Don't look now

By Andrew Joscelyne
Photos by Ed van der Elsken

With the government and the electronics behemoths funding dozens of machine translation projects, to the tune of billions of yen (two or three times the number and funding of all the other projects in the world combined), the Japanese are positioning themselves to dominate yet another strategic growth industry.

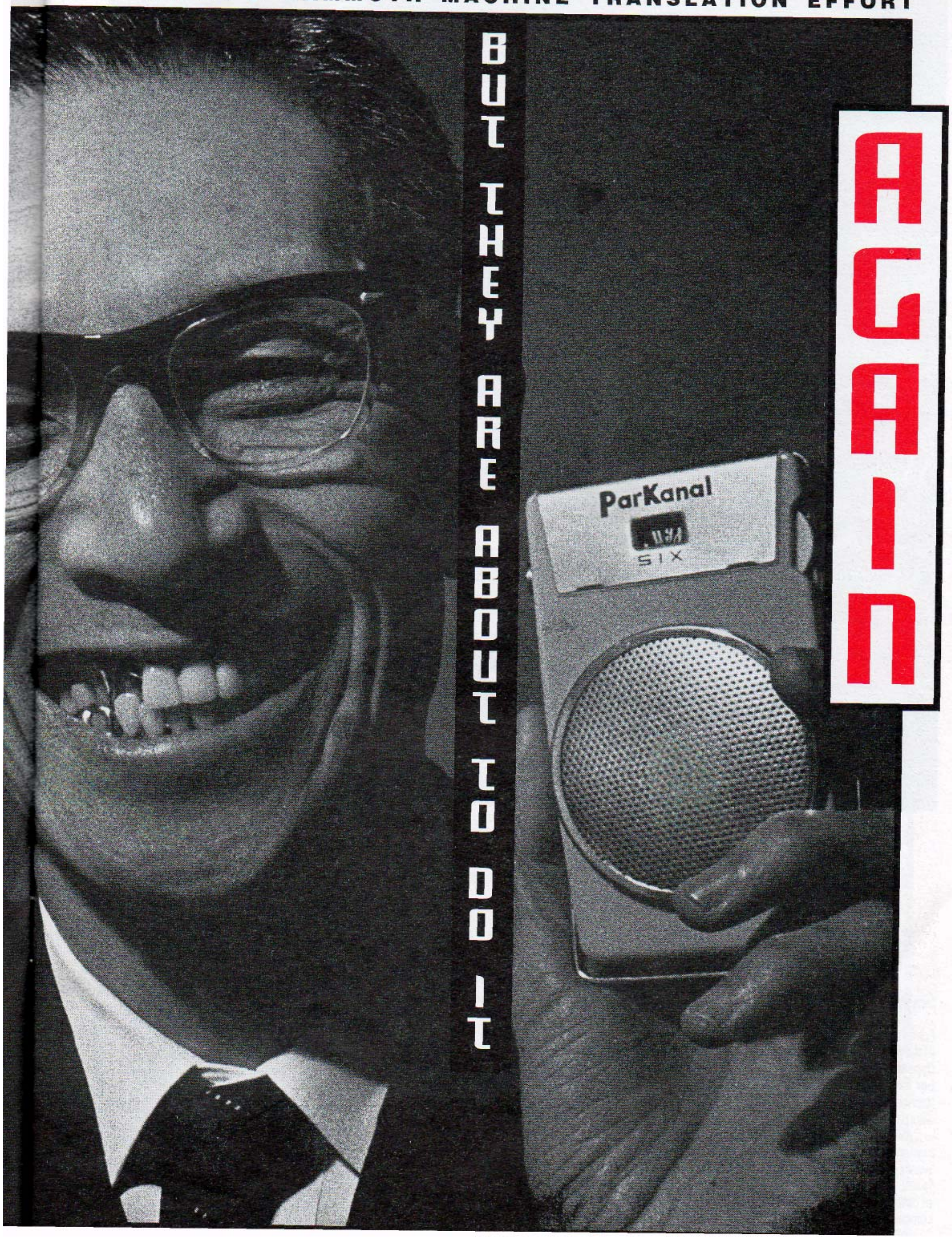
Their need is obvious. As world competition in the hightech industries heats up, Japanese R & D departments in space, biotechnology, robotics, opto-electronics and medical technology need more and more access to Western market intelligence and research – and fast. A recent study by the Japan Electronic Industry Development Association estimated the translation market at a mind-numbing 800 bn yen (US$7 bn) per year, of which 43% was Japanese-to-English and 47% English-to-Japanese.

