-related documents, for instance, and the word "disk" is used, one can be fairly certain that the meaning is the specialized term "computer disk" and not a record or simply a thin, circular object. However, even in specialized areas, the technique of using semantic features and compatibility testing to screen parses is faulty and certainly has limits. Overall, the problem of filtering sets of syntactically well-formed parses on the basis of semantic information seems to be the largest single obstacle to MT becoming an unequivocal technical success. Hence, this filtering problem is the motivating factor for exploring alternative approaches, such as the use of deep reasoning in interlingual systems (see Section 9.1) or the incorporation of probabilistic information based on specific corpora into conventional analysis procedures.

Figure 2-10, an example from IBM Japan's JETS system, illustrates the use of both inherent semantic features on nouns and selectional features on verbs, in conjunction with a scoring metric, to determine the most highly valued parse. Figure 2-11 illustrates in more detail the scoring mechanism used in JETS. MT systems differ in the number and kinds of features used, and they also vary according to whether they use syntactic-oriented case relations such as subject or direct object; semantically oriented case relations such agent or theme; or morphological cases such as, for Japanese, *ga* or *ni* (see Figure 2-10).

Figures 2-12 [Nagao 89], 2-13 [Nagao 85], and 2-14 [Nagao 89] give some indication of the variation in specific cases and relations used in various systems. However, regardless of the specific realization, the basic process is the same: (1) mark nouns listed in a dictionary with inherent semantic features such as "human" or "location" and (2) specify for each sense of every verb an *argument frame* or *case frame*, which shows what semantic features the various verbal arguments take, e.g., the agent of "know" must be a human; the agent of "eat" must be an animal; the intransitive "break" as in, "Glass breaks," takes a patient argument (realized as the subject), but the transitive "break" as in, "John broke the glass," takes both an agent argument and a patient argument.

There are two other types of grammar frameworks often used for parsing: transformational grammar, as exemplified by the MU/JICST system, and augmented transition networks, as exemplified by Toshiba's and CSK's systems. Figure 2-15 summarizes the major approaches to analysis.

Kare ni heya wo deru youni me de aizushi ta
("he") ("room") ("leave") ("eye") ("make a sign")
("I made a sign to him with my eyes to leave the room")

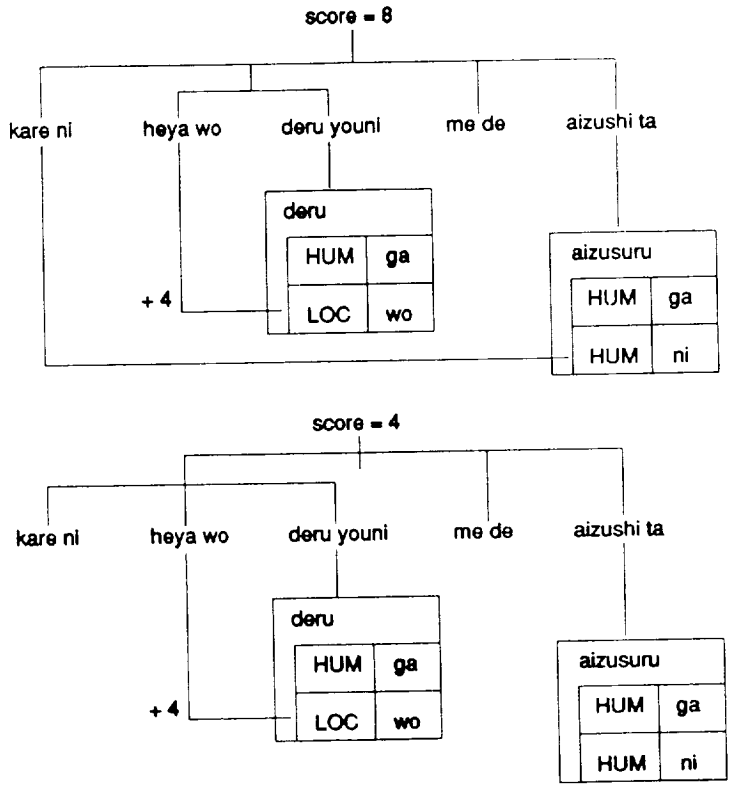

Figure 2-10: Using Selectional Restrictions and a Scoring Metric

	Sm = Sp	Sm ≠ Sp	Sm = Unknown or Sp = Unknown*
Cm = Cp	+4	+1	+2
Cm = Unknown+	+2	+0	+1
Otherwise	+0	+0	+0

Sm: Semantic feature of modifier
Sp: Semantic feature of predicate
Cm: Case particle of modifier
Cp: Case particle of predicate
+ Information not found in the lexicon
* Case article missing

Figure 2-11: The JETS Scoring Procedure

2.4 Transfer

The transfer stage consists, conceptually, of two subprocesses: (1) lexical transfer and (2) structural transfer. (These may be interleaved in practice.) Figure 2-16 illustrates lexical transfer for the simple Japanese sentence, *"Boku ga sakana o tabeta"* (I ate fish). The top of Figure 2-16 shows the mapping from the Japanese phrase-structure tree to the dependency structure, headed by the Japanese verb

(1) Subject	*The boy* walked home.
(2) Object	She found *the book.*
(3) Recipient	gave *to her.*
(4) Origin	received *from him.*
(5) Partner	to consult with ...
(6) Opponent	to protect from ...
(7) Time	in 1980 ...
(8) Time-from	from May of last year, ...
(9) Time-to	until next year, ...
(10) Duration	over a period of five minutes, ...
(11) Space	... is located at ...
(12) Space-from	to return from ...
(13) Space-to	to send to ...
(14) Space-through	to pass through ...
(15) Source	to translate from Japanese
(16) Goal	to translate into English
(17) Attribute	to be rich in ...
(18) Cause	to be due to ...
(19) Tool	... with a hammer ...
(20) Material	to be made out of ...
(21) Component	to consist of ...
(22) Manner	at a rate of ...
(23) Condition	to determine under the conditions of ...
(24) Purpose	adapted to ...
(25) Role	to use as ...
(26) Content	to be seen as ...
(27) Range	with regard to ...
(28) Topic	as for the topic of ...
(29) Viewpoint	from the perspective of ...
(30) Comparison	better than ...
(31) Accompaniment	together with ...
(32) Degree	an increase of 5%
(33) Predication	... is ...

Figure 2-12: The 33 Cases Used in the Analysis of Japanese in MU

1. AGenT	17. RANge
2. Causal-POtency	18. COmpaRison
3. EXPeriencer	19. TOOl
4. OBJect	20. PURpose
5. RECipient	21. Space-FRom
6. ORIgin	22. Space-AT
7. SOUrce	23. Space-TO
8. GOAl	24. Space-THrough
9. CONtent	25. Time-FRom
10. PARtner	26. Time-AT
11. OPPonent	27. Time-TO
12. BENificiary	28. DURation
13. ACCompaniment	29. CAUse
14. ROLe	30. CONdition
15. DEGree	31. RESult
16. MANner	21. ConCessive

Figure 2-13: The English Cases Used in MU

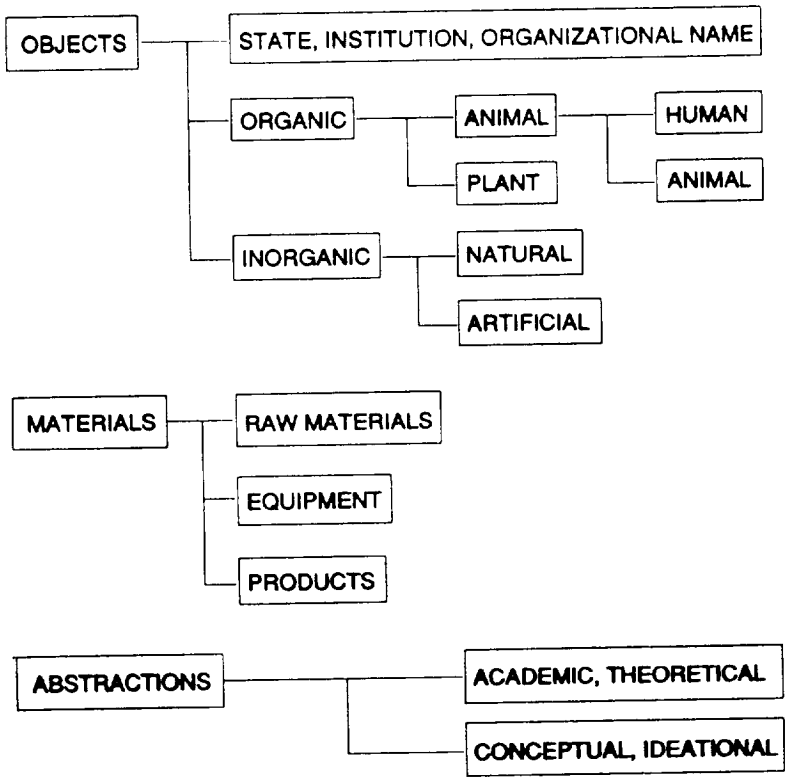

Figure 2-14: Example Semantic Primitives Used in MU

- Transformation Grammar
  - Subgrammars
  - Heuristic Rules
    - Ordered by reliability
  - Lexically triggered rules
- Augmented transition networks
- Augmented context-free phrase structure grammars
  - Syntactic/semantic features
  - GPSG, LFG
- PS ==> DS mapping
  - Semantic features
  - Parse scoring

Figure 2-15: Summary of Approaches to Analysis

tabeta (ate). The valence relations (superficial case relations) are encoded as features on the arguments; e.g., *boku* (I) is marked for *wa* (topic) and *ga* (subject), and *sakana* (fish) for *o* (direct object). Semantic interpretation maps the valence representation to a deep-case dependency structure in which *boku* (I) is marked as the agent and *sakana*(fish) as the theme of the clause.

The dictionary information needed to map superficial cases into deep semantic cases in this example is shown in Figure 2-17 (taken from [Nagao 86]). Next, lexical transfer replaces *tabe* with "eat", *boku* with "I", and *sakana* with "fish." Structural changes are not required in this atypically simple example.

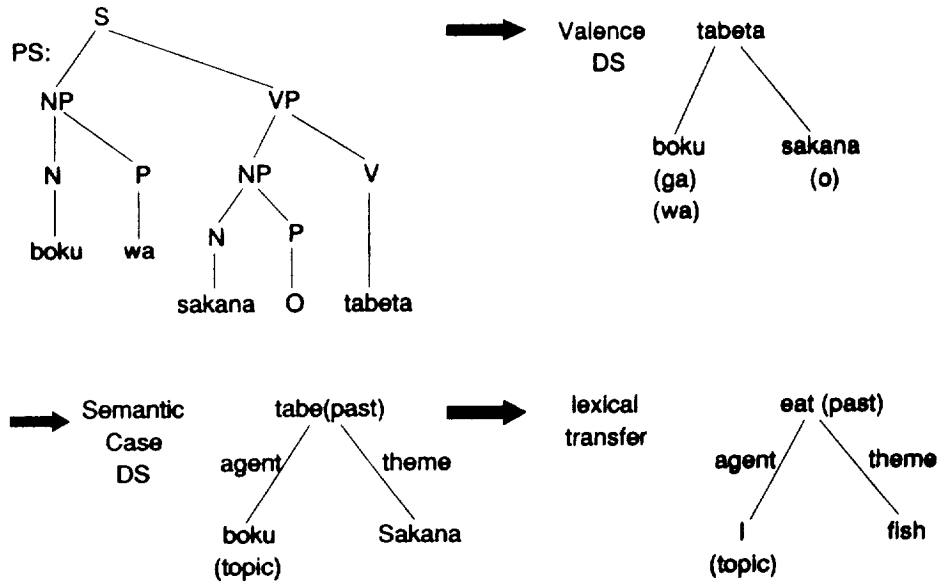

Figure 2-16: An Example of the Transfer Process

	Surface Case	Semantic Primitives of N	Deep Case
Eat	N-GA	Eatable Material	OBJECT
	N-GA	Animal	AGENT
	N-0	Thing	OBJECT

Figure 2-17: Example of a Dictionary Entry for "Eat"

J-Surface-Case	J-Deep-Case	E-Deep-Case	Default Preposition
ni	RECipient	REC,BENeficiary	to(REC--to,BEN--for)
	ORIgin	ORI	from
	PARticipant	PAR	with
	TIMe	Time-AT	in
	ROLe	ROL	as
	GOAl	GOA	to
	...	...	...

Figure 2-18: Default Rule for Assigning English Case to the Japanese Postposition *ni*

Lexical transfer can be a very difficult process, even without obviously tricky cases such as idioms and metaphors. For instance, Figure 2-18 shows the complexity inherent in translating the Japanese postposition *ni* into its English counterparts [Nagao 85]. As shown, the Japanese postposition *ni* can signal a variety of semantic relations that require different prepositions to be used in the corresponding English translations. The correct translation of particles requires a semantic analysis of sentences. That is to say, the complexity of translating even minor function words can be challenging.

Japanese Sentential Connective	Deep Case	English Sentential Connective
Renyo	tool	by -ing...
(-shi)te	tool	by -ing...
Renyo	cause	because...
(-shi)te	"	"
-tame	"	"
-node	"	"
-kara	"	"
-to	time	when...
-toki	"	"
-te	"	"
-tame	purpose	so-that-may
-noni	"	"
-you	"	"
-you	manner	as if
-kotonaku	"	without -ing...
-nagara	accompany	while -ing
-ba	circumstance	when...
...	...	...

Figure 2-19: Correspondence of Sentential Connectives between Japanese and English in the MU System

A similar problem arises in the translation of sentential connectives, as shown in Figure 2-19 (taken from [Nagao 86]). These tables do, however, encode in an accessible and easily maintainable format some of the basic linguistic knowledge necessary for transferring so-called function words such as postpositions/prepositions and sentence connectives. This sort of table-driven processing cleanly separates data from processing algorithms, and, although insufficent to encode all the knowledge needed for correct choices, represents a sensible beginning toward solving the general problem.

Structural transfer is a process by which the internal representation of the Japanese source sentence is restructured to provide a simple and sound basis for generating an appropriate English correspondent. Figure 2-20 shows an example of this process from Hitachi's HICATS-J/E system. The Japanese sentence, *"Kono sisutemu wa ouyouhan'i ga hiroi,"* means literally, "As for this system, application range

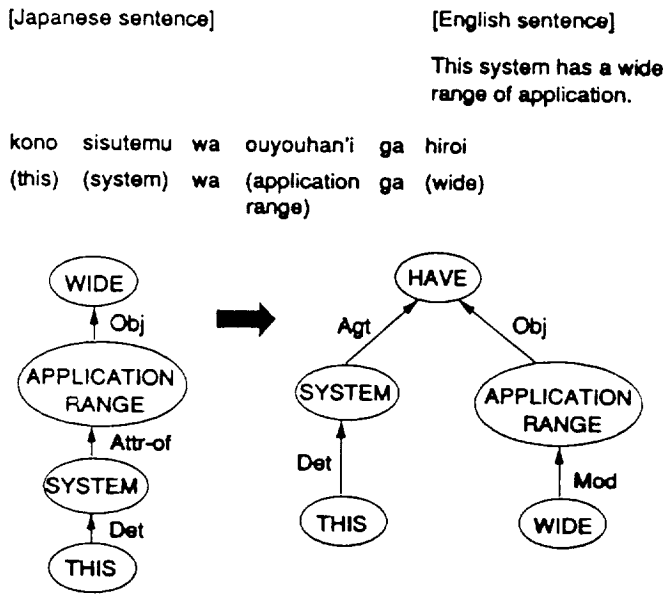

Figure 2-20: Transformation of Semantic Representations in HICATS

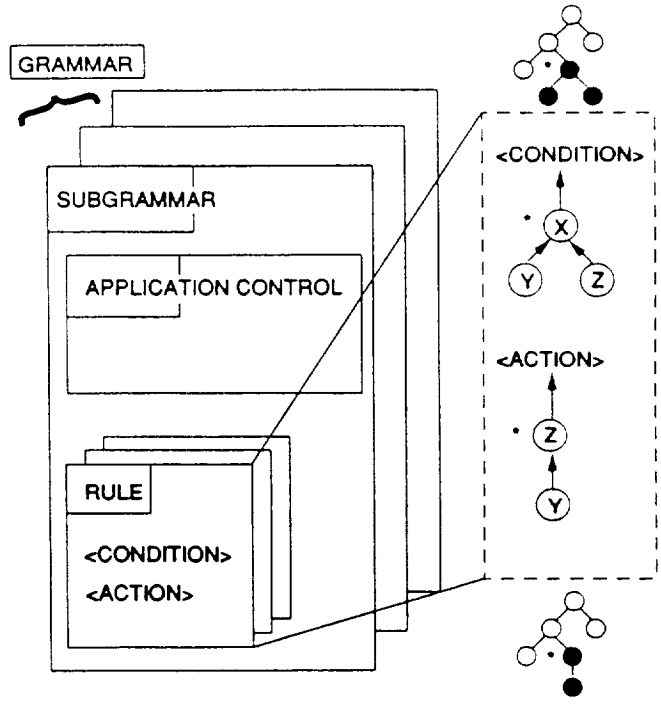

Figure 2-21: HICATS Grammar Description Language

is wide." The desired English correspondent is, "This system has a wide range of application." In Figure 2-20, a transfer rule — a tree transformation — restructures the Japanese internal structure (whose main predicate is "WIDE") into an English-oriented one (whose main predicate is "HAVE"). The English-oriented structure is then passed to the English generation grammar and results in the sentence, "This system has a wide range of application," as desired. HICATS uses a single grammar description language in which rules are organized into subgrammars and each rule is specified as a condition/action pair. (See Figure 2-21.) In addition, conditions controlling the application of rules can be specified, e.g., whether the rule is optional or obligatory. This high-level grammar organization seems fairly common, at least wherever transformational grammars are used.

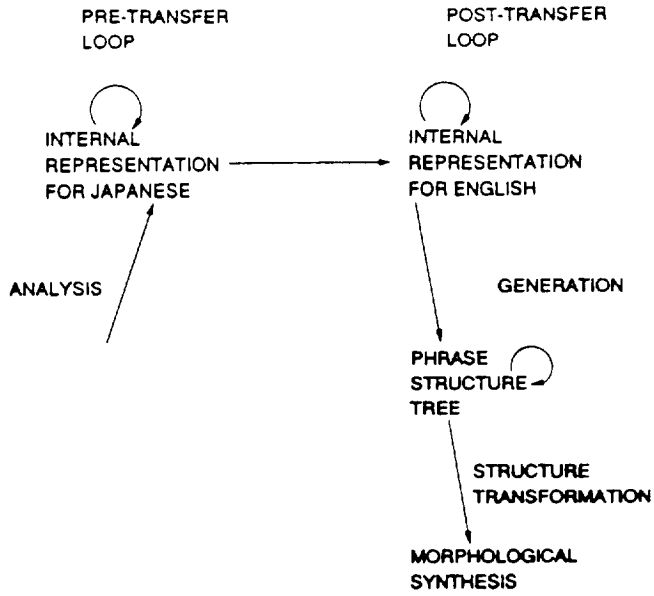

Figure 2-22: Transfer and Generation in MU

- Transformational Grammar (Tree Transduction)
- Stages
  - Pre-transfer: Structure Changing
  - Transfer Proper
    - Lexical Transfer
    - Structure Changing
  - Post-transfer: Structure Changing
- PS==>DS (Valence Representation)
- DS==>DS (Semantic Case Representation)

Figure 2-23: Summary of the Transfer Phase

Transfer grammars typically consist of a large group of transformational rules that restructure dependency trees. Transfer grammars can be very complicated. One aspect of this complexity is the triggering of subgrammars by specific lexical items. This data-driven organization, while flexible, makes it difficult to understand the processing flow, since changing the dictionary can have far-ranging effects on which rules are applied. It is also not atypical for the transfer component to have three stages involving structural transfer: (1) pre-transfer, (2) transfer proper, and (3) post-transfer (as shown for the MU system in Figure 2-22 [Nagao 85]). The distinction among these phases is somewhat arbitrary but provides a productive distinction in practice. The main stages and functions of the transfer stage are summarized in Figure 2-23.

2.5 Generation

As discussed, the final stage, generation (or synthesis), takes the output of transfer and ideally produces a well-formed sentence in the target language. In most systems, transfer output is converted to a phrase-structure tree, which includes the proper word order for the target language sentence.

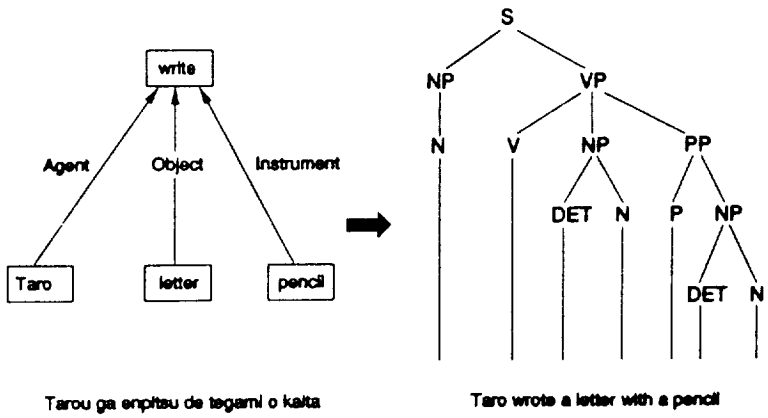

Figure 2-24: A Generation Example from HICATS

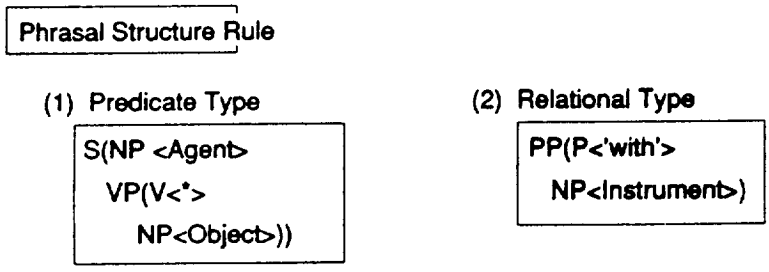

Figure 2-25: A Generation Rule from HICATS

This is illustrated in Figure 2-24. This transduction is carried out either by applying a set of transformational rules or a set of phrase-structure templates, such as those shown in Figure 2-25. For example, the predicate type rule in Figure 2-25 encodes the fact that a dependency pattern headed by a

verb and containing an agent and object can be mapped into the constituent structure pattern (S NP (VP V NP)) such that the agent corresponds to the first NP and the object to the second one. The relational type rule, also shown in Figure 2-25, maps prepositions and their objects into PPs (prepositional phrases), where the objects are identified by their semantic relations (e.g., "instrument"). The two generation rules in Figure 2-25 encapsulate the linguistic knowledge needed to map the dependency structure shown on the left in Figure 2-24 into the constituent-structure tree shown on the right.

In many MT systems, the generation component is, compared to analysis and transfer components, relatively small; that is, much of the work needed to construct a target language sentence is done in the structure-changing part of the transfer process. Generation is basically viewed as a clean up stage, and as such will often consist of ad hoc routines rather than robust grammars that accurately reflect the linguistic facts of the target language. Hence these grammars are not bidirectional, that is, they cannot be used for both parsing and generation. Figure 2-26 summarizes the major techniques and grammatical frameworks used in generation.

- Transformational Grammar
- Augmented Transition Network
- Augmented Context Free Phrase Structure Grammar
- Ad-Hoc Program (not Grammar-Based)
- Typically:
  - Minor Component
  - No Independently Justified Grammar

Figure 2-26: Summary of Techniques Used in Generation

3. Languages and Application Domains

Muriel Vasconcellos

3.1 Current Range of Source and Target Languages

Reflecting the nation's political, economic, and social imperatives, Japan's machine translation activities have focused largely on English and Japanese. There are at least 20 MT systems in Japan whose developers are addressing the complex challenge posed by translation between these two languages. The extent to which the corresponding knowledge sources have been developed in English and Japanese varies in proportion to the history of system development. Several of the projects date back 8, 10, and 12 years. The present discussion refers to all systems---long-standing, new on the market, and research prototype---since what is of interest here are Japan's priorities and an appreciation of where investments are being made.

Figure 3-1 shows the respective language combinations being developed under the 20 MT initiatives for which the JTEC team had information.[1] Figure 3-2 shows the number of sites at which each combination is being developed, thus giving a rough indication of the relative importance of the different languages for Japan. The intent of both of these tables is to show the diversity that was observed. But it should be kept in mind that they do not accurately reflect the distribution of actual effort on the various language pairs, since some entries correspond to well-developed systems, while others represent small, experimental prototypes.

Of the sites visited by the JTEC team, 17 have already developed or are in the process of developing MT systems that translate from Japanese into English (J/E), from English into Japanese (E/J), or in both directions. In the majority of cases, initial efforts were concentrated on Japanese into English because of the far greater demand for that combination. At the same time, however, there has also been considerable motivation to develop English into Japanese. The difference, of course, is that in the case of J/E the target market is foreign with Japanese information being disseminated to wider audiences; with E/J it is national and information from overseas is being assimilated for the people's own use. Work on this combination is attractive for several reasons: (1) in Japan there is an important demand for information translated from English; (2) it is the "easier" of the two combinations to develop, since, by comparison, Japanese source analysis is a far more daunting challenge; and (3) posteditors are available in much larger numbers and are less expensive to hire and train. In fact, despite the greater demand for J/E, there are nearly as many offerings for E/J: from Japanese into English there are 17 systems, and from English into Japanese there are 15 (see Figure 3-2). Thirteen sites have systems in both directions. In addition, the JEIDA Report [JEIDA 89] mentions CBU's HANTRAN, for E/J, and Mitsubishi's MELTRAN, for J/E.

After years of concentration on English and Japanese, some of Japan's MT developers have recently ventured to add other languages. A major effort is being undertaken by the Center for the International Cooperation in Computerization (CICC), a MITI-organized international consortium that has the participation of seven industry giants, for developing an interlingua-based system in Japanese, Chinese,

---

[1]The information in this table is drawn primarily from the panel's visits to the sites mentioned. The exceptions to this are the lines corresponding to CBU's HANTRAN system and Mitsubishi's MELTRAN system, which are taken from [JEIDA 89].

Developer	J/E	E/J	Other
ATR	•	•	
Bravice	•	•	E/C, /K, K/E, E/F, /G, /I, /P, /S, F/E, S/E
Catena		•	F/J
CBU		•	
CICC			J/C/Indonesian/Malay/Thai (all pairs)
CSK	•	•	
Fujitsu	•	•	E/C, /K, /F, /G, /S, /Innuit, /Swahili, J/C, /K, /F, /G, /S
Hitachi	•	•	
IBM	•	•	E/C, /K
JICST	•		
Matsushita	•		J/C
Mitsubishi	•		
NEC	•	•	K/S/Thai (all pairs)
NTT	•		
Oki	•	•	J/C
Ricoh	•	•	
Sanyo	•	•	
Sharp	•	•	
Systran Corp.	•	•	
Toshiba	•	•	

Key: C=Chinese, E=English, F=French, G=German, I=Italian, J=Japanese,
K=Korean, P=Portuguese, S=Spanish.
E/J means English to Japanese.

Figure 3-1: Source and Target Language Combinations in Japanese MT Systems, by Site

Thai, Malay, and Indonesian. This undertaking is described in greater detail in Section 9.9.

Efforts at adding new languages to existing interlingua-based systems are being pursued at Fujitsu and NEC. Fujitsu is developing experimental versions of the ATLAS-II system that accept English or Japanese as source languages and can generate target text in German, French, Spanish, Chinese, Korean, Japanese, and English. There have also been experiments with Swahili and Innuit. NEC's PIVOT project has started research on Korean, Thai, and Spanish. ATLAS-II and PIVOT are likely to be headed for commercial markets in the West, and in fact ATLAS-II is already known in Europe and was recently launched in the United States.

Matsushita, collaborating with researchers in China, has started work on a Chinese target for the Japanese source component of PAROLE. Japanese-to-Chinese is also the subject of development

Target / Source	Chinese	English	French	German	Indonesian	Innuit	Italian	Japanese	Korean	Malay	Portugese	Spanish	Swahili	Thai
Chinese					1			1		1				1
English	4		2	2		1	1	15	4		1	3	1	1
French		1						1						
German														
Indonesian	1							1		1				1
Innuit														
Italian														
Japanese	4	17	1	1	1				3	1		2		2
Korean		2						1						
Malay	1				1			1						1
Portugese														
Spanish		2						1	1					1
Swahili														
Thai	1	1			1			2	1	1		1		

Figure 3-2: Source and Target Language Combinations, by Languages

efforts by Oki Electric, which is adding Chinese as a target language to PENSEE, under a project being carried out in conjunction with Nanjing University [Wang 90]. Catena's STAR system is being enhanced with a capability of translating French into Japanese [Aizawa 90].

The two other companies whose list of MT languages goes beyond English/Japanese have a history of MT involvement that began outside Japan, and their objectives do not necessarily parallel those of Japanese industry. IBM is developing English/Chinese and English/Korean for the translation of documentation to support the sale of U.S. products. Bravice (which appears to have gone out of business in early 1991) informed the JTEC team that it had seven pairs of European languages in various stages of development for the personal computer (English into French, Spanish, Portuguese, Italian, and German; French into English; and Spanish into English), as well as a PC system translating from English

into Chinese. In addition, Bravice subcontracted with Executive Communication Systems in Provo, Utah, for the development of a bidirectional English/Korean prototype, which was delivered through TRW's Federal Systems Group to the U.S. Signal Corps. Bravice had also started work on a Japanese/Korean translation capability.

3.2 Addition of New Source and Target Languages

A few of the companies that now have one direction only, either E/J or J/E, have immediate or future plans to add the other direction. In the near term, ATR and JICST plan to expand into English/Japanese. Fujitsu intends to upgrade the Japanese source component of ATLAS-II so that it will be sufficiently robust to stand alongside English as a multitarget source for its other languages.

In addition to the languages mentioned in the previous section, a few others are in the wings, mainly for the interlingua-based systems. ATLAS-II, which already has an impressive inventory, will be expanded to cover additional European languages. CICC, in turn, may eventually include English in its suite of Asian languages. CSK plans to expand its ARGO system into the languages of Western Europe and Southeast Asia.

Except as noted so far, it would appear that the typical transfer-based developer visited by JTEC, rather than adding new languages, would prefer to focus efforts on improving the accuracy of their J/E and/or E/J systems, expanding into new domains for existing combinations, building up knowledge sources, integrating into the electronic publication chain, and developing user-friendly interfaces, customized tools, and other nonlinguistic enhancements.

Clearly cost is an important factor in determining the rate at which systems are extended to new language combinations. There is good reason to expect that it will be relatively less costly to add new target languages to systems that have a well-developed source analysis module coupled with a robust set of transfer rules or an interlingua. It should be kept in mind, however, that with interlingual systems considerable effort is required in order to bring a source language up to full multitarget capability.

Several developers provided the JTEC team with information on the cost of adding new languages. Systran quoted a cost of US$100,000 to add a new target language to an existing multitarget source, regardless of the language. Dictionary-building is extra.[2] On the other hand, a new, fully operational multitarget source language may cost anywhere from 10 to 50 times as much---US$1,000,000 for an Indo-European language such as Czech or Norwegian, US$2,000,000 for a Roman alphabet non-Indo-European language such as Hungarian or Turkish, US$3,000,000 for a non-Roman alphabet non-Indo-European language for which existing work in Japanese could be utilized (e.g., Chinese or Korean), and US$5,000,000 for an entirely new project such as Arabic. A second estimate came from Bravice, where the company's president, Takehito Yamamoto, estimated that to add a new language combination if one of the languages is English takes about 250 person-months, and if both languages are new, at least 480 person-months. Matsushita indicated that to add a new source or target language to PAROLE would take about 20 person-years. Fujitsu reported that the addition of a new language takes three to five years at 5

[2]Figures supplied by Denis Gachot, president of Systran Translation Systems, Inc. According to Gachot, dictionary development costs US$3.00 for a stem entry and US$6.00 for a multiword expression. Entries of the latter kind represent about 20% of the total dictionary. For an application in a limited domain, the dictionary should have from 40,000 to 60,000 entries; for a general-purpose application it should have from 100,000 to 150,000 entries.

to 10 people per year plus additional resources for dictionary development.

