[*Language Industry Monitor* 15, May-June 1993]

TM/2: Tip of the Iceberg?

TM/2 heralds the long-awaited arrival of IBM's formidable language processing technologies.

By now, most people in the translation and linguistic software world will have heard of Translation Manager/2, IBM's entry into the translation software market. In a nutshell, this OS/2 package offers a text editor, a translation memory facility, and tools for managing monolingual and bilingual lexicons. Regardless of its particular merits or shortcomings, TM/2 is a milestone product. Why? With its MT-aware design, TM/2 goes further than any translation package hitherto available in reflecting a fundamental shift in translation technology, best embodied in the concept of the"translator's workbench." The batch processing approach of yesterday is slowly being superceded by the interactive paradigm of today, with the previously distinct line between machine translation and machine-*aided* translation slowly being blurred.

There is probably no one in IBM who is more eager to talk about TM/2 than Edward Lippmann, currently based at IBM's German Software Development Lab (GSDL) in Böblingen, on the outskirts of Stuttgart. For TM/2 is in many ways Ed Lippmann's baby. Lippmann, whose career at IBM spans several decades, has played a major role in turning IBM's inhouse translation system into a product. Lippmann seems like the right person for the job, for this naturalized American, born in Germany, clearly feels equally at ease in both cultures, and he has been thinking about translation technology for many years. Two papers of his on the subject of translator's aids, which date from the early 1970s when he was at the Watson Labs, present seminal concepts later expanded upon by Martin Kay, Alan Melby, and others. While the translator's workbench concept has been around for many years, it is clearly an idea whose time has come.

"TM/2 is not marketed as an end product," explains Lippmann. Rather, he says that IBM currently positions it as a "host-like" package, and it is priced, packaged, and marketed accordingly, with a sixty-day trial period and free updates. Lippmann points out that TM/2 incorporates "nontrivial" linguistic capabilities, namely sentence boundary recognition, morphological reduction, and spellchecking for the nineteen source languages the package currently supports. (For target languages, TM/2 can also handle Chinese, Japanese, Korean, or other OS/2 compatible character code sets.) TM/2 originated at IBM's European Language Services (ELS) in Copenhagen, where translation and localization services are coordinated within IBM worldwide. The system grew out of IBM's experiences developing a version of the alps tss software for IBM mainframes in the early eighties. IBM ELS's Flemming Svanholm spent a year in Provo in pursuit of this endeavour; he later contributed to the design of TM/2, which was originally intended solely for inhouse use at IBM.

The development of TM/2 provides some fascinating insights into the research-development-product trajectory within IBM (and we hope that TM/2 isn't the last of IBM's formidable language processing technologies to survive this long and arduous route). In most if not all of its large national markets, IBM has established "science centers," partly out of corporate largesse. While these centers have historically been free to pursue a wide variety of research topics, one obvious activity has been to address language- specific computing problems associated with their respective national languages. In Germany, IBM has software development labs in Hannover and Berlin, in addition to the one in Büblingen, the largest. It also has a development lab in Vienna, where the German version of SpeechServer, IBM's speech recognition system, was developed. IBM's German science center is in Heidelberg.

Algorithms for such things as sentence segmentation and morphological reduction typically emanate from these scientific centers and are turned into standard programming libraries by IBM's NLP at its software development lab in Cary, North Carolina. The Cary lab

“repackages” and standardizes these routines, converting them, for example, from Lisp to C, thereby making it possible for them to be incorporated within any application being developed at IBM. “If someone needs sorting for Hebrew in a database application,” says Lippmann, “they don’t have to write it themselves. Someone else in IBM has probably already done it.” The translation memory component of TM/2, for example, came from the IBM science center in Haifa, Israel.

The NLP group has been primarily responsible for two distinct NLP systems, dictionary technology and the plnlp system (with its showcase Critique application); the Cary lab inherited the latter from the Research Division. The NLP group has been refining its dictionary technology for the past twelve years, and it offers functions ranging from spellchecking and hyphenation to thesaurus lookup, lemmatization and generation of full inflectional paradigms. These last two functions are used by TM/2 in its translation memory. Other IBM products use these functions in information retrieval and handwriting recognition applications, among other things.

Thanks in part to the efforts of the national scientific centers, the NLP group is able to offer spellchecking for such diverse languages as Icelandic, Afrikaans, and Greek. As the NLP group’s Ian Hersey explains, “in the late 80s, we had full-time linguists at science centers in every country working on these dictionaries and doing lots of testing. The results have paid off.” He says that what sets the NLP group’s technology apart from others available on the oem market is a combination of compaction, speed, accuracy and language coverage. “Our competitive analysis has shown that we do a better job on the languages that we have in common with the oem vendors,” claims Hersey. He adds that they have acquired or compiled themselves synonym data for all of the languages they support, except Icelandic.

Up until recently, however, this technology has been “captive” within IBM. As long as Big Blue had its own wordprocessors, the NLP group couldn’t dream of licencing it to third parties, because it was regarded as a competitive advantage. However, products which were to incorporate this technology never saw the light of day. But that is changing now. TM/2 is the first of what will hopefully be more packages exploiting IBM’s extensive NLP technology. Moreover, the technology groups in IBM are now being encouraged to enter the oem market. “We will be seeing more and more of this great technology that has been hidden for so long,” prophesies Hersey.

IBM decided to develop translation support software, taking advantage in the process of these substantial inhouse resources, for one very concrete reason: to satisfy its own huge translation requirements. As Ed Lippmann points out, “If you look at IBM’s software business separately, we are the largest software company in the world, with sales of US$7 billion a year. That means a lot of localization.” To put things into perspective, Lippmann says that IBM spent US$100 million on the localization of its products in 1990. In looking at ways to make the localization process more efficient, IBM has three priorities: the highest is making the production cycle faster, the second is improving quality through consistency, and the third is saving money.