What hasn't been obvious in the West is exactly what's going in Japan. To paraphrase Raymond Chandler, what we don't know about Japanese machine translation fforts could fill Tokyo Bay. Like, who is doing what with whom, and how far have they gotten?

Wonder no more. LT/Electric Word's roving correspondent Andrew Joscelyne is just back from apan with the first comprehensive report on the Japanese machine translation effort. An LT/ lectric Word world exclusive.

e photos are published in "De ontdekking van Japan" by Ed van der Elsken, Publisher Fragment, Amsterdam '88

LOOK AT JAPAN'S MAMMOTH MACHINE TRANSLATION EFFORT



BUT THEY ARE ABOUT TO DO IT AGAIN

PHOTO: ED VAN DER ELSKEN

Keizo Sakurai runs a bustling translation company in Hachioji, just half an hour by bullet train from central Tokyo. His inhouse team of 24 translators, backed up by 90 modem-owning freelancers, net Sakurai a respectable annual turnover of around 200 mn yen (US$1.5 mn). Nothing to holler from the rooftops about there, you might say.

However, Sakurai's is a firm with a difference. For this relatively small translation company is one of the few in Japan to use *kikai honyaku*, or machine translation. Sakurai pays 200,000 yen (US$1400) per month to rent Duet, a heavy-duty English-to-Japanese hard/software package developed by Sharp to crunch industrial scale text in batch mode.

With its 32-bit processor, Sakurai's Duet workstation yields an optimum raw output of 30,000 words per hour (wph). Of course, he'd like it to be faster – ideally 10,000 pages per month. But speed of output is determined by quality of input. And the company has to crunch a lot of sloppy source text.

All the same, Keizo Sakurai maintains that MT has more than doubled his company's productivity, and he wonders why so few of his competitors seem to follow suit.

To understand how much of an exception Sakurai is, consider this fact. Of the 80 *owners* and leasers of the Sharp Duet, there is only one other real *user* : the Japanese Civil Defense Force (or "army," as it's sometimes bluntly called). The other 78 presumably find it either too complicated or too unprofitable to handle.

MITI AND THE NAGAO GENERATION

Potential users may be slow on the uptake. But grand-scale MT research has been proceeding apace in Japan since the early 1980s, when the Ministry for International Trade and Industry (MITI) mapped out the future of advanced information technology in its ongoing Fifth Generation computer technology program.

Charged with the machine translation component of this project was Makoto Nagao, who started teaching computational linguistics in Kyoto University's department of electrical engineering in the late 1970s.

More than any other individual, Nagao has personally shaped the current state of Japanese machine translation. Not least thanks to the fact that he himself trained most of the MT researchers working for the "Big Eight" – those eight leading electronics firms which, in the true Japanese collectivist tradition, themselves still contribute both funds and staff to the Fifth Generation project: Fujitsu, Toshiba, Hitachi, Mitsubishi, NEC, OKI, Matsushita, and Sharp.

For the companies, the chief benefit of such costly participation is that the know-how they gain from the Fifth Generation project can be constantly plowed back home, enabling them to develop and launch their own – albeit only "first" or "second" generation – MT products as they proceed. There is a high degree of uniformity among the Big Eight's approaches to MT. Thanks to Nagao's overwhelming influence, Japan has been remarkably free of the theoretical infighting among researchers that have so dogged the European MT scene.

One result is that the architecture of the various commercial systems now available tends to be remarkably similar.

Their uniform focus tends to be one of simple design: how do you build a system with a dictionary entry ceiling of around 40,000 general, and another 50,000 technical, words which will offer rapid – 60,000 wph – raw output from a mainframe host computer?

Almost all the systems use Nagao's preferred case grammar parsing technique. Their more recent versions have been written in the LISP AI language – some having Prolog applications suitable for workstation use. And all eight adhere to Nagao's evaluation methods: a five-point scale for intelligibility and a seven-degree scale for accuracy.