3.3 Application Domains, Domain Adaptability

It is useful to distinguish between special-purpose and general-purpose MT systems. Special-purpose, or domain-specific, systems are designed to handle text in a limited subject area that has fairly predictable linguistic structures and vocabulary. Depending on the area covered in the domain, there may be very few ambiguities to resolve (i.e., few meanings and readings to make decisions about), and these decisions can be facilitated by the use of a knowledge base that gives a full range of attributes for the agents, objects, etc., in the text. The lexicon need not be very large (up to 60,000 stem entries). Development costs are relatively low by MT standards. From the user's perspective, the output is fairly stable, requiring considerably less human intervention than a general "try-anything" application.

The general-purpose system is a much greater challenge for MT: there are many ambiguities, the MT dictionary must be rich with coding if it is to support the choices that are required, and it must have far more entries---at least 100,000 to 150,000. The MT product, because it is less predictable, requires greater human intervention.

Paralleling these contrasts, the domain-specific system is usually for the dissemination of information and therefore calls for high standards of quality. It is often used for translation from a single source to multiple target languages. The general-purpose system, on the other hand, may be used for the assimilation of information over a broad range of topics and can be useful even if quality standards are relaxed---as indeed they must be if the cost of human intervention is to be minimized. In such applications there may be several source languages translated into a single target.

There is a clear market for both domain-specific and general-purpose systems. Domain-specific systems can be used in many applications that are of significant technical and commercial interest. General-purpose systems are also important. They are more interesting commercially because they can attract a broad range of clients. As a result, many companies are working with this goal in mind. Sometimes general-purpose systems can be successfully customized for special applications as well. They are amenable to such downsizing when they have the capacity to process the syntactic and semantic information needed for eliciting context-sensitive translations---i.e., they are domain-adaptable.

In Japan, a number of systems began by being domain-specific, focusing for the most part on the area of computer manual translation, which is said to represent 80% of all MT use in the country. Often their developers were hardware manufacturers who saw MT first as a tool for helping them to reach overseas markets and second as a potential commercial product. Typically, a system begins as domain-specific and gradually progresses to be domain-adaptable and ultimately general-purpose. (See Chapter 5.) This was the case with CSK's ARGO (expanding from tightly-worded financial bulletins to economic texts and now branching out into political and social areas and biotechnology): Fujitsu's ATLAS-II (banking, pharmaceuticals, chemicals): Hitachi's HICATS (information processing, electronics, computers, civil engineering, construction, transportation, natural science, biology, machinery, chemicals, metals); NEC's PIVOT (aviation E/J, navigation J/E): Oki Electric's PENSEE (medicine, finance, electronic communication), and Toshiba's ASTRANSAC (general-purpose, with domains such as information science, electronics).

On the other hand, some systems have been designed and developed for more general-purpose

translation from the start. The JICST system, for example, translates database abstracts in a variety of scientific and technical fields. Since this was its goal from the beginning, a substantial amount of work has gone into building up its technical lexicon. Some of the preceding systems are also being used for general purposes in translation bureaus. Another that was developed from the start as a general-purpose system is Sharp's DUET E/J. Perhaps the most challenged MT system in Japan is Catena's STAR at NHK (Japan's public television station), which is being used to write subtitles for excerpts from English-language news stories, which may be on any topic that is worthy of headlines. STAR is also being used on an experimental basis to monitor incoming newswire bulletins on a near-real time basis. It is not a coincidence that all the projects mentioned have been engaged in MT development for more than a decade.

The trend in Japan of starting with a specific application and working up to a general-purpose system contrasts somewhat with patterns in the West, where there seems to be more of a dichotomy between special- and general-purpose systems. However, commercial general-purpose systems are often domain-adapted for a specialized application---for example, SYSTRAN at Xerox Corporation---and perform sufficiently well to meet the needs of the user.

The range of language combinations and domains, many of which have been the subject of intensive dictionary work, show that Japan is moving ahead persistently on a broad front. This strategy, when followed appropriately, cannot fail to contribute to the performance of Japanese MT systems.

4. Knowledge Sources for Machine Translation

Yorick Wilks

4.1 Overview of Knowledge Sources

The knowledge sources needed to perform MT depend at least to a limited extent on the MT method used. For example, some current U.S. projects (such as the work at IBM on English to French MT [Brown 89]) make use of very large-scale statistical information from texts, while Japanese systems do not. Conversely, an experimental MT system at Kyoto University uses large lists of sample sentences against which a sentence to be translated is matched [Sato 90], whereas currently no U.S. systems do this. Most MT systems, however, make use of at least some of the following kinds of knowledge sources:

- Morphology tables
- Grammar rules (including analysis, generation, and transfer rules)
- Lexicons
- Representations of world knowledge

Sometimes the first, second, or fourth of these knowledge sources may be absent. For instance, it is possible both to analyze and to represent the English language without the use of morphology tables, since it is inflected only to a small degree; for the analysis of a highly inflected language like Japanese, on the other hand, they are almost essential. Some analysis systems do not use an independent set of identifiable grammar rules, but these systems must somewhere contain syntactic information, such as the fact that in English an article precedes a noun. Although there is room for doubt as to whether certain items of linguistic knowledge belong in morphology or grammar (in Italian, for example, forms like pronouns may stand alone, but can also act as suffixes to verbs: e.g., *dagliela*), in Japanese and English this ambiguity of type is very unlikely. The fourth category is even more uncommon. Only some MT systems (usually those that owe some allegiance to artificial intelligence methods) claim to contain world knowledge representations.

The third form of knowledge (lexical) appears in virtually every MT system, however, except for the purely statistical type of system referred to earlier. Unfortunately, the distinction between lexical and world knowledge can also be tricky. In a German lexicon, for example, *Das Fraulein* is marked as neuter in gender but, in the real world, it must be marked feminine, since the word means "a young woman." We should deduce from this that a lexicon is a rag-bag of information, containing more than just semantic information about meanings.

Although there are many ways in which the development of these knowledge sources can take place, we will mention one example. Toshiba developed its knowledge sources in the chronological order shown in Figure 4-1. Because the role of knowledge sources within an MT system depends so heavily on the overall structure of the system, it is interesting to look also at Toshiba's overall translation procedure, which is shown in Figure 4-2. Without committing to a specific view of what "semantic transfer" means, we can infer that the bolder arrows represent the translation tasks to be performed, while the lighter arrows indicate Toshiba's view of where the knowledge forms they emphasize distribute across those tasks.

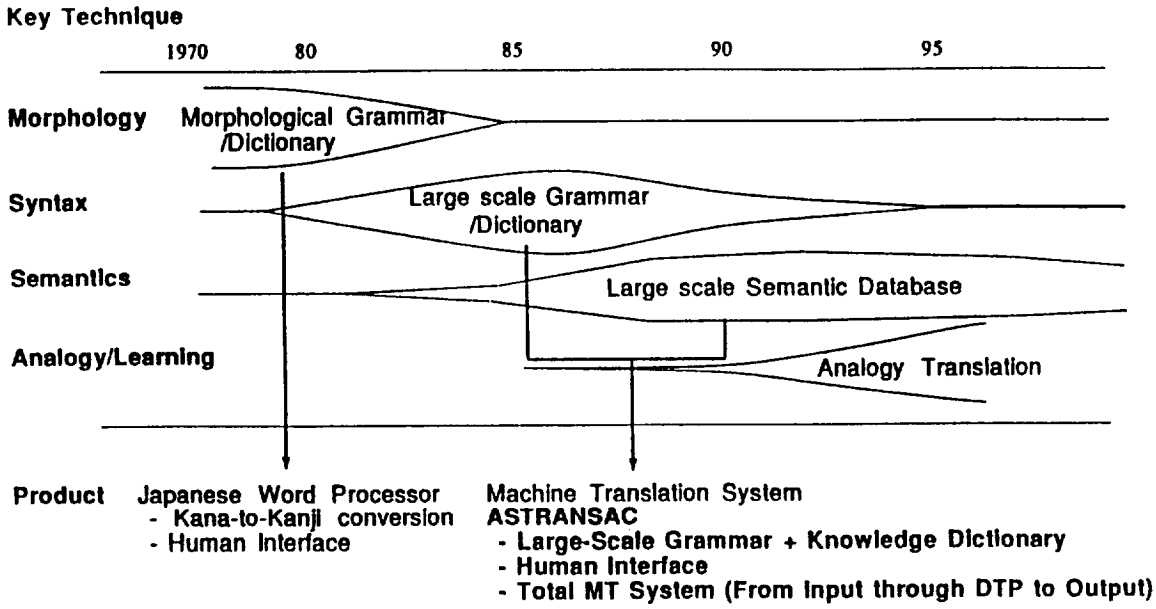

Figure 4-1: Toshiba's Development of Knowledge Sources

4.2 Use of Knowledge Sources in Specific Japanese MT Systems

Much of this chapter's content could be summed up by the tables shown in Figures 4-3 and 4-4, which list 22 systems by their major features, such as the type of MT system (direct, transfer, or interlingual[3]), the major language directions expressed as letter pairs (e.g. J/E for Japanese to English), the type of grammar (ATN's, case-frame, etc. - see Chapter 2), the number of rules (if available), the lexicon's size and type (also if available), and any kind of knowledge representation that is used.

One noticeable feature of the table is that only one MT system explicitly claims to use a knowledge representation: IBM Japan's SHALT2 uses the Framekit-based system. Also note that although the EDR Electronic Dictionaries have been included on the chart, they are not an MT system, but a very large scale set of lexical and conceptual tools, as described below.

What NTT (in the system type column) describes as J/J transfer means extensive, automatic, pre-editing to (a) remove character combinations known to cause problems for analysis programs and (b) insert segmentation boundaries into sentences to break them into sections, making analysis easier [Ikehara 91]. This process is also called source-to-source translation. Variations of these methods were found at other sites (e.g., Sharp and NHK), and although (b) originated in earlier MT practice, these methods constitute a practical heuristic that has almost certainly improved translation quality.

[3]In Chapter 1, these terms were defined. But we also presented there a caveat about their use, since this terminology has not been standardized. Here, as in Chapter 1, we use the developer's words to describe each system.

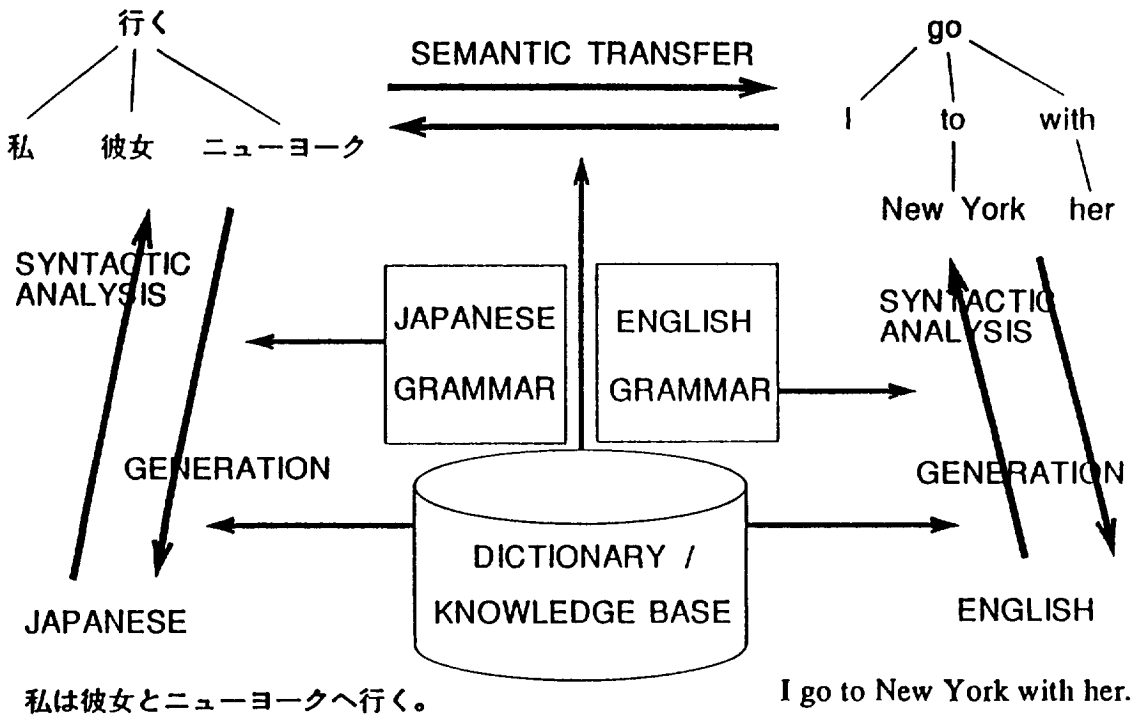

Figure 4-2: The Translation Process in the Toshiba System

4.3 Knowledge Sources and Linguistic Theory

Techniques such as source-to-source translation are interesting because they fall under the rubric of what Bar-Hillel (wrongly believed by many to be the arch-enemy of MT) described when he wrote that "MT research should restrict itself, in my opinion, to the development of what I called before 'bags of tricks' and follow the general linguistic research only to such a degree as is necessary without losing itself in Utopian ideas [Bar-Hillel 71]."

More than U.S. projects, and much more than European projects like EUROTRA, Japanese MT work has arrived at the same conclusion as Bar-Hillel. Very little Japanese work owes much to Western-style linguistic theory beyond some general use of "case frame" and some concepts taken from unification grammar. Instead, it has developed its own indigenous, Japanese tradition of linguistic description, as exemplified in the work at NTT.

If accurate, this observation is a reason to re-examine the Western notion of "knowledge sources." Given that the list at the beginning of this chapter took its categories directly from Western linguistics and does not tailor itself very well to Japanese MT work (if one agrees that the most successful Japanese systems are mainly driven by the information in their lexicons, as is SYSTRAN), then our very

Company	System Type	Grammar	Lexicon	Knowledge Representation
ATR	Semantic Transfer E/J	Lexically-Based Unification Grammar (JPSG) 130 Rules	Case-Roles Thesaurus	—
Bravice	Syntactic Transfer J/E & E/J	J/E 4K Rules E/J LFG/UNIFIC 8K Rules	J/E: 70K Basic 240K Technical E/J: 40K Basic	—
Catena STAR	Syntactic Transfer E/J	2,000 Context-Free Rules	20K Basic 55K Technical	—
The Translator	Syntactic Transfer E/J	3,000 Context-Free Rules	25K Basic 35K Technical	—
CICC	Interlingual (J,C,TH,IN,MAL)	—	50K Basic 25K Technical	—
CSK	Interlingual J/E & E/J	ATNS	50K	—
EDR (Not an MT system)	Implicitly Interlingual	—	J: 300K E: 300K	400K Concepts in Concept Dictionary
Fujitsu ATLAS-I	Syntactic Transfer J/E	—	—	—
ATLAS-II	Interlingual J/E	5K J Rules 5K E Rules 500 Transfer Rules	70K Each Way +300K Technical in subdictionary	—
Hitachi	Semantic Transfer J/E & E/J	Case-Based J/E: 5,000 Rules E/J: 3,000 Rules	J/E 50K E/J 50K	—
IBM SHALT	Syntactic Transfer E/J	Phrase Structure 200 E Rules 800 Transfer Rules 900 J Rules	E/J 90K	—
IBM SHALT2	Semantic Transfer E/J	—	LDOCE	Framekit-Based Representation

Figure 4-3: Knowledge Sources in Japanese Systems

assumptions of what knowledge sources actually drive MT should be reassessed.

For the sake of clarity, it may be profitable to return to the notion of a knowledge source, and to throw some additional light on it by contrasting it with MT without knowledge sources.

Earlier we mentioned that current work at IBM/Yorktown Heights performs English to French (E/F) MT without help from any of the knowledge sources we listed above. Even its definition of what constitutes a word is derived from frequent collocation measures of other "words," and therefore is not a priority. That

Company	System Type	Grammar	Lexicon	Knowledge Representation
JICST	Semantic Transfer J/E	1500 J Rules 500 Transfer Rules 450 E Rules	350K J 250K E	----
Matsushita	Syntactic Transfer J/E	800 J Rules 300 Transfer Rules	31K Each Direction	----
NEC	Interlingual J/E & E/J	Case-Frame Grammar	90K J/E, 70K E/J +600K Technical Lexicons	----
NTT	Syntactic Transfer J/E (J/J Transfer)	----	400K (Includes 300K Proper Nouns)	----
Oki	Syntactic Transfer J/E & E/J	Context-Free Rules 1K Rules Each Direction	J/E: 90K E/J: 60K	----
Ricoh	Syntactic Transfer E/J	2500 Rules 300 Transfer Rules	55K	----
Sanyo	Syntactic Transfer J/E & E/J	Context-Free Phrase-Structure 650 Rules Each Way	50K Each J/E & E/J	----
Sharp	Semantic Transfer J/E & E/J	Augmented Context-Free and Case Frames	J/E: 70K E/J: 79K	----
Systran Japan	Transfer J/E & E/J <->	Multi-Pass Phase-Finding	E/J 200K (Interpress) J/E 50K	----
Toshiba	Semantic Transfer J/E & E/J	ATN+Lexical Rules 100K Rules Each Way	50K General <200K Technical <200K Users	----

Figure 4-4: Knowledge Sources in Japanese Systems

system generates word strings that connect to form statistically "natural strings," frequently at the expense of their relation to anything in the source text. This is accomplished without any knowledge sources: that is, without any analytic, combinatory, or symbolic structures.

An interesting example at the other end of the spectrum is SYSTRAN [Toma 76]. Although SYSTRAN is primarily an American system and thus really outside the scope of this report, its J/E and E/J modules are still owned in Japan by Iona International Corporation. The contrast between SYSTRAN and the IBM E/F system is instructive here, partly because the systems' goals are to translate by symbolic and statistical methods, respectively,and partly because it is SYSTRAN's "sentence correct percentage" that IBM would have to beat to be successful although it is nowhere near doing so at the present time.

SYSTRAN has also been described at least in parody as utilizing no knowledge sources; it has been

thought of by some as having, in effect, a mere sentence dictionary of source and target languages. Nor is this notion as absurd as linguists used to think: the number of English sentences under fifteen words long, for instance, is very large, but not infinite. So, based on the preceding definition, an MT system that did MT by such a method of direct one-to-one sentence pairing would definitely not have a knowledge source. But, although part of the success of the SYSTRAN Russian/English system installed at the U.S. Air Force's Foreign Technology Divison is certainly due to its 350,000-word lexicon of phrases, idioms, and semi-sentences [Wilks 91], SYSTRAN does not really conform to this parody of it [Toma 76]. Moreover, the new version of SYSTRAN, according to their president, is being re-engineerd with a more conventional modular structure and explicit knowledge sources.

One might say that while U.S. and European systems tend to fall toward the extremes of a spectrum (running from linguistically-motivated systems at one end to those with no knowledge sources at the other), Japanese systems tend to fall in between, and to have *sui generis* knowledge sources, as does SYSTRAN itself.

Another way of thinking about knowledge sources for MT is that they are never completely pure data in the way that linguistic theory sometimes supposes. That is to say that the role and content of a knowledge source cannot really be understood without some consideration of the processes that make use of it.

4.4 Lexicon Samples

Figure 4-5 shows an example from the ATR lexicon, and is for the verb *kudasai.* It is unusual in that it is a lexical entry for a strongly linguistically-motivated system; indeed, one can deduce from its structure that it is almost certainly intended to fit within an HPSG[4] grammar system. This confirms that such knowledge sources are not independent of the processes that apply to them.

It is important to emphasize once more the paramount role of lexicons in many Japanese systems, their substantial size (and the manpower required to construct them), and the wealth of specialized technical lexicons available in some of these systems. For example, Figure 4-6 shows the set of 13 technical lexicons available for the Fujitsu ATLAS system. These are in addition to the basic dictionary, which contains about 70,000 entries. The effort required to build an MT dictionary depends on several factors. We were given several different estimates for the rate at which system builders could add new terms to the dictionary, ranging from five entries/hour (Matsushita) to six person years to customize a dictionary for a new application (Hitachi).

As noted, SYSTRAN is a strongly lexically-dependent MT system. SYSTRAN's J/E and E/J modules have three types of dictionaries described by the company in [SYSTRAN 91] as:

- A "word boundary" dictionary for matching words and establishing word boundaries in Japanese text, where each word is not clearly bounded by spaces (as in English and other European languages).
- A "stem" dictionary containing source language words and their most frequently used target language equivalents. This dictionary also contains morphological, syntactic, and semantic information about each entry word.

[4]Head-Drive Phrase Structure Grammar [Pollard 87].

ɔ.3.4 依頼

人に動作をするように頼む場合のムード。直接依頼の「下さい」「てもらう」のみ登録されている。

直接依頼形式 直接相手に動作の依頼をする。

間接依頼形式 自分の実情を述べて、相手に間接的に動作の依頼をする。

- ～て / て下さい / てちょうだい
- ～てくれませんか / てもらえますか / てもらえませんか
- ～てほしい / てもらいたい / てほしいんだけど
- ～てくれるといいんだが / てくれるとありがたいんだけど

「下さい」の語彙記述の例

「下さい」は、動詞テ形(「送って下さい」)、または、サ変名詞(「御参加下さい」「お送り下さい」)を下位範疇化する補助動詞として記述されている。

```
([[PHON (:DLIST        kudasai
                |?X07| )]
  [SYN [[SLASH [DLIST[IN ?X04[]]
                     [OUT ?X04]]]
       [HEAD [[POS  V]
              [GRFS [[SUBJ [[SYN [[SUBCAT (:LIST  )]
                                  [HEAD [[POS  P]
                                         [FORM  ga]
                                         [COMPLEMENT  +]]]]]]
                           [SEM ?X02[]]]]]
              [ASPECT  +]
              [VASP [[CHNG  +]
                     [HOME  -]
                     [ACTV  +]]]
              [SUBV  +]]]
       [WH [DLIST[IN ?X05[]]
                 [OUT ?X05]]]
       [MORPH [[CTYPE  HONC]]]]
  [SEM [[RELN  下さい-REQUEST]
        [AGEN ?X03[]]
        [RECP ?X02]]]
  [PRAG [[RESTR (:DLIST        [[RELN  RESPECT]
                                [AGEN ?X03]
                                [RECP ?X02]]
                               [[RELN  POLITE]
                                [AGEN ?X03]
                                [RECP ?X02]]
                         |?X01| )]
         [SPEAKER ?X03]
         [HEARER ?X02]]]
  [ORTH (:DLIST        下さい
               |?X06| )]]
 ([[SYN [[MORPH [[CFORM  INFN]]]]]]) ; 連用形(「下さいますか」)
  ([[SYN [[MORPH [[CFORM  IMPR]]]]]]; 命令形(通常の用法)
  ([[SYN [[SUBCAT (:LIST        ?X11[[SYN [[SUBCAT (:LIST        ?X10[[SYN [[SUBCAT (:LIST  )]
                                                                             [HEAD [[POS  P]
                                                                                    [FORM  ga]
                                                                                    [COMPLEMENT  +]]]]]
                                                                   [SEM ?X08[]]]
                                                             )]
                                                       [HEAD [[POS  ADV]
                                                              [GRFS [[SUBJ ?X10]]]]]]]
                                                 [SEM ?X09[]]]
                                    )]
          [HEAD [[MODL [[EVID  DIRC]]]
                 [GRFS [[SUBJ [[SEM ?X08]]]
                        [COMP ?X11]]]]]]]
    [SEM [[RECP ?X08]
          [OBJE ?X09]]]
```

Figure 4-5: An Example Entry from the ATR Dictionary

Field	Number of Entries
Biology and Medicine	9,200 words
Industrial chemistry	14,400 words
Meteorology, Seismology, and Astronomy	13,500 words
Mechanical engineering	28,100 words
Civil engineering and Construction	14,400 words
Physics and Atomic Energy	15,000 words
Transportation	21,800 words
Plant	36,000 words
Automobile	18,000 words
Biochemistry	15,000 words
Information processing	26,000 words
Electircity and Electronics	17,100 words
Mathematics and Information	31,900 words

Figure 4-6: Technical Lexicons Available for the ATLAS System and Used to Supplement the Basic General-Purpose Lexicon

- A "limited semantics" (LS) dictionary of expressions, special collocations, and macro instructions for handling translation problems of low to medium complexity.

These are accessed within the main SYSTRAN framework, as shown in Figure 4-7. SYSTRAN's dictionary list for its newer multitarget systems is shown in Figure 4-8. The English-Japanese component (at 150,000 source items) is about three times the size of the corresponding J/E dictionary, which alone would account for its superior quality in the sample test conducted during the JTEC visit. This newer system is called multitarget because SYSTRAN has now fully integrated its earlier methodology of detaching and reusing chunks of older programs for new languages. SYSTRAN is now described as a transfer rather than a direct system. This is an interesting evolutionary, bottom-up approach to design development.

A sample of SYSTRAN's small J/E dictionary is shown in Figure 4-9. This dictionary is at an early stage of development but already displays the standard and successful SYSTRAN trend towards long source strings, within the now well-understood limits, in its approach to other languages.

Finally, the adventurous EDR dictionary project [EDR 90] (see Section 9.6) provides a formal,

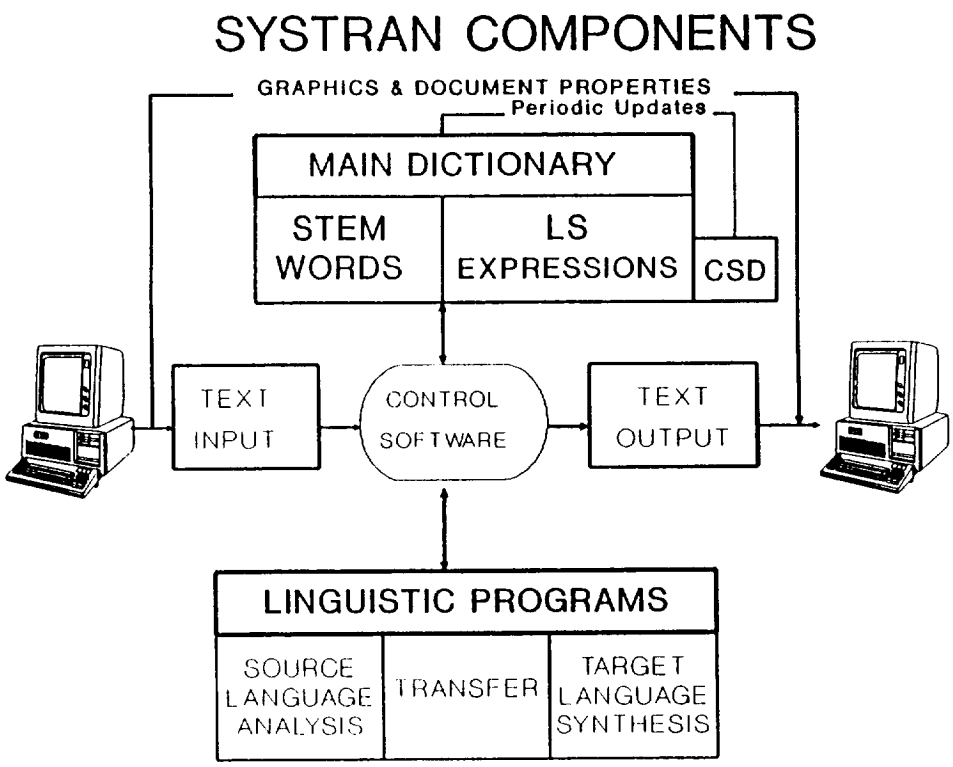

Figure 4-7: The Use of Dictionaries in SYSTRAN

E-Multitarget	Source	Target
ENGLISH SOURCE	172,056	
English-French		200,166
English-German		129,916
English-Italian		149,387
English-Portugese		42,130
English-Spanish		103,337
English-Korean		6,412
English-Arabic	162,640	150,147
English-Dutch	97,994	79,075
English-Japanese	156,866	66,384
English-Russian	19,329	34,773

Figure 4-8: SYSTRAN Dictionary Size
(As of 6/30/89)

```
1NEW DICTIONARY RECORDS
1DC STEM/ID/EXPRESSION          POS  JAPANESE - ENGLISH      140   W T H MEANING    02-18-91  PAGE     1  D SYN AA GR CC WC   DATE
                                BPQ                               N G I                                  P     RN AE 22 D2   LAST
                                                                    D                                    Q     T  DF 01 13  UPDATED
031  #ATAMA#WARI .BCPRT Q#DE                                      1 0 0 HEAD                             1 000 00 00   0  01-24-91
                                                                  2 0 0 PER                              0 000 00 00   0
                                NO1 01-24-91 ASSIGN MNG TO 'ATAMAWARI DE'; ASK-7 (KE)
 81  #GEN#BA Q#DE                                                 1 0 0 SITE                             1 000 00 00   0  10-29-90
                                                                  2 0 0 ON                               0 000 00 30   0
                                NO1 10-29-90 GENBA DE = 'ON SITE'; J3INTEC2-88  (LG)
 41  #HAI#KA .AD Q#DE                                             1 0 0 DIRECTION                        1 000 20 00   0  09-21-83
                                                                  2 0 0 UNDER                            0 000 00 30   0
                                NO1 09-21-83 JFAC3C-235 (HA)
                                NO2 09-21-83 SET TR= 'UNDER THE DIRECTION OF' (HA)
 82  #HURUI Q#TOKORO Q#DE Q#HA                                    1 0 0 $$$                              0 000 00 20   0  10-10-90
                                                                  2 0 0 LOOKING BACK INTO HISTORY        0 000 00 30   0
                                                                  3 0 0 $$$                              0 000 00 20   0
                                                                  4 0 0 $$$                              0 000 00 20   0
                                NO1 10-10-90 ASSIGN MNG., TOKORO MULT MN PROJECT.  MHWDTOKO-132 (LG)
 41  #I#HI Q#DE                                                   1 0 0 SENSE                            1 004 00 00   0  06-09-89
                                                                  2 0 0 IN                               0 000 00 30   0
0                               NO1 06-09-89 JABAF-42; JFIFTHC-22,46; JFIFTHB-7 (HA/LG)
 41  #I#TAKU .POS=10 .AD/MODL                                     1 0 0 REQUEST                          1 004 00 00   0  04-08-83
       .IF .B28 .$Q#YORU                                          2 0 0 AT                               1 000 00 30   0
       .OR .BCPRT Q#DE
                                NO1 04-08-83 VOX79-4 (HA)
 84  #ITI #KOU.KUTI Q#DE #IEBA                                    1 0 0 $$$                              0 000 00 20   0  10-11-89
                                                                  2 0 0 $$$                              0 000 00 20   0
                                                                  3 0 0 $$$                              0 000 00 20   0
                                                                  4 0 0 PUT SIMPLY                       0 000 00 30   0
                                NO1 10-11-89 BITI (MC)
 41  #ITU#POU .BL+SSU .PW,CW .BR Q#DE .BR .$Q#HA                  1 0 0 OTHER HAND                       1 000 20 00   0  03-17-87
                                                                  2 0 0 ON                               0 000 00 30   0
                                NO1 03-17-87 TRANSLATE EXPRESSION 'ON THE OTHER HAND' (YD/HA)
 418 #KAN#REN Q#DE .PW,CW .BMODL+BPQ=28 .PW,CW,B20,B30            1 0 0 RELATION                         1 000 00 00   0  02-22-84
       .Z-LMOD .PW,CW .Z-MODL                                       CDS-GENTO
                                                                  2 0 0 IN                               0 000 00 30   0
                                NO1 02-22-84 RESET 16/26 TO 30/20 BETWEEN PW AND BPQ28 MODIFIER AND
0                               NO2 02-22-84 SET ON SPMNCD GENTO (HA)
 41  #KEI.KATATI Q#DE .PW,CW                                      1 0 0 WAY                              1 004 00 00   0  01-07-91
       .IF .BANTEC+ANSUB .MID,0                                     1 FORM                               1 004 00 00   0
       .OR .ANTEC+N-LINKGVR .MID,1                                2 0 0 IN                               0 000 00 30   0
       .OR .Z-MODL .AD .MID,1                                       1 IN                                 0 000 00 30   0
       .OR .B26 .POS=20 .PW,CW .MID,0
                                NO1 10-29-90 ALLOW CONSISTENTLY ADVERBIAL FUNCTION OF KATATI DE (LG)
                                NO2 10-29-90 CASE MARKER DE PROJECT.  (LG)
 B18 #KIYOU#DOU .BR Q#DE .POS=50 .PW,CW .S-POS=30                 1 0 0 JOINTLY                          0 000 00 30   0  10-29-90
                                                                  2 0 0 $$$                              0 000 00 20   0
                                NO1 10-29-90 KIYOUDOU IS ADV WHEN FLLWD BY DE.  DE IS MADE CSMKR
                                NO2 10-29-90 IN HONOR.  J3ENERGY-25 (LG)
 41  #KOU#SEI .PSV .BOBJ .BCPRT                                   1 0 0 CONSIST                          3 004 00 01   0  09-08-83
       .IFNB Q#KARA                                               2 0 0 OF                               0 000 00 30   0
       .OR Q#DE                                                   3 0 0 OF                               0 000 00 30   0
                                NO1 09-08-83 XJEOBJ2-34 (HA)
                                NO2 09-08-83 EXPAND TO INCL #KARA CPRT AS WELL AS Q#DE (HA)
 41  #KUTI#UTUSI .N-LINKGVR .BCPRT                                1 0 0 MOUTH TO MOUTH                   0 000 00 00   0  04-08-83
0      .IFNB Q#N1                                                 2 0 0 $$$                              0 000 00 20   0
       .OR Q#DE                                                   3 0 0 $$$                              0 000 00 20   0
                                NO1 04-08-83 SET TRANSLATE Q#N1/Q#DE WHEN PW IS NOT A PRED. NOM. (MO)
                                NO2 04-08-83 SSTD1-136 (MO)
 82  #SE#KAI #KI#BO Q#DE                                          1 0 0 WORLDWIDE                        2 000 80 00   0  01-02-90
                                                                  2 0 0 SCALE                            1 000 00 00   0
                                                                  3 0 0 ON                               0 000 00 30   0
                                NO1 01-02-90 AGRMNT-2 (HA/MS)
 41  #SIYU#DOU .BCPRT Q#DE                                        1 0 0 HAND                             1 004 70 00   0  05-07-81
                                                                  2 0 0 BY                               0 000 00 30   0
                                NO1 05-07-81 NIS4-27 (BO)
 418 #SOBA .AD Q#DE .PW,CW .Z-LOCAT+S-PROX                        1 0 0 VICINITY                         1 000 20 00   0  09-21-83
```

Figure 4-9: A Sample from the SYSTRAN J/E Dictionary

conceptual description of at least 400,000 head items (roughly corresponding to word senses), with an interface to sense definitions in English and Japanese. A sample of the English interface is shown in Figure 4-10. The dictionaries are designed to be a knowledge source in the pure sense, free of implied process, although, in fact, their conceptual coding scheme will most likely appeal to a lexically-driven, interlingual MT system. This is an enormous enterprise; it is both manpower- and computation-intensive. It is not yet clear how much of the conceptual coding has been completed, even though both language interfaces are available.