Why did IBM decide to turn the package into a product? Lippmann: “Our customers kept asking us, ‘IBM, how do you deal with your multilingual documentation?’ Some were disappointed that we hadn’t automated it fully,” says Lippmann with bemusement, “we had to break them the bad news that that wasn’t possible yet.” Before being formally introduced, TM/2 enjoyed a protracted testing period at customers in the chemical, banking, and automotive sectors, together with an online news service and, of course, ELS. As Lippmann points out, “we can’t very well test an air traffic control system. But we can test a translation system. That makes a big difference.” Svanholm adds that TM/2 is one of those happy cases where IBM can live up to that familiar maxim, “Use what you sell and sell what you use.”

As you might expect, before IBM jumps into a given arena, it does its homework. IBM estimates that the global translation market amounts to some US$30 billion a year and enjoys an annual growth rate of fifteen percent. It knows that there are 21,500 subsidiaries of foreign companies in the US, that there are 17,000 US companies operating internationally, and that there are, for example, 1100 US companies active in Korea.

With a market like this under its nose, you might think IBM would go the whole route and offer an MT system. According to Siegfried Ester, senior development manager at GSDL's Office Applications group, IBM had indeed embarked on a parallel effort to launch an MT system alongside TM/2, and there was much discussion about it internally. "We could have offered the engine first, like, for example, Siemens," explains Ester. "But we realized that was not the right way for us to go about it. We thought we had better concentrate on the end-user first and fully understand what their requirements are. Our strategy has therefore been to offer an end-user platform first, then build on that." TM/2 has been designed with MT in mind; it offers a built-in interface to an external translation system. That could be Metal or Logos; it could also be an IBM MT system. Lippmann stresses that with TM/2 the user is at the center of the application, that he or she remains in control. "We're not selling artificial anything," he says reassuringly.

MT at IBM

Over the years, IBM has been home to a variety of experimental MT systems, including one, tomcats (based on the plnlp grammar of the former Heidorn group), which is currently being used in-house by IBM Japan to translate manuals from English into Japanese. At the moment, however, all hands appear to be on deck on behalf of one system in particular, Logic-Based Machine Translation (LMT), which originated in the work of Michael McCord at the Thomas J. Watson Laboratory in Yorktown Heights, New York. Currently, the language pair English-German (both ways) is most advanced in development, followed by English- Spanish, English-Danish, and English-French. A number of other language pairs are also being worked on at various sites, including Hebrew combinations.

LMT currently runs on IBM mainframes and RS/6000 Unix systems. It's a transfer-based system but has a language- independent kernel and is written in Prolog, meaning that developing a new language pair for LMT is less arduous than starting from scratch. The declarative programming style of Prolog closely resembles the logic-based grammar formalisms in vogue today, of which McCord's dependency-like Slot Grammar is one. LMT has a strong lexical orientation, with most of the linguistic information the system uses coded in the dictionaries rather than in the grammar rules. This shifts, for better or for worse, the burden of development from the grammar to the dictionary. LMT is not state-of-the-art machine translation; as Svanholm puts it, "it's pragmatic and open-ended, designed with the goal of showing some results quickly." LMT offers detailed linguistic coverage for several specific technical domains that partly reflect IBM's internal priorities, with further development being planned on a domain by domain basis.

Much of the work on the German LMT language pairs has been done at IBM Deutschland's Science Center in Heidelberg, and Heidelberg's Hubert Lehmann, who knows the system intimately, was demonstrating LMT at CeBIT last March. Lehmann was using TM/2 as a front end to LMT, which was running remotely on a mainframe in Heidelberg. LMT can be smoothly integrated within TM/2, functioning as one service among several at the disposal of the translator. With LMT active, TM/2 passes sentences to it which it does not find in its translation memory, its bilingual database of parallel, previously translated sentences. TM/2 then returns the sentence translated by LMT with an <m> marker in the left margin, indicating it was translated by the machine. The translation memory will retain these translations, but crucially, in TM/2's so-called post-editing mode, the changes the user makes are automatically updated in the translation memory. As such, LMT can be considered a useful way to build translation memory files semi-automatically. As used internally, IBM configures TM/2's translation memory to retain both the original and user modified sentences, thereby providing developers with feedback concerning the system's performance which they can use to tune the grammar rules. The big question remains: What are IBM's plans with LMT? The general answer seems to be: We haven't decided yet.

Tm/2 reflects the current trend epitomized by the translator's workbench concept; it also reflects the specific needs of ELS. At ELS, TM/2 is just one tool among many in IBM's huge application development framework, where it is used in conjunction with product, file, and project management tools to handle the localization of IBM's vast product line. ELS

routes requests from IBM development labs for localization services to the national language services in each of the respective countries. These in turn work with the local IBM sales staff, who, being responsible for selling the product, need to stand behind the localization job. According to Svanholm, all translation within IBM is done in the target countries in conjunction with local IBM offices; ELS handles all "non-translation" chores, i.e., updating and recompiling source files, testing new versions, and multilingual document production. To streamline the localization process, ELS assists IBM's development labs in ensuring that their programs are as linguistically flexible as possible. ELS also maintains IBM's multilingual terminology database software and a hotline for translators and ibmers. "ELS offers the same kinds of services to external companies that it supplies to internal development labs," points out Svanholm, "including translation planning, preparation, project management, quality assurance, testing, manufacturing, and distribution of localized products."

Meanwhile