No wonder Nagao himself considers the quality of all their completed commercial products "almost equivalent." Well, he would say that, wouldn't he?

STANDALONE VS ONLINE

Two separate trends are currently changing the face of the Japanese MT scene: on the one hand, the arrival of workstation versions of previously mainframe software; and on the other, the growing implementation of data-transmission networks to extend remote users' access to mainframe packages. On the face of it, these trends would seem to tug in opposite directions. But taken together, they have the potential to fuel a level of growth in *real* use that has eluded Japanese MT in its first near-decade.

This year, a number of MT systems have already appeared for the workstation, marketed as cheap solutions to heavy translation loads. The makers are hoping that the resulting increase in use will also incite more useful customer feedback. So far, there have been too few real users to badger the

MANAGING THE SILICON ARCHIPELAGO

MITI'S FIFTH GENERATION

LANGUAGE TECHNOLOGY PROJECTS

In September 1979, Keiichi Konaga, Japan's then top man at the Ministry for International Trade and Industry (MITI), published his "Vision for the 1980s," an R&D blueprint for combining the resources of industry and academia for the greater glory – and profit – of Japan Inc.

The result: the launch that same year of the much-vaunted MITI Fifth Generation project at the Institute for New Generation Computer Technology. Its aim: to discover the basic silicon technology that would underwrite all future computer applications.

MITI's longterm objective was to develop Knowledge Information Processing Systems, or KIPS: machines to handle conceptual material in intelligent ways, such as talking, understanding, and translating.

Boffins at the ministry's electrotechnical laboratory, where major research projects in the field are coordinated, saw natural language processing as having a key role not just for the decade ahead, but for communications between men, machines, and cultures well inro the 21st century.

A frenzy of research work gripped labs across the archipelago, from the Big Eight manufacturers to small universities. In the MT field alone, a staggering 18 research programs were being conducted in 1982 – two of them at Kyoto University alone.

As it happened, the initial euphoria didn't last long, and most of the early research bit the dust. However, two survivors still rise like twin Mount Fujis above a forest of cherry blossoms: the EDR and CICC projects.

GETTYSBURG ACCESS

The nine-year electronic dictionary, or *EDR Project* is being conducted at the EDR Institute Ltd., Tokyo. This body was set up in 1982 by the Japan Key Technology Center. Its funding: 14 bn yen, MITI footing most of the bill and the Big Eight the rest.

EDR's objective is grandly stated as "to build dictionaries *of computers, by computers, and for computers*" - in other words, to make fully automatic language knowledge bases that can eventually be used in every other natural language application from spoken information retrieval through machine translation.

There are two interlinked routes to this grandiose ambition. On the one hand, EDR is compiling a set of dictionaries; on the other, it is developing a software environment to handle them. The dictionaries consist of a 200,000-entry Japanese-and-English wordlist and a 100,000-entry information technology dictionary, from which a 500,000-entry "concept dictionary" will be derived. The concept dictionary will consist of conceptual descriptions of terms that will be mapped together onto a concept classification chart – a sort of semantic network of conceptual relations.

The idea is that, through *understanding* dictionary entries via the conceptual relations behind them, computers will learn how to tackle natural language tasks.

WORD WORK AHEAD

The environment EDR is developing to handle its dictionaries consists of software to search for new words in a textbase, editing entries, analyzing and generating sentences, and verifying conceptual structures.

According to EDR chief Yoshio Toshio, after three years of compiling data, the research team has realized that electronic dictionaries will never be "complete" and will need permanent updating and maintenance work. The most useful aspect of EDR's research for the outside world lies in its exploration of the technology used in dictionary development, where factors such as speed and accuracy of input can offer longterm advantages - provided, of course, the basic architecture is sound.

Certainly, EDR's infrastructure is exemplary: the eight laboratories scattered around Japan are linked by a DDX-P packet-switched network, offering rapid access to computing power and information exchange, as well as potential openings to other participants.

Toshio echoes other Japanese infotech decision makers when he claims that EDR's plan to handle vast volumes of language (sci-tech terminology could reach 50 million terms) makes demands on basic architecture that render natural language processing the key technology to the future.

When Kiyonori Konishi, head of telecoms giant NTT, was asked what he thought of the Fifth Generation project, he replied that there were two main results - one was to have stimulated competition in Europe and the US, and the other was EDR.

