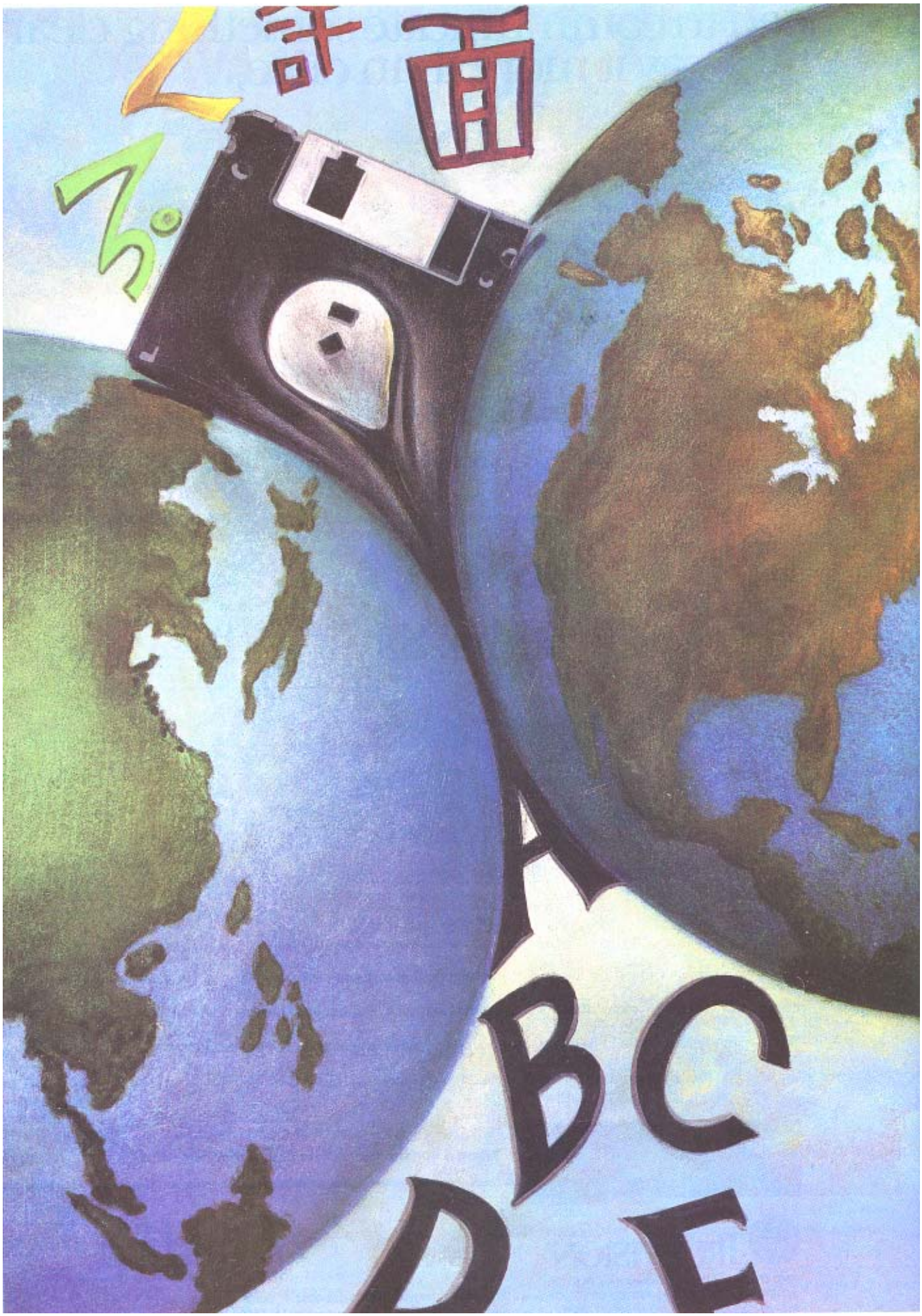
計画
A
BC
D E

MACHINE TRANSLATION

Translating the languages of the world on a desktop computer comes of age

MURIEL VASCONCELLOS

he process of converting information from one language to another with a computer, MT (machine translation), is an increasingly important technology. International economic and political stability and well-being are dependent on shared information. Nev-e in history has there been a more urgent need to top-linguistic barriers that divide the people of the world. common markets and stepped-up trade throughout ld have created overwhelming demands for linguistic . Just communicating in the nine official languages uropean Community means translating in 72 different ns.