Funds for the project have been provided both by the government and by major companies with MT activity [Fujitsu, NEC, Hitachi, etc.]. While these companies all plan to make use of the EDR dictionary's

【royal】 《00C7405》 very fine and costly
【royal】 《0d48d2》 a member of the royal family
【royal】 《03F6944》 of a person noble and refined in mind and character
【royal】 《0d48d3》 a small mast, sail, or yard, set above the topgallant
【royal】 《0d48d4》 a size of writing paper
【royal】 《0d48d5》 a size of printing paper
【royal】 《0d48d6》 any one of various coins in former times
【royal】 《0ea551》 to become holy and sacred／尊くおごそかなさまになる
【royal】 《00F7FF4》 of a condition of a thing, excellent／すぐれてよいさま
【royal】 《0fa960》 of a condition of a view, magnificent／眺めが壮大であるさま
【royal】 《1086d6》 precious things／得がたいもの
【royal】 《03BD198》 a facility built by a king or his family／国王や王族が設立した施設
【royal】 《03CE55B》 of a condition, excellent／すばらしいさま
【royal】 《03CE649》 of a condition, satisfactory／満足がいく状態であるさま
【royal】 《03CF00A》 a state of being excellent and noble／優れて気高いさま
【royal】 《03CF119》 luxurious and magnificent／規模が大きく，りっぱで美しいさま
【royal】 《03CF2E5》 a condition of someone having a noble position／身分が高く，尊いさま
【royal】 《03CF6A7》 a state of being solemn and respectable／おごそかで立派なさま
【royal antler】 《26d6a3》 the third tine above the base of a stag's antler
【royal blue】 《03C6272》 a vivid bright indigo named royal blue／ロイヤルブルーという，あざやかな明るい藍色
【royal blue】 《03EE202》 a colour named royal blue
【royal coachman】 《026D6A5》 a fishing fly named royal coachman
【royal coachman】 《03C34FB》 a fishing fly with a mosquite-shaped feather attached to it／羽毛で蚊の形に作ったつり針
【royal commission】 《26d6a6》 a group of people commissioned by the Crown
【royal commission】 《26d6a7》 the inquiry conducted by royal commission
【royal demesne】 《26d6a8》 the private property of the Crown
【royal fern】 《026D6A9》 a fern named royal fern
【royal fern】 《03C0EBF》 a fern named osmund／ゼンマイというシダ植物
【royal flush】 《26d6aa》 a straight flush in poker, named royal flush
【royal jelly】 《026D6AB》 a nutritious secretion of the pharyngeal glands of the honeybee, named royal jelly
【royal jelly】 《03C6271》 a nutritious secretion of the honeybee named royal jelly／ロイヤルゼリーという，ミツバチの栄養になる分泌物
【royal mast】 《26d6ac》 the mast next above the topgallant
【royal moth】 《0d5293》 a moth named saturniid
【royal moth】 《03BCDB7》 an insect named io moth／山繭蛾という昆虫
【royal palm】 《26d6ae》 a palm tree named royal palm
【royal poinciana】 《26d6af》 a tree named royal poinciana
【royal purple】 《03C5BA7》 a color named royal purple／青紫という色
【royal purple】 《03EDB37》 a dark reddish purple
【royal tennis】 《26d1f2》 court tennis
【royal tern】 《26d6b2》 a tern named royal tern
【royal water】 《03C3172》 a mixture of nitric acid and hydrochloric acid／濃塩酸と濃硝酸の混合液
【royalism】 《03E99D8》 the condition of adhering to monarchism
【royalist】 《0d48d8》 someone who supports a king or queen, as in a civil war, or who believes that a country should be ruled by a king or queen
【royalist】 《0d48d9》 typical of someone who supports a king or queen, as in a civil war, or who believes that a country should be ruled by a king or queen
【royalize】 《0d48da》 to make royal
【royalize】 《0d48db》 to assume royal power
【royally】 《0d48dd》 with the pomp and ceremony due a sovereign
【royally】 《0d48dc》 by the crown
【royally】 《0d48de》 with the utmost care and consideration
【royally】 《0d48e0》 on a large scale; gloriously
【royally】 《0d48df》 in a splendid manner; magnificently
【royalty】 《00D48E2》 people of the royal family
【royalty】 《00D48E3》 a payment made to an author or composer for each copy of his or her work sold, or to an inventer for each article sold
【royalty】 《03F6436》 the rank of a king or queen
【royalty】 《0d48e4》 a share of the product or profit kept by the grantor of especially an oil or mining lease
【royalty】 《026BDB7》 a payment made to the mineral content of a certain area of land
【royalty】 《03C1A4A》 a payment made as a fee for a copyright／著作権の使用料
【royalty】 《03C1A4C》 authority of the king／王の威光
【royalty】 《03CEBCC》 the rank of king or queen／王としての位
【royalty】 《0d48e1》 royal power and rank
【royalty】 《03E99DC》 the condition of having regal charactor or bearing
【royster】 《00DB85E》 to swagger
【rozzer】 《03CE627》 a person whose occupation is called policeman
【rozzer】 《03F5E91》 a person who belongs to the police
【rpm】 《0d48e8》 = r.p.m.
【rps】 《0d48e9》 = r.p.s.
【rpt】 《00F0C4C》 to announce something publically／（何かを）公表する
【rpt】 《03F60F0》 to repeat; an act of repeating
【rpt.】 《03CF4EF》 to announce something publically by a paper or orally／何かを書類や口頭で公表する
【rpt.】 《00F0C4C》 to announce something publically／（何かを）公表する

Figure 4-10: An Example of the English Interface to EDR's Concept Dictionary

final form, the intention is also to make the entire system available everywhere for a reasonable price.

EDR strives to be maximally cooperative with researchers world-wide, both in terms of joint effort on the project itself (where they already have a collaboration agreement with UMIST and UK and exchanges of technical information with a French team), and on subsequent use of the material for MT.

5. Life Cycle of Machine Translation Systems

Masaru Tomita

This chapter describes how MT systems are developed in Japan. During the JTEC visits, few of the sites provided specific information on the development of their MT systems. JTEC team members were also not given precise information about the amount of money spent on MT development in Japan. Therefore, what follows is based on informal conversations with their researchers and project leaders, with some speculation on our part.

As with most or all MT systems, the projects in Japan tend to have the following four stages:

- Research Prototype — a "toy" system to demonstrate feasibility of the approach and framework,
- Operational Prototype — for public demonstration and to validate the system,
- Practical System (Special-Purpose) — for actual day-to-day use,
- Commercial System (General-Purpose) — to generate revenue.

Of course, each project is different. Some systems do not evolve into commercial systems. Some projects have stepped back to prior stages to redesign or reimplement their systems. Some systems (as will be described in Chapter 6) are intended for different kinds of environments. What is described in this chapter is a generalization of all the projects.

5.1 Research Prototype

The first step in developing an MT system is to design its theoretical framework and build a small laboratory prototype system to demonstrate the feasibility of the framework. The number of researchers per system at this stage is very small. Usually a principal researcher coordinates the entire system design and a few other researchers assist in designing details or implementing a prototype system. The typical duration of this stage is one or two years.

In Japan, the results of this stage are typically published in the following technical journals and professional meetings:

- Domestic Journals
  - JOUHOU SHORI GAKKAISHI *(Journal of IPSJ: Information Processing Society of Japan)*
  - JOUHOU SHORI GAKKAI RONBUNSHI *(Transaction of IPSJ)*
  - DENKI TSUUSHIN GAKKAISHI *(Journal of IECEJ: The Institute of Electronics and Communication Engineers of Japan)*
- Domestic Meetings
  - SHIZENGENGO SHORI KENKYUUKAI (IPSJ Working Group in Natural Language Processing)
  - ZENKOKU TAIKAI (IPSJ semi-annual National meeting)
- International Journals
  - *Computational Linguistics*
  - *Machine Translation*

- International Meetings
  - COLING: International Conference on Computational Linguistics
  - ACL: Association of Computational Linguistics
  - MT Summits

It is worth noting that even private companies do not hesitate to publish their results at this stage (technical approaches, grammar formalism, dictionary configuration, theoretical framework, etc.). Possible explanations are:

- Prestige is one of the important factors. The more publication, the more recognition.
- Competitors do not like to adopt other approaches anyway. Each project wants to maintain originality. If some competitors adopt your approach, this means more prestige to you, and less prestige to them.
- The information does not help the competitors very much, after all. Many believe that the difficulties of MT lie in the development stage, not in the design stage. It is fun to design an MT system, but very hard to develop it into a fully operational environment.

The research prototype system is usually very primitive; it may have only a few hundred words in its dictionary, and may be very slow with little consideration for efficiency. It can translate only a small set of sample sentences, and does not work for most other sentences. The research prototype is clearly not sufficient for public demonstration.

MT projects at ICOT and ETL stop here; their objectives are to demonstrate specific theories of language and not to develop operational MT systems. Most academic projects at universities also stop here, with the exception of the MU project, which aimed at operational system development.

5.2 Operational Prototype

Once a research prototype has been implemented and its framework is proven feasible, the next stage is to develop an operational MT system based on the research prototype. A typical operational prototype has:

- broad grammar coverage to handle most input sentences,
- a dictionary with 10,000 - 100,000 entries, and
- modules to cope with practical problems such as idiomatic expressions, proper nouns, segmentation, punctuation, etc.

At this stage, the number of researchers increases to around 10 ~ 30. Development of the operational prototype takes place in two steps: first, developing an initial version of the system; and second, testing and debugging it. For the initial system development, three major tasks are necessary:

- Creating Dictionary Entries. It takes many people many months just to enter dictionary entries within the specifications defined during the research stage. This well-defined task is often considered tedious, and most projects find it convenient to subcontract the task to an outside software house.
- Writing Grammar Rules for Analysis, Transfer and Generation. The task of grammar-rule writing also may be tedious, but it is hard to give to outside subcontractors because:
  - The task of grammar-writing is not as well-defined as dictionary development.
  - The task is quite difficult to divide into smaller subtasks.
  - The task requires special skills. Grammar writers must be familiar not only with the

syntactic structures of the source and target languages, but also with the system's implementation.

- The task requires interaction with other members of the project. One of the important missions of the grammar writers at this stage is to give feedback to the designers of the grammar formalism and the system implementers.

- System Programming. The task of system implementation may be easier to distribute to several researchers. An MT system can be divided into modules such as the analysis module, the transfer module, the generation module, modules for morphological analysis and synthesis, a module to handle idioms, and a module to handle proper nouns. Each of the modules can be assigned to a single programmer; and some of them could be subcontracted.

When the dictionary entries and an initial version of the grammar rules have been completed, and when the system modules have been programmed, then the real enterprise, testing and debugging, begins. Usually each project has a large corpus of sentences to use to test its system. The typical debugging cycle is:

1. Running the system through (a part of) the corpus.
2. Evaluating translation output produced by the system.
3. Analyzing the cause of each error/mistranslation.
4. Notifying appropriate researchers responsible for the bugs.
5. Returning to step one after all bugs are fixed.

This cycle continues until the system's performance becomes satisfactory. This is a very difficult process because the cause of errors may be in the grammars, dictionaries, or system programs, as well as in the fundamental design or framework of the system. Different people are responsible for maintaining different parts of the system.

In this way, a research prototype evolves into an operational prototype system. At the end of this stage, the system is usually announced publicly, and demonstrations of the system are given at press conferences and at technical meetings, such as the MT Summits (Hakone - 1987, Munich - 1989, Washington D.C. - 1991), the EDR Symposium and Workshop (Tokyo - 1988, Kanagawa - 1990), and other conferences and trade shows.

ATR and CICC will stop here. The MU project also stopped here. However, the university's hope was that somebody else would pick up their operational prototype and develop it further to make it practical. In fact, the MU project was picked up by JICST and used as a basis for the JICST Machine Translation system.

5.3 Practical System (Special-Purpose)

After an operational prototype has been developed and public demonstrations have been given, the next step is to make the system usable in a day-to-day translation operation. The following three improvements are usually required:

- Specialized grammar and dictionary. The grammar rules and dictionary entries have to be adapted for the particular task domain.
- System Robustness. The system must endure under heavy daily operation. For example, it cannot afford to crash under any circumstance.

- Better User Interface. The system must be sufficiently user-friendly to be used by nonproject members.
- Peripheral Software. Pre- and postediting tools, user dictionary development tools, etc., must be developed.

Here are some examples of internal and external use of MT systems:

- Internal Use
  - IBM Japan (SHALT) — Translating IBM manuals
  - JICST — Translating scientific abstracts
- External Use
  - Fujitsu (ATLAS-II) — Mazda Motor Corporation
  - NEC (PIVOT) — Japan Convention Services, Inc.
  - HITACHI (HICATS/JE) — Japan Patent Information Organization
  - etc.

At this stage, external use tends to be like a joint venture project; the users are usually very cooperative. Let us elaborate just one of the examples of external use — Fujitsu's ATLAS-II at the Mazda Motor Corporation. Mazda started a joint venture after Fujitsu completed its operational prototype in 1985.

- First year (85/86) — Feasibility study
  - Evaluation by Translation Service Department.
  - Preparation of Mazda's basic dictionaries and accumulation of pre-editing know-how.
  - Summary of evaluation and planning of system tuning.
- Second year (86/87) — Trial and system tuning
  - System trial with service manuals.
  - Systematic maintenance of dictionaries.
  - Tuning of processing function for translation.
- Third year (87/88) — Business use
  - Application to various overseas product manuals.
  - Gradual maintenance of dictionaries and accumulation of know-how for business use.
  - Preparation of expansion plan.
- Fourth year (88/89) — Expansion of application
  - Expansion of application to technical documents.
  - Gradual integration into integrated document processing system.

5.4 Commercial System (General-Purpose)

If a system has proven useful in a specialized task domain, the next and final step is to make it general purpose, so that it can be used by many customers for multiple purposes. One of the biggest motivations for developing a general-purpose commercial system is to earn revenue directly from the system. There are three ways for a general purpose MT system to generate income:

- Sales as a software package (US$5,000 to US$30,000 per copy),
- Monthly lease (software and hardware), and

- Online service (charge for CPU time or by the word or page of text to be translated).

Unlike users at the prior stage, users at this stage are not necessarily cooperative; in fact, they are often critical and impatient. The number of people on the project at this stage is large --- as many as 100 and many of them are customer-support personnel.

In the last two or three years, several translation service companies have started systematic use of MT systems. (See Chapter 6.) Some users of these systems have said that the quality of MT-produced translation, even with human assistance in pre- or postediting, is lower than that of full human translation without MT. However, in some cases the service bureau will charge less (often around 30% less) for MT-produced translation, and some customers appear to be happy with lower-quality translation at lower cost. Other users demand higher quality translation.

5.5 Ongoing Use

Several factors determine the fate of an MT system once it has been put into production. For example, experience has shown that if a system is not operated by appropriately trained people then the results will be unsatisfactory, which will lead to the system being abandoned. The level of support provided by the manufacturer is also very important, particularly since some kinds of modifications to MT systems, such as expansion of the grammars, cannot be done by the users; they must be implemented by the developers.

6. The Uses of Machine Translation in Japan

Muriel Vasconcellos

6.1 Introduction

As might be expected, the large number of machine translation systems originating in Japan is matched by a broad variety of roles that have been found for the technology. The JTEC team was able to observe not only the full range of uses for MT that have been tried in other countries but also some variations and innovative applications that are being pioneered for the first time anywhere. Undoubtedly the heavy demand for translation between Japanese and English, the relative scarcity of highly qualified translators in these combinations, and the impressive linguistic distance between the two languages have stimulated the search for creative ways to enlist MT in the service of communication with the West.

Despite the many different MT modalities, domains, and translation purposes that are being tried, there is in fact a dearth of hard data about how Japan's more than 20 systems are actually being used. An authoritative summary of all the applications would be impossible. For this reason, the present chapter has been limited to a discussion of types of applications in the broad sense, with illustrative examples where appropriate. The rest of this section examines some of the general issues. Section 6.2 offers "case studies" drawn from the sites visited and other data available to the JTEC team, and Section 6.3 summarizes the reports of usage by the commercial vendors contacted during the course of the mission.

6.1.1 Modalities of Implementation

Based on form of input files

For any source language, the use of electronic files as input for MT makes a major difference in its overall cost-effectiveness by cutting down on human intervention at the front end. Since the input of Japanese is exceptionally complex, given the tedious problem of typing in kana and kanji, the savings to be realized are naturally even greater. It comes as no surprise, therefore, that the Japanese are exploring a variety of MT applications that take advantage of the fact that the input text has already been captured in machine-readable form.

The electronic files used for MT input in Japan come from a range of sources that include desktop publishing processes, public telecommunication networks, newswire services, and databases. Optical character recognition (OCR) has recently started to achieve sufficient accuracy to be considered a practical way to convert paper documents in high quality print into electronic ones, as discussed below.

As in the West, MT is often embedded in the publishing chain. Hardware manufacturers, the major users of MT in Japan, have introduced the technology into a process that starts at the point where specially trained writers draft their texts in-house and generate files that can be used both for publication directly in Japanese and for input to MT. Desktop publishing is the typical mode of operation for hardware manufacturers that have developed MT systems to translate their own product manuals. (See Section 6.1.3.)

For companies that do not have their own MT system, the translation of manuals is sometimes subcontracted to translation service bureaus that use MT. In a number of cases, the files are transmitted

from the customer to the translation agency through a computer service network. At CSK, a major computer service company, customers' files are received on-line for machine translation into English by the firm's proprietary system, ARGO. (See Section 6.2.1.)

At the time of the JTEC visit, CSK was developing an international network that would permit ARGO to be accessed on-line in the United States. The general public can already tap into Fujitsu's ATLAS-II via NIFTY-Serve, the Japanese network modeled after CompuServe, and obtain a machine translation from Japanese into English. Any of NIFTY-Serve's 200,000 subscribers can exercise this option; the cost is ¥10 per minute of connect time and ¥1 for each word of English output. The customer can either submit a file or key in the text directly. NEC has followed suit with its PIVOT system, which since December 1990 has been available on PC-VAN, Japan's other major computer service network with a similar number of subscribers and 104 access points in Japan, and, in addition, through GEnie it has the equivalent of more than 550 access points in North America [Sakurai 91].

Internal electronic mail (email) can be a vehicle for MT input as well. At Oki Electric, for example, staff can access the company's PENSEE system as a menu option on their desktop terminals. Fujitsu also offers email access to its MT system, ATLAS-II.

At Nippon Hoso Kyokai (NHK), the nation's public television station, Associated Press wire reports are being monitored and translated on an experimental basis by Catena's STAR system. STAR keeps up with the incoming stream on a near real-time basis. In a reverse application, CSK uses ARGO to pick up data from the Japanese securities market and supply the corresponding English translation to the Nikkei Telecom network, which in turn links up with Reuters and is beamed to the world's 40 major financial centers.

Databases provide another form of ready electronic input. For this reason, plus a number of others, they are seen to be a natural application for MT. And in fact MT is already being exercised on the databases at the Japan Information Center of Science and Technology (JICST) and the Japan Patent Information Organization (JAPIO). The importance of translating databases is discussed in Section 6.1.2 below.

When electronic files are not available, several factors enter into deciding whether or not it is worthwhile to use labor-intensive means to create a machine-readable document. If the investment of time and manpower is justified, the question that remains is whether optical character recognition (OCR) can be of assistance. Both in Japan and elsewhere, OCR has been used for some time as a tool to facilitate the task of MT input. However, the technology is not the panacea that some had hoped it would be. Even with English, which presents fewer challenges than languages with accents and diacritics, not to mention those with non-Roman alphabets, OCRs still produce misread characters. For purposes of MT input, where "close doesn't count," the OCR-generated file needs to be reviewed for errors—a problem in Japan, where speakers and writers of English are at a premium. The relative usefulness of OCR for inputting English in Japan can be deduced from the fees charged for raw MT supplied by IBS, one of the translation bureaus visited: input using OCR is charged at ¥200 per page, versus ¥360 when manual input is required.[5] From these figures it may be assumed that OCR costs 44% less but by no means

[5]The figures of ¥200 and ¥360 are based on the fee schedule circulated by IBS at the time of the JTEC visit, according to which raw MT cost ¥290 per page with input in the form of an electronic file, ¥490 per page for OCR input, and ¥630 per page for manual input. At this writing, $US1 = ¥130.

obviates the need for human intervention.

OCR for the Japanese language, of course, is more challenging. Several companies (most notably Fuji Electric, Toshiba, and Sharp, among others) have developed products that read kana and kanji, but they still have limitations. There appears to be little use of OCR as a front end for Japanese to English MT systems, although the technology is poised to make rapid inroads.

Based on Degree of Human Intervention

For several reasons, human intervention becomes an extremely important issue in machine translation between Japanese and English. To begin with, the two languages are linguistically quite distant from one another,[6] which means that the challenge for the computer is greater than it is when translating among European languages. This inevitably leaves more work to be done by human beings either upstream, downstream, or midstream. The use of highly paid professional translators to work with MT quickly adds to the cost of the process. Moreover, for translation into English, such professionals are in relatively short supply in Japan. These are all disadvantages for MT acceptance, where the value of an application is measured in terms of cost, speed of turnaround, and ease of implementation---always in light of the purpose of the particular translation. It is clear that MT will meet with the greatest acceptance wherever human intervention can be minimized.

The manual aspects of text input have already been mentioned. In addition, when the source text is in Japanese, many MT applications rely on pre-editing to facilitate the linguistic task that confronts the machine. Given the distribution of available human resources in Japan, it is more practical and economical to hire non-translator native speakers of Japanese to massage the input than to induct Japanese speakers of (often shaky) English into the complexities of postediting. Sophisticated user-friendly interfaces for pre-editing have been or are being developed for a number of the systems that the JTEC team saw --- for example, ALT-J/E (NTT), ASTRANSAC (Toshiba), ATLAS-II (Fujitsu), DUET-E/J (Sharp), HICATS (Hitachi), PAROLE (Matsushita), STAR (Catena), and JICST (scheduled for 1991). Such interfaces speed up the translation task by providing ready criteria for simplification of the grammatical structure so that the input is more in line with the capabilities of the MT system itself.

Pre-editing does not necessarily eliminate the need for postediting. The degree of intervention required in the output depends on the purpose for which the translation is to be used. If postediting can be dispensed with---whether because the quality is good to begin with, or because problems have been solved in pre-editing upstream, or because the application does not require a high level of quality---then MT becomes much more economically attractive. As a general rule, postediting costs are higher than pre-editing, in part because postediting requires a highly trained bilingual translator who can compare the source and target texts, rather than two monolingual readers, as is the case for pre-editing.

The ultimate MT application is the one that uses raw MT directly and requires no human intervention whatsoever. Raw MT is currently being sold in Japan. As noted above, it can be purchased via NIFTY-Serve, which reports 50-60 accesses a day, and on PC-VAN, a new application. It is also available from the IBS translation service bureau, which at the time of the JTEC visit was selling raw DUET-E/J and is now selling raw ASTRANSAC via PC-VAN. In addition to the processing of input and output, for

[6]See [Becker 84] on translation between distant languages.

which a small fee is charged, pre-editing is offered for an additional ¥500 per page. These prices are at most less than half the cost of finished translation, but even so, clients usually prefer for IBS to do the postediting and provide them with a final product. The main obstacle to sales of raw translations is accuracy. More accurate raw translations would lead to much greater sales.

6.1.2 Translation for Assimilation: Domains and Applications

MT is considered to be well suited for applications in which the purpose of the translation is to gather information, convey the gist of a text, or perhaps answer specific questions that the end-user has in mind---in other words, applications for *assimilation*, in which large volumes of foreign text are scanned and translation quality is not a high priority. Often, with applications of this kind, quick turnaround is essential because the information has a limited shelf life, after which it is worthless or of little value. When human intervention can be minimized downstream as well, MT becomes very attractive because of its speed. If the input text is already machine-readable, MT becomes an especially attractive option.

In Japan, even more than elsewhere, attention has been focused on databases as prime candidates for information-only translation using MT. The fact that the files are already in electronic form sets the stage for an effective application, but other considerations are even more compelling: they contain valuable, sometimes critical, information that is not normally translated. Several factors militate against translating databases in the traditional way. To begin with, the task is immense, while human translators are costly and in short supply. Even if there were enough translators, there is no organized customer base to support such an effort. From the standpoint of the consumer, especially in the West where goals must be met in the near term, there has been little impetus so far for investing in the capture of information from such sources; consumers are unwilling to set a price for information before they know how valuable it will be to them. This is a classic problem in information science.

The situation can be remedied to a large degree, however, with the help of robust, general-purpose MT systems, which can be a powerful tool for screening information. As a first pass, MT can provide quick translations of titles and descriptors, alerting analysts to potential areas of interest [Bostad 90]. Once the analyst has spotted material of interest, MT can then provide additional information through translations of the corresponding abstracts. Often the unedited raw output is sufficient for the purpose. This tool is especially valuable in bridging the gap between Japanese and English, where linguistic differences hide all clues to the concepts that analysts may be seeking. This scanning function of MT can greatly reduce the need for high-cost human involvement.

It should be emphasized that the use of MT for database searching requires powerful general-purpose systems with very large and carefully refined knowledge sources---in other words, systems that have had the benefit of long-term investments of manpower to build up the "know-how" needed to produce translations in a wide variety of technical domains.

In Japan, the first database operation to enlist the aid of machine translation was the Japan Patent Information Organization (JAPIO), an auxiliary arm of MITI and the Japanese Patent Office which has been working with Hitachi's HICATS/JE since 1985 to facilitate the translation of patent titles and abstracts for distribution around the world. JAPIO's database---more than 10 million entries of domestic data and another 16 million entries of foreign data---is the backbone of its service, which offers information retrieval for the public in the areas of patents, utility models, designs, and trademarks. With the aid of HICATS, each year some 300,000 titles and 270,000 abstracts (averaging 4.4 sentences in

length) are translated into English.

More recently the Japan Information Center of Science and Technology (JICST) launched a massive program using its own MT system to translate the content of its database into English. (See Section 6.2.6.)

Another database-type application is CSK's use of ARGO to translate information on the Japanese securities market into English for use by Reuters. (See Section 6.2.1.)

MT is also being used in the J/E direction to feed the JAPINFO database supported by the European Commission in Luxembourg, which is available through the DataStar network and has about 300 regular users in nine countries including the United States. In this case, however, the input documents are in the form of hard copy and need to be keyed in manually in order to submit them to MT. (Section 6.2.5.)

A different and innovative application for information-only translation (i.e., not high enough quality for dissemination) is STAR's around-the-clock translation into Japanese of incoming Associated Press news bulletins. This service is primarily used to identify incoming stories that are of interest so that they may receive more careful translation. (See Section 6.2.7).

Despite the Japanese tradition of information-gathering and the current interest in databases, the JTEC team did not identify any MT application in Japan comparable to the use of SYSTRAN in the United States, which has been being used to monitor foreign technology for more than 20 years [Bostad 90], [Vasconcellos 91].[7]

6.1.3 Translation for Dissemination: Domains and Applications

In the case of translation for *dissemination*, quality is usually more important, so MT is typically more appropriate for well-defined domains. However, applications differ: one application may call on MT to deal with many topics, whereas another may limit translations to a single domain (a specialized subject area that has a sublanguage with a relatively small, unambiguous vocabulary and simple, predictable syntactic structures) while a third one may fall somewhere between these two extremes. The first type of application calls for *general-purpose* or "try-anything" systems similar to, and at least as robust as, those that handle translation for assimilation. At the other end of the scale, *specialized* systems for domain-specific applications may require less MT development and are not as costly to implement. Texts in specialized subject areas make good grist for MT because they tend to yield predictable, uniform results

[7]The prime example of an MT installation that produces information-only translations is the U.S. Air Force's Foreign Technology Division (FTD), Wright-Patterson AFB, Ohio. FTD provides translations to scientific and technical analysts whose job is to stay abreast of foreign developments and prevent technological surprise that could threaten the United States. SYSTRAN has been in continuous operation at this site since 1969, generating translations from Russian, and more recently German, French, and Spanish, into English at a rate of 50,000 to 60,000 pages a year. In 1978, Russian/English output was deemed mature enough to be delivered to consumers with only partial postediting. In this semi-automated process, which involves intervention in about 20% of the text, a software module identifies known potential problem areas in the MT output and brings them to the attention of the posteditor. More recently, MT was made directly available to analysts through a gateway to the mainframe from their desktop PCs. They can use this connection to obtain raw MT with immediate turnaround. Since the texts have to be input by hand, the mode is best used for rapid translation of book titles, tables of contents, captions of tables and graphs, and isolated sentences and paragraphs. Despite this limitation, however, the system is tapped as often as 600 times a month [Bostad 90], and some of the analysts have indicated that they would be willing to accept raw MT for full-length documents, forsaking partial postediting, if the manual entry of the input could be done for them [Vasconcellos 91].

that sometimes require very little human intervention (e.g., METEO [Chandioux 89], [Grimaila 91][8]); they require relatively less development effort to bring them to a fully functional level; and they often produce higher accuracy translations.