At present, EDR dictionary research is focusing on Japanese and English – the language pair processed by 99% of all Japanese natural language work. In the more distant future, though, EDR plans to test the validity of their conceptual framework against other – especially – Asian languages.

RACE FOR THE RIM

MITI's other great survivor is the Center for International Cooperation in Computerization's *CICC Project* for an interlingua-based multilingual East-Asian-languages MT system. CITT brings together researchers both from universities and the Big Eight.

The languages targeted by CICC are Indonesian and Malay, Chinese and Thai. Curiously, CICC has no plans for Korean, perhaps because Korea is seen as a rival in the race to dominate the economies of both countries' emerging Pacific neighbors.

According to CICC boss Akiko Uehara, Japan and the Pacific Rim present "a whole new market for technical documentation, for which there is no available translation practice. In fact, we are going to create a new type of information society by the widespread use of our developing translation systems."

A techno-watcher at the French embassy in Tokyo goes so far as to see Japanese as the future "world interlingua," with Japan poised crucially between the developing countries of Asia and the West. "If the Japanese complete their dictionary and MT projects as planned, they'll be in a privileged position to transfer know-how in both directions."

No one at MITI could have put it better.

makers about low-quality utilities or flawed design.

At the same time, mainframe vendors are hoping that the growing use of value-added data-transmission networks (VANs) among translation companies will expand the market for extant but underused mainframe software. Some are already offering rapid online machine translation in combination with the postediting services of a human translation company.

Some manufacturers, such as NEC and Fujistu, already provide their main users with VANs. And when the NTT (the semi-governmental Japanese telecoms concern) extends the ISDN version of its existing Information Network System, ISNET 64,- this will increase generalized access not only to online databases but also to processing services such as MT. Combined with the upcoming G4 (400 dpi) fax standard, the technology will even offer laser quality document printouts.

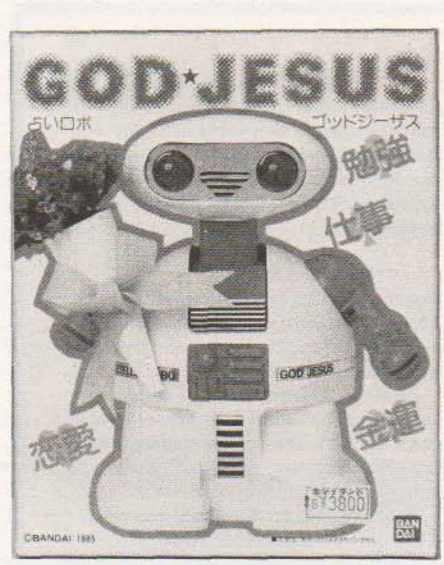


PHOTO: RENDA VAN DER BURG

FUJITSU STILL ON TOP

Japan's first ever homegrown commercial MT product was launched by Fujitsu Ltd. in 1984. Called Atlas I, it was a mainframe-based first generation English-to-Japanese transfer package – and a prototype for the most widely used range of mainframe-based MT products in the world.

Fujitsu is the biggest player in the Japanese MT game. To its pioneering Atlas I, the company has added a Japanese-to-English module, called Atlas II, which has been operational for four years, outputting 60,000 "raw" words per hour. (By contrast, Systran, the main MT supplier in Europe and the US, claims 500,000 wph when running on a IBM 4381 mainframe.)

Fujitsu is currently working on a multilingual workstation version of Atlas II, with German, French, and Chinese as potential target languages, using an "interlingua"-based architecture.

BEYOND FUJITSU

Japan's commercial MT scene is about far more than just Fujitsu's Atlas. Despite the country's low number of MT users, competition among suppliers is keen. 1989's key trend: the shift from the mainframe to the workstation.

Here's an across-the-board guide to what's happening outside Fujitsu in the Japanese MT marketplace:

Hitachi Ltd. is one company that's busily porting MT software from mainframe computers to workstations.

As mainframe products, Hitachi's English-to-Japanese program, _EJ_, and its more recent Japanese-to-English counterpart, _JE_, are capable of churning out 60,000 words per hour.

But the large number of support staff required to manage mainframe-based dictionaries and finetune the products' other software have made these programs prohibitively expensive.