rough translation meticulously captures all the nu-f the original text. Sometimes, though, a rough trans-all that is needed. Most translations are still performed le, but computers are shouldering part of the burden (see box "An International Network" on page 156).

reters—translators who deal only with spoken lan--don't have to worry about problems with input and But translators who must produce written output need f transferring their results to hard copy. Whether they tating machine, a typewriter, or a word processing pro-ie process is slow and costly. Computers can take on the drudgery of this process and free the human trans-the more creative aspects of the task.

hat Is It?

nes under the generic heading of NLP (natural-lan-rocessing). At the same time, because the technolo-ves many complex tasks, it's often seen as a category elf. MT's special status may also stem from the fact as the earliest kind of NLP. The first translation ma-vere designed in the early 1930s, and serious efforts op MT were under way soon after the ENIAC (Elec-lumerical Integrator and Calculator) made its debut

e software that merely looks up words, MT analyzes in the original language (the source language) and automatically generates sentences in the target language in which you want the translation. Input to a computer for translation is machine-readable text written in the source language. Output consists of text in the target language, which may be displayed on-screen or printed. Hard copy often shows the source and target texts side by side (see "How MT Works" on page 167).

MT can involve human assistance, but it shouldn't be confused with MAT (machine-assisted translation), a related but different mode. In MAT, a human translator prepares the target version using a word processing program and musters the aid of automatic terminology managers, on-line multilingual term banks, text-critiquing software, repetitions processing, and other computer-based tools that help to boost productivity.

The difference between MT and MAT is becoming less clear. Innovative systems in the research stage are blurring the distinction between the two by providing pieces of text that can serve as translation building blocks. Computer-based tools have become standard components of the translator's workstation, which may include full MT as well.

What Does It Do?

The dream is to build the equivalent of the babblefish of Douglas Adams' book *The Hitchhiker's Guide to the Galaxy*—a wearable device