In Japan, MT is being used to disseminate texts not only from Japanese into English but also from English into Japanese. By far the most common MT application is for internal J/E translation in hardware companies that sell their computer products overseas. Providing product documentation in English is crucial to their capturing markets in the United States, Europe, and other parts of the world. Indeed, most of the firms visited--- Fujitsu (ATLAS-II), Hitachi (HICATS), Matsushita (PAROLE), NEC (PIVOT), NTT (ALT-J/E), Oki Electric (PENSEE), Ricoh (RMT), Sanyo (SWP-7800), Sharp (DUET-E/J), and Toshiba (ASTRANSAC)---gave this as their main reason for being involved in MT, or at least cited it as a major application and probably their original one. Hardware firms not visited that have developed MT systems are Canon (LAMB) and Mitsubishi (MELTRAN), both still in the research stage. Eight of these companies---Fujitsu, Hitachi, NEC, Oki, Ricoh, Sanyo, Sharp, and Toshiba---have gone on to develop a commercial product.

While the translation of product documentation certainly keeps the MT systems busy in Japan (said to represent 80% of all MT use), companies that have developed their own systems are using them increasingly for other in-house tasks as well. At the same time, translation service bureaus are finding that MT helps to shave costs, and they are using the technology as an aid to the production of translations of high quality in a broad range of applications, limited mainly by whether or not clients present their texts in machine-readable form.

Bravice's MICROPAK, in addition to being used for product documentation, is said to be popular among researchers at universities and hospitals, who use it to produce English-language articles and reports for publication abroad.

At NHK, the use of MT to assist in creating Japanese subtitles for television introduces an entirely new type of MT application --- one that is highly demanding. (See Section 6.2.7.)

6.2 User Sites Visited

Given the short duration of the JTEC mission, the team was only able to see a few of the MT user sites in Japan. Different members of the team were able to visit a total of seven sites where MT was in practical use. The circumstances varied and the sample was broad, but it in no way purports to be complete. The team is aware of a number of interesting applications that could not be included because of the shortness of the visit. This section summarizes highlights from the seven installations.

---

[8]One of the most successful cases of MT for dissemination is the translation of Canadian weather forecasts around the clock by METEO, which in the last 15 years has processed more than 100,000,000 words for the Canadian public [Chandioux 89]. This application involves repetitive text with a limited vocabulary, although the input comes in free syntax from a variety of sources. Very little intervention is needed in the machine output (about 4% [Grimaila 91]).

6.2.1 CSK

CSK Corporation, Japan's leading computer multiservice company, provides computer programming and software development services, and it also sells and leases computers. CSK's internal work in artificial intelligence provided synergy for the development of a Japanese/English MT system (TEE), which was placed in service in September 1986. Through this capability, in which CSK had the cooperation of Nihon Keizai Shimbun, Inc., vital data on the Japanese securities market is currently supplied to the Japan News Retrieval service via the Nikkei Telecom network, which, as mentioned earlier, links up with Reuters and from there is broadcast worldwide. In April 1988, CSK introduced ARGO, a newer system that produces faster and better translations in the areas of both finance and economics. The following year a prototype Japanese/English version was also introduced.

The input texts for ARGO are prepared by specialized writers. Both pre- and postediting are done, with emphasis on the latter. A native Japanese translator and a native English editor are integral parts of the CSK team that produces their daily output. CSK also offers Japanese/English MT service to at least 10 customers in Japan, for which ARGO processes some 37,500 pages a year.

The clients, most of them with texts in fields relating to securities and economics, have generally reported that ARGO has been useful for them. They are free to extend the source and target dictionaries but not to modify the grammar, which is done by CSK at least in part on the basis of their feedback. The developers are constantly seeking to improve the quality of translation.

CSK is in an expansion mode. In addition to its international network (Section 6.1.1), the company is developing a more convenient user interface, and it has plans to broaden its client base over the next five years by adding new domains, more languages, and large specialized dictionaries in support of these efforts.

6.2.2 DEC

Digital Equipment Corporation (DEC), a U.S. company with operations in Japan, uses MT software developed by another firm to translate its user documentation. A small percentage of DEC's total translation production from English into Japanese is done with the help of Toshiba's Translation Accelerator, ASTRANSAC.

This application began in March 1989. Currently ASTRANSAC supports the translation of about 100 pages of product documentation a month. It is also used occasionally for information purposes only or for first drafts. The experience to date with the product manuals is that MT has cut translation time by half and that costs are considerably reduced. Whereas the output of a traditional human translator is 5 pages a day, ASTRANSAC makes it possible to produce 11.4 pages a day. Cost savings are greatest when the input file contains SGML[9] markup tags and the codes are automatically ported into the target text. (See Section 9.11.) This feature eliminates the need to reintroduce the codes and reformat the text. The cost of traditional translation is ¥6,000 per page plus ¥3,000 for formatting, whereas with MT the total cost of the two steps together is ¥5,300 per page. Posteditors are paid between 50% and 70% of the regular translation fee, depending on the type of text.

---

[9]Standard Generalized Markup Language (ISO Standard 8879).

Translation quality leaves something to be desired, and postediting, in which technical writers get involved, is rather heavy. Pre-editing, with dictionary updating for unknown words, is also necessary. The pre-editing interface is deemed to be excellent, and the dictionary updating interface is user-friendly. Postediting is done on the company's own VAX equipment, although the Toshiba text editor is considered to be good.

Toshiba is responsive to DEC's requests for improvements and customer-specific adaptations.

6.2.3 IBM

Unlike DEC, IBM Japan opted to develop its own MT software to aid in the translation of product manuals from English into Japanese. The System for Human-Assisted Language Translation (SHALT) was placed in service at IBM's Japan Translation Center in July 1988, and from that date until the time of the JTEC visit it had been used to produce 60 manuals. SHALT also facilitates the translation of other computer-related documents, including memoranda to customers. Although no actual figures were supplied to the JTEC team, IBM has stated elsewhere that it is counting on productivity gains from the use of SHALT on the order of 150% to 300% [Smith 89].

SHALT does not rely on pre-editing in this application, but the user is expected to begin the process by running an interactive search for unknown words. Dictionary entries are prepared for all the missing words using interactive software. Once the target output is produced, postediting is undertaken using IBM's proprietary Translation Support System (TSS). SHALT occupies an important place in the document production chain, in which every effort is made to automate the publication process.

A very different and much-improved English-to-Japanese system, SHALT2 (see Chapter 9), is already in the wings. SHALT2, in addition to producing translations of better quality, will have machine-learning capability and will be expanded into other domains and language combinations, including Korean and Chinese as target languages.

6.2.4 IBS

International Business Service (IBS), a translation service bureau that receives work from a wide range of sources, was using Sharp's DUET-E/J for translating from 10% to 20% of its volume from English into Japanese at the time of the JTEC visit. They had also used Bravice's MICROPAK and Oki's PENSEE, and and recently had acquired NEC's PIVOT system (J/E and E/J). In December 1990, IBS inaugurated MT service on PC-VAN, a major computer service network, using PIVOT in both directions. IBS is the official translation bureau for this service, and in the first six months of operations they have reached a volume of more than 2,000 pages a month [Kazunori, personal communication]. In their own words, they have now introduced MT into their translation business "in earnest" [Sakurai 91]. It is now understood to be their predominant mode of operation.

Through PC-VAN, clients can send their files to IBS via modem. Previously the company did not use MT for the J/E combination because of the problem of inputting Japanese text, but with this new network capability IBS has gained the possibility of working from Japanese to English. The charge for texts that are pre-edited but not postedited is ¥1,200 (both directions), or about half the charge for human translation from English to Japanese (¥2,100- ¥2,900 per page) and an even smaller fraction of the rate for human translation from Japanese to English (¥4,100-¥4,900).

Prior to the PC-VAN connection, input for MT was optically scanned. The OCR output was verified by a native English speaker using the WordStar or WordPerfect spelling checker.

Since as a translation service bureau IBS receives texts in a wide range of subject areas, productivity gains with MT are bound to be less impressive than figures reported for more homogenous and circumscribed applications. Nevertheless, the process is economical because the overall task can be divided into small steps, as in a production line, taking advantage of operators who are not as highly paid as translators wherever possible.

With DUET-E/J, from 40% to 60% of the English input was being pre-edited at the time of the JTEC visit, while the rest was considered adequate to submit directly to the computer. Pre-editing is done by a native speaker of Japanese using interactive software that assists in such areas as filling in ellipses, marking source words for their part of speech in the particular context, bracketing, expanding reduced relative clauses, and breaking up long sentences. Postediting, of course, must be done by Japanese native speakers, and sometimes two passes are required. The postedited MT output usually becomes input for the next step in the desktop publishing chain.

With PIVOT on PC-VAN, the upstream process has three steps: entry of not-found words (even though IBS's version of PIVOT has a lexicon of 600,000 entries), pre-editing, and correction of mistakes in syntactic analysis. Six rules are applied for pre-editing, which can be done by a monolingual person and takes only 0.6 minutes per sentence (average length 21 words). The most time-consuming task is analysis and correction of syntactic errors, which averages 5.1 minutes per sentence. Final rewriting of the output takes only 1.6 minutes per sentence [Sakurai 91]. Even though MT with PIVOT is currently taking on average 27% longer than traditional hand translation, IBS still considers it to be more economical because of the utilization of lower paid operators. They also like MT because jobs can be split up, terminology is uniform, and the work can be more easily managed with people working in teams. In any case, IBS hopes to improve the slowness of turnaround (caused largely by the fact that this is a general-purpose application with random input) through dictionary-building and adjustments in their manning structure.

The IBS application is interesting because of the new on-line capability, the fact that raw pre-edited MT is available for purchase, the wide variety of texts being handled, the regular use of OCR to capture input, and the number of different MT systems that have been tried.

6.2.5 Inter Group

This company has been offering a wide array of language services since 1966: technical translation, simultaneous interpretation for international meetings, translator and interpreter training, planning and support for meetings, and, more recently, MT pre-editing, postediting, dictionary-building, posteditor training for translators, and user support.

At the time of the JTEC visit, Inter Group was using Fujitsu's ATLAS-II for about 20% of their total translation volume. They also had a contract with JICST to pre-edit and postedit technical abstracts translated by the JICST MT system. So far, all the MT work done at Inter Group has been from Japanese to English. The company had a sizable roster of personnel engaged in various aspects of MT. More than 100 subcontractors were being used for postediting alone. In the company's biggest MT application, ATLAS-II has been used since 1987 to translate some 1,000 technical abstracts a month for the

JAPINFO project. (See Section 6.1.2.) The JICST system was also being used to translate technical abstracts. (See Section 6.2.6). Inter Group was using and providing translation support for the ATLAS-II on-line MT service on NIFTY-Serve. (See Section 6.1.1.)

Translations with ATLAS-II may be received and dispatched as electronic files, which resolves the problem of text input, although in the JAPINFO project, manual input is still required because the materials have been gathered from a broad range of hard-copy sources. With ATLAS-II, pre-editing is minimal; the emphasis is on postediting. With the JICST system, on the other hand, pre-editing is stressed rather than postediting.

MT has brought improvements in productivity for Inter Group. Calculation of all the steps in the process has shown that MT, when used on general applications, takes from 68% to 76% of the time required for traditional human translation (HT). On the basis of these figures, translators are paid 30% less for postediting than for HT. Thus, Inter Group is able to produce translations both faster and at a lower cost.

In 1987, shortly after MT Summit I, Fujitsu approached Inter Group and proposed that the company enter the business of training people to use MT. By the end of 1988, Inter Group had introduced MT postediting into the regular curriculum of its training institute, Inter School. The course is given twice a year for 18 weeks at four hours a week. At the time of the JTEC visit it already had 50 alumni, of whom 35 were working for the firm in the capacity for which they had been trained. In addition to training posteditors, Inter Group has been offering a consultation service for Fujitsu customers who are faced with MT for the first time. Four days of intensive introductory sessions are followed by three months of consultation during which the new user practices document input, pre-editing, postediting, and dictionary-building.

In yet another project, Inter Group serves as a subcontractor for the Electronic Dictionary Research (EDR) project. Inter Group's job includes providing definitions for new words as they come up in the corpus. The definitions are largely adapted from commercial hard-cover dictionaries.

The company's future plans call for building specialized dictionaries for ATLAS users. The company sees text input as one of its major problems and looks forward to greater integration of MT into the total document production chain.

6.2.6 JICST

The Japan Information Center of Science and Technology (JICST) employs its own system to translate Japanese scientific abstracts for its English database, JICST-E. JICST provides information in Japan via the JICST On-Line Information Service (JOIS) and the Scientific and Technical Information Network (STN) [Ashizaki 89]. The initiative to translate the JICST database into English, prompted by the desire to promote worldwide distribution of Japanese scientific and technological information, dates from 1984. Since 1986, JICST has been developing its own practical MT operation based on the results of the MU project, which was supported by the Science and Technology Agency (STA) Promotion Coordination Fund, during the period 1982-1985. Total investments as of mid-1991 are estimated at ¥900 million.

Following three years of intensive dictionary development, beta-testing of JICST began in 1989, and full production got under way in July 1990. Since then, about 29,000 titles and 9,000 abstracts (including

titles) have been translated using MT. This amounts to about 17% of all the records in JICST. The target for 1991 is 70,000 titles and 20,000 abstracts, representing 41% of all the records. The abstracts, while they may be from any scientific or technical domain, tend to be concentrated in the area of electrical engineering.

Pre-editing is relied on, and the system performs best on short sentences. Currently the turnaround using MT is approximately the same as for conventional manual translation. The computer takes about three days to translate 1,000 abstracts. A job this size also entails a week of pre-editing, two weeks of postediting, one week of proofreading, and two weeks of manual keyboarding to input the final translation. A new interactive pre-editing tool for flagging 10 types of problems in the input is being introduced, and attention is now focused on reducing the time spent on postediting. Pre-editing and postediting are farmed out to four different translation bureaus.

While savings from faster turnaround have yet to be demonstrated, the JICST MT system has already been shown to be less expensive, since its total cost amounts to only 59% of that for conventional translation. JICST does not feel, however, that the present MT system is appropriate for mass production. There are plans to increase the percentage of MT-produced abstracts in JICST-E, shorten the lag time, and reduce costs even further.

Plans for the immediate future include building the store of technical terminology to a level of 500,000 noun entries plus 12,000 verbs by the end of 1991. Very soon JICST will be offering an on-line MT network service similar to that offered on NIFTY-Serve and PC-VAN. (See Sections 6.1.1, 6.2.4, and 6.2.5.) JICST plans to make raw MT available to the public to be used either for information scanning or for final translations to be postedited by professional translators. JICST also plans to develop E/J capability in order to provide Japanese translations in the national database of abstracts from foreign academic journals and scientific and medical reports.

6.2.7 NHK

Once every day, Japan's public broadcasting corporation, Nippon Hoso Kyokai (NHK), produces a program of news from around the world that is considered to be the core of its services [Aizawa 90]. "World News" is fed by video, audio, and wire reports arriving constantly in English, French, German, Italian, Russian, Korean, and Chinese. All this material must be edited and woven into a professional show for broadcast via satellite to the Japanese public. A team of some 50 interpreters and translators work in shifts around the clock to sort through incoming communiques and provide NHK's audiences with Japanese-language versions in the form of simultaneous interpretation or subtitles. As is the case with any news broadcasting, time is of the essence, so the work has to be done as quickly as possible. With news generated in the United States, the clock ticks even faster because of the time difference between Japan and the Americas.

Enter machine translation. In 1986, NHK and Catena-resource Institute, Inc. became partners in development of the English/Japanese STAR system, which is now being used to prepare subtitles for a daily 5-minute broadcast segment and to provide real-time raw translation of incoming Associated Press wire reports on an experimental basis. It should be kept in mind that news text is an inherently difficult challenge for MT because of its broad coverage. There is no limit to the subject areas to which news headlines can refer.

In the subtitle production system, the first step is for the operator/translator to listen to the news in English and transcribe it into a summarized, simplified form, which is then submitted for MT processing. This version of STAR produces output sentence by sentence. For each sentence, three different Japanese versions are generated in descending order of desirability based on a preferential weighting of possible outcomes (a switch can be set to display more or fewer versions). The weighting criteria include "comprehensibility," "complexity," and "rareness." Faced with the three options, the operator/translator chooses the best and does any further postediting that may be required. If the translation is correct but "awkward and charming," it may be left as the machine produced it. (A Japanese viewer has commented that the subtitles seem very good compared with raw MT but not as good as traditional human translation.)

The AP wire reports, on the other hand, are machine-translated using a batch version of STAR without any pre- or postediting. The purpose is to monitor the news; if an item is picked up for rebroadcasting, it then goes through the regular programming cycle, with translation provided as appropriate.

Strategies for improving system performance take their cue from the very different conditions in the application environment. With the subtitle production system, the focus is on finding ways to cut down on human intervention—to save time in this case, as well as money. The subtitles are a challenge for MT for several reasons. At the input end, the audio version of the program is transcribed in a painstaking process that involves capturing each word on the tape. After this step, there is considerable pre-editing. Postediting is also necessary, because NHK's responsibility for the information it provides to the public requires that there be no mistakes in meaning. And there are other constraints as well: the Japanese subtitle has to be brief; it has to be synchronized with the corresponding action on the screen; and it has to be as informative as the original English text. With the AP wire service news, on the other hand, translation quality is not so important, which makes it possible for the operation to be totally independent of human intervention from start to finish. Since the main goal of this application has already been met—that is, to attain sufficient speed so that the translation can keep up with the bulletins as they come in—NHK is free to concentrate on improving the translation engine itself.

There are three sources of NHK feedback for researchers. First, the operators flag the hard copy to call attention to errors in the dictionary or grammatical problems in the MT system, a large portion of which can be corrected by NHK R&D staff. Second, the R&D staff works directly with the developers, Catena-resource Institute. And third, troublesome areas are detected based on a review of the log and the time spent on human intervention for the different steps in the process. This information helps to set priorities for R&D and action to be taken.

The main challenge at the moment is to streamline human intervention in the subtitle production system, which, so far, has not been able to improve on the time taken for traditional translation. The current operation is very labor-intensive, with tasks not necessarily being performed by the most suitable personnel. NHK staff feel that improved deployment of human resources would make it possible to speed up the process and reduce the number of operators required.

6.3 MT Users: The Vendor Perspective

The previous section highlighted the user sites visited by members of the JTEC team. The team also heard about MT applications directly from MT vendors, who reported on ways in which their customers were using their systems. Many of these vendors are also users (as was pointed out in Section 6.1.3), so many of the commercial developers had gotten into the MT business in large part to satisfy their own translation needs. This information is summarized in the present section. Again, it does not purport to be exhaustive.

Bravice,[10] at the time of the JTEC visit, was claiming brisk sales of MICROPAK, its PC-based Japanese-English system. Some 4,800 units were said to be in the hands of users. This figure would undoubtedly make MICROPAK the MT system with the widest market penetration in the world.[11] Among the purchasers cited were Tokyo University (27 copies), Toshiba (30 copies) and Honda (30 copies). Company President Takehiko Yamamoto stated that 60% of the 4,800 units had been purchased for use in corporate settings and 40% for use in academic environments and hospitals for the preparation of technical papers in English. Honda harnesses its MICROPAK units into a network to facilitate product documentation. Use of MICROPAK normally entails pre-editing as well as postediting. Not all purchasers become active users—perhaps only 65%. Mr. Yamamoto emphasized that customization early on---i.e., the addition of 3,000 to 7,000 application-specific dictionary entries during the first months of operation---is an important factor in the client becoming a regular user. Feedback has been collected by the sales staff on an ongoing basis and incorporated into new releases.

Fujitsu was the first of the Japanese companies to become involved in MT. Its Automatic TransLAtion System (ATLAS), now in its second version, is a general-purpose system which has been tried on a wide variety of applications. The company reported that there were 200 users of ATLAS-I, 130 users of the mainframe version of ATLAS-II, and 150 users of ATLAS-II on workstations. Inter Group (Section 6.2.5) has been their service bureau. As mentioned earlier, ATLAS-II is available on-line to subscribers of NIFTY-Serve. Within Fujitsu, ATLAS-II has been used to produce 40 manuals for the company's own products. Mazda uses it for similar purposes and has incorporated it into a total document management system. They report that their productivity has increased by 30%. They are pleased with the role that MT has played in standardizing source texts and developing a lexicon of corporate terminology. On average, Fujitsu's users are experiencing a 50% reduction in costs, as well as shortened delivery times [Sato 89]. In customized applications, Fujitsu claims 80% accuracy, but this performance level declines somewhat on random general text [Valigra 91].

Hitachi, another early entrant into the MT arena, has developed HICATS, a general-purpose system. Three products are now on the market: J/E and E/J on mainframe, and J/E on workstation. The E/J workstation version was described as being nearly ready at the time of the JTEC visit. The mainframe versions have each been sold to about 100 customers and the J/E workstation version to another 100. One customer is Toin Corporation, a translation service bureau that uses HICATS to translate product documentation. Another customer for the J/E mainframe system is the Japan Patent Information Organization (JAPIO --- See Section 6.1.1 above), which has built the dictionary to a level of 300,000 entries. A JAPIO evaluation performed on a corpus of 5,000 titles showed that 40% of them could be

[10]According to information received in early 1991, Bravice is no longer in business.

[11]However, no independent confirmation could be obtained for this sales figure, nor could we find reliable user testimony for it.

used directly without postediting, 26% required slight postediting, 32% required substantial revision, and 2% were not translated at all. So far, turnaround does not represent any improvement over human translation. This application is still in the beta-testing phase and is expected to be fully operational by the end of 1991. For its other applications, most of which are in specialized domains, Hitachi estimates that it takes six person-years to fine-tune a dictionary. Once this is done, HICATS handles 60% of the sentences correctly, which means that human intervention is greatly reduced and productivity is more than doubled. Some users were said to be disappointed in the quality of output and the effort required to build the dictionaries. An experiment showed that pre-editing speeds up the process slightly (on a given job, 25.8 minutes without pre-editing vs. 23.6 minutes with pre-editing) [Kaji 88].

NEC has been using its PIVOT system for internal documentation and launched it as a commercial product in July 1990 (J/E and E/J). In mid-1991 the company reported more than 120 users outside NEC [Kazunori, personal communication]. As noted earlier, PIVOT was introduced as an on-line service to subscribers of PC-VAN in early 1991. (See Section 6.2.4.) IBS is their service bureau for all these applications. In production mode, pre-editing is done by monolingual operators with a view to reducing the time spent by professional translators. The company expects that the future of the product will be strongly user-driven. Another service bureau that uses PIVOT is Subaru International, a company that provides translations, among other services.

Oki Electric's PENSEE (J/E and E/J) is used internally and is also a commercial product. It has several active users. One unit has been purchased by MCC in the United States. PENSEE is not currently being used by any translation bureaus. Osaka Gas Co., a partner, tests the system and aids in refining the dictionary. Pre-editing and postediting are required, but they are kept to a minimum. Users are encouraged to make entries in the dictionary (except verbs). As noted earlier, PENSEE is integrated into the company's internal electronic mail service.

Ricoh's RMT/EJ, a system designed for the general office environment, has been in beta-test on the Japanese market since March 1989, and the company was in the process of incorporating user feedback before introducing it as a commercial product. Pre- and postediting are recommended but not mandatory. Pre-editing includes interactive updating of the dictionary with new entries, plus the addition of new features to existing entries. The system also has interactive postediting software that permits efficient manipulation of the output and displays up to five possible translation results, alternate translations, and the original source word or phrase for any Japanese character generated in the target [Yamauchi 88]. Ricoh's J/E effort, still in the research stage, will be limited to special-purpose applications, since the company is skeptical about the capability of current MT to produce a robust general-purpose system that translates from Japanese to English.

Sanyo has a J/E product, SWP-7800, already on the market, and the JTEC team was told about research prototypes in both directions, which have since been announced as commercial products (under the name HEAVEN JE/EJ). With SWP-7800, pre-editing consisted mainly of spell-checking and text reformatting, while an interactive mode permited selection of the target word from among several choices offered, as well as dictionary consultation. There were also facilities for postediting, and the user could update the dictionary in simple cases. There was no information available on user experience.

Sharp's DUET-E/J has been on the market since 1988. DUET-E/J II (the second generation of DUET) began to be marketed in 1990. The two systems together are reported to have an installed user base of

600 clients, mostly for the English/Japanese combination. About 45% of the applications are for general or "miscellaneous" purposes, while approximately 30% are for reports, 25% for manuals, and 2% for patents. This distribution differs somewhat from the needs of prospective users, who would like to use MT more for patents and contracts. Internally, DUET supports 90% of the company's translation needs. DUET has also been used by: (1) public agencies such as MITI (Policy Planning Information System Department) and the Australian Embassy in Tokyo; (2) manufacturers such as Nissan and Digital Equipment in Japan; and (3) printing companies and translation bureaus such as Nikkei Printing, Toppan Printing, Nagase Co., Ltd, and Subaru International. Several books were recently translated from English to Japanese using DUET to produce the first draft.

Toshiba states that they have about 80 customers who use their Translation Accelerator, ASTRANSAC[12], from English into Japanese, plus a few who work in the other direction. These are general-purpose systems, although most of their clients are makers of computers and other hardware. One customer is the Digital Equipment Corporation (see 6.2.2). Toshiba also has translation service bureaus and a trading company as users. Typically, ASTRANSAC is incorporated into the overall text production process, from input to desktop publishing, but it is also used on a personal basis. Four-fold increases in productivity have been reported with product documentation. The company emphasizes pre-editing for the J/E version and postediting for the E/J version. Dictionary updating is user-friendly. Toshiba's development team is anxious to respond to feedback from users, not only on the quality of translation but also with regard to facilities for the human interface. In a comparative test that considered each step in the translation process, a series of texts that were human-translated in 20 to 50 minutes were translated with the aid of ASTRANSAC in 8 to 27 minutes [Amano 89].

In 1987, ASTRANSAC was used in an experimental on-line satellite hookup between Japan and Switzerland as part of the 5th World Telecommunications Exhibition (Geneva, 20-27 October 1987). For purposes of this demonstration, the system was linked up to the input and output modules of an interactive dialogue system [Amano 88], so that conference participants could converse via keyboard with the Toshiba laboratory in Japan. The translations were considered to be of good quality. The main factor that impaired system performance was sloppy input, which could have been avoided in a more controlled environment.

6.4 The Broader Outlook

The fact that MT use is taking hold in Japan may be more important than measurable gains in productivity and cost (as described in the following chapter). It was seen, for example, that the IBS users of PIVOT, despite lower productivity, were satisfied with the reduced requirement for high-cost translators, unified terminology, and easier project management [Sakurai 91]. Moreover, IBS cited an advantage that might not occur to a Western manager --- namely, "teamwork is built" [Sakurai 91].

There seems to be a general sense that the steady growth of successful MT applications, with dissemination of these experiences, is the way to consolidate the technology. It is also recognized as the way to find out what MT can and cannot do --- for the Japanese are anxious to understand and deal with these realities.

---

[12]ASTRANSAC was preceded by TAURAS [Amano 89], an experimental model that is not being marketed.

7. Acceptance of MT: Quality and Productivity

Muriel Vasconcellos and Elaine Rich

It is extremely difficult to calculate the extent to which MT is in actual use in Japan. Not all the MT units that have been sold are currently in service. The JEIDA report estimates total sales at 4,000 units at the time of their survey [JEIDA 89], but goes on to add: "many are said to have been returned to the seller and some are not used and are idle." Bravice, for example, estimated the proportion of "sleeping" users at 35% of those who had purchased the system. Sharp, on the other hand, reported that they have sold 600 copies of the DUET E/J system, and so far none has been returned. A questionnaire sent to DUET's users showed that 82.6% of the total 600 units are in frequent use.

These numbers suggest that there is, at least in some cases, a difference between buying an MT system and using it effectively. This suggests that it is worthwhile to take a look at the expectations versus the reality of MT, and to see which factors have the greatest bearing on MT acceptance.

7.1 Productivity and Cost

Productivity is a function of many things, including quality of the machine-produced translations, throughput, and ease-of-use of the tools that are provided for such tasks as pre- and postediting and dictionary updating. But raw translation quality is by the far the most important of these because, as quality goes down, the amount of human intervention required goes up. This chapter explores the state of the art with respect to both quality and throughput and summarizes the overall measures of productivity obtained by the JTEC team. Several cases in which MT improved productivity were cited in the last chapter. With product documentation and other specialized applications, there have been reports of productivity gains ranging up to 400%, with concomitant improvements in turnaround and savings in cost. On the other hand, productivity gains of about 30% are typical of general-purpose applications. It is mostly in these latter areas that there have been some reports of user frustration and, in some cases, actual loss of productivity.

Increased productivity means faster turnaround. Winning the race to get products to market, especially the U.S. market, can be vital to corporate profits, and even survival. In many cases MT, by speeding up the translation of essential documentation, has definitely made it possible to accelerate product delivery, thereby achieving the goal for which it was developed.

Increased productivity also means lower translation costs. For a fully postedited product, the cost is typically between 65% and 75% of the cost of traditional human translation. While this is certainly a motivating factor, it does not appear to be the primary one in Japan, where there is general recognition of the fact that investments have to be made in order to get past the language barrier.

The up-front cost of MT systems is apparently not an issue. (Although it is worth noting that the software cost of MT systems is quite low. Almost all sell for under US$15,000.) At the JEIDA/JTEC meeting in Japan, the JEIDA members voiced a nearly unanimous lack of concern about the price of their product. On the contrary, they said, the average corporate purchaser of MT in Japan may be suspicious of a low-priced product.

7.2 Translation Quality

Both users and vendors of MT systems in Japan realize that the quality of the translations produced is overwhelmingly the most important factor in determining the usefulness of MT. As part of a survey for the JEIDA Report, 27 companies commented on their ideas and expectations about machine translation [JEIDA 89]. The sample had a heavy proportion of general companies that had not yet had direct contact with MT, but it also included translation bureaus and a few firms that are already using MT. By far the most frequent objection to MT was that the quality was poor. Only 17% of all respondents felt that MT was usable for rough translation (14% of the general companies, 25% of the translation bureaus, and 50% of the firms already using MT). Each group, of course, brings a different perspective. From the responses to these and other questions, it may be concluded that the general companies (mainly firms that do not yet use MT) want an MT system that does not require human intervention. In other words, they want a system that is fully automatic and that produces high-quality machine translation. They are apparently not interested in MT as a tool for translation support. Many professional translators, on the other hand, even when they are not yet using MT, already recognize its potential, and those who are using it (albeit a much smaller proportion of the sample surveyed) are more willing to accept the output. Of course, it should be kept in mind that the groups of respondents are to some extent self-selected: if they do not use MT, whether in general companies or translation bureaus, it may well be that they are predisposed against it, since they certainly have not lacked for opportunities to try it out.

1	Productivity in science itself is often measured in terms of papers produced, presented, and/or published.	科学自身のＰｒｏｄｕｃｔｉｖｉｔｙは、たびたび、生み出されて、提出された論文について測定されて、および／または、出版される。
2	Another frequently used measure of creativity is the number of patents applied for, granted, etc.	創造的の別の頻繁に使用された物差しは、申し込まれた特許権の数であるとか、与えたなど。
3	According to The National Science Board, the U.S. share of the world scientific and technical articles in engineering and technology dropped 10% from 1981 to 1986.	国民科学ボードに従って、エンジニアリングと技術の世界科学および技術記事のアメリカ合衆国分け前は、１９８１から１９８６まで１０％を低下させた。

Figure 7-1: Example 1: One English to Japanese Translation

The opinions that the JTEC team heard during its visit echoed the JEIDA report in their emphasis on the importance of translation quality. When the JEIDA group itself met with the JTEC team, most of the JEIDA members stated that their own view was that quality is the most important factor in increasing user acceptance of MT. They also said that their own highest priority was to improve translation quality in their MT systems. This is not surprising, since some of the developers visited by JTEC mentioned explicitly that they had received complaints from their users about raw MT quality and/or the effort needed to customize the systems.