So Hitachi has opted for a standalone solution. In April 1989, the company announced that it would port its MT software to its Engineering Work Station (EWS). An EWS version of JE will be released in October under the name HICATS (Hitachi Computer-Aided Translation System). Price: 3.8 million yen (US$34,400).

HICATS will bundle a number of software support tools for pre-and postediting. These will include CD-ROM dictionary assistance and multimedia document production utilities for handling diagrams, tables and charts, as well as dictionary maintenance facilities adapted for both end and expert users.

The CD-ROM dictionaries – 75,000-entry English-to-Japanese and 50,000-entry Japanese-to-English – are both based on the famous Kenkyusha bilingual tomes. And Hitachi plans to digitize a further suite of specialized scientific and technical wordlists to CD-ROM, including the standard Iwanami medical dictionary.

HICATS' icing on the cake is that it runs on Hitachi's superfast HI-UX operating system, whose multitasking function allows both pre- and postediting while translation itself is also underway.

Another workstation contender is the _NEC Corp._, whose _Pivot_ MT system is interlingua-based.

Pivot is an ACOS-based integrated MT system already available on NEC mainframes via a VAN. One of its customers is Japan Convention Services Ltd., the country's biggest translation and interpreting company.

This autumn, NEC is to launch Pivot for workstations in both standalone and cluster models, the latter allegedly more cost-effective since its disk file can be shared by more than two operators handling dictionary support and editing functions.

In the short term, NEC intends to extend Pivot's language pairs to include a Korean-Japanese module, and in the long term, to integrate speech recognition input and synthesis output.

Toshiba Corp.'s new ten-to-fifteen-million-yen (US$111,000) _AS-TRANSAC_ package is seen by many Japan-watchers as _the_ state-of-the-art English-to-Japanese MT rig. Like many Japanese MT packages, AS-TRANSAC has been through several mutations in its still young development life.

Two years ago, known simply as the Toshiba MT System, it was running on an old UX-700 32-bit minicomputer but doing a respectable job translating two-way Japanese-and-English scientific texts.

As such, it was integrated into the company's futuristic Automatic Translation Typing Phone (ATTP), in which machine translation was stitched onto a datacoms network in order to allow realtime written communication between different language users seated at remote terminals.

Then in 1988, came an update and the first name change. The system was redubbed TAURAS (Toshiba Automatic Translation System Reinforced by Semantics) and beefed up with the addition of lexical rules for semantic processing.

However, TAURAS was shortlived. This year, the package was once again renamed – to AS-TRANSAC. This now commercially available Japanese-to-English product runs with its dictionaries and support tools on Toshiba's 32-bit AS 300C series workstation.

A more recent entry in the MT workstation race is the _OKI Electric Industry Company Ltd._, which has been selling its _PENSEE_ Japanese-to-English package since 1987.

Like most of the other operators, OKI too will be bringing out a new version of its product this autumn: a two-way rig running on a 1000 Unitopia UNIX-based workstation with 8 Mb of main memory plus 80 Mb of auxiliary storage and a output rate of a modest 3,000 words per hour.

OKI's sales department claims over 100 owners. But just like other manufacturers, the company bundles PENSEE with its brand new workstation. Real user, please stand up and take a bow.

Sanyo Corp. has also produced a prototype Japanese-to-English MT package running on its SWP-7800 workstation. A commercial version is due out at year's end, aimed at clearing up some of the original gobbledygook raw output.

Last but not least, _Mitsubishi Electric Corp._ also bundles a Japanese-and-English MT product with its Melcom PSI II workstation under the name _Meltran J/E_. After five years of development, the company claims a sizzling optimum output rate of 10,000 wph. – more than double that of most workstation-based rigs. Of course, "optimum" implies perfect machine-ready input – the age-old MT dream.

All the same, Meltran will soon be able to read MS-DOS text files – a rarity among Japanese MT rigs. It will also keep records of unknown words for later dictionary processing, be clusterable into a multi-workstation environment, and offer a "word-learning" facility whereby it will absorb postedited translations and select preferred lexical values for future use.

Can this be the translation company's ideal semi-intelligent tool?

JAPAN LAYS ALPAC TO REST

Landing in Japan's hightech mailboxes in July 1989 will be a new fifteen-chapter report on the current state of Japanese machine translation, 23 years after America's ALPAC Kiss of Death.