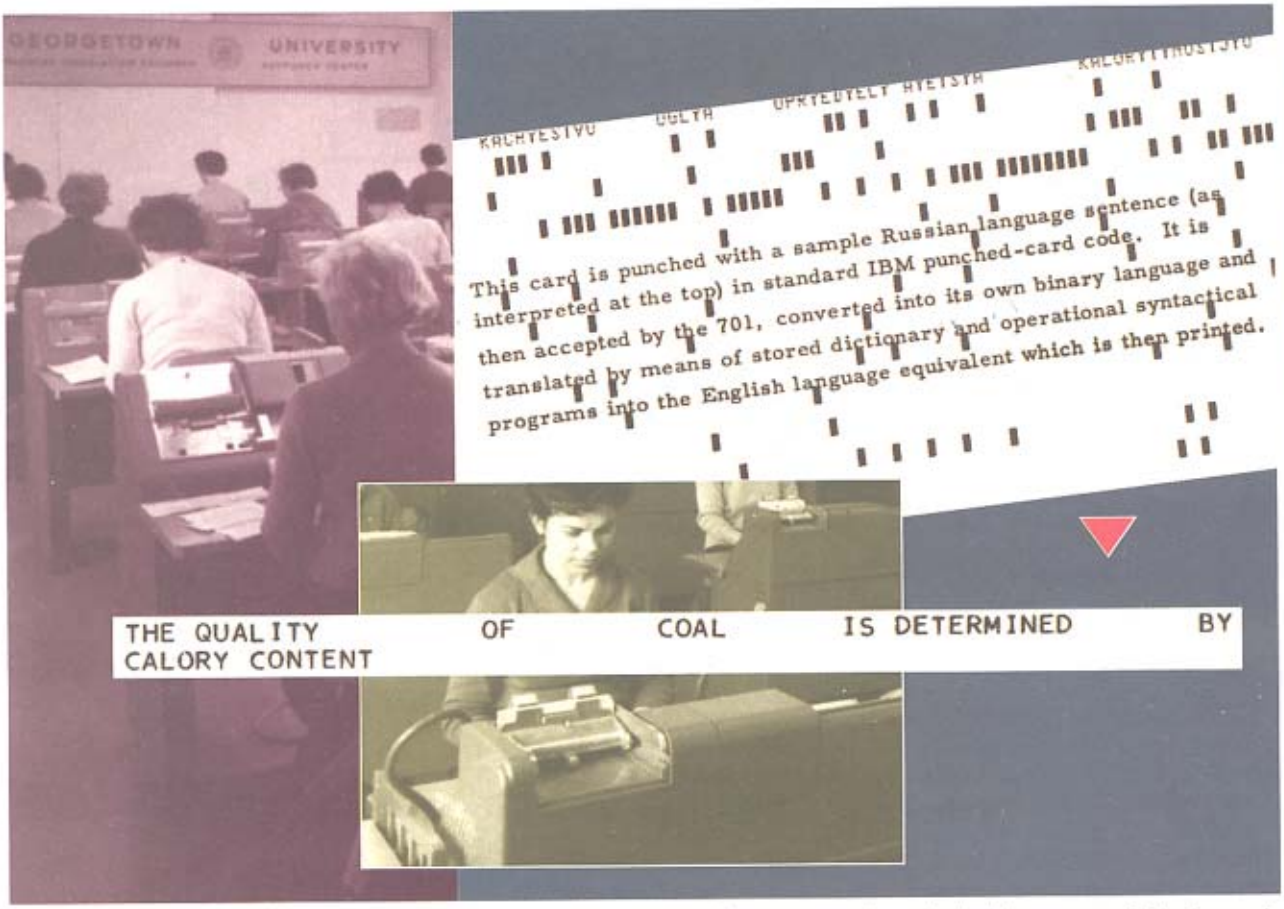

Photo 1: *The first known trial of MT took place in January 1954. Shown here are a card punched with a sentence in Russian and a printout of the translation in English.*

that simultaneously interprets from and into any language of the world. This concept sounds like science fiction, but in reality, speech-to-speech technology, in limited forms, is already in the wings. In the meantime, MT of written text is proving its mettle in a respectable range of settings.

MT works best if the subject matter is specific or restricted (e.g., maintenance manuals). The results are even better when the original text is straightforward and devoid of ambiguities.

Car manuals, for example, are consistent in style and vocabulary. Peter Wheeler of Antler Translation Services (Sparta, NJ) uses MT to translate automobile manuals from English to French for General Motors. "Automobile manuals are ideal MT texts—very dry, very objective, very factual, extremely repetitive, and very boring. That's not the sort of stuff a human translator works with well. With MT, I've achieved a threefold increase in throughput."

Progress in MT is measured by a system's ability to gradually handle more difficult text types and language combinations, with as little human assistance as possible. Another key goal is to be able to translate between European languages and languages that have non-Roman alphabets and structures (e.g., Japanese, Korean, Chinese, and Arabic). Finally, progress in the field is also gauged by how flexibly the system fits into the user's operation.

Key Features

- fills a huge and growing need for its technology
- is moving to desktop systems
- can double human output
- is cost-efficient
- offers dial-up services

Future Enhancements

- provide better-quality systems
- add more applications

Two key factors have come together to make MT easier to use. For a long time, the primary obstacle to more widespread use of MT was the cost and difficulty of getting text into the computer (see photo 1). Now there are large volumes of text in electronic files ready to serve as fodder for MT.

But the most dramatic difference is that personal computers and workstations now offer enough processing power to take on the MT functions that have been mainframe-dependent for nearly 40 years. Downsizing from the Goliaths to the Davids of computing has produced a new generation of devices that will soon be able to perform MT applications on the fly.