1	Productivity in science itself is often measured in terms of papers produced, presented, and/or published.	科学自体における生産力は、製造された新聞に関してしばしば測定されます（示される、かつ、または、公表される）。
2	Another frequently used measure of creativity is the number of patents applied for, granted, etc.	創造性の他の頻繁に使用された方法は、〔に〕適用できられる承諾された特許の数、などです。
3	According to The National Science Board, the U.S. share of the world scientific and technical articles in engineering and technology dropped 10% from 1981 to 1986.	The National Science Boardによれば、工学、及び、技術における世界の科学の、そして、技術的な品物の米国シェアは、１９８１年から１９８６年まで１０％減少しました。

Figure 7-2: Example 2: A Second English to Japanese Translation of the Same Text

The pressing need for quality translation stems from two major concerns: poor quality implies increased human intervention in the best of cases, and in the worst of cases it means that the MT output is unintelligible and/or irreparable. Human intervention is costly, and the more it can be minimized or rationalized through the assignment of less skilled operators, the more the technology will show savings for its users in terms of both time and money. If the output is totally unusable, the technology has failed. Because of the linguistic distance between Japanese and English, the language pair to which most of these comments implicitly refer, these MT systems are especially vulnerable: a poor machine translation system can produce unintelligible output no matter how enthusiastic and forgiving the consumer is. Although some of the companies were reluctant to disclose the criteria they use for evaluating translation quality, one of them offered the following five-level scale:

1. Failure or error in syntactic analysis.
2. Intended meaning not conveyed because of inappropriate word choice.
3. Basically meaning conveyed, but with minor errors.
4. Literal translation: no grammatical mistakes but expressions lack refinement.
5. Naturally expressed correct translation.

Much depends on the application, of course. Specialized applications produce more reliable results and therefore require less postediting. For example, HICATS was found to handle 60% of its sentences correctly in one of its applications. In the case of IBS (Section 6.2.4), 40% to 60% of the DUET E/J output required pre-editing, but the rest of the input was considered adequate to submit directly.

One aspect of quality to which MT makes a positive contribution is the standardization of terminology. A number of users in Japan have reported that they are pleased with the role MT has played in this regard, and they have even remarked that the presence of MT has helped to improve consistency in original technical writing.

1

Productivity in science itself is often measured in terms of papers produced, presented, and/or publish ed.

科学それ自体の生産性は、生産される、そして、／を贈られる、または、出版された新聞の点からしばしば測られる。

2

Another frequently used measure of creativity is the number of patents applied for, granted, etc.

創造性の別のしばしば中古の寸法は、かなえてやるその他られる思い出られる特許の数である。

3

According to the national science board, the U.S. share of the world scientific and technical articles in engineering and technology dropped 10% from 1981 to 1986.

合衆国は、世界科学のと技術的な品で１９８１から１９８６まで１０％を落ちられるテクノロジーをそして設計することにおいて国立科学委員会によって分け合う。

4

The Japanese scientific position, measured by papers produced, has been rising 0.5% per year.

日本の科学の位置は、生産される新聞によって測られると、年について０．５％を上っている。

Figure 7-3: Example 3: A Third English to Japanese Translation of the Same Text

Before embarking on its site visits, the JTEC team selected one short passage in English and one in Japanese. At each of the sites where there were operational MT systems we asked our hosts to try our sample texts on their system, going in which ever direction(s) they supported. Almost all were willing to try. It is, of course, difficult to conduct a quantitative evaluation of the results of this informal experiment. For one thing, not all the systems came at the task equally well prepared. Some of the systems are intended to be general purpose and so should do well with a randomly selected text. Others are intended to be customized for particular domains; they in general did not have the necessary vocabulary to handle our examples without some tuning first. But Figures 7-1 through 7-6 show examples of the results we observed. As a basis for comparison, we asked a professional translator to translate the Japanese text into English. He produced two translations for each sentence, one fairly literal (labeled A) and one that he described as "natural" or "idiomatic" (labeled B). His translations were:

Sentence 1

A. High-tech industry grew rapidly, and one day, Japan suddenly appeared to have become a world super-nation.

B. Japan's high-tech industry grew quickly. Virtually overnight, it seemed, the nation had become an economic superpower.

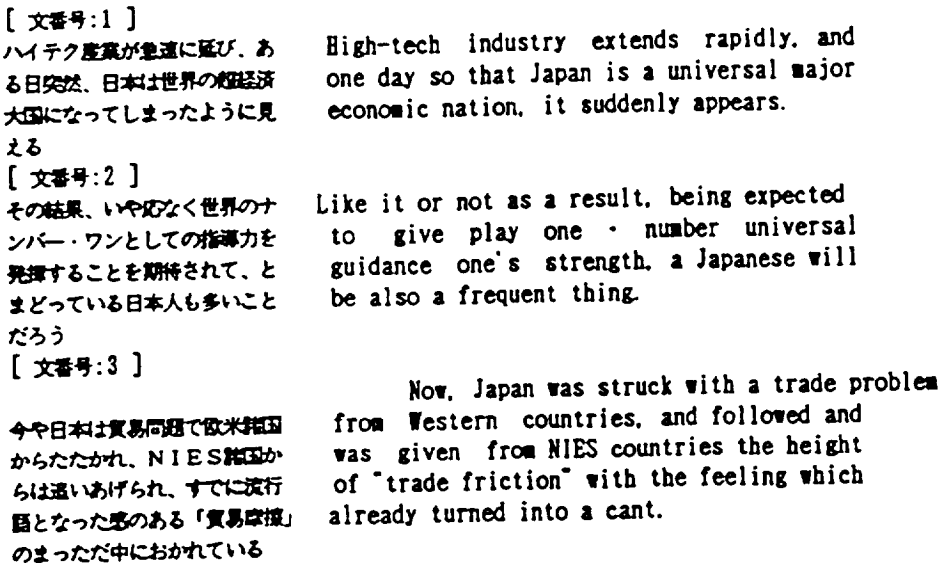

Figure 7-4: Example 4: One Japanese to English Translation

Sentence 2

A. As a result, whether they like it or not, Japanese people are expected to display leadership as No. 1 in the world, and there are probably many confused Japanese.

B. Like it or not, Japan is now expected to display the kind of leadership befitting its status as "No. 1" in the world. This new role undoubtedly leaves many Japanese people perplexed.

Sentence 3

A. Nowadays, Japan is being bashed by the European and American countries over trade problems, chased by the newly industrialized economies, and has become the center of trade friction, which has already acquired the atmosphere of an "in" expression.

B. Japan has become the whipping-boy of Europe and America in countless trade disputes, and the target of competitive pressures to catch up with it from the newly industrializing economies of East Asia. Japan finds itself in the midst of a veritable storm of "trade friction" – a phrase that has already become a household word throughout the land.

The translator also made the following note: The author appears to stretch the meaning of trade friction (written in kanji in a literal translation from English to Japanese) to include purely competitive pressures from SE Asia. Japanese often adopts English words and then modifies their meaning – a real pitfall for translators, machine and human alike, if they are not truly bilingual.

7.3 Throughput

Although it is difficult to compare throughput numbers across systems because of differences in both hardware platforms and kinds of texts, we include some representative throughput figures in Figure 7-7 to provide a rough idea of the order of magnitude of speeds at which current systems perform. All the numbers are given in words per hour. When we could, we included in parentheses after the speed the hardware platform on which the speed was obtained. These numbers came from two different sources.

Raw unedited text:

SENTENCE NAME = 111
ハイテク産業が急速に伸び，ある日突然，日本は世界の超経済大国になって
ARTC NODE FOUND/SKIPPED(@S1192)
しまったように見える。

Japan is visible in order to high tech manufacturing improve rapidly, and in order to become super-economy large country in world suddenly some.

SENTENCE NAME = 112
その結果，いや応なく世界のナンバー・ワンとしての指導力を発揮すること

ARTC NODE FOUND/SKIPPED(@S1192)
を期待されて，とまどっている日本人も多いことだろう。

As a result, an abounding thing will also be the Japanese whom it has been at loss by expecting that there is and that there is not 応, and that it demonstrates the leadership as no. ワン of world.

During pre-editing, sentence 1 was divided up and was eventually translated as:

High-technology industry in Japan developed rapidly.

Then Japan became super-economy large country.

Figure 7-5: Example 5: A Second Japanese to English Translation of the Same Text

Those labeled (1) were taken from literature supplied by the vendors at MT Summit III in July 1991. Those labeled (2) were taken from [JEIDA 91].

It is important to keep in mind that these numbers reflect only the time required to run the texts through the MT system. As we have described elsewhere, the total time required to produce usable translations will be much larger, since other things, such as pre- and postediting are also required.

7.4 Customization

There is increasing awareness in Japan of the importance of user-friendly tools for customizing an MT system. Many developers rank this factor second only to quality in importance, and one of them felt it was even more essential in terms of gaining user acceptance. All MT applications require some fine-tuning, and users will have a sense of control over their systems when they can readily elicit usable translations that incorporate their own terminology and other preferences. Mainly, this means developing easy and efficient interfaces for pre-editing, postediting, and dictionary updating. But it can also involve interactive variations such as those described for Ricoh, as well as basic tools including corpus extraction, indexes of key words in context, etc.

1 ハイテク産業が急速に伸び、ある日突然、日本は世界の超大国になってしまったようにみえる。

2 その結果、いや応なく世界のナンバーワンとしての指導力を発揮することを期待されて、とまどっている日本人も多いことだろう。

As a result, the expected and confused Japanese will be many things of showing guide power as ワン number in the world where いや応 quacks.

3 いまや日本は貿易問題で欧米諸国からたたかれ、NIES諸国からは追いあげられ、すでに流行語となった感のある「貿易摩擦」のまっただ中におかれている。

On Japan of doing having まや is beat from a European and American various countries by a trade question, it finishes being chased from NIES various countries and the "trade friction" placing of it which has a sense of becoming a popularity word already is done in まっ 's being free.

Figure 7-6: Example 6: A Third Japanese to English Translation of the Same Text

Company	System	Languages	Speed (w/h)	Source
Catena	STAR	E/J	15,000	(1)
Catena	The Translator	E/J	10,000-20,000 (MAC IIcx)	(1)
Fujitsu	ATLAS-II	J/E	60,000 (FACOM M380)	(2)
Hitachi	HICATS	E/J	30,000-60,000 (HITACM-680)	(2)
Hitachi	HICATS	J/E	20,000-60,000 (HITACM-680)	(2)
Matsushita	PAROLE	J/E	30,000 (Solbourne)	(2)
Mitsubishi	MELTRAN	J/E	5,000-10,000 (MELCOM PSI/UX)	(1)
NEC	PIVOT	J/E	>30,000 (EWS4800)	(1)
NEC	PIVOT	E/J	>20,000 (EWS4800)	(1)
NTT	ALT-J/E	J/E	5,000 (VAX8800)	(2)
Oki	PENSEE	E/J and J/E	15,000	(1)
Ricoh	RMT-E/J	E/J	4,500	(2)
Sharp	DUET	E/J	12,000	(1)
Toshiba	ASTRANSAC	E/J and J/E	10,000-20,000	(2)

Figure 7-7: Throughput Rates for Selected MT Systems

7.5 Integration

MT brings greater savings in terms of both time and money when it is fully integrated into the publication chain. The two key factors in this regard are input of the source text and generation of MT output files that retain typesetting codes. The problems of text input were discussed in Section 6.1.1. Time is at least as important as cost. Manual keyboarding of MT source text can cut deeply into the advantages to be gained by using the technology. This problem is being solved to some extent with OCR, but even more important, the widespread use of word processing, coupled with the capacity to transmit electronic files via network, means that a steadily increasing proportion of input text is already going to be machine-readable. Up to now, OCR for such languages as Japanese, Chinese, and Arabic has posed a major challenge. The availability of OCR for Japanese will be especially important for U.S. efforts to access information in that language because of the difficulties that Westerners face in inputting kana and kanji. At the output end, many MT installations in Japan are already linked to desktop publication. The remaining challenge is to be able to capture all formatting codes, because this capability will greatly reduce time and effort spent.

7.6 Open Systems and Software Portability

Increasingly in the United States, software is built to be as independent of particular hardware platforms as possible. The idea of open systems, in which customers can buy pieces of hardware and software from multiple vendors and be assured that they will work together, has begun to dominate the computer industry. At least with respect to MT, this has not yet happened in Japan. Most of the commercial MT systems have been developed by hardware platform vendors. Their MT systems run only on their hardware. Thus users must have access to the appropriate hardware before they can run the MT system of their choice. The availability of some MT systems over commercial networks (as described in Section 6.2) reduces this problem somewhat from the user's point of view. Nevertheless, the close connection of the main MT systems to particular hardware platforms suggests that many of the vendors may be looking ahead to a time when MT will be widespread and will create a new demand for their hardware. We expect, however, that at least some of the MT vendors will make their MT software available on internationally standard operating systems, independent of hardware. The increasing acceptance of the UNIX family of operating systems, for example, makes this trend inevitable even in Japan.

8. MT Contrasts between the United States and Europe

Yorick Wilks

Although this report focuses on state of the art of MT in Japan, it is useful to look briefly also at the MT picture in the United States and in Europe for a broader perspective. For a more detailed description of some of the efforts that are mentioned here, see [Hutchins 86].

To compare MT in Japan with that in the United States and Europe, one must distinguish between the true and mythical histories of the technology in these last two locations. The widely-believed mythical history, shown in Figure 8-1, states that MT began in the U.S. in the late fifties. Funding peaked when the ALPAC report recommended that federal funding be withdrawn [ALPAC 66]. Europe and Canada then continued their MT work and, since 1977 or so, have been joined by Japan, now the main contributor to MT research and development.

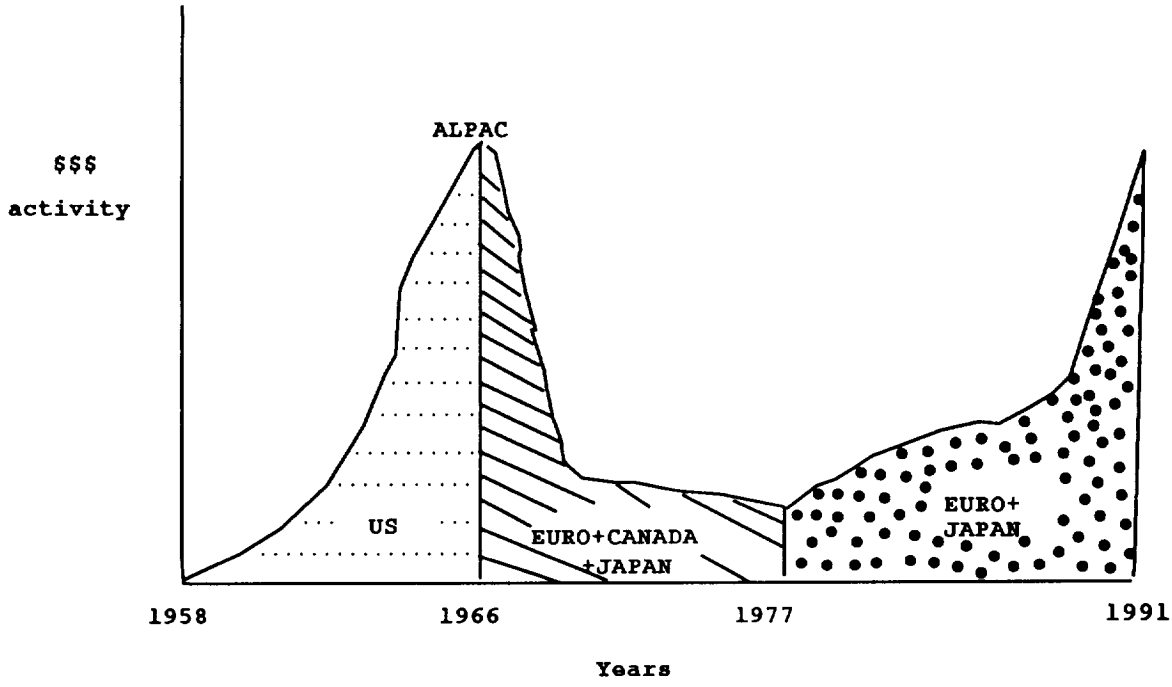

Figure 8-1: An Incorrect Model of the History of MT Development

But this picture is misleading in several ways (see Chapter 1). First, it ignores the substantial contribution made to the earliest MT development by the USSR and Western Europe. Second, it ignores the fact that some Europeans (particularly the British, who had developed several major systems in the sixties) also stopped government funding of MT as a result of the ALPAC report. That the French are widely believed to have made the major European MT effort is largely a result of their refusal to be influenced by U.S. trends, while the Canadian decision to start MT development in the mid-sixties was not so much a result of ignoring U.S. influence as Canada's enormous translation needs, imposed by legislation that equalized the federal status of English and French. Third, this picture ignores the fact that MT work got going in Japan almost as soon as it did in the U.S., although the Japanese effort did not

scale up seriously until the 1980s.

A final fallacy in the mythical history of MT is that U.S. work ceased shortly after the publication of the ALPAC report. In fact, U.S. defense funds continued (although at a lower level) to support MT work at Texas, Berkeley, and other sites. The use and development of the SYSTRAN system at the U.S. Air Force's Foreign Technology Division in Dayton, Ohio was begun during that period and continues to the present. MT has recently undergone a serious resurgence in the U.S., particularly with respect to work on interlingual systems.

Thus the true history of MT is better captured diagramatically in Figure 8-2.

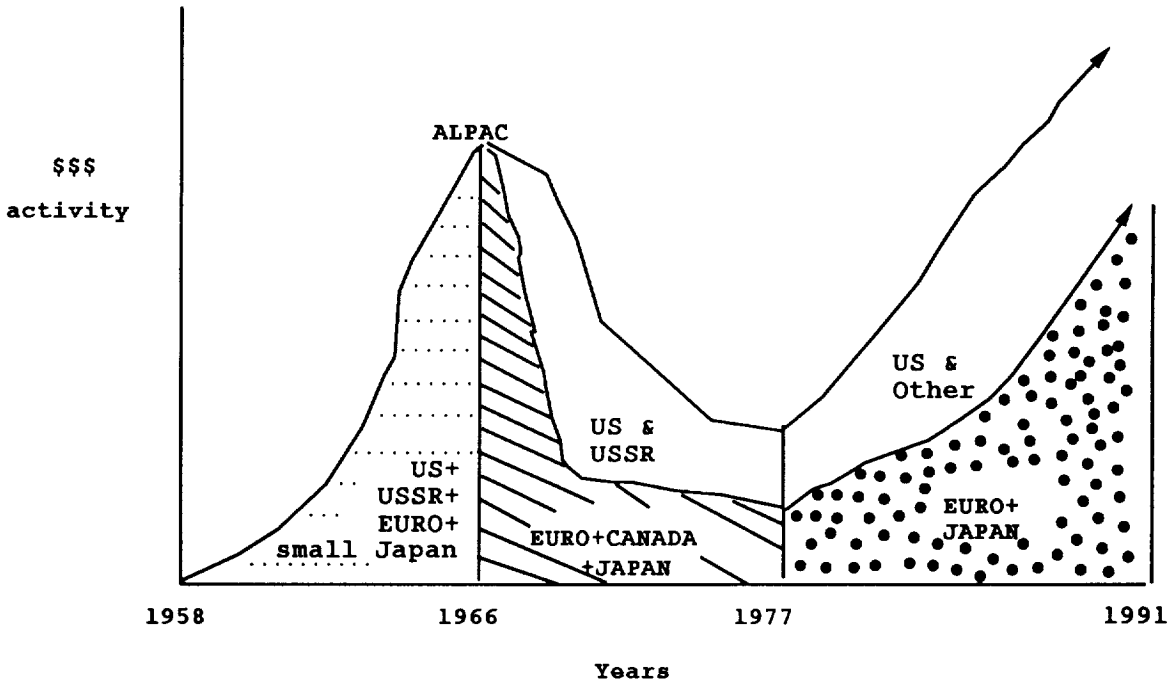

Figure 8-2: A Better Model of the History of MT Development

One mysterious aspect of MT history is the Westerner's nearly total ignorance of Soviet research, although there is a great deal of unclassified material on it, much of which is available in translation (e.g., [Melchuk 63]). Paradoxically, the USSR's greatest influence on MT has probably been exporting of researchers whose first spoken and research language is Russian (e.g., Raskin and Nirenburg in the U.S.; Perschke in the European Community).

While it is difficult to obtain precise figures for government-funded MT in the U.S., Japan, and Europe during the last 15 years (the approximate period of Japan's MT growth), rule-of-thumb figures would probably be US$20 million, US$200 million and US$70 million respectively. The European figure is largely accounted for by the EUROTRA project and the U.S. figure by support of the FTD work with SYSTRAN.

It would be a great mistake, however, to assume that those figures offer simple mapping of the MT quality R&D in the three zones. Even though this report is not devoted to the topic, an impressionistic

C-2.

verdict (based in part on the JTEC survey) is that the EUROTRA project has not yet produced a system beyond a very limited demonstration (see below), and the huge expense in Japan has produced a handful of systems that compare in breadth, speed, and quality with the best of SYSTRAN's language couples (but keep in mind that J/E and E/J are not among SYSTRAN'S best pairs).

Interestingly, EUROTRA's goal was explicitly to equal and then surpass SYSTRAN through research advances --- precisely what it has not achieved. This goal was imposed by the EC Information Science Directorate, which had also purchased SYSTRAN for trial use in the seventies, four years before EUROTRA funding began. SYSTRAN has been substantially extended in Luxembourg, and is now being used for rough internal translation of certain classes of memoranda. Its use is increasing.

If further correction were needed of the mythical view that Japanese (if not European) MT has already taken over the world scene, one could turn to Figure 8-3 (taken from [JEIDA 89]). The table lists uses of MT systems by organizations in the U.S., Europe, and Canada. Notice that only one of the systems listed (ATLAS) originated outside North America.

a) Utilization in Canada

Canadian government	TAUM-METEO
Canada GM	SYSTRAN

b) Utilization in the United States

U.S. government	SYSTRAN
NASA	SYSTRAN
U.S. Air Force	SYSTRAN
XEROX	SYSTRAN
Caterpillar	SMART
PAHO	PAHO

c) Utilization in Europe

CEC	SYSTRAN
	ATLAS-II
NATO	SYSTRAN
KFKS	SYSTRAN
Minitel	SYSTRAN

Figure 8-3: Origination of MT Systems Used in Europe and North America

8.1 Major MT Centers and Systems in the US

A partial list of MT R&D groups in the U.S. follows. It is a list subject to rapid change (e.g., the CMU, CRL, ISI, and IBM groups have only recently been strengthened by the new DARPA initiative in MT). The list also excludes smaller commercial MT groups such as Alps, Smart, Globalink, etc., as well as smaller university-based groups at Hunter College, Monmouth College etc.:

- Carnegie Mellon University's Center for Machine Translation.
- Computing Research Laboratory at New Mexico State/Tradux.
- Linguistic Research Center at the University of Texas at Austin (originators of METAL).

- IBM/Yorktown Heights. One group is pursuing the statistical approach, another a traditional one.
- Pan American Health Organization (PAHO).
- SYSTRAN development group.
- LOGOS development group.
- New York University (sublanguage translation).
- University of Southern California's Information Sciences Institute (ISI).
- MCC.

Other than CMU and CRL, none of the groups has more than ten people. Although the list is incomplete, it should be clear that levels of personnel and funding are considerably lower in the U.S. than in Japan. More details about system type, languages and materials treated, etc., can be seen from the table of four major U.S. MT products, shown in Figure 8-4.

	Languages	Type	Topics	Status	Organization
LOGOS	E,G,S,F	Transfer Some semantics	Manuals and general	Commercial U.S. owned	Logos Corp.
SYSTRAN (at FTD)	E,R	Direct	Technical literature	U.S. Govt.-funded	LATSEC, Inc.
SYSTRAN (multitarget)	E,F,S, etc.	Transfer Some semantics	Manuals and general	Commercial	SYSTRAN
METAL	E,G	Transfer Case-frame semantics	Manuals and general	Commercial	Siemens/ Nixdorf

Figure 8-4: Major US MT Products

However, we should not confuse progress and the state of the art only with dollars spent and researchers and/or developers employed. No comparison between Japanese and American or European work would be balanced unless we stressed the pivotal role of SYSTRAN [Toma 76]. At this time, SYSTRAN remains the international benchmark for MT, one that Japanese (or any new U.S. or European efforts) would have to improve upon demonstrably in order to have pulled ahead of the U.S. in MT. Conveniently, SYSTRAN is also a benchmark in that its levels of achievement are fairly well fixed, and have barely shifted upwards in terms of percentage of correctly translated sentences for its principal languages for many years [Wilks 91]. SYSTRAN proved not only that MT really works, in the sense of satisfying a substantial class of users but also that stamina (i.e., long-term commitment of funds and effort to MT) pays off. The Japanese have certainly taken this lesson to heart. The Europeans have done so to a lesser extent, as exemplified by their commitment to the EUROTRA program (see below).

8.2 Influences among MT Groups

Simply listing current MT groups in the U.S. leaves out many important effects at work: effects and relationships that are essential to understanding how international MT has become, as has all of science, engineering and commerce.

Many of the influences among MT groups are explicit, and are related to migrations and sabbaticals of researchers. Others are the effect of the close proximity of groups, or of the decay of one and the rise of another. Sometimes influences are connected to the movement of key personnel. Others are simply academic and intellectual. Still others are the result of commercial sale (e.g., Weidner to Bravice and Texas' METAL to Siemens/Nixdorf). These influences are summarized in Figure 8-5. Notice that both theoretical and software connections are intercontinental.

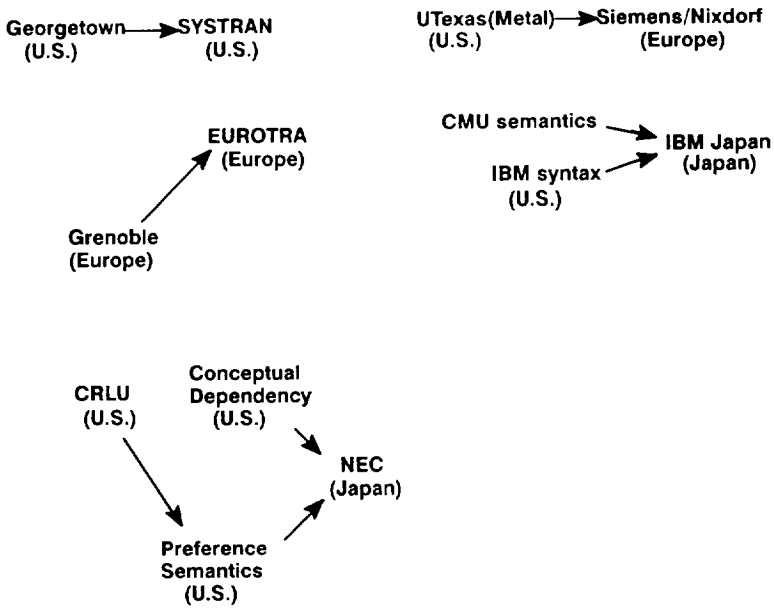

Figure 8-5: Influences on MT Efforts

These same connections can be illustrated by mapping major systems on a time line, as shown in Figure 8-6, which is a more detailed version of Figure 1-1. Here the systems are divided into three main system types: direct, transfer, and interlingual. Figure 8-6 shows that there is a gradual movement towards methods that exploit semantic analysis and this shift transcends continental boundaries.

Other factors also make it increasingly difficult to classify MT work as simply belonging to a particular continent or country. The following illustrates a trend away from such simple identifications:

- IBM Japan's new MT system (JETS) has had substantial research contributed by the United States.
- Systran's J/E and E/J systems are currently owned by a Japanese company (Iona), although that may change.
- SYSTRAN itself is closely tied to French business interests.
- Fujitsu's J/S system is under development at Fujitsu Espana in Barcelona.

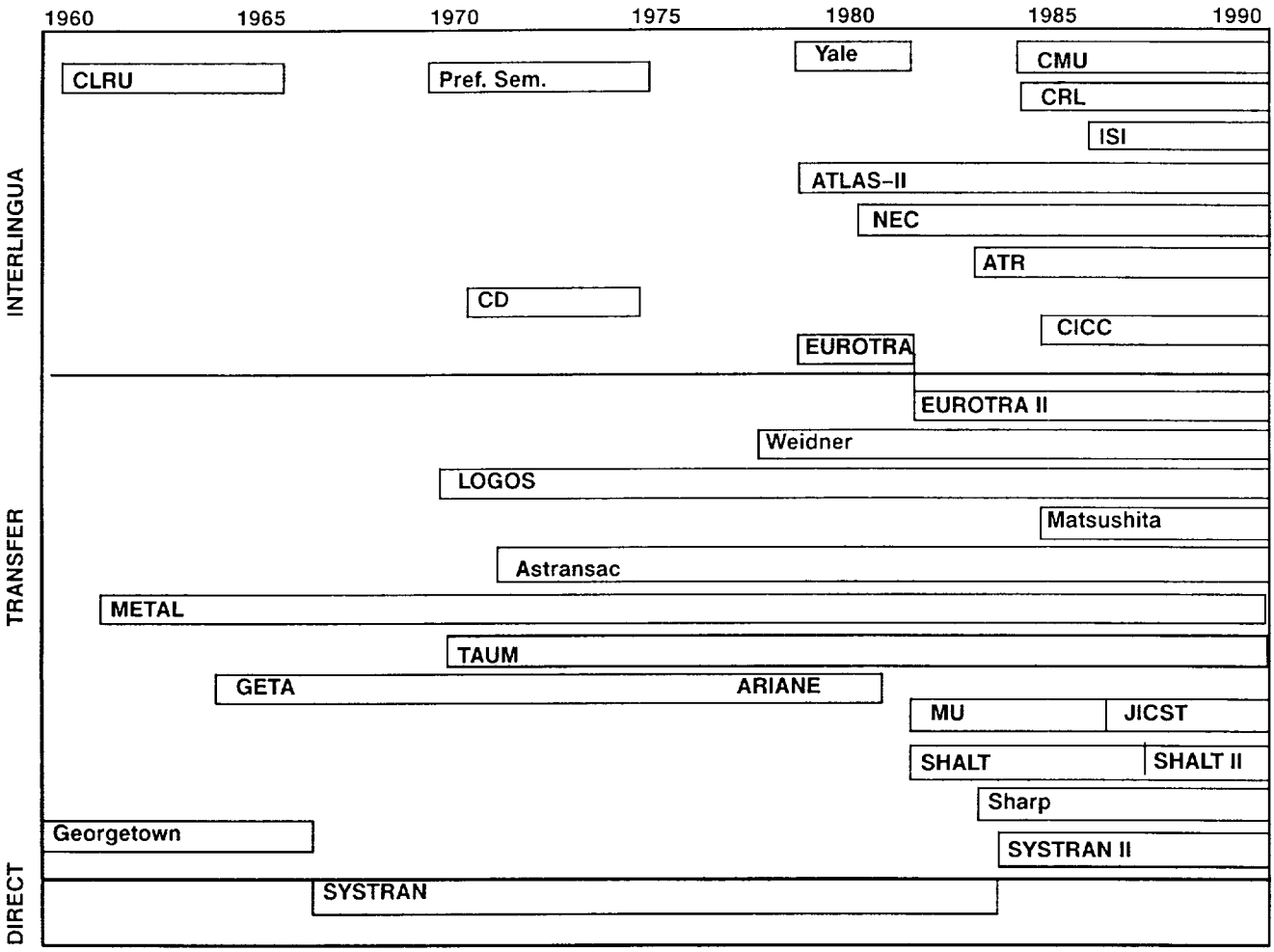

Figure 8-6: Past MT Systems: A Time Line

- This year, both Canon and Sharp have set up new MT R&D labs in the United Kingdom.
- The Texas METAL system is now owned by Germany (Siemens/Nixdorf), although development continues in both the U.S. and Germany.