In 1966 – if you need reminding – the fateful Automatic Language Processing Advisory Committee (ALPAC) Report presented the US National Academy of Sciences with the grim recommendation that government funds for MT research be terminated with extreme prejudice. The US government duly complied, and American MT went belly-up for fifteen years.

1989's Japanese report is a whole other story.

Nicknamed JALPAC – counter-intuitively, in view of its overall glowing optimism – it was written by a study committee appointed by MITI and the Japan Electronic Industry Development Association (JEIDA), which had already sponsored a report on Japanese documentation translation in 1982.

RUMORS

Presented by Hirosato Nomura of the Kyushu Institute of Technology at April's International Forum for Translation Technology in Oisu, the report is a comprehensive roundup of Japan's MT scene.

It contains the results of a questionnaire about translation demand, supply and general practice, plus a point-by-point comparison of today's technological environment with that of the 1960s.

It also includes a survey of current MT systems, accompanied by an evaluation method based on a 600,000-word English-to-Japanese benchtext, plus a set of proposals for improving Japanese technical writing skills and developing MT systems with discourse – as well as sentence-level – analysis. All very predictable.

However, what you won't find in the JALPAC Report is any confirmation of the rumor about its origins which is rife in Japanese MT circles. The word is that the decision to produce this glowing report was taken by the inner caucus of JEIDA's MT group after one of their number had heard US MT expert Martin Kay hinting darkly that American MT funding was about to receive another government thumbs-down.

After all the big yen the Japanese had sunk into natural language processing – so goes the rumor – they were naturally anxious to pre-empt another ALPAC and the poisoning of the MT atmosphere worldwide that such a negative judgement would bring.

RAISING STANDARDS

Firmly patting its compatriots on the back, JALPAC is a sterling morale-booster for the Japanese MT scene. You'll scour its pages in vain for serious criticism.

Foreign observers, however, are quick to point out certain shortcomings, such as the need for more shared international standards in developing MT applications and for avoiding the duplication of basic work.

Both Alan Melby (Brigham Young University) and Muriel Vasconcellos (the Pan American Health Organization MT Center) have called for the establishment of a standard method of electronic dictionary entry, so that in the long term Japan's EDR project (see MITI sidebar) will be able to benefit from and contribute to other similar projects worldwide.

Document formatting is another area in which international standards will be increasingly significant, especially once MT has entered the information flow as a node in a vast ISDN network.

In this respect, Fujitsu Ltd. deserves a mention. Along with Systran International (the French-owned MT supplier) and Systran Corp. (Japan's Systran operation – and an entirely separate company from its French namesake), Fujitsu is participating in the EC's Esprit-based program to establish document interchange on the ODA (Office Document Architecture) standard.

GETTING TOGETHER

JALPAC's bottom-line message is that the success of Japanese MT depends more than anything on the organizational strategy behind the country's various projects: serious long-term goals must be formulated; manufacturers, as well as government agencies, have to be persuaded to invest money; industrial rivals need to be grouped into cooperative R & D teams and given state-of-the-art equipment to work on.

It came as no surprise that Bud Scott, chief MT scientist to America's struggling Logos Corp., sounded bitter when asked to compare the Japanese and Western MT scenes: "The US government has singularly failed to synergize national MT resources over the 25 years since the ALPAC report sent MT underground."

And like other observers, Scott showed obvious admiration for how corporate and supra-corporate cooperation in Japan had produced tangible results. "Maybe it's because they're all engineers," he mused. "It's always linguists that never agree on how to get the MT job done."

Japanese source text will first be analyzed into semantic segments and translated into a language-independent formal code, or interlingua. Then, if it does the job, there will be no variation in the ease with which the various target language translations are built up.

Information about what Atlas customers really do with their systems is less readily forthcoming than that on products. Any Fujistu salesman will tell you deadpan how some 300 Japanese manufacturing companies have purchased the Atlas mainframe system. And that's far more than any other MT developer would even dare to claim and about twice as many as the total number of MT users in the rest of the world put together.

However, when it comes to real use, the truth is less spectacular. Atlas MT software is bundled to Fujitsu's mainframe customers. But the customers don't necessarily bother to use them. Says Hiroshi Uchida, head of research at Fujitsu: "Only about 10% of those who have Atlas actually use it, mainly because inhouse translators can't adapt to it."