MT systems spend a lot of time looking at the various ways in which a sentence can be parsed and considering the roles and meanings that each word can have. Most of this time is spent mulling over possible choices. For example, the mainframe-based Systran system from Systran Translation Systems (La Jolla, CA) processes about 10,000 rules per second. If

you are performing MT on a desktop system and your document has many pages, the process may tie up your computer for quite a while.

Junior Babel-Busters

As the technology gravitates to smaller and more personalized computers, MT is becoming accessible to a larger public. The first personal computer–based system—the MicroCAT (which is no longer produced)—appeared in 1983. Today, the Sun Microsystem Sparcstation and other midrange Unix workstations are host to many commercial MT systems, as are virtually all the laboratory prototypes (see "Babelware for the Desktop" on page 177). Unix workstations, 386 and 486 PCs, and high-end Macs all provide sufficient power on the desktop to run the biggest MT systems. The challenge is to adapt the software to the new environments.

A recent example of a system designed for the capabilities of the 386 is the Engspan, which was developed by the Pan American Health Organization in Washington, D.C. In late 1992, this system, which translates from English to Spanish, was ported from a mainframe computer and runs efficiently on a 33-MHz 386 with DOS, 2 MB of RAM, and an 80-MB hard disk.

MT on Your Desktop

Being able to tap into MT from your desktop has several advantages. For example, you can use OCR (optical character recognition), CD-ROM, and internal modems and faxes to capture text and graphics, download databases, and exchange electronic files with clients anywhere in the world. Many databases offer information in languages other than English. For example, you might search other countries' patents or a body of legal decisions or update your client on the latest Japanese advances in superconductivity.

With database management tools for retrieving terminology and previously translated text, style checkers, and desktop publishing software, you have everything you need to set up your own multilingual operation. Executive Communication Systems (Provo, UT) makes MT ToolKit, which enables you to create your own dictionaries, write your own linguistic rules, and customize the basic architecture of an MT system. It has been used to develop systems for Korean and Norwegian translation.

LANs offer large groups of users the potential to centralize some of the more time-consuming tasks. You can farm out a CPU-intensive translation to a less-used machine and receive the results back as a file on a server. You can store the large main lexicon and specialized glossaries in one location and make them available to all. Use of a centralized dictionary makes it possible for managers and terminologists to control the introduction of updates.

You can also incorporate MT into the desktop publishing process. By the time it reaches the MT phase, input text will have already been tagged with the publisher's markup codes. Here MT can offer considerable savings, because the reintroduction of markup codes can double the cost of translating a text.

Graphics and tables, which are expensive and painstaking to translate by hand, can be reproduced exactly as they appear in the original. MT can not only speed up the task but also prevent errors that could slip in if the data were rekeyed. These savings, of course, are multiplied by the number of target versions generated.

An alternative way of bringing MT to the average personal computer user is through a dial-up service. From your computer, you can send a file by modem to a mainframe host. In the U.S., you can call up Systran and access a smorgasbord of languages. In France, you can get Systran translations through the nationwide network Minitel. And in Japan, you have a choice of Fujitsu's Atlas-II on NiftyServe or NEC's Pivot on PC-VAN, another large network. CompuServe will soon be offering similar services.

How MT Works

The philosopher I. A. Richards once wrote that translation is "probably the most complex type of event yet produced in the evolution of the cosmos." It's no wonder, then, that the architectures of MT systems vary in seemingly infinite ways. Certain elements, however, are common to the process.

In any MT system, the computer uses three sets of data: the input text, the translation program (including I/O routines), and the permanent resident knowledge sources. The most essential of the knowledge sources is the dictionary—a file of records containing the words and phrases of the source language against which the input text must be matched. Knowledge sources also include the sets of rules that are fired at various points in the translation process. Finally, many systems store a bank of information about the concepts invoked by the dictionary.