8.3 Current European Systems

For more than a decade after 1966, GETA [Vauquois 84] at Grenoble was the best known European system, and since 1980, EUROTRA has occupied that position. They have both been supported by enormous quantities of government funds and were defensible in terms of the linguistic theories of their time, but have never worked very well.

EUROTRA [Johnson 85] was a bold initiative formally launched in 1982, financed by Directorate XIII of the CEC (the European Commission) in Luxembourg and by the Trade and Industry ministries of most of the European Community (EC) states. It was set up to meet the translation needs of the Commission and Community which, for some documents, required translation into all nine languages of the EC (implying the ability for 72 pair-wise translations). Some documents appear only in the six principal languages, and many central documents appear only in the three core languages (English, French, German) since there is an assumption that all "eurocrats" can read one of these. But even these latter two methods imply enormous volumes of translation, given that the EC is a larger entity than the U.S., both in population and

GNP.

The EUROTRA program for a multilingual MT system, to be constructed by multistate teams, was initially impressive. It has already cost over US$50 million, but has only a very small demonstration program to show for it at the present time (even though the formal funding is now coming to an end and is continuing only for programs to extend the lexicon). Its initial aim was to translate CEC documents within the Luxembourg bureaucracy, but this has now been scaled back to the mere treatment of examples.

The best features of the EUROTRA methodology were the separation of software methodology from linguistic specifications, and the separation of the work into modules concerned with particular languages (though not necessarily based on particular states). These features have been incorporated into other systems (e.g., at CMU and in ULTRA [Farwell 90] in the U.S., and in SWETRA, and SWISSTRA [Estival 90] in Europe.)

The initial design in the early 1980s was a compromise between those who wanted the system to be basically interlingual and semantics-driven, and those who wanted to preserve as much as possible of the GETA representational structure (multilevel dependency trees to encode semantic, syntactic, and morphological information). This compromise held until about 1984 when detailed implementation was due to begin.

Around that time, the whole design team was reconstructed, although the top-level project leadership was unchanged, and EUROTRA went through a series of changes motivated almost entirely by considerations of linguistic fashion. This led to the present situation in which EUROTRA is transfer-based, and the representation is a form of unification grammar, the system being entirely syntax-driven.

That EUROTRA has produced so little for so much investment is significant and instructive for efforts elsewhere. If it is a failure, it has been almost entirely a management failure. Its concentration on linguistic issues, at the expense of engineering and implementation ones, contrasts with Japanese and benchmark U.S. approaches. It ignored what one might call a golden rule for almost any prototype and product: to settle on a representation, stick to it, and develop it to its maximum capabilities.

The problems EUROTRA has experienced to date should not cause one to ignore other approaches in Europe, such as SUSY (at Saarbruecken), SWETRA and SWISSTRA, all of which have contributed to and benefited from EUROTRA, despite the recent cancellation of the EUROTRA effort.

One might end with the following remarkable example, shown in Figure 8-7, which in a way, sums up the MT relationship between the U.S. and Europe. The figure is the first part of a substantial public document released by the European Commission at a language trade fair in 1989. It describes the EC's investment in EUROTRA and, since such documents must appear in at least English and French, it adds on the right a translation provided by the SYSTRAN E/F version the EC has worked on for over 12 years, with the added header "TRADUCTION BRUTE SYSTRAN", indicating that it is raw SYSTRAN translation, although the document may have been postedited without removing that label. It is clear though, that somewhere in the EC, someone has a sense of humor...

Direction Générale
Télécommunications, Industries de l'Information et Innovation

ORIGINAL

Contribution for DG XIII brochure
EUROTRA

EUROTRA is a Community research and development programme for the creation of a machine translation system of advanced design capable of dealing with all the official languages of the EC. It was adopted by Council Decision 82/752/EEC of 4 November 1982 and extended by Council Decision 86/591/EEC of 26 November 1986 to include Spanish and Portuguese following the accession of Spain and Portugal.

The programme is jointly financed by the Community and its member States. Its objective is the creation of a prototype system which would be operational for a limited subject field and for a limited number of

TRADUCTION BRUTE SYSTRAN

Contribution pour la brochure de DG XIII
EUROTRA

Eurotra est un programme communautaire de recherches et de développement pour la création d'un système de traduction automatique de conception avancée capable de traiter de toutes les langues officielles de la CE. Il a été adopté par la décision 82/752/EEC du Conseil du 4 novembre 1982 et élargi par la décision 86/591/EEC du Conseil du 26 novembre 1986 pour comprendre espagnol et portugais après l'adhésion de l'Espagne et du Portugal.

Le programme est conjointement financé par la Communauté et ses Etats membres. Son objectif est la création d'un système prototype qui serait opérationnel pour un domaine limité et pour un nombre limité de types de texte

Figure 8-7: An Example of EUROTRA Output

9. Research and Development

Elaine Rich

There is a strong and longstanding commitment in Japan both to MT research and to a technology transfer process that has been very successful in moving results from research labs into development organizations. A relatively early example of this was the government-sponsored project that led to the development of the MU system. This project required the close collaboration of four research organizations. Its structure was described in [Nagao 85] as follows:

> At Kyoto University, we have the responsibility of developing the software system for the core part of the machine translation process (grammar writing system and execution system); grammar systems for analysis, transfer and synthesis; detailed specification of what information is written in the word dictionaries (all the parts of speech in the analysis, transfer, and generation dictionaries), and the working manuals for constructing these dictionaries. The Electrotechnical Laboratories (ETL) are responsible for the machine translation text input and output, morphological analysis and synthesis, and the construction of the verb and adjective dictionaries based on the working manuals prepared at Kyoto. The Japan Information Center of Science and Technology (JICST) is in charge of the noun dictionary and the compiling of special technical terms in scientific and technical fields. The Research Information Processing System (RIPS) under the Agency of Engineering Technology is responsible for completing the machine translation system, including the man-machine interfaces to the system developed at Kyoto, which allow pre- and post-editing, access to grammar rules, and dictionary maintenance.

This research prototype has since been used as the basis for the production MT system currently in use by JICST.

As this example suggests, part of the reason that the Japanese have been so successful at making use of their MT research results may be the diversity of organizations within which the research is being done. Current R&D efforts are taking place within four kinds of institutions in Japan: academic, industrial, government, and consortium labs. Historically, the most visible academic lab has been the one at Kyoto University under the direction of Professor Makoto Nagao. Other labs also exist at Tokyo Institute of Technology, Osaka University, and Kyushu University, among others. All of the industrial sites that have previously developed production MT systems also have ongoing research and development projects (although some are primarily focused on development rather than on research). These include NEC, Fujitsu, Hitachi, IBM Japan, CSK, Oki, Sanyo, Toshiba, NHK/Catena, Sharp, and Bravice. In addition, Matsushita, Ricoh, and NTT have begun research efforts. Substantial MT efforts are underway at two government-supported labs, CICC and JICST, and two consortia, ATR and EDR, are doing work in MT or related technologies. In addition, there is a substantial amount of basic research in natural language processing at many of those same institutions, as well as at ICOT, ETL (Electrotechnical Laboratory), and many universities.

These efforts are focused on the following major areas:

- New overall approaches to translation, including:
  - Interlingua-based translation
  - Example-based translation
  - Transfer-based translation
- New grammatical frameworks
- New approaches to target text generation

- Development of dictionaries
- Treatment of discourse-level phenomena
- Better tools for users
- Extension of the Japanese and English systems to additional languages
- Speech-to-speech translation
- Embedding MT into larger information-processing systems

This chapter will describe the work in each of these areas in more detail.

9.1 Interlingua-Based Translation

Chapter 1 showed that MT systems could be described as falling along a continuum defined by when the transfer from the source to the target is performed. At one extreme would be systems that translate sentences directly (i.e., with no prior analysis to determine their internal structure). This overall approach is not used in any current system but it can be used as part of a more comprehensive MT architecture. At the other extreme are systems that map the source text into a complete meaning representation and then map that back out to target text, with no actual transfer between the two languages at all. In the middle are systems (including most of the ones now in existence) in which some analysis is performed. Next a transfer step maps the analyzed structure into a corresponding structure in the target language. Finally target generation takes the remaining steps toward constructing the final target text. Figure 1-15, repeated here as Figure 9-1, shows some of the major points along this spectrum.

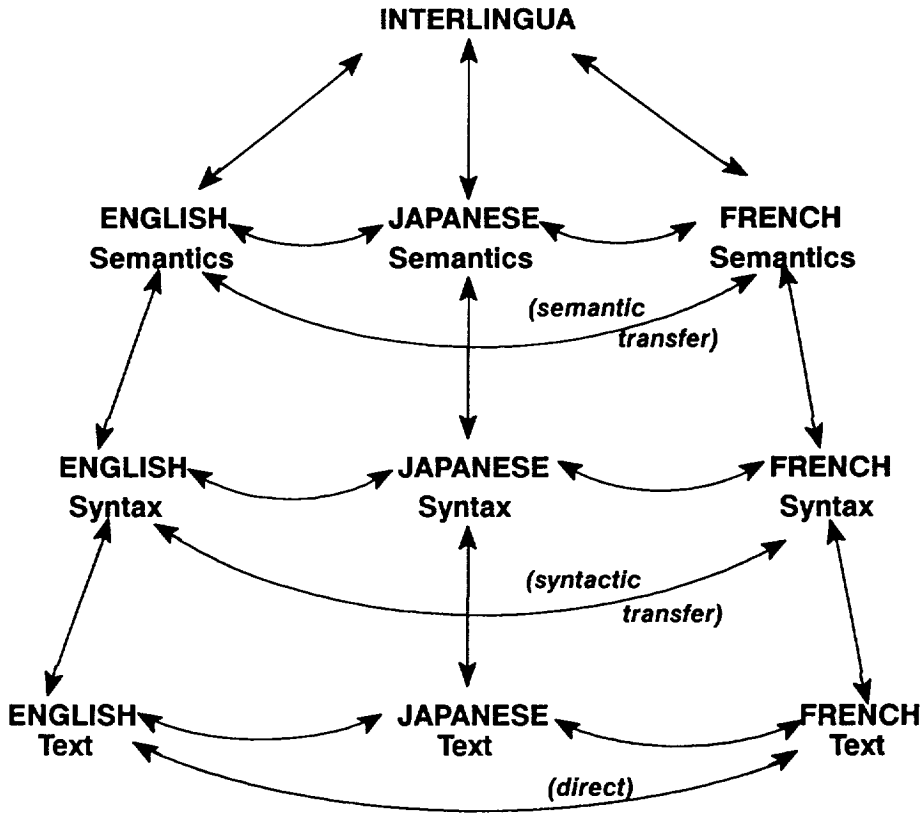

Figure 9-1: The Transfer-Interlingua Dimension

For a long time, almost all of the MT systems built in Japan could best be described as syntactic transfer systems. They followed the line of arrows shown second from the bottom in the figure; that is, they analyzed the source text into a syntactic form, then applied transfer rules, sometimes to transform the source syntactic form itself, and then, in any case, to derive a target form, which then served as the basis for generation. Some systems began to move higher up the diagram, including Fujitsu's ATLAS-II [Uchida 89a], NEC's PIVOT [Muraki 89, Ichiyama 89], and CSK's ARGO. These systems build a partial semantic representation of the source sentences, and use that either as the basis for a transfer operation or as a form of interlingua. The semantic representations they use are based on the idea of a case-frame structure, in which each sentence is represented as a major predicate and a set of semantic roles, each of which is filled by some major constituent of the sentence. These constituents, in turn, are represented as structures that are built out of a set of semantic primitives that describe the entities that can exist in the domain of discourse. But, as shown in Figure 9-1, an explicit transfer step is still required because case frames remain quite close to the linguistic surface form. Almost all Japanese MT systems today exploit this basic approach, although they vary in several ways, including the depth of the semantic representation and the extent to which surface linguistic facts, such as word order, are extracted from the source text and used to guide target generation.

The idea of moving further toward interlingua-based systems has been advocated by some Japanese researchers for a long time (for example, see [Uchida 89b]). But other very influential figures (for example [Nagao 87]) argue that it is not yet practical to build deep, meaning-based systems, and there has been less work in this area in Japan than in the U.S. In the last several years, however, interest in moving further toward meaning-based systems has increased considerably in Japan. The main reasons for this are:

- Transfer rules must be written for every language *pair*. In contrast, the rules for mapping to and from the interlingua need only be specified for each language once. So, particularly for an MT system that is intended to handle several languages, the amount of development work should decrease as the need for transfer rules goes down.
- Since, in an interlingual system, each language can be specified almost entirely on its own with little consideration of the others, the various languages can be developed relatively independently. This means, in particular, that each language can be developed in its native country by native speakers.
- There appears to be a limit to how good a translation system can be if it has no representation of meaning. The problem is that many sentences are ambiguous. But often it is not possible just to pass the ambiguity on to the target text, since languages differ on the ambiguities they allow. In these cases, it is necessary to decide on the intended meaning of the source text in order to translate it correctly into the target. Sometimes this can be done without any actual representation of meaning (for example by using selectional restrictions as described in Section 2.3, or by using the example-based approach that will be described in Section 9.2), but in some cases it appears to be necessary to reason about the meaning of the sentence in its discourse context in order to choose the correct translation. To do this requires a meaning representation of the texts that can be used in conjunction with one or more knowledge bases that describe the domain(s) of the texts that are being translated.

Several efforts are moving in the direction of deeper, interlingua-based systems. The first is the work at EDR on building a widely available dictionary that maps from words (in English and Japanese) to a semantic conceptual structure. This dictionary is intended to be used as a basis for MT systems that make use of the conceptual structure as an interlingua. We will describe this effort in more detail in Section 9.6. There is already one major MT effort underway that uses the EDR dictionary and its

conceptual structure. This is the multilingual (Japanese, Chinese, Thai, Malay, and Indonesian) system that is being built at CICC. See Section 9.9 for a discussion of this effort.

Another important effort is the SHALT2 project at IBM Japan. SHALT2 is an E/J system (although extensions to other languages are planned; see Section 9.9), that is based on the earlier work at IBM on SHALT, a transfer-based system, as well as on collaborations with Carnegie Mellon University and their work on interlingua-based systems.

One idea that has come up in several places in Japan is to view each sentence as having more than one component, corresponding to the propositional content of the sentence as well as various other properties, including the speaker's attitude and intent in producing it. Then all the components need not be treated the same way; some can be translated using transfer rules, while others may be mapped to an interlingua (or relatively language-independent form). For example, the NADINE system at ATR (see Section 9.7), divides the meaning of a sentence into two parts, the propositional content and the illocutionary force. The illocutionary force component is what distinguishes between questions, commands, and declarative statements. It is mapped into a language-independent interlingua, while the propositional content is translated using a transfer-based system that is driven by a set of rules associated with individual words in the lexicon [Hasegawa 90]. Another example of this basic approach is the MLMT (Multi-Level Machine Translation) method that has been developed for the ALT-J/E system at NTT [Ikehara 89, Ikehara 91]. In this approach, a Japanese sentence is first decomposed into an objective component and a subjective one (which includes the speaker's emotions and intentions). The objective part is translated using a transfer-based system. Then the subjective part is rearranged using a table-driven method and recombined with the objective component. A third example is the work at ETL [Ikeda 89], which divides each sentence into three parts, the propositional content and two kinds of attitude descriptors. But in this work, the representations of the three parts taken together are viewed as an interlingual representation of the sentence and all are translated through the interlingua.

Despite this increased interest in interlingua-based systems, however, there is still no clear consensus that that is the right way to go. For example, several relatively new projects, including ones at IBM (JETS), Matsushita, Sanyo, Ricoh [Yamauchi 88], and Toshiba, are based on the more traditional transfer approach.

9.2 Example-Based Translation

Interlingual MT systems are based on the idea that the *meaning*, rather than just the *form* of the source text can be used to drive the translation process. One interesting alternative is to go back to the idea of relying at least partially on the form and to appeal to a large database of previously translated forms to guide each new translation. The idea that one can solve a problem by appealing to a knowledge base of prior problems and their solutions is beginning to show promise in a large variety of computerized problem-solving contexts [Stanfill 86], so it is not surprising that it is being applied to MT. Work on example-based machine translation (EBMT) was pioneered at Kyoto University [Nagao 84, Sato 90] and is now being conducted at several industrial labs, including Hitachi and ATR.

A schematic representation of the process in the ATR system [Sumita 90] is shown in Figure 9-2. The idea here is to retain the traditional first step in MT, namely an analysis of the source text, as well as the traditional last step, generation into the target language. And the idea of an explicit transfer step that maps from the source to the target is also retained. But instead of basing the transfer step on a set of

rules and dictionary entries, all of which were carefully crafted by the MT system builder, an example-based system, as its name implies, drives the transfer process by a database of translated examples. This database is augmented by a thesaurus that is used to enable the system to find examples that, although they do not exactly match the current text, are close to it in the sense that the words are related in the thesaurus and thus can be expected to be translated in analogous ways.

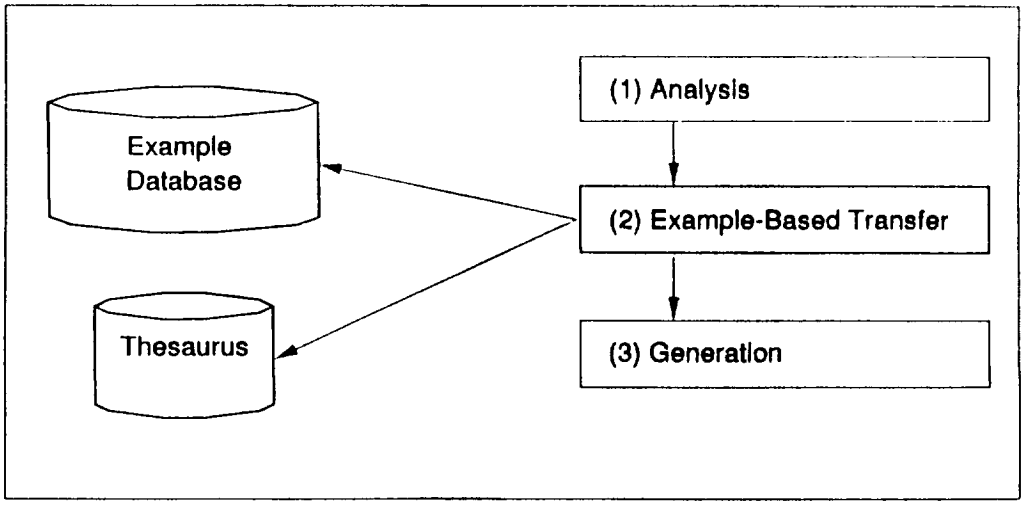

Figure 9-2: Example-Based Machine Translation

Japanese	English	Translation Pattern
yooka no gogo [8th, afternoon]	the afternoon of the 8th	B of A
kaigi no sankaryoo [conference, application fee]	the application fee for the conference ? the application fee of the conference	B for A
kyoto no kaigi [Kyoto, conference]	the conference in Kyoto ? the conference of Kyoto	B in A
issyuu no kyuuka [a week, holiday]	a week's holiday ? the holiday of a week	A's B
hoteru no yoyaku [hotel, reservation]	the hotel reservation ? the reservation of the hotel	AB
mitsu no hoteru [three, hotel]	three hotels ? hotels of three	AB

Figure 9-3: Examples for Use in EBMT

Figure 9-3 shows a simplified fragment [Sumita 90] of an EBMT database. It also illustrates the kind of problem that this approach is trying to solve. Japanese uses particles, such as *no*, to indicate many kinds of relationships among sentence constituents (as shown in the first column of the figure). English also

has mechanisms for indicating those relationships (as shown in the third column). The problem is that there is not a one-to-one relationship between the Japanese method and the English, so the Japanese particle *no* can have several different translations into English, depending on the meanings of the constituents that it is connecting.

The traditional way to solve this problem is for the MT system builder to encode a set of rules that attempt to describe the circumstances under which each of the various translations should be used. In the EBMT approach, such rules are not necessary. Instead, examples of the various translations are collected from actual texts and stored in a database such as the one in the figure. When a new sentence is to be translated, it is compared against the examples. If there is an exact match, then it is clear what to do. But usually there is not a precise match. In fact, the key to the success of this approach is the ability of the system to find the right close match. This is done by using the thesaurus. So, for translating a Japanese phrase corresponding to [Tokyo,conference], the third example offers the best choice because Tokyo and Kyoto are both cities. To make this idea more precise, the EBMT system exploits the notion of *semantic distance*, which is defined by a formula that can be calculated for any (input,example) pair.

Experimental results [Sumita 91] on the *no* example show a correct translation rate, using a database of about 2000 examples, of 78%, as compared with an estimated success rate of 20%, which would be achieved if the single most common translation ("B of A") were used all the time.

In some ways, this work on example-based MT is moving in a very different direction than is the work on a meaning-based interlingua described in Section 9.1. But there is also a view, expressed to us, for example by Professor Hajime Narita of Osaka Univeristy, that EBMT can serve as a bridge technology until the necessary framework for genuine A.I.-based MT systems is developed. The increasing availability of machine-readable, bilingual corpora -- on which this technique depends -- means that this may be a useful approach.

9.3 Transfer-Driven Translation

A third interesting new approach to the design of an overall MT architecture is the idea of transfer-driven machine translation (TDMT), which is also being pursued at ATR. TDMT can be thought of as contrasting with the more traditional approach to MT, which is analysis-driven in the sense that the bulk of the effort in the system is devoted to producing an analysis of the source text that is as accurate and complete as possible given the level of description on which the system is based (be it syntactic constituent structure, semantic case frames, or a deep interlingua). This analysis is then used to drive the transfer (if necessary) and generation processes. In contrast, in the TDMT approach, as little analysis as possible is done, and then only as it is needed. So some sentences or phrases might be translated directly by matching them against stored patterns that are tied to the appropriate translation. If that does not work, then syntactic analysis is done, and again transfer occurs if it is clear what to do. If not, semantic analysis is done and transfer is tried again.

Several empirical observations underlie this approach. On the one hand, there are texts that cannot be translated correctly without recourse to the meaning of each sentence, as well as the discourse context provided by surrounding parts of the text. In these cases, a deep analysis is clearly required. But there are many other examples in which surface pattern matching works and is fast. In particular, in at least some domains, there may be a small set of distinct sentences that account for a very large percentage of the sentences or utterances that are encountered. If this happens, then it makes a lot of sense simply to

store the translations of those sentences and look them up as they are needed. For example, in the telephone conference registration domain, the 10 most frequent utterances account for 22% of the total utterances out of corpus of 15,811 sentences accumulated at ATR. Not surprisingly, the top four utterances are *hai*, *moshimoshi*, *wakarimashita*, and *soudesuka*. It is likely that most texts (as opposed to the interactive dialogues studied at ATR) do not have quite this concentration of very common sentences, and as the size of the domain increases, the number of different sentences needed to cover a significant fraction of the corpus also will increase. Nevertheless, the idea that some very common sentences may be able to be translated with very little analysis seems powerful. It is particularly so in combination with some of the ideas that are being pursued in the EBMT work, including the fact that large bilingual corpora, which are necessary as the basis for any surface form-based translation system, are becoming available.

9.4 Grammars

Two traditional approaches to grammar development have provided the basis for most of the MT system development in Japan – case frames and phrase structure (possibly incorporating transformations) grammars. Most ongoing efforts continue to use these traditional frameworks, since there is a fairly widely held belief that these techniques provide the best available foundation for large sytems. But there is also some research on alternative approaches.

9.4.1 Constraint Dependency Grammars

One alternative approach is to view grammars as sets of constraints on how a sentence can be put together. Then constraint propagation can be used gradually to narrow the set of possible interpretations until, ideally, only one that satisfies all the constraints remains. This approach is being pursued by several different groups. For example, at ICOT it is being implemented using a form of logic programming [Sugimura 88]. But as long as this process is internal to the parser, it may produce no noticeable change in system behavior from the point of view of an MT system user.

However, it is also possible to make this approach visible to the user and to enlist the user's aid during source sentence analysis. This is being explored in the JETS J/E system at IBM. A schematic example of their interactive parser, JAWB [Maruyama 90a, Maruyama 90b], is shown in Figure 9-4. The input to the constraint-based parser is a list of the phrases (*bunsetsu*) contained in the sentence. The grammar rules are viewed as providing a set of constraints on the way that the phrases can be combined to form the overall syntactic structure of a sentence. Each time a new sentence is encountered, the grammar rules fire, and their constraints are propagated. The result of this step is usually a set (sometimes very large) of possible interpretations for the sentence. The grammar rules themselves are usually inadequate for reducing the size of the set because syntactic knowledge alone cannot provide the basis for selecting among competing interpretations. Fortunately, it is not necessary to represent all of the interpretations explicitly. Instead, they are represented implicitly using the constraints. The approximate size of the set of candidate interpretations is displayed to the user, who can exploit additional knowledge about the meaning of the sentence and thus has the ability to add constraints that the grammar alone could not provide. To make it easy for nonlinguists to use this system, users are allowed the opportunity to add only one kind of constraint. The system displays the phrases it has found, and for each dependent phrase, the candidate phrases on which it might depend are shown. The system also displays its first choice for the dependency. The user can look at the alternatives and then choose one.

Since choices at several different points in the parse all interact to produce a large number of complete

parses, it often happens that the user, by adding only a small number of additional constraints, can substantially reduce the total number of possible interpretations. The system allows the user to add a constraint and then see the result of propagating it through the existing constraint set. This enables the user to add constraints interactively, checking to make sure to avoid inadvertently ruling out the correct interpretation, until only a single consistent interpretation remains. Although this may appear to be solving the parsing problem by throwing it back on the user, it is important to keep in mind that the system is still doing all the bookkeeping and often all the user needs to do is to make a couple of decisions, each of which involves only a small number of alternatives. In one evaluation study, JAWB's best guesses were right only 47% of the time, but with user assistance the correct interpretation was found 99.8% of the time. (The remaining cases were ones where the initial constraints provided by the grammar ruled out the correct interpretation before it could be selected by the user.)

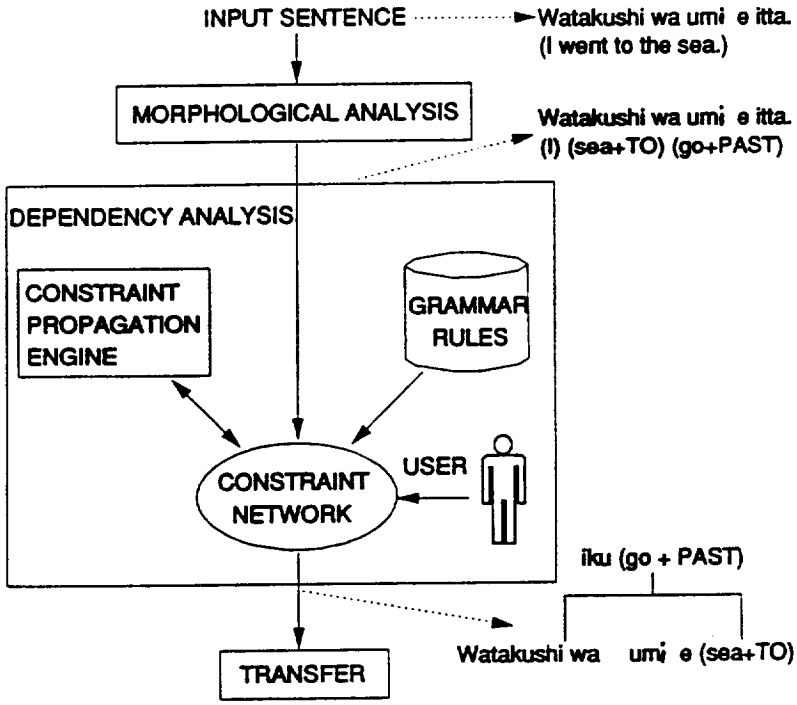

Figure 9-4: The Use of Constraint Dependency Grammar in JAWB

9.4.2 Alternative Grammatical Frameworks

The constraint dependency approach is novel both in how the grammar rules are represented and in the overall control regime that applies during the parsing process. Some other research efforts are directed at the more restricted issue of the grammar itself, without proposing a new control scheme.

The JTEC team did not see any work on totally new linguistic frameworks. But we did see work that uses frameworks other than the traditional ones described above. For example, ideas from the English Head-Driven Phrase Structure Grammar system (HPSG) [Pollard 87] and the Japanese JPSG [Gunji

87][13] form the linguistic basis for the syntactic analysis component [Kogure 89] of the speech-to-speech translation system being developed at ATR (see Section 9.10).

Another example of a newer syntactic framework being used for MT research is LFG [Bresnan 82, Kaplan 89], which is being used as a basis for research on the English/Japanese system at Bravice; on SHALT2 at IBM; and for some of the experimental work at ATR [Kudo 90]. There is also a substantial amount of work in many of the labs that do basic research on natural language processing on the use of various unification-based systems as the basis for sentence analysis, and, to a lesser extent, sentence generation. Although this work is novel and interesting from a linguistic point of view, its results will probably not have a substantial impact on the performance of MT systems viewed from the outside (i.e., by looking just at translation results rather than at the mechanisms that produced them).

But a final category of grammatical research may have such an impact, and that is work on bidirectional grammars. Such grammars allow a single linguistic description of a language to be used both for analysis (when the language is used as the source) and for generation (when the language is used as the target.) This approach contrasts with the more traditional one in which different grammars, usually written in different frameworks, are used for the two processes. The advantage of bidirectional grammars is that they have the potential to reduce the total cost of adding a new language (as both source and target) to an MT system. They can also reduce maintenance costs, since information is only represented in a single place. Bidirectional grammars are an important research topic within the larger international natural language processing community (for example, there was a workshop on bidirectional grammars at the 1991 ACL meeting in Berkeley), but they are only beginning to be explored in Japan, for example, by the SHALT2 and JETS projects at IBM [Takeda 90].

9.5 Generation

Most of the formal work on the use of linguistic knowledge in MT systems has focused on the use of that knowledge to aid in analyzing the source text. Generation into the target is usually considered a much easier problem. For example, at IBM we heard a second-hand quote from Professor Nagao that in doing Japanese to English translation, 80% of the errors occur in analyzing the Japanese source sentences. As a result, there is much less concern with improving the performance of the generation side of most MT systems than there is in improving the results of the analysis process. One consequence of this is that in most of the systems we saw, the English generation rules were not written by native English speakers, with the result that much of the output did not seem natural to us. There is beginning to be an increased concern with this issue, however, and many of the sites we visited expressed an interest in collaborating with the U.S. particularly in the area of developing English dictionaries and grammars.

There is also some work in the research labs on other aspects of the generation problem. The work on bidirectional grammars that we mentioned above is one example of this. So is some work, also at IBM Japan, on the use of the same chart mechanism for generation that is usually used in parsing. The advantage of the chart is that it prevents the same edge (constituent) from being generated more than once, so the overall of the efficiency of the system can be expected to increase.

---

[13]The English system GPSG [Gazdar 85] provided the initial basis for JPSG, so it too should be listed as a major influence on this work.

An interesting idea for the generation component of an interlingua-based system is to share a generator across similar languages and to represent the differences between the languages in the knowledge base and dictionaries. This approach is being pursued in the PIVOT system [Okumura 91].

9.6 Dictionaries

As shown in Chapter 4, comprehensive dictionaries play a very important role in the effectiveness of MT systems. Thus it is not surprising that all ongoing MT projects are devoting some of their R&D efforts to work on dictionaries. These efforts can be divided into four classes:

- Development of dictionaries (possibly in new subject domains) within the overall framework that has been defined for their existing MT system.
- Research with a goal of discovering new, more powerful dictionary structures.
- Tools that increase the productivity of dictionary builders.
- Construction of dictionaries for new languages.

The first of these was discussed in Chapter 4. The last was described in Chapter 3. This section will analyze the second effort --- research that will lead to more powerful dictionary structures.