His colleague Shigeru Sato, board director at Fujitsu Laboratories, is even more candid: "Ten percent have applied to develop their own customer-specific dictionaries in Atlas. Of those, only one or two will actually use the product to perform machine translation."

Two Atlas users are often cited as exemplary: industrial vehicles manufacturer Matsuda and Nippon Steel, both of which use Atlas to put maintenance documentation into English.

MT IN THE VAN-GUARD

In order to boost real use and counter anxiety about how to handle the system, Fujitsu has opted for networking. The company has recently opened a VAN center for its customers, designed both to broaden the market and improve Atlas translation quality.

Called Atlas-Mail, this service allows users access to Atlas via the Fenics e-mail network. Network service supplier Facom Information Processing Corp. (FIP) has joined forces with Inter Group Ltd., a leading translation company, to offer two-way networked Japanese and English translation. FIP provides the network support via Fujitsu Oasys or FMR PC terminals, while Inter Group supplies online postediting.

Atlas-Mail's grand design is for Inter Group translators to contribute to dictionary development by feeding terminology obtained during postediting back into Atlas's central dictionaries. In this way, it is hoped that customer lexical know-how can be used to upgrade Atlas.

Customers can enter their text on terminals or via fax machines, track the progress of ongoing translation via the network, and retrieve the final postedited version from their e-mail box. The cost: around 3000 yen (US$25) per A4 page. Access costs, however, are as yet prohibitive for any but very large companies: between four and five million yen ($45,289) per year.

Fujitsu realizes that the only way to turn its system into a manageable business tool is to improve both upstream and downstream processing. Sticking Atlas into an aeronautics manufacturer's mainframe and saying "get on with it" will just get Fujistu - and MT - a bad name.

However, all is not yet cherry-blossom time for Atlas-Mail. FIP's Tsumomu Tanaka reckons that only 15% of the networked translation performed by Atlas is actually entered by e-mail.

"Most of it still comes either on disk or simply as manuscript. The problem is that we haven't yet agreed on standards for document file structure. Until this problem has been solved, we won't be able to offer the full benefits of an MT networking service."

BACK TO SCHOOL

Atlas-Mail faces a number of other miscellaneous problems.

One such is the endemic MT stumbling block of input text quality - the system provides no pre-editing. And as the saying goes: "Garbage in . . ."

One solution to this problem might be Tanaka's own recent suggestion that FIP offer customers a scale of charges based on how easily Atlas can handle the source text. The better written – i.e. optimally unambiguous – the document, the cheaper the MT rate.

Another difficulty for the FIP-Inter Group partnership is that of boosting Atlas's thirteen domain-specific technical dictionaries beyond their current 300,000 terms. This is taking far longer than expected and is very expensive. "We need five to six million technical terms," says Tanaka. "And you can't do that without government support."

Despite these local difficulties, Fujitsu was last year contracted by the European Community to supply it with online English translations of Japanese conference proceedings and technical reports from such vast sources as the Japan Information Center for Science and Technology (JICST) databases.

Adapting human beings to technology presents another set of problems for Atlas-Mail. Using a wordprocessor is a fairly recent innovation in Japan and by no means as widespread as in the West – and slow onscreen document handling would make a mockery of Atlas's translation speed.

Postediting is of course a crucial – and the only uniquely human component in Atlas-Mail's service. So Shogo Iwashita, head of MT at Inter Group, has set up what may well be the world's first all-comers postediting seminar: a training course to turn translator-revisors into machine-output bilingual editors.

Iwashita claims to be swamped with applications, and after a fifteen-week 32-lesson course, he offers jobs to his best students.

Back in Hachioji, Keizo Sakurai knows all about the importance of postediting. In his office, an employee who has almost certainly never followed rival Inter Group's post-ed seminar, is re-working the garbage-out text of a particularly bad MT run.

Ironically, she has found that resorting to good old dictaphone methods is the best way to get quality revision. So she shuts herself up in a booth, speaks the corrected text into the mike, and then gets an entry operator to wordprocess it later. Very 70s . . . .

Like everything else, machine translation bounds two steps forward with resounding hightech élan and then skids one step back as human beings catch up with the snags. Even in Japan.