The largest MT systems work with dictionaries containing several hundred thousand words. For each word, a record holds formalized representations of information about how the word functions. Even when condensed, the record for each word can be as much as 100 bytes long. With a heavy-duty system, the dictionary is measured in tens of megabytes.

The first task of any MT system is to match the words of the input text against those stored in the dictionary. It can use either a binary or hash search strategy. When it needs to look up a word, it first goes to the index residing in memory and locates the appropriate page of memory. For each word that it matches, it retrieves a complete record that includes information about the possible functions of the word

An International Network

MT users and would-be users, as well as researchers and commercial developers, have recently joined in a common endeavor to improve and promote the technology as well as share information about MT. They have formed the IAMT (International Association for Machine Translation) and, within its overall framework, the AMTA (Association for Machine Translation in the Americas).

IAMT and AMTA publish *MT News International* every four months and the *MT Yellow Pages* once a year. In November 1992, IAMT/AMTA held a workshop that evaluated MT technologies and provided a showcase of MT systems.

For further information, write to the Association for Machine Translation in the Americas, 655 15th St. NW, Suite 310, Washington, D.C. 20005.

and its relationship to words that may occur with it.

Translation Quality

Translation quality generally improves as the systems acquire more rules and larger and more detailed lexicons. But there is a trade-off.

In the long run, systems that have robust dictionaries and rule bases demand less human intervention, but they are more costly to develop or tailor to a particular environment. On the other hand, systems that do not have these resources require more human labor to turn out a finished product.

Other factors of consideration are the structural proximity of the two languages, the domain, and the type of text. Most important, the quality required of the translation depends on how it will be used. Can the raw output be delivered without further polishing? How much human intervention is needed to make it acceptable to the client?

The quality of MT is also closely tied to the amount of human assistance the user is willing and able to provide. The raw MT product—direct from the machine—may be usable for certain purposes, but some human participation is usually involved.

You can intervene at any point along the way: before, during, or after the automatic translation process. People's time costs more than that of computers. The idea is to keep the number of human steps to a minimum by choosing the form of intervention that offers the best mileage for the application.

When to Edit

If the operator intervenes before a source text is translated, the step is called *preediting*. The idea is to eliminate lexical and structural ambiguities before a translation program takes over.

Preediting comes in two flavors. In the first instance, you revise a text that already exists. Sometimes, there is easy-to-use interactive software to help you with the task. For instance, The Smart Expert Editor by Smart Communications (New York, NY) is designed to serve as MT preprocessing software.

In the second kind of preediting, you prepare the text for the machine. It may be a new version of an existing text, or it may be a text that was drafted for the purpose, according to preestablished rules and vocabulary.

Although preediting makes the job easier for the machine, you often have to edit the output. Preediting is worthwhile when you are translating from one language to many, because it reduces the need for

human assistance downstream. This step can also be justified when the source language poses major linguistic problems at the input level.

When an operator responds to questions posed by the computer during the translation, the mode is called *interactive editing*. The operator is asked to resolve ambiguities that the program has identified. The computer offers various alternatives, and the operator clicks on the most appropriate choices.

By making these decisions before the target-generation phase, interactive editing reduces the manual editing required after the translation. An early product that offered interactive editing was Transactive by Alpnet (Provo, UT). And making its debut is the Augmentor, developed at Carnegie Mellon University (Pittsburgh, PA). Carnegie Mellon hopes that the combination of a rich interlingua, a domain-specific application, and an interactive interrogation component will eliminate manual preediting or postediting.

The most common form of human assistance is *postediting*. In this mode, you add the finishing touches to the machine-translated output after the computer has finished its job. Postediting is more labor-intensive than the other forms of editing, but it gives you control over the quality of the text. You can rarely avoid this stage when the translation is intended for a large number of readers.

In most situations, the posteditor, who is ordinarily a professional translator, thoroughly reviews the output and makes any necessary changes. The standards and purposes of the user will affect how long the process takes.

Typically the posteditor works at the computer, using an off-the-shelf word processing package. Macros designed for MT can speed up the process. Depending on the text, posteditors can double the output of traditional human translation, turning out between 3000 and 10,000 words in an 8-hour day.