The Electronic Dictionary Project (EDR), mentioned briefly in Section 4.4, is conducting a large-scale, nine-year research effort whose goal is the construction of a set of machine-usable dictionaries that, taken together, will be able to serve as the basis for a large range of natural language processing applications, including MT [EDR 91].

Figure 9-5 (taken frm [EDR 91]) contains a schematic description of the EDR dictionary set, which is composed of the following four main pieces:

- A dictionary of individual words. Both general and technical terms are included. The current area of focus for the technical dictionary is information processing. The only languages that are being considered at EDR are English and Japanese, but CICC is extending this system to include other Asian languages as well. (See Section 9.9.) The word dictionary contains basic linguistic information about each word, as well as a set of pointers that describe the meanings of the words in terms of the concept dictionary. Each component of the word dictionary is monolingual; no bilingual information is stored there.
- A concept dictionary, which is not a dictionary in the conventional sense, since it is not a list of words. Instead, it contains a set of semantic concepts and a set of relationships among them. The concepts in this dictionary provide the basis for representing the meanings of sentences. The concepts in this dictionary range from very general ones, such as *physical-object* and *action*, to very specific ones, such as *sparrow*.
- A co-occurrence dictionary, which provides information about surface co-occurrence relations between words. For example, since it is okay to say, "He drives a car," but not, "He drives a bike," this dictionary will contain the information that *car* can occur as the object of *drive* but *bike* cannot (while, on the other hand, *bike* can occur as the object of *ride* while *car* cannot). This information can be used by a natural language generator to enable it to choose the correct wording for common concepts, such as, "control a vehicle so it goes to the right place," many of which have several different surface realizations (such as *drive*, *ride*, and *pilot*). It is worth pointing out that the need for a dictionary of this sort as part of an MT system seems to be quite widely recognized, particularly for transfer-based systems. So there are other similar efforts under way at other places including Mitsubishi [Suzuki 91] and NHK [Tanaka 91].
- A bilingual dictionary, which defines the correspondence between words in the Japanese

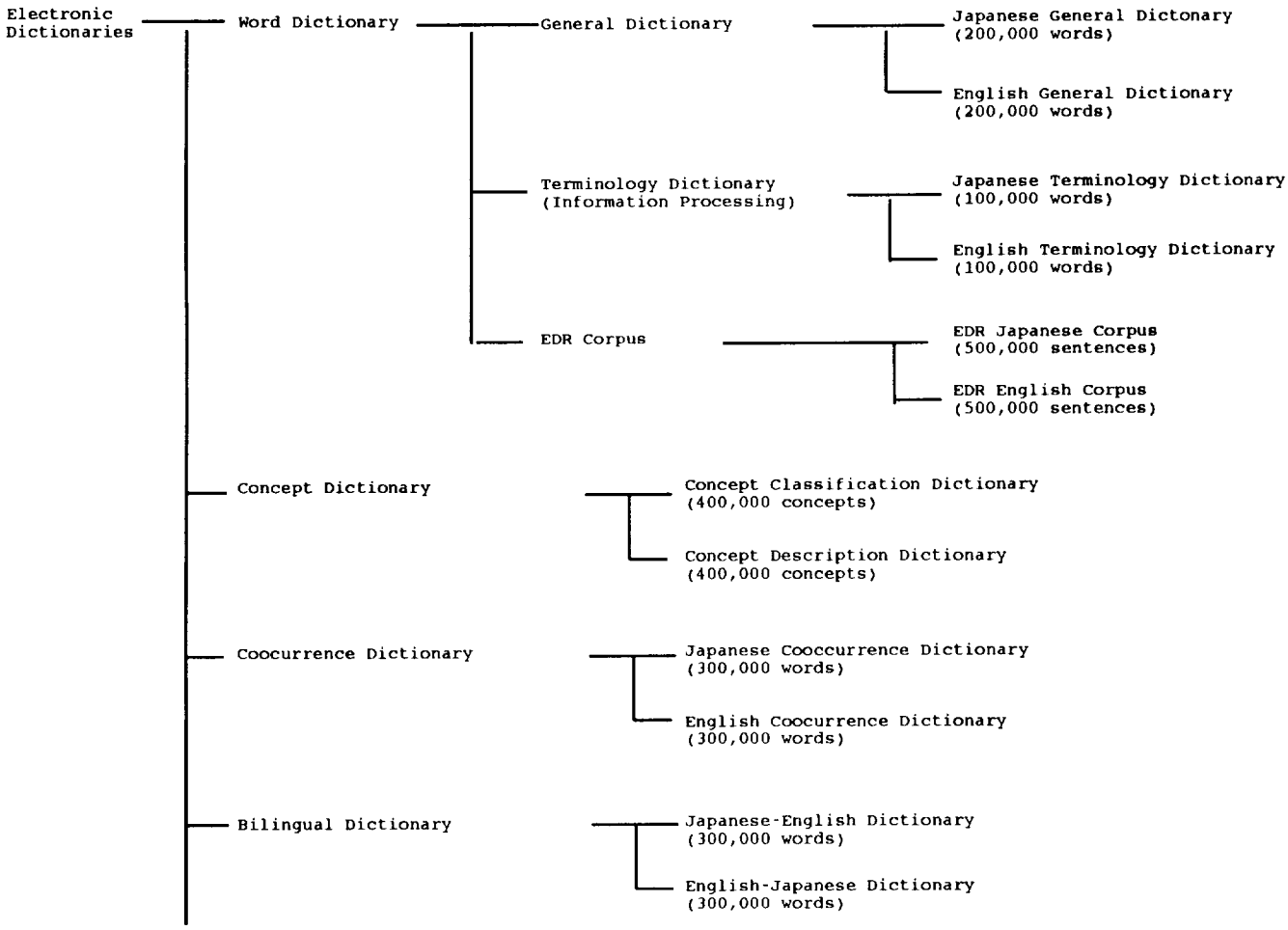

Figure 9-5: Structure of the EDR Electronic Dictionaries

word dictionary and words in the English word dictionary. Sometimes these correspondences relate words whose meanings are identical. But sometimes two languages do not have identical words. For example, the English word *horizon* has two corresponding words in Japanese, depending on whether the boundary is between earth and sky or between sea and sky. So the correspondences in this dictionary are marked as falling into one of four categories: equivalent, synonymous, broader (more general) than, and narrower (more specific) than.

Constructing dictionaries such as these so that they accurately reflect the languages that they are supposed to describe requires reliance on a large corpus of real texts. This is particulary important in the case of the co-occurrence dictionary. As a result, a side effect of EDR's dictionary building effort has been the construction of the EDR corpus, which is intended eventually to contain a half million sentences (consisting of both English and Japanese sentences).

The EDR dictionary is already being used in at least one important MT project, the Asian language effort at CICC (which will be described in more detail in Section 9.9). Thus the initial work in English and Japanese is being extended to Chinese [Zhu 89], Thai, Malay, and Indonesian.

Since the goal of the EDR effort is to support a wide range of natural language activities, EDR has stated the following policy on the distribution of their work:

> In principle, all the results of the EDR project will be sold at reasonable prices. The same conditions regarding the usage of the EDR Electronic Dictionaries will be applied to all users no matter whether they are domestic or overseas users. It is expected that the prices will be somewhat lower than those of machine-readable dictionaries that are currently on sale. Special measures will be arranged for those users for academic purposes, such as universities and public research institutions. [Yokoi 91]

The dictionary interface description was published in January 1991. The interface itself, including the word list for both English and Japanese, has been announced as being available for the price of copying and shipping, and we know of some U.S. institutions that have received it. The first editions of the word and bilingual dictionaries are expected to be completed soon, but their release dates have not yet been determined.

9.7 Discourse-Level Issues

Japanese natural language processing researchers have been concerned for a long time with analyzing properties of Japanese texts and dialogues above the sentence level. Much of this work is based on discourse theories that were originally developed for English, but because Japanese discourses are structured very differently from English ones, a substantial amount of original work must be done. Examples include: work at ICOT on a model of the use of honorifics in Japanese [Sugimura 86] (based on Situation Semantics [Barwise 83]); an alternative treatment at ATR [Maeda 88] of the same phenomenon based on HPSG [Pollard 87] and Discourse Representation Theory [Kamp 84]; work at ETL [Ishizaki 88] on quantitative measures of the complexity of Japanese sentences (which is intended to serve as a basis for the evaluation of MT systems); and work at NTT [Shimazu 90] on analyzing Japanese sentences using an argumentation system based on defeasible reasoning [Konolige 88].

But in the more specific area of MT, there is much less work on issues that cross sentence boundaries. All the production MT systems that the JTEC team saw translate a single sentence at a time. In the research labs, there is also very little work on discourse-level phenomena. Instead, MT is largely viewed as a single sentence at a time process. For example, [Ikehara 89] claims that about 90% of Japanese written sentences in practical use can be translated in isolation and without any domain knowledge. As a result, he argues for ignoring both discourse issues and world knowledge. Despite this view, however, there is some work on discourse phenomena.

The one MT system we know about that is not organized primarily around the translation of individual sentences is the CONTRAST [Ishizaki 89a, Ishizaki 90] system at ETL. CONTRAST's task is to translate short newspaper stories. It makes use of a set of stored, script-like objects called contextual representational structures. These structures correspond to common, newspaper story situations, such as hijacking and kidnapping. They differ from scripts in that they do not depend solely on a small number of very low-level primitives as their basic units. The analysis part of CONTRAST uses the title of the story to find the correct contextual structure. Then it analyzes the rest of the story and fills in values that correspond to the details of the particular incident that is being described. The generation component then takes the complete instantiated structure and generates a description of it in the target language. Thus there is no guarantee that the structure of the target paragraph will be the same as the structure of the input, and in fact one goal of this effort is to enable translation between languages in situations where the conventions for organizing information are very different. Although CONTRAST is very ambitious in its structure, it is still a small prototype, which has only five stored contextual representational structures and which works only on short newspaper stories about those five things.

One of the most widely studied discourse phenomena in Japan is ellipsis, which is the process by which a necessary sentential element is omitted from the surface form with the expectation that it can be recovered from the discourse context. A simple example of ellipsis occurs in the English sentence, "Let's try again to loosen the screw, this time using a wrench." The phrase, "using a wrench" modifies a verb phrase that is missing. But it is intended to be "to loosen the screw," which can be picked up from an earlier phrase in the same sentence. Often, though, ellipsis can only be resolved by appeal to earlier sentences, so a general treatment of ellipsis must consider a larger discourse context. Ellipsis is even more common in Japanese than it is in English (for example, see the dialogue in Figure 9-6), so it is a very important issue, particularly for J/E systems. As a result, there is a long tradition of Japanese work on this problem. (See, for example, [Nagao 76].) The most common general-purpose approach to ellipsis resolution is to look back through prior sentences to try to find a constituent that is of the appropriate type (both syntactic and semantic) to fill the current gap. In the specific case of elided subject, which is very common in Japanese, the most common approach is to use the English passive in the translation. This can be done without any recourse to discourse context since the missing constituent is omitted in the English sentence also. This technique is used in many J/E MT systems, but it often leads to unnatural sounding translations.

Research on other techniques for the treatment of ellipsis is being done in several places, but most of it is primarily theoretical rather than implementation oriented; several of the efforts do not use any extrasentential context even for this problem. Work at NTT focuses on the use of rules to avoid generating unnatural translations for elliptical constructions [Nakaiwa 90]. And work at ATR takes advantage of the fact that they are working in a limited domain (telephone conference registrations, see Section 9.10) and so the form of each sentence can be used to determine what dialogue function it is serving (e.g., a request for information or a promise to do something) [Dohsaka 91]. The dialogue function then is used as a basis for filling in the missing constituents. In particular, facts about the way in which honorifics have been used can sometimes help to determine whether the speaker or the hearer is the intended subject of the sentence.

Another area in which some discourse-oriented work is being done is the use of task and domain knowledge as a basis for building a model of a task-oriented discourse. We discuss ATR's work in this area later, in Section 9.10, in the context of ATR's larger goal of building a translating telephone. But we should point out here that although there is work at ATR on learning about the structure of task-oriented dialogues, the primary MT system that is being built there still operates on a single sentence at a time.

A final area we will mention is the treatment of anaphora (expressions, such as pronouns, that necessarily derive at least some part of their meaning from some other linguistic expression on which they depend). We found very little work on anaphora, although some groups, such as the SHALT2 team at IBM and the group at NTT, indicated that they intend to extend their system to handle anaphora in the future.

9.8 Better Tools for Users

The JTEC team saw a substantial amount of work on the development of better tools for users. The bulk of this was aimed at reducing the cost of pre-editing. For example, JICST is just completing a system that performs about ten kinds of style checks automatically during pre-editing. See Section 6.1.1 for a more complete list of sites that are doing work in this area. There is much less work on tools for

postediting, but we saw one system under development at Ricoh (as described in Section 6.3.)

One idea that appears to be gaining popularity is a move away from a black-box, batch-mode translation system and toward a system that interacts incrementally with a user throughout the translation process. This may be an improvement over the traditional approach in at least two ways. It may beat pre-editing, in which users have to anticipate where the problems are likely to occur. As a result, they may waste time in places where there would not have been a problem anyway, or they may miss a place where there is a serious problem and they could easily have helped. The interactive approach can also beat postediting because it allows the user to give some advice. Then the system picks up the advice and propagates it through the rest of the translation process. This contrasts with postediting, where, if users make changes, they must fix the entire translation if it depends on the change. This interactive approach to the overall translation process is being investigated in several labs, including Matsushita and Fujitsu. The JAWB interactive parser described in Section 9.4 is another example of this approach.

Another way to solve the imperfect translation problem is to give up on the idea that the MT system should output a single "best" translation. Instead, it can display several alternatives to the user, whose job is just to click on the right one. This is a lot simpler than the standard postediting task in which some of the MT system's output must be rewritten. This idea is being exploited in Ricoh's English/Japanese MT system, as well as in Catena's STAR.

9.9 Extension to Other Languages

As described in Chapter 3, the first generation of Japanese-built MT systems focused on the problem of translating between Japanese and English. We are now beginning to see, however, a broadening of the scope of the Japanese MT effort to include other Asian languages as well as European ones. Most of the efforts represent extensions of existing systems to new language pairs. The work at CICC is different, though, since it is an entirely new MT effort.

CICC began a six year project in 1987 with Overseas Development Assistance (ODA) funding. The technical goal of the project is to build a demonstration prototype of an MT system for Japanese, Chinese, Thai, Malay, and Indonesian. This system is aimed at technical texts, with an initial focus on information processing. The political goal is to develop strong connections between Japan and the emerging Asian marketplace. As a result, this is a strongly collaborative effort, with most of the work on the non-Japanese languages being done in the native countries.

The CICC system [Tanaka 89, Tsuji 90] is firmly grounded in prior Japanese work on MT. The syntactic analysis component of the system currently contains three parsers taken from three existing commercial MT systems: ATLAS-II (from Fujitsu), HICATS (from Hitachi), and PIVOT (from NEC). The plan is to integrate them into a single system. The system is interlingual [Ishizaki 89b], with the EDR concept dictionary serving as the basis of the interlingua. This project is thus an interesting test case for the EDR concept structure, which was initially developed to support processing Japanese and English. Of course, if the concept structure is truly language-independent, then no changes will be necessary to use it to process a new language. But no one expected that it would be language-independent in this extreme sense. The CICC team estimates that the size of the change to the concept structure required to add a new language is 10%. Of all the advantages of the interlingual approach described in Section 9.1, one is of particular significance for this effort. Because each language is defined separately, it is possible to spread out the development effort, with each language being worked on in its home country.

Because of the important structural differences that exist among the languages addressed by this project, a wide range of technical problems will need to be solved. One example of this that was mentioned during the JTEC visit is the fact that it is very difficult in Chinese to determine the part of speech of a word since there are no inflectional markers, as there are in many European languages, nor are there postpositional markers as there are in Japanese. Another example is that Thai not only has no word boundary markers; it has no sentence boundary markers either.

The prototype system is written in the programming language C. The performance goal for the system is to be able to translate more than 5,000 words per hour with an accuracy of more than 90% on grammatically correct texts all of whose words are in the dictionary. The dictionary is expected to contain 50,000 basic words in each language, plus 25,000 additional words in the domain of information processing. At the time of the JTEC visit, the system could translate a corpus of 500 sentences between the five languages, using dictionaries with 20,000 word vocabularies. The CICC team also expressed a strong interest in adding English to their system.

9.10 Speech-to-Speech Translation

The vision of a speech-to-speech translation system is widespread in Japan. Work on various aspects of this problem is being done in several labs, such as Matsushita, where they are building a Japanese/English system that includes a speech system based on the widely used idea of a Hidden Markov Model (HMM) [Rabiner 89].

The two largest efforts, though, are at NEC and ATR. NEC's vision is of a hand-held, speech-to-speech MT system. They have already produced a prototype E/J and J/E workstation-based system that can recognize a 500 word vocabulary for continuous speech and a 5,000 word vocabulary for isolated words. The typical time required for a short sentence is about five seconds for recognition and another ten seconds for translation.

But the most substantial effort in this area is being conducted at ATR, which has been working since 1986 on a planned 15-year project. One system is currently under development, and others are planned for the future.

The goal for the current effort is a 1500-word, speaker-independent, real-time, limited-domain, English/Japanese and Japanese/English system with greater than 75% accuracy. The task domain for the system is international conference registration. The prototype implementation, SL-TRANS [Kurematsu 91], is operational on a small set of example dialogues, such as the one shown in Figure 9-6 (taken from [Kogure 90]). Figure 9-7 shows the top-level architecture of SL-TRANS.

The speech recognition part of SL-TRANS is based on an HMM, with some enhancements (e.g., [Hanazawa 89]) that have emerged from this project. An LR phrasal parser is used to predict the next phoneme in the speech input [Kita 89]. Unfortunately, this combination is not powerful enough to determine a unique interpretation for each phrase, so a semantic filtering technique [Morimoto 90] is applied and early experiments suggest that it can reduce the number of candidate interpretations to less than a third of the original number. Once this hybrid procedure has produced an interpretation of the spoken input, that interpretation is passed to an analysis procedure, just as the typed input is in a conventional MT system. As described in Section 9.4, this analysis procedure is based on an HPSG/JPSG grammar and is performed using an active chart parser. The transfer component of this

	Japanese Input Utterances	English Output Utterances
1	Moshimoshi Sochira wa kaigijimukyoku desu ka.	Hello. Is this the office for the conference?
2	Hai Sou desu.	Yes. That is right.
3	(Watashi wa) kaigi ni moushikomi tai no desu ga.	I would like to apply for the conference.
4	(Anata wa) tourokuyoushi o sudeni o-mochi deshou ka.	Do you already have a registration form?
5	Iie mada desu.	No. Not yet.
6	Wakari mashi ta. Soredewa (watashi wa anata ni) tourokuyoushi wa o-okuri itashi masu. (Anata no) go-jusho to o-namae o onegai shi masu.	All right Then, I will send you a registration form. Your name and your address, please.
7	Juusho wa Oosaka-shi Kita-ku Chaya-machi nijuu-san desu. Namae wa Suzuki Mayumi desu.	The address is 23 Chaya-machi Kita-ku, Osaka. The name is Mayumi Suzuki.
8	Wakari mashi ta. Torokuyoushi wa shikyu okura se te itadaki masu. Wakara nai ten ga gozai mashi tara, itsudemo (watashi ni) o-kiki kudasai.	All right. I will send you the registration form immediately. If there will be a question, please ask me at any time.
9	Arigatou gozai masu. Soredewa shitsurei itashi masu.	Thank you. Good-bye.
10	Doumo shitsurei itashi masu.	Good-bye.

The odd numbered utterances come from the questioner; the even numbered ones come from the conference secretary. Parenthesized phrases are not expressed explicitly.

Figure 9-6: An Example of a Task-Oriented Dialogue

system is interesting because, as we mentioned in Section 9.1, it is a hybrid transfer/interlingua-based system.

The work at ATR differs from most of the other MT efforts that we looked at in its focus on two-person, interactive dialogues rather than on static text. This difference manifests itself both at the individual sentence level (where, for example, one sees sentences such as the ones in blocks 1, 2, 5, 9, and 10 above, which would be very unlikely to show up in a written text) and at the level of larger discourse units. As we said in Section 9.7, however, very little attention is currently being paid to phenomena at this larger level. But there is some work at ATR on such issues as analyzing the larger structure of task-oriented dialogues (e.g., in the NADINE system [Kogure 90]) and looking at patterns of interruptions in real telephone conversations [Myers 90], presumably to lay the groundwork for later efforts to exploit the larger dialogue context during the translation process.

Other work at ATR is aimed at exploring alternative approaches to parts of the speech-to-speech MT problem that may serve as the basis for future systems. For example, there is work on a neural net-based speech recognition system and another effort that is exploring the use of a hybrid symbolic/subsymbolic, massively parallel system for both speech and natural language processing [Tomabechi 90]. And, as described in Section 9.2, there is work on a new approach to translation that is driven by a large database of translated examples. Some of the work at ATR relies on international collaboration with institutions such as Carnegie Mellon University in the U.S. and the University of Manchester in the U.K.

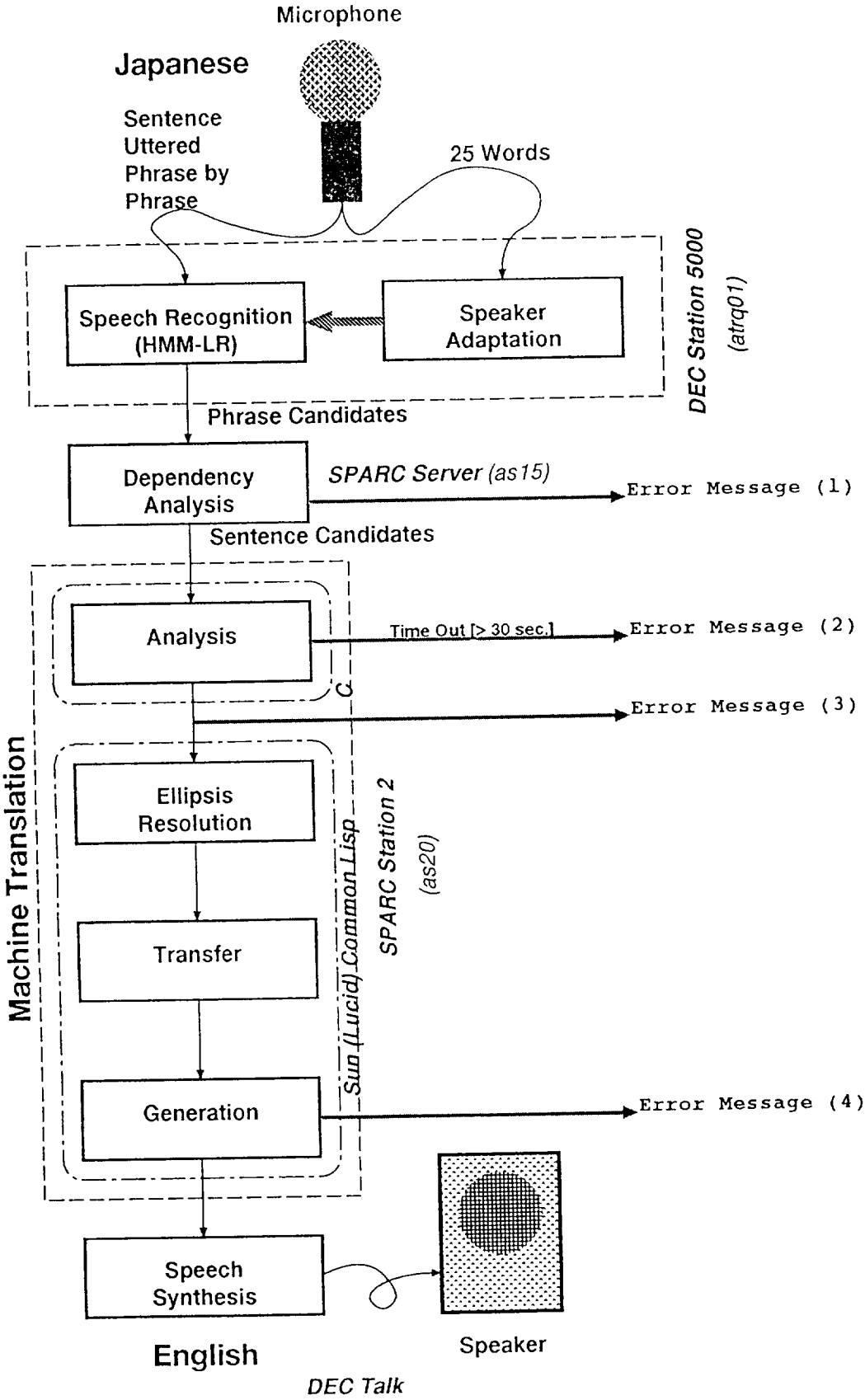

Figure 9-7: The Architecture of the SL-TRANS Speech-to-Speech MT System

9.11 Embedded MT Systems

As we pointed out in Chapter 6, there is a widespread appreciation in Japan of the potential role of MT as a component of larger information processing systems. So although it can be very useful as a stand-alone translation engine, it can also be embedded in database systems, electronic mail programs, and document production environments. Given this perspective, it is not surprising that many of the MT sites are working on prototypes of these kinds of MT systems. We described some of those applications in Chapter 6. We mention a few others here.

For example, at Fujitsu the JTEC team saw a demonstration of a system that allows German- and English-speaking users to retrieve texts from a Japanese text database. The user can enter key words in German or English, and they are translated into Japanese. The retrieval against the Japanese database is then done, and the titles of the articles that are returned are translated into German or English and displayed. Entire articles can also be translated at the user's request. We also saw demonstrations of an MT system embedded into an electronic mail system at both Fujitsu and Oki.

Toshiba has developed a system that embeds ASTRANSAC [Hirakawa 91], their J/E and E/J MT system, into a larger system that includes an OCR front end and an interface to a document processing system. The OCR front end scans the source document and creates a structural description of it (which means, for example, that pictures can be separated out and reintroduced later after translation has occurred.) The translation process then takes place using the structured document, and the software deals appropriately with the structure codes. The goal of this work on an embedded MT system was partly to make MT available to a wide array of end users. To make the system even more accessible to these users, ASTRANSAC offers several end-user customization capabilities, including: the ability to specify default values for various linguistic parameters in the system (including formality, rules for translating articles, and rules for choosing between active and passive voice); the ability to modify the dictionary; a tool for using the set of lexical items that occur in user-selected texts in the target language to bias lexical choice during translation; and the ability to use multiple occurrences of an ambiguous phrase in the source text to help resolve the ambiguity. One particularly interesting thing about all of these features is that they have been designed so that it will be easy to merge end-user customizations with vendor supplied upgrades without one invalidating the other.

The JTEC team was also informed of work on integrating MT into larger, document-processing environments at Hitachi, NEC, Toshiba, Matsushita, Oki, and others.

9.12 Massively Parallel Hardware

There is a substantial amount of research in Japan on massively parallel hardware. MITI's New Information Processing Technology (NIPT) plan (sometimes known as the Sixth Generation Project) is backing this effort. Also, most of the major computer companies are building high performance parallel machines, including Fujitsu, Hitachi, IBM Japan, Matsushita, Mitsubishi, NEC, NTT, Oki Electric, Sony, and Toshiba. There is also work at ICOT, ETL, and several university laboratories. Although this work is not directly related to MT, we mention it here because of the potential for exploiting this technology as a platform for future MT systems.

Partly as a legacy from the ICOT Fifth Generation Project, some of these machines are specifically targeted toward artificial intelligence applications, including representing and manipulating knowledge

structures, such as semantic nets, that are playing an increasing role in MT systems as they move toward the use of conceptual structures as their intermediate representations (as described in Section 9.1). ICOT is now focusing a substantial amount of effort on the development of parallel natural language processing systems, and there is already some work on the use of massive parallelism for MT, for example at ATR, which bought a Connection Machine (from *Thinking Machines* in the U.S.), and is using it for MT research. Particularly because of the overlap between the companies that are involved in MT and those that are building parallel machines, there will be excellent opportunities to exploit this combination of technologies.

9.13 The Future

A very substantial amount of research, both on basic natural language processing and on machine translation, is being conducted in Japan. This section has provided an outline of some of the major trends. One key one, that was articulated by Professor Nagao, is that there will be a gradual shift toward knowledge-based (A.I.-based) systems for MT. On the other hand, some interesting possible trends are significantly absent. For example, other fundamental representational techniques, such as subsymbolic (or connectionist/neural net) systems, appear to be receiving very little attention. Despite all the research effort in MT, there was a consensus at the JEIDA roundtable held at the end of the JTEC visit that there are unlikely to be any significant research breakthroughs in MT during the next five years. Instead, it is reasonable to expect steady progress toward more powerful systems. This is not surprising in view of the fact that most of the work we saw is being conducted within well-established research traditions. There does, however, seem to be a clear consensus that the delivered MT systems of five years from now will be more robust, more usable, more accurate, and more widely exploited than they are today.

10. Future Directions in Machine Translation

Elaine Rich

The fundamental basis for any projection of future MT development in Japan lies in the recognition of MT as a key technology, particularly within the larger context of the information processing society of the future. Most of the major vendors of information processing technology (including both computer and communication companies) are involved in MT efforts in Japan. There is a clear symbiosis between these two efforts. The MT systems of today make available information that facilitates the development of the technology of the future. And that future technology will be the platform on which more powerful MT systems will run.

Perhaps because of the large distance between Japanese and most other major world languages, or because of their roots as an island culture, the Japanese well recognize that isolation is not in their long-term interest. Reducing isolation and increasing the ability to communicate globally will give the Japanese enhanced ability to achieve major national objectives. Not only is their exporting of goods and services facilitated by MT, but also their ability to import strategic information is enhanced. Thus, there is a consensus of industry, academic, and government leaders that the significant MT achievements of the 1980s must be continued through the 1990s and beyond the year 2000. Indeed, the vision of automated speech-to-speech interpretation is targeted for achievement in 2015.

While MT technology will continue to be pushed, the pull of thousands of Japanese users will ensure market development. These users already exist, both within the major companies that are participating in MT efforts, as well as within the wide range of organizations that make use of the MT services that are offered by commercial translation service bureaus. And this base of users is increasing. Just in the one week that the JTEC team was in Japan, we heard about several new MT services that had been or were about to be announced both by the major vendors and by the service bureaus. Each of them has the potential to bring in substantial numbers of new users. This is clearly a very dynamic arena.

Integration of MT with related technologies will probably accelerate so that optical character recognition (OCR), voice recognition, word processing, desktop publishing, office automation, document management, database use, electronic mail, language instruction, and other such technologies will be increasingly seen in products in the 1990s. Newer MT applications of existing MT technology, such as gisting and scanning large text databases to find relevant documents, will also become more common. Spin-offs such as software for pre-editing and small bilingual dictionaries for word processors and pocket translators are already on the market as separate products, and this trend is expected to continue. Since short-term profitability is not the determining factor, these longer term investments in innovative products will continue.

While no major technology breakthroughs are expected in the next five years, steady improvement will be seen in vital areas. At the wrap-up session at JEIDA at the end of our visit, we mentioned four areas of potential improvement: translation quality, better integration of MT systems with other applications, lower cost, and better user interfaces. We asked the Japanese representatives at the meeting (most of whom represented the MT manufacturers and a few of whom came from universities) which of these areas they felt was most important from the point of view of MT users. Of the 17 people, 13 said higher quality, three said a better user interface, one said better integration, and no one mentioned lower cost.

These views appear to be driving ongoing research and development efforts, most of which are focused on the creation of systems that exploit greater amounts of knowledge (in the form of grammars, dictionaries, examples, context models, and domain knowledge bases) in an attempt to produce higher quality translations.

As knowledge bases grow in quantity, quality, and comprehensiveness, the sharing of these intellectual properties will become more common, both for research and commercial purposes. International collaboration on MT research and development will be enhanced by this knowledge-sharing. There is a widespread recognition in Japan that international collaboration is particularly important in the development of knowledge bases for MT systems, since dictionaries and grammars can best be developed by native speakers of the languages that are being described. So, although most of the current Japanese MT systems have been developed with very little involvement by native speakers of languages other than Japanese, this situation will probably change considerably over the next five years.

