[*Language Industry Monitor* 17, September-October 1993]

Focus on Japan

Why there might be a Japanese MT system in your future.

If there was proportionally the same level of MT activity in America as there is in Japan, a country of 130 million people, every major hardware manufacturer and a handful of adven-turous system houses would have an in-house MT development project. There would be at least twenty commercial workstation-based systems on the market, with another dozen or so lurking in the wings. The major players would be co-funding the development of an ambitious lexical resource for an important language pair. And a variety of organizations, such as the Library of Congress and one or two major broadcasting networks, would have developed their own in-house systems. While the many obvious differences between these two countries preclude pursuing this admittedly facile comparison any further, this exercise should nonetheless provide a crude indication of the level of interest in MT in Japan.

There is probably no single country in the world with the need for translation—and, by association, machine translation—that Japan has, except perhaps Canada. It comes as no surprise, therefore, to discover that half of the world's MT research is found on that densely populated archipelago. This country's insatiable appetite for information, the commercial imperatives of its export activities, and its population's comparative lack of foreign language skills drive the Japanese to build and experiment with MT systems relentlessly.

The Japanese MT experience didn't materialize out of thin air of course. Realizing that no one else was going to do it for them, the Japanese have been grappling with the sticky matter of mating their exceedingly complex language with computers for decades. The introduction of the first *wapuro* by Toshiba in 1978, with its sophisticated Kana-kanji conversion routines, has had a profound effect on Japanese culture, giving Kanji a new lease on life and producing a generation of *wapurobaka* (literally "wordprocessor idiots"), young office workers who are suspected of barely being able to write Japanese by hand. nec, Fujitsu, and others quickly followed Toshiba into the wapuro arena. It is now a ¥300 billion a year business; the market leader is currently Sharp, with a market share of twenty-five percent. With the wapuro, Japanese writing accoutrements vaulted in one leap from the pre-industrial age to the digital era. Few Japanese ever mastered the Japanese typewriter, an ungainly behemoth with two thousand keys and seven shift states.

As if that wasn't enough, the Japanese have also been singularly successful in exploiting beyond their borders the computers and peripherals they designed to handle their language. Having had the need for bit-mapped, high output resolution output technology, Japan unleashed twenty-four pin dot matrix printers on an unsuspecting world. In its quest to build a smaller wapuro, Toshiba more or less invented the laptop computer. And of course the fax was also a response to the language processing challenge. It likewise had a revolutionary impact on Japanese business (the Japanese telex system never caught on) while spreading like wildfire throughout the rest of the world. Kana-kanji conversion is a non-trivial task (current systems get it right about ninety percent of the time) and calls for morphological, syntactic, and even semantic analysis to ascertain the correct kanji character represented by a given string of kana characters. Japanese companies therefore acquired experience in this new field simply out of necessity. But this necessity has stood them well; these pragmatic origins have proven fertile ground for subsequent experimentation with more advanced language processing technologies like MT.

This pragmatism may also go part way to explaining a more utilitarian, even opportunistic approach to MT than is found in the West. No doubt more than a few companies cherish the desire to replicate the smashing success of the wapuro. Unfortunately, many Western researchers disdain Japanese MT because of this utilitarian approach. For them, the only "good" MT research being done in the world is in the us, where it is protected and sanctified by the guardians of theoretical purity. That Japanese MT by definition lacks firm theoretical foundations is nonsense. Rather, Japanese MT developers are not dogmatic, something that cannot always be said of their American counterparts. As Kyoto University's Makoto Nagao pointed out at the previous Summit in Washington, "[Japanese] manufacturers know well that a single linguistic theory cannot lead to a good MT system. They realize that a huge amount of language phenomena must be processed in an ad hoc manner. " CMU's Masaru Tomita once put it another way: "it is fun to design an MT system, but very hard to develop it into a fully operational environment. " Today, a system like ASTRANSAC proves that it isn't where you start that matters but how far you get. ASTRANSAC is based on the ATN formalism, something out of vogue in academic circles, but years of tinkering, improvising, and a hundred thousand lexical rules later it is a working, commercial system. Likewise, the strength of the mainframe jicst system is probably the 500,000 technical terms in the system's lexicon.

Four or five years ago, you might have looked around Japan, seen the systems being offered, acknowledged their existence, but wondered: where, though, are the users? There was indeed a time when Japanese MT systems were "bundled with the mainframe. " acquired for reasons of corporate prestige, or given away as presents, like baskets of ¥10,000 melons. Meeting a genuine Japanese MT user was once as rare as encountering a redhead on the Tokyo metro. Have times changed? Consider these examples: Nikkei Printing uses NEC's PIVOT and Sharp's DUET to translate 18,000 pages of computer manuals into Japanese per year. The Japan Information Center of Science and Technology (JICST) translates 15,000 abstracts and 70,000 citations per year with the Japanese-English system it developed in-house. And Mazda churns out 1200 pages of English automotive service manuals a month using ATLAS. At the other end of the spectrum, four hundred users each paid ¥50,000 within the past year to upgrade to version five of Bravice J/E, a PC-based MT system. The bottom line is that Japan is quietly acquiring formidable experience *using* MT. That boils down to exploring the constraints and compromises required to be able to exploit this generation of MT systems in areas where human translators are currently deployed and using creative thinking to discover new domains where materials are not being translated at the moment.