Gathering vs. Disseminating Data

How you use MT depends on whether you want to gather or disseminate information. When gathering information, you translate text from a foreign language into your own. When you disseminate information, you translate it from your language into another.

Often, the usefulness of the information you gather is time-dependent (e.g., weather reports, job listings, and patent information). And at times, only a few people will see an information-only translation. For this reason, the quality doesn't have to be perfect. Because you can rarely predict what the style and subject matter of source text will be like, you need an MT system that is robust enough to deal with whatever it encounters. This is known as a general-purpose system.

The demands of general-purpose MT place a heavy burden on the system's analysis component: The grammar must cover a broad range of situations, and the dictionaries and knowledge sources must be large and detailed. Even with the best linguistic preparation, however, the quality of the output will not be as smooth as that produced by a system tailored to a specific domain.

In information-gathering operations, the input documents usually come from a wide range of sources and are available only in hard copy. The cost of converting the input into an electronic file may be prohibitive. And the use of OCR in combination with automatic postprocessing and human monitoring might not make enough of a difference to warrant the introduction

Circle 141 on Inquiry Card.

of MT. However, what is making MT more feasible for information-gathering purposes is the widespread availability of text in digital form.

General-purpose MT systems can speed up the work of in-house translators who have to produce publication-quality copy in various subject areas. For example, the Logos system, developed by Logos (Mt. Arlington, NJ), supports translators that perform this kind of work in the Canadian Department of the Secretary of State and in a number of translation service bureaus. Similarly, the translation team at the Union Bank of Switzerland uses Metal, marketed in the Americas by SieTech of Siemens Nixdorf (Munich, Germany).

The most widespread use of MT is in the translation of texts in limited domains (e.g., customer support manuals). This application allows companies to launch their products in several countries simultaneously. Here the users call the shots: They reduce input ambiguity by having a single domain, and they can predict, and even control, the style of the source text (see the text box "Is MT Right for You?" on page 180). In these applications, MT also helps to keep the terminology consistent throughout a firm's branches—an important feature in large projects, where product manuals can be thousands of pages long.

Now What?

The written-text MT systems of today will give way to the voice-based systems of tomorrow. Soon special-purpose, speaker-dependent applications will begin to emerge (see the text box "MT at Your Service" on page 160). Progress in this area will depend not only on advances in the MT environment but also on breakthroughs in speech-recognition technology.

On a broader scale, the research that has gone into developing knowledge sources and internal representations for MT is useful in other areas. Progress in MT foreshadows a bigger step toward the general availability of NLP applications. Natural-language analyzers and text generators—key components of MT systems—will be standard software.

The results of MT research are also being used to explore better ways of capturing, representing, and storing knowledge. The basic step that must taken before anyone can use text is to parse it. As general-purpose parsers become available, it will be possible for computers to parse the entire body of knowledge that is stored in the world's libraries. The establishment of an archive holding parsed information available to all would be a boon to scientists who build large knowledge bases—and ultimately to you.

MT has never enjoyed greater public awareness or a more favorable climate of opinion than it does today. If you can't conquer Babel, at least, thanks to MT, you can have a better idea of the knowledge that's available in the world and how you can tap into it. ■

ACKNOWLEDGMENT

Cris A. Fitch, vice president of engineering for Systran Translation Systems; Marjorie León, of the Pan American Health Organization; and Mark Clarkson, a freelance science writer from Wichita, Kansas, contributed to this article.

Muriel Vasconcellos is president of the Association for Machine Translation in the Americas and is a Washington, D.C.–based consultant in translation and machine translation. You can reach her on BIX c/o "editors" or on CompuServe at 71024,123.

Circle 79 on Inquiry Card (RESELLERS: 80).

Is MT Right for You?

MURIEL VASCONCELLOS

You need to keep several points in mind when deciding if MT is right for you. First, you must determine if you have an application for which MT is appropriate. It's important to pick your application and then decide on your system rather than vice versa.