User interfaces are also improving, partially as a result of feedback from the growing community of MT system users. This is, of course, a positive feedback loop, and as the interfaces improve, the user community grows, more feedback is available, and so forth. As a result, the Japanese fully expect to see a return on the substantial investment that they have made and are continuing to make in MT.

11. References

[Aizawa 90] Aizawa, T., T. Ehara, N. Uratani, H. Tanaka, N. Kato, S. Nakase, N. Aruga, & T. Matsuda.
A Machine Translation System for Foreign News in Satellite Broadcasting.
In *Proceedings of COLING '90*, pages 308-310. 1990.

[ALPAC 66] Automatic Language Processing Advisory Committee (ALPAC).
Language and Machines: Computers in Translation and Linguistics.
Division of Behavioral Sciences, National Academy of Sciences, National Research Council Publication 1416, Washington, 1966.

[Amano 88] Amano, S., H. Nogami, and S. Miike.
A Step Towards Telecommunication with Machine Interpreter.
In *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Natural Language Processing.* 1988.

[Amano 89] Amano, S., H. Hirakawa, and H. Nogami.
The TAURAS Design Philosophy.
In *Proceedings of MT Summit II*, pages 36-41. 1989.

[Ashizaki 89] Ashizaki, T.
Outline of the JICST Machine Translation System.
In *Proceedings of MT Summit II*, pages 44-49. 1989.

[Bar-Hillel 71] Bar-Hillel, Y.
Some Reflections on the Present Outlook for High-Quality Machine Translation.
In W. Lehmann and R. Stachowitz (editors), *Feasibility Study on Fully Automated High Quality Translation.* Rome Air Development Center, Rome AFB, Rome, N.Y., 1971.

[Barwise 83] Barwise, J. and Perry, J.
Situations and Attitudes.
M.I.T. Press, 1983.

[Becker 84] Becker, A.L.
Biography of a Sentence: A Burmese Proverb.
In E. M. Bruner (editor), *Text, Play, and Story: The Construction and Reconstruction of Self and Society.* American Ethnological Society, Washington, D.C., 1984.

[Bennett 85] Bennett, W. S. & Slocum, J.
The LRC Machine Translation System.
Computational Linguistics 11(2-3), 1985.

[Bostad 90] Bostad, D. A.
Aspects of Machine Translation in the United States Air Force.
In *Benefits of Computer Assisted Translation to Information Managers and End-Users, AGARD Lecture Series No. 171.* North Atlantic Treaty Organization, Advisory Group for Aerospace Research and Development, Neuilly-sur-Seine (France), 1990.

[Bresnan 82] Bresnan, J. (editor).
The Mental Representation of Grammatical Relations.
MIT Press, Cambridge, Mass., 1982.

[Brown 89] Brown, et. al.
A Statistical Approach to Machine Translation.
Technical Report, IBM Research Division Technical Report in Computer Science RC 14773 (#66226), T. J. Watson Research Center, Yorktown Heights, N.Y., 1989.

[Chandioux 89] Chandioux, J.
METEO: 100 Million Words Later.
In D. L. Hammond (editor), *Coming of Age: Proceedings of the 30th Annual Conference of the American Translators Association.* Learned Information, Medford, N.J., 1989.

[Dohsaka 91] Dohsaka, K.
Interpretation and Generation through Efficient Representations.
Natural Language SIG Reports of the IPSJ, March 15, 1991.
(in Japanese).

[EDR 90] EDR.
An Overview of the EDR Electronic Dictionaries.
Technical Report, Japan Electronic Dictionary Research Institute TR-024, 1990.

[EDR 91] EDR.
Proceedings of the International Workshop on Electronic Dictionaries.
Technical Report, Japan Electronic Dictionary Research Institute TR-031, 1991.

[Estival 90] Estival, D.
Generating French with a Reversible Unification Grammar.
In *Proceedings of COLING '90.* 1990.

[Farwell 90] Farwell, D., & Wilks, Y.
ULTRA: a multilingual machine translator.
Technical Report, Computing Research Laboratory, Las Cruces, New Mexico MCCS-90-202, 1990.

[Gazdar 85] Gazdar, G., E. Klein, G. K. Pullum, & I. Sag.
Generalized Phrase Structure Grammar.
Harvard University Press, Cambridge, Mass., 1985.

[Grimaila 91] Grimaila, A. with J. Chandioux.
Machine Translation in the Real World.
In J. Newton (editor), *Computers in Translation: A Practical Appraisal.* Routledge, London, 1991.
(In press).

[Gunji 87] Gunji, T.
Japanese Phrase Structure Grammar.
D. Reidel, Dordrecht, 1987.

[Hanazawa 89] Hanazawa, T., T. Kawabata, & K. Shikano.
Recognition of Japanese Voiced Stops Using Hidden Markov Models.
Journal of the Acoustical Society of Japan 10:776-785, 1989.

[Hasegawa 90] Hasegawa, T.
A Rule Application Control Method in a Lexicon-Driven Transfer Model of a Dialogue Translation System.
In *Proceedings of the 9th European Conference on Artificial Intelligence.* 1990.

[Hirakawa 91] Hirakawa, H., H. Nogami, S. Amano.
EJ/JE Machine Translation System ASTRANSAC - Extensions toward Personalization.
In *Proceedings of MT Summit III.* 1991.

[Hutchins 86] Hutchins, W. J.
Machine Translation: Past, Present, Future.
Ellis Horwood, Chichester, England, 1986.

[Ichiyama 89] Ichiyama, S.
Multi-lingual Machine Translation System.
Office Equipment and Products 18(131):46-48, August, 1989.

[Ikeda 89] Ikeda, T.
On an Interlingua Representation.
Natural Language SIG Reports of the IPSJ , June 30, 1989.
(in Japanese).

[Ikehara 89] Ikehara, S.
Multi-Level Machine Translation Method.
Future Computing Systems 2(3), 1989.

[Ikehara 91] Ikehara, S., S. Shirai, A. Yokoo, and H. Nakaiwa.
Toward an MT system without Pre-Editing – Effects of New Methods in ALT-J/E.
In *Proceedings of MT Summit III.* 1991.

[Ishizaki 88] Ishizaki, S. & H. Isahara.
Extraction of Qualitative and Quantitative Characteristics of Complexity Included in Japanese Sentences.
Natural Language SIG Reports of the IPSJ , July 22, 1988.
(in Japanese).

[Ishizaki 89a] Ishizaki, S.
Machine Translation System Using Contextual Information.
In M. Nagao, H. Tanaka, T. Makino, H. Nomura, H. Uchida, & S. Ishizaki (editors), *Proceedings of Machine Translation Summit I.* Ohmsha, Tokyo, 1989.

[Ishizaki 89b] Ishizaki, S. & H. Uchida.
On Interlingua for Machine Translation.
Natural Language SIG Reports of the IPSJ , January 20, 1989.
(in Japanese).

[Ishizaki 90] Ishizaki, S., H. Isahara, T. Tokunaga, & H. Tanaka.
Steps toward a Machine Translation Using Context and World Model.
Japanese Artificial Intelligence Journal 4(6), 1990.
(in Japanese).

[JEIDA 89] Japan Electronic Industry Development Association.
A Japanese View of Machine Translation in Light of the Considerations and Recommendations Reported by ALPAC, U.S.A.
JEIDA, Machine Translation System Research Committee, Tokyo, 1989.

[JEIDA 91] JEIDA.
Revised Outlines of MT Systems.
1991.

[Johnson 85] Johnson, R.L., M. King, and L. des Tombe.
EUROTRA: A Multi-lingual System under Development.
Computational Linguistics 11:155-169, 1985.

[Kaji 88] Kaji, H.
Language Control for Effective Utilization of HICATS/JE.
In *Proceedings of MT Summit II*, pages 72-77. 1988.

[Kamp 84] Kamp, H.
A Theory of Truth and Semantic Representation.
In J. Groenendijk, T. Janssen, & M. Stokhof (editors), *Truth, Interpretation, and Information.* Foris, Dordrecht, 1984.

[Kaplan 89] Kaplan, R. M., K. Netter, J. Wedekind, and A. Zaenen.
Translation by Structural Correspondences.
In *Proceedings of the 4th Conference of the European ACL, Manchester.* 1989.

[Kita 89] Kita, K., T. Kawabata, & H. Saito.
HMM Continuous Speech Recognition Using Predictive LR Parsing.
In *Proceedings of the International Conference on Acoustics and Speech Signal Processing.* 1989.

[Kogure 89] Kogure, K.
Parsing Japanese Spoken Sentences Based on HPSG.
In *Proceedings of the Int. Workshop on Parsing Technology.* 1989.

[Kogure 90] Kogure, K., H. Iida, T. Hasegawa, & K. Ogura.
NADINE: An Experimental Dialogue Translation System from Japanese to English.
In *Proceedings of the InfoJapan'90 Computer Conference Organized by IPSJ to Commemorate the 30th Anniversary.* 1990.

[Konolige 88] Konolige, K.
Defeasible Argumentation in Reasoning about Events.
In *Proceedings of the International Symposium on Machine Intelligence and Systems.* 1988.

[Kudo 90] Kudo, I.
Local Cohesive Knowledge for A Dialogue-Machine Translation System.
In *Proceedings of the 28th Annual Meeting of the ACL.* 1990.

[Kurematsu 91] Kurematsu, A., H. Iida, T. Morimoto, & K. Shikano.
Language Processing in Connection with Speech Translation at ATR Interpreting Telephony Research Laboratories.
Speech Communication Journal 10:1-9, 1991.

[Maeda 88] Maeda, H., S. Kato, K. Kogure, & H. Iida.
Parsing Japanese Honorifics in Unification-Based Grammar.
In *Proceedings of the 26th Annual Meeting of the ACL.* 1988.

[Maruyama 90a] Maruyama, H.
Structural Disambiguation with Constraint Propagation.
In *Proceedings of the 28th Annual Meeting of the ACL.* 1990.

[Maruyama 90b] Maruyama, H., H. Watanabe, & S. Ogino.
An Interactive Japanese Parser for Machine Translation.
In *Proceedings of COLING '90.* 1990.

[Melchuk 63] Melchuk, I.
Machine Translation and Linguistics.
In O. Akhmanova, I. Melchuk, R. Frumkina and E. Paducheva (editors), *Exact Methods in Linguistic Research. R-397-PR.* The RAND Corporation, Santa Monica, 1963.

[Morimoto 90] Morimoto, T., K. Shikano, H. Iida, & A. Kurematsu.
Integration of Speech Recognition and Language Processing in Spoken Language Translation System (SL-TRANS).
In *Proceedings of the International Conference on Spoken Language Processing.* 1990.

[Muraki 89] Muraki, K.
PIVOT: Two-Phase Machine Translation System.
In M. Nagao, H. Tanaka, T. Makino, H. Nomura, H. Uchida, & S. Ishizaki (editors), *Proceedings of Machine Translation Summit I.* Ohmsha, Tokyo, 1989.

[Myers 90] Myers, J. K.
Methods for Handling Spoken Interruptions for an Interpreting Telephone.
Natural Language and Communication SIG Reports of the Institute of Electronis, Information and Communication Engineers 90(44), 1990.

[Nagao 76] Nagao, M., J. Tsujii, and K. Tanaka.
Analysis of Japanese Sentences, by Using Semantic and Contextual Information–Content Analysis.
Joho Shori 17(1), January, 1976.
(in Japanese).

[Nagao 84] Nagao, M.
A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.
In A. Elithorn & R. Banerji (editors), *Artificial and Human Intelligence*, pages 173-180. North Holland, 1984.

[Nagao 85] Nagao, M., Tsujii, J. & Nakamura, J.
The Japanese Government Project for Machine Translation.
Computational Linguistics 11(2-3), 1985.

[Nagao 86] Nagao, M. J. Tsujii, & J. Nakamura.
Machine Translation from Japanese into English.
In *Proceedings of the IEEE*. 1986.

[Nagao 87] Nagao, M.
Role of Structural Transformation in a Machine Translation System.
In S. Nirenburg (editor), *Machine Translation: Theoretical and Methodological Issues*. Cambridge University Press, Cambridge, 1987.

[Nagao 89] Nagao, M.
Machine Translation: How Far Can It Go?
Oxford University Press, Oxford, 1989.

[Nakaiwa 90] Nakaiwa, H.
Natural Form English Generation of Supplemented Case Elements in Japanese to English Machine Translation System.
In *Proceedings of the 4th National (Japanese) Conference on Artificial Intelligence*. 1990.
(in Japanese).

[Okumura 91] Okumura, A., K. Muraki, and S. Akamine.
Mult-lingual Sentence Generation from the PIVOT Interlingua.
In *Proceedings of MT Summit III*. 1991.

[Pollard 87] Pollard, C. & I. A. Sag.
An Information-Based Syntax and Semantics, Vol. 1.
Technical Report, CSLI Lecture Note Number 13, 1987.

[Rabiner 89] Rabiner, L.
A Tutorial on HMMs and Selected Applications in Speech Recognition.
Proc. of IEEE 77, 1989.

[Sakurai 91] Sakurai, K., M. Ozeki and Y. Nishihara.
MT Application for a Translation Agency.
In *Proceedings of MT Summit III*. 1991.

[Sato 89] Sato, S.
Practical Experience in the Application of MT Systems.
In *Proceedings of MT Summit II*, pages 125-127. 1989.

[Sato 90] Sato, S. & M. Nagao.
Toward Memory-Based Translation.
In *Proceedings COLING '90*, pages 247-252. 1990.

[Shimazu 90] Shimazu, A.
Japanese Sentence Analysis as Argumentation.
In *Proceedings of the 28th Annual Meeting of the ACL*. 1990.

[Smith 89] Smith, B.
The Art of Machine Translation; Japanese to English.
IBM Research Magazine 27(4), 1989.

[Stanfill 86] Stanfill, C. & D. Waltz.
Toward Memory-Based Reasoning.
Communications of the ACM 29(12):1213-1228, 1986.

[Sugimura 86] Sugimura, R.
Japanese Honorifics and Situation Semantics.
In *Proceedings of COLING 86*. 1986.

[Sugimura 88] Sugimura, R. H. Miyoshi, & K. Mukai.
Constraint Analysis on Japanese Modification.
In V. Dahl & P. Saint-Dizier (editors), *Natural Language Understanding and Logic Programming, II*. Elsevier Science Publishers B.V., 1988.

[Sumita 90] Sumita, E., H. Iida, & H. Kohyama.
Translating with Examples: A New Approach to Machine Translation.
In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*. 1990.

[Sumita 91] Sumita, E. & H. Iida.
Experiments and Prospects of Example-Based Machine Translation.
Natural Language SIG Reports of the IPSJ :82-85, March 15, 1991.

[Suzuki 91] Suzuki, K. & T. Dasai.
A Processing of Co-Occurrence in Japanese English Machine Translation.
Natural Language SIG Reports of the IPSJ , March 15, 1991.
(in Japanese).

[SYSTRAN 91] Systran.
SYSTRAN Company Handout.
1991.

[Takeda 90] Takeda, K.
Bi-Directional Grammars for Machine Translation.
In *Proceedings of Seoul International Conference on Natural Language Processing*. 1990.

[Tanaka 89] Tanaka, H., S. Ishizaki, A. Uehara, & H. Uchida.
Research and Development of Cooperation Project on a Machine Translation System for Japan and Its Neighboring Countries.
In *Proceedings of Machine Translation Summit II*. 1989.

[Tanaka 91] Tanaka, H. & T. Aizawa.
A Method for Translation of English Delexical Verb-Deverbal Noun Phrases into Japanese.
Natural Language SIG Reports of the IPSJ , March 15, 1991.
(in Japanese).

[Toma 76] Toma, P.
An Operational Machine Translation System.
In R. W. Brislin (editor), *Translation: Applications and Research*, pages 247-259. Gardner, New York, 1976.

[Tomabechi 90] Tomabechi, H.
Symbolic and Subsymbolic Massive-Parallelism for Speech-to-Speech Translation: Hybrid Time-Delay, Recurrent, and Constraint Propagation Connectionist Architecture.
In *Proceedings of an International Conference Organized by the IPSJ to Commemorate the 30th Anniversary*. 1990.

[Tsuji 90] Tsuji, Y.
Multi-Language Translation System at Using Interlingua for Asian Languages.
In *Proceedings of an International Conference Organized by the IPSJ to Commemorate the 30th Anniversary*. 1990.

[Uchida 89a] Uchida, H.
ATLAS.
In *Proceedings of Machine Translation Summit II*, pages 152-157. 1989.

[Uchida 89b] Uchida, H.
Interlingua: Necessity of Interlingua for Multilingual Translation.
In M. Nagao, H. Tanaka, T. Makino, H. Nomura, H. Uchida, & S. Ishizaki (editors), *Proceedings of Machine Translation Summit I*. Ohmsha, Tokyo, 1989.

[Uemura 86] Uemura, S.
On Early Japanese Research Activities on Mechanical Translation.
ETL Bulletin 50(7):78-89, 1986.
(in Japanese).

[Valigra 91] Valigra, L.
Japan Shows Progress in Machine Translation.
The Institute (news supplement to IEEE Spectrum) , May/June, 1991.

[Vasconcellos 91] Vasconcellos, M. and D. A. Bostad.
MT in a High-Volume Translation Environment.
In J. Newton (editor), *Computers in Translation: A Practical Appraisal*. Routledge, London, 1991.
(In press).

[Vauquois 84] Vauquois, B.
Automated Translation at GETA.
Technical Report, Grenoble: GETA, 1984.

[Wang 90] Wang, Q., X. Wang, Y. Huang, & H. Yasuhara.
The Generation of Chinese Text from the Case Relations.
Natural Language SIG Reports of the IPSJ , November, 22, 1990.
(in Japanese).

[Weaver 55] Weaver, W.
Translation.
In W. N. Locke and A. D. Booth (editors), *Machine Translation of Languages*. MIT Press, Cambridge, Mass., 1955.
Originally published in 1949 and later reprinted in this collection.

[Wilks 91] Wilks, Y.
SYSTRAN: It Obviously Works, but How Much Can it be Improved?.
Technical Report, Memorandum in Computer and Cognitive Science, MCCS-91-215. Computing Research Laboratory, New Mexico State University, Las Cruces, N.M., 1991.

[Yamauchi 88] Yamauchi, S.
Ricoh English-Japanese Machine Translation System RMT/EJ.
BIT, September, 1988.
(in Japanese).

[Yokoi 91] Yokoi, T.
Collaboration and Cooperation for Development of Electronic Dictionaries - Case of the EDR Electronic Dictionary Project.
In *Proceedings of the International Workship on Electronic Dictionaries.* Japan Electronic Dictionary Research Institute TR-031, 1991.

[Zhu 89] Zhu, M, & H. Uchida.
Chinese Dictionary for Multilingual Machine Translation.
Natural Language SIG Reports of the IPSJ, March 18, 1989.
(in Japanese).

I. Appendix: Japanese Sites Mentioned in the Report

We list here all of the major Japanese sites that are mentioned in this report. There is a standard form for each entry. The name of the institution is given first, followed by the name(s) of the major MT system(s) that have been developed there. Notice that sometimes the name of the system is just the name of the company that developed it. Not all of the sites that we mention have developed a system, so the list will be empty for them. For sites that are not developers or that are unusual in some other way, there is an additional comment field that explains the site's role in this report. Most of the sites on this list were visited by one or more panel members. There are, however, a few sites that were not visited but that are mentioned in this report. Asterisks appear at the end of the entry line for those sites.

Advanced Telecommunications Research Institute International (ATR)

SL-TRANS, NADINE

Bravice International

MICROPAK

The panel has been told that Bravice went out of business in the first quarter of 1991.

Canon

LAMB ****

Catena-resource Institute

STAR, The Translator (Macintosh version of STAR)

CBU

HANTRAN ****

Center of the International Cooperation for Computerization (CICC)

CICC

CSK

ARGO

Digital Equipment Corporation (DEC), Japan

The panel looked at DEC only as a user of MT systems.

Electronic Dictionary Research Institute (EDR)

Not doing any MT development but they are producing a dictionary that is intended to support MT.

Electro Technical Laboratory (ETL)

CONTRAST ****

Fuji Electric

We visited Fuji to see their OCR system.

Fujitsu

ATLAS-I, ATLAS-II

Hitachi HICATS

IBM Japan, Ltd.

SHALT, SHALT2, JETS

International Business Service (IBS)

IBS is a translation service bureau that uses some MT systems.

Institution for New Generation Computer Technology (ICOT)

ICOT is doing work in general NL processing but has no active MT project.

Inter Group

Inter Group uses MT systems.

Japan Electronics Industry Development Association (JEIDA)

The Japan Electronic Industry Development Association has an active committee on MT. Its members are drawn from companies, academia, and government organizations.

Japan Information Center of Science and Technology (JICST)

JICST

Kyoto University

Research on several MT and NLP projects, including a major contribution to the MU system.

Matsushita Electric Industrial Company

PAROLE

Ministry of International Trade and Industry (MITI)

The Japanese Ministry of Trade and Industry has shown a great deal of interest in MT.

Mitsubishi

MELTRAN ****

NEC

PIVOT

Nippon Hoso Kyokai (NHK)

NHK uses the Catena STAR system.

Nippon Telegraph and Telephone (NTT)

ALT-J/E

Oki Electric

PENSEE

Ricoh

RMT

Sanyo Electric

SWP-7800 Translation Word Processor

Sharp

DUET

Systran

SYSTRAN

Toshiba

ASTRANSAC

II. Appendix: Biographies of Panel Members

Jaime Carbonell

Jaime G. Carbonell is Professor of Computer Science and Director of the Center for Machine Translation at Carnegie-Mellon University. He received his B.S. degrees in Physics and in Mathematics from MIT in 1975, and his M.S. and Ph.D. degrees in Computer Science from Yale University in 1976 and 1979, respectively. Dr. Carbonell has authored some 140 technical papers, and has edited or authored several books, including *Machine Learning: An Artificial Intelligence Approach*, volumes 1 and 2, and *Machine Learning: Paradigms and Methods and Knowledge-Based Machine Translation*. He is executive editor of the international journal, *Machine Learning*, and serves on several editorial boards, including that of *Artifical Intelligence*. He has also served as chair of SIGART (1983-1985), the special interest group on A.I. of the ACM, served on several government advisory committees, including that of the NIH human genome project, and is a founder and director of Carnegie Group, Inc.

Dr. Carbonell's research interests span several areas of artificial intelligence, including: machine learning, natural language processing, planning and problem-solving, knowledge-based machine translation, analogical reasoning, knowledge representation, and very large knowledge bases. In particular, Dr. Carbonell leads the PRODIGY project, an integrated architecture for planning and learning in complex domains, and also leads a multilingual high-accuracy machine translation research project.

David E. Johnson

David E. Johnson is a research scientist in the Theoretical and Computational Linguistics group, Mathematical Sciences department, IBM T. J. Watson Research Center, Yorktown Heights, New York. He received his Ph.D. in linguistics from the University of Illinois (Champaign-Urbana) in 1974 and has taught linguistics at the University of Illinois and at Yale University. His numerous publications include two books on linguistic theory: *Toward a Theory of Relationally-Based Grammar* (Garland Publishing: 1979) and *Arc Pair Grammar* (Princeton University Press: 1980) [with Paul M. Postal].

Dr. Johnson has had extensive experience in natural language processing. From 1974 to 1978 and again from 1982 to 1987, he was involved in the development of IBM's prototype natural-language database query system, TQA, whose leading-edge linguistic processing technology contributed significantly to IBM's product LanguageAccess. From 1987 to 1989, he was a staff manager in the Japanese Processing group at the IBM Japan Tokyo Research Laboratory, where he designed and oversaw the development of the English generator used in the JETS Japanese-English machine translation system, and participated in the development of the transfer component.

Elaine Rich

Elaine Rich is Director of the Artificial Intelligence Lab in MCC's Advanced Computing Technology (ACT) Program, where she has been responsible for the development of a knowledge-based natural language processing system. This work is now being used as the basis of an interlingual MT system. She was an assistant professor of computer sciences at the University of Texas at Austin prior to joining MCC.

Dr. Rich received an A.B. in linguistics and applied mathematics from Brown University in 1972 and a Ph.D. in computer science from Carnegie-Mellon University in 1979. She has published extensively,

including the best-selling textbook *Artificial Intelligence* (McGraw-Hill: 1983, 1991). She also has extensive experience as a consultant to several major corporations in the areas of AI.

Dr. Rich is a Fellow of the American Association for Artificial Intelligence. She serves on advisory committees for such government organizations as the National Science Foundation and the Office of Technology Assessment. She is Editor of *AI Magazine*, and serves on the editorial boards of several other AI journals.

Masaru Tomita

Masaru Tomita is an Associate Professor in the Computer Science Department at Carnegie Mellon University, where he is also the Associate Director of the Center for Machine Translation.

He holds a Ph.D. and a Master's Degree in Computer Science from Carnegie Mellon University (1985 and 1983, respectively) and a Bachelor's Degree in Mathematics from Kelo University (Yokohama, Japan, 1981.) During 1984, he was a Visiting Scientist in the Electrical Engineering Department at Kyoto University (Kyoto, Japan).

Dr. Tomita's research interests are in the area of natural language processing, including machine translation, parsing, natural language interfaces, computational linguistics and speech recognition. He has published three books, and authored or co-authored over 40 reference papers. In 1988 he received a Presidential Young Investigators Award from the National Science Foundation. He is an editorial board member of two international journals: *Computational Linguistics* and *Machine Translation.*

Muriel Vasconcellos

Muriel Vasconcellos has been professionally involved in translation, with focus on machine translation, for 27 years. As chief of the translation program at the Pan American Health Organization, a UN-family agency, she has directed the development and practical implementation of MT since 1977.

Dr. Vasconcellos' studies have been in linguistics, in which she holds the Bachelor's (1958), Master's (1982), and Ph.D. (1985) degrees from Georgetown University. Her graduate specialization was in theoretical linguistics, and her thesis was on translation theory. For 11 years she lectured on translation at Georgetown, where she taught a course on machine translation (1980-1988) and was co-presenter of intensive workshops on MT in 1985 and 1987.

Dr. Vasconcellos has more than 50 articles on machine translation and translation theory to her credit, as well as the book *Technology as Translation Strategy* (SUNY Press: 1988), and she serves on the editorial boards of several journals, including *Machine Translation.*

She has been active in professional translator associations throughout her career. She is currently president of the Association for Machine Translation in the Americas and secretary of the International Association for Machine Translation.

Yorick Wilks

Yorick Wilks is Director of the Computing Research Laboratory at New Mexico State University, a center for research in artificial intelligence and its applications. He received his doctorate from Cambridge University in 1968 for work on computer programs that understand written English in terms of a theory

later called "preference semantics": the claim that language is to be understood by means of a search for semantic "gists," combined with a coherence function over such structures that minimizes effort in the analyser.

This has continued as the focus of his work, and has had applications in the areas of machine translation, the use of English as a "front end" for users of databases, and the computation of belief structures. He was a researcher at Stanford AI Laboratory, and then Professor of Computer Science and Linguistics at the University of Essex in England before coming to New Mexico. He has published numerous articles and five books in the area of artificial intelligence, of which the most recent is *Artificial Believers* (Lawrence Erlbaum Associates: 1991) [with Afzal Ballim].

Dr. Wilks is also a Fellow of the American Association for Artificial Intelligence, on advisory committees for the National Science Foundation, and on the boards of some fifteen AI-related journals.

III. Appendix: Abbreviations Used in This Report

ATN	Augmented Transition Network
ATR	Advanced Telecommunications Research Institute International
CICC	Center of the International Cooperation for Computerization
DARPA	Defense Advanced Research Projects Agency (in the U.S.)
DEC	Digital Equipment Corporation
DS	Dependency Structure
EBMT	Example-Based Machine Translation
EC	European Community
EDR	Electronic Dictionary Research Institute
ETL	Electro Technical Laboratory
GPSG	Generalized Phrase Structure Grammar
HPSG	Head-driven Phrase Structure Grammar
HT	Human Translation
IBS	International Business Service
ICOT	Institution for New Generation Computer Technology
JEIDA	Japan Electronics Industry Development Association
JICST	Japan Information Center of Science and Technology
JPSG	Japanese Phrase Structure Grammar, a Japanese version of the English HPSG
JTEC	Japanese Technology Evaluation Center
LFG	Lexical Functional Grammar
MAT	Machine Aided Translation
MITI	Ministry of International Trade and Industry (in Japan)
MT	Machine Translation
NHK	Nippon Hoso Kyokai
NL	Natural Language
NLP	Natural Language Processing
NP	Noun Phrase
NTT	Nippon Telegraph and Telephone
OCR	Optical Character Recognition
PP	Prepositional Phrase
PS	Phrase Structure
S	Sentence
SL	Source Language
TDMT	Transfer-Driven Machine Translation
TL	Target Language
VP	Verb Phrase

Index

JTEC\WTEC reports are available from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650. Prices are as of 1/92 and subject to change. Add $3.00 for mailing per order, not per report. Add $7.50 for billing if order is not prepaid. These prices are for the U.S., Canada and Mexico. For information via Fax (703) 321-8547.

Title/Order Number

JTECH Panel Report on Computer Science in Japan (12/84) PB85-216760
E06/E01 ($24.00/11.00)*

JTECH Panel Report on Opto-and Microelectronics (5/85) PB85-242402
E10/E01 ($36.00/11.00)

JTECH Panel Report on Mechatronics in Japan (6/85) PB85-249019
E04/E01 ($19.00/11.00)

JTECH Panel Report on Biotechnology in Japan (5/86) PB85-249241
E07/E01 ($27.00/11.00)

JTECH Panel Report on Telecommunications Technology in Japan (5/86) PB86-202330/XAB
E08/E01 ($30.00/11.00)

JTECH Panel Report on Advanced Materials (5/86) PB86-229929/XAB
E08/E01 ($30.00/11.00)

JTECH Panel Report on Advanced Computing in Japan (12/87) PB88-153572/XAB
E04/A01 ($19.00/9.00)

JTECH Panel Report on CIM and CAD for the Semiconductor Industry in Japan (12/88)
PB89-138259/XAB
E07/A01 ($27.00/9.00)

JTECH Panel Report on the Japanese Exploratory Research for Advanced Technology (ERATO) Program (12/88)
PB89-133946/XAB
E09/A01 ($33.00/9.00)

JTECH Panel Report on Advanced Sensors in Japan (1/89) PB89-158760/XAB
E11/A01 ($39.00/9.00)

JTEC Panel Report on High Temperature Superconductivity in Japan (11/89)
PB90-123126
E10/A02 ($36.00/12.50)

JTEC Panel Report on Space Propulsion in Japan (8/90) PB90-215732
E10/A02 ($36.00/12.50)

JTEC Panel Report on Nuclear Power in Japan (10/90) PB90-215724
A14/A02 ($43.00/12.50)

JTEC Panel Report on Advanced Computing in Japan (10/90) PB90-215765
A10/A02 ($35.00/12.50)

JTEC Panel Report on Space Robotics in Japan (1/91) PB91-100040
E14/E01 ($50.00/11.00)

JTEC Panel Report on High Definition Systems in Japan (2/91) PB91-100032
E14/E01 ($50.00/11.00)

JTEC Panel Report on Advanced Composites in Japan (3/91) PB90-215740
E10/E01 ($36.00/11.00)

JTEC Panel Report on Construction Technologies in Japan (6/91) PB91-100057
E14/E01 ($50.00/11.00)

JTEC Panel Report on X-Ray Lithography in Japan (10/91) PB92-100205
E10/E01 ($36.00/11.00)

WTEC Panel Report on European Nuclear Instrumentation and Controls (12/91)
PB92-100197

JTEC Panel Report on Machine Translation in Japan (1/92) PB92-100239

*The first code and price are for a hardcopy. The second set is for microfiche.