But don't be mistaken. Japanese MT developers are not getting rich; MT is still a long way from the mainstream of computing in Japan and in that regard does not differ from the situation in North America or Europe. Some suppliers are at the point where they are funding their ongoing development work from revenues generated, but they are all a long way from recouping their huge initial invest- ment. "Are you kidding?", exclaimed Fujitsu's Michael Beirne when asked whether his company had recovered its R&D investment in MT with ATLAS. "No way. But our company is in the telecoms business. We know that if we don't stay with MT inevitably somebody else will do it. " As Nagao pointed out at the last Summit in Washington, "manufacturers have already invested huge amounts of money and manpower and therefore cannot withdraw from MT easily. A manufacturer cannot stop their R&D efforts unless other competing companies do so at the same time. Dropping out of this competition spells defeat in the future big markets of the information society. " Unless you suspect some miti-induced conspiracy or fit of collective madness the people in Japanese companies who are paid to think ahead *do* appear to be thinking ahead. They must see the logical outcome of Japan's economic evolution, as this manufacturing giant moves from the industrial into the information age in which communications play a vital role.

Ironically, as meticulous as the Japanese are in so many aspects of life, they appear remarkably tolerant about bad translations. Commented one Western observer, "in a country were very, very few people have a command of English, Japanese companies are desperate when it comes to export documentation and will put up with anything, as long as it looks like English. " But the problem runs deeper. There is no technical writing tradition in Japan. Technical writing is not taught in schools or universities, possibly because students at that level are still mastering the complexities of the general language. It is not uncommon for Japanese companies to teach young recruits how to write.

Moreover, the Japanese use ellipsis frequently, resulting in sentences without subjects or verbs. The implications for MT are obvious: *lots* of pre-editing. Nearly all MT suppliers try to impress upon the customers the sizable return they get by thinking in terms of MT-friendly texts. No deep secrets here. "Writing shorter and clearer sentences. " says

Fujitsu’s Kenji Sugiyama, “and ensuring that sentences have both a subject and main verb improve the quality of the MT output. ” He adds that writing with MT in mind has had some welcome side- effects: “the Japanese manuals have gotten better too. ” something that must have Japanese language purists wringing their hands in despair. Relates Toshiba’s Amano, “We used to say, ’Japanese hardware is great, the software is good, but the documentation is bad.’” But this is changing, he hastens to add.

Not all Japanese MT systems have been created equal. While many of the systems have been developed by protégés of Nagao, the doyen of Japanese MT, the systems in practice do vary, from the nearly direct Pensée system of Oki to sophisticated, semantically-rich systems like Toshiba’s ASTRANSAC and Fujitsu’s ATLAS. Other systems boast a wholly different lineage, notably newcomer LogoVista. Based on the linguistic theories of Harvard University professor Susumo Kuno, LogoVista is being developed by Language Engineering Corp. of Belmont, Mass. (USA), with the support of a Japanese consortium which includes Catena-resource, the developer of STAR.

Today, there may not be one Japanese MT system that stands head and shoulders above the rest, but many of them do have noteworthy features and characteristics. While some originated on the mainframe, nearly all now run on workstations, albeit in some cases proprietary systems. Building (or downscaling) MT systems for workstations has brought with it some important gains; it also makes certain demands. The gains include better integration with existing document production software and, correspondingly, with the entire document production environment. Packages like ASTRANSAC and Argo are now designed to work with industry standard publishing packages such as Interleaf and FrameMaker, important issues which can even outweigh linguistic matters. However, this greater accessibility makes correspondingly larger demands on the developer in terms of the system’s front end the user interface. Japanese MT developers realize this and have directed considerable effort in this direction in recent years. The Matsushita system, for example, which has not been commercially released, is an machine- *aided* translation system proper, whose strong point is its manual translating facilities and sophisticated online bilingual dictionaries. Most of the other Japanese MT systems, including Hitachi’s HICATS, NEC’s PIVOT, and Catena’s STAR, offer at the very least rudimentary bilingual editing facilities.

While initial prognostications (and over-eager sales pitches) may have led potential users to hope otherwise, the MT systems currently available in Japan are primarily (but not exclusively) suited to technical documentation; in that respect, the situation is again no different here than in North America or Europe. The suppliers are now quite candid about this, making explicit for which the domains their systems are suited. There are, after all, a lot of instruction and service manuals in the world. To help new users get up to speed, almost all of the suppliers offer supplementary lexicons for their systems, containing terminology for various technical domains.

However, whether MT will ever become more than an extension of the manufacturing process is not at all certain. A frequent complaint of MT systems in general is that MT output is grammatical and correct but stated in a way that no human translator would express it. A very promising new technique which is receiving a lot of interest in Japan is Example-Based Machine Translation (EBMT). Very simply, it is based on the notion of using a corpus of bilingual phrases and sentences, together with a thesaurus system for substituting words, to generate translations. While it is unlikely that a system could be built solely using EBMT techniques, EBMT could be employed for both extending the coverage of MT systems into new domains and producing more natural sounding output.

In the meanwhile, maybe we should simply get used to Japanese-style English: no articles, few determiners, and everything in the singular. Says Toshiba’s Shin-ya Amano, “since the Meiji Restoration, Japanese people have grown accustomed to English-sounding Japanese. As a result, we have a higher tolerance for less than perfect MT output. ” Is it now our turn?