Costs soar when the input isn't in machine-readable form, and an OCR (optical character recognition) device, while helpful, isn't a panacea. If your documents aren't in electronic form, you may want to think twice about using MT. In addition, there should be a large volume of material (e.g., 100,000 words per month) to be translated, with the expectation that more will be coming from the same source. In the beginning, there should be only one domain (i.e., subject matter); you can branch out later as you become more familiar with all the ins and outs.

Your decision to use MT will hinge, in large part, on the format, volume, and linguistic characteristics of the source-language text. The text should contain no ambiguities.

Hardware and Human Factors

You need a hardware platform that an MT system will run on. In a multiuser setting, you must be sure that there is good word processing support and that all the users have strong word processing skills. Multitasking workstations designed for translators are helpful.

Be sure to recruit people who have a positive attitude about using MT. This is especially important during the first few months while you are getting your system up and running. This stage involves customizing the dictionaries and gaining proficiency in postediting.

Your choice of a system will depend on the characteristics of your application, so it's important to identify criteria that are specific to your needs. And don't be tempted to buy software just because it's inexpensive. As with a house pet, the price you pay up front is a drop in the bucket compared to the cost of the care and feeding for the rest of your mutual lives. For example, a less expensive system might cost more in terms of support personnel and customer support (some MT companies charge you for it). Also, if your time frame for translations is tight and if your budget for human intervention is limited, it's crucial that you test the system's performance on randomly selected texts.

Despite decades of scientific study, the evaluation of translations is an uncertain exercise. The definition of an error will vary, depending on the purpose of the translation and the values of the end users. Errors in raw output are important mainly to the analyst, who knows the inner workings of MT systems and can classify the error types according to their causes. Such an analysis can tell something about the system's potential and the effort that will be required to fix and to maintain it. You should make sure that you compare outputs from different systems produced under the same conditions.

The value of a system depends on its potential to grow and to improve its performance, as well as how easy it is to use and to maintain. It's important to know the language combinations that have been developed for the system, the size of the dictionaries or knowledge bases, the ease with which you can add to the dictionaries, and the possibilities of extending the system to include the domain that you are interested in.

Muriel Vasconcellos is president of the Association for Machine Translation in the Americas and is a Washington, D.C.–based consultant in translation and machine translation. You can reach her on BIX c/o "editors" or on CompuServe at 71024,123.

is essential with any MT product for personal computers, because it lets you add your own terminology to the program. PC-Translator simplifies the creation of your own dictionaries by importing lists of terms in ASCII format directly into the software. In addition, MT software generates lists of words not found in a given text to help you customize your software. You decide which words and phrases to add to the dictionaries. The ability to add, delete, or modify dictionary entries dramatically improves the quality of a translation and reduces the time spent postediting an output. You'll find that it can take from two to four weeks to customize a system.

All these systems ask you to insert the part of speech of the word you are adding and to provide its translation. With extensive dictionary coding, the system can deal with ambiguities that arise from the use of words that can take the form of multiple parts of speech. For example, the program will recognize the different translations of a homograph (i.e., a word that is spelled like another but has a different meaning or pronunciation) used as a verb and as a noun in the same sentence (e.g., "The can can explode"). Because Globalink's GTS-Professional can classify *can* as both a verb and a noun, it's better able to translate the sentence than a product that requires less dictionary coding.

Workstation-Based MT Products

MT workstation products are designed to handle heavy volume—when you have to translate 2000 or more pages of text per year. Translation speeds range from 20,000 to 1 million words per hour.

A workstation MT system is a large investment. Software prices start at $10,000. A system can cost several hundred thousand dollars, and pricing structures are as diverse as the possible configurations.

Socatra (Quebec, Canada) spent over 12 years preparing its XLT computer-assisted translation system for the commercial market. Access to XLT is uniquely controlled by the company. Socatra rents software for a specific number of words. After you pay an initial subscription, Socatra provides you with the software and an access card, which resembles a credit card. The card contains a microprocessor that counts the words translated and acts as a security device. You can obtain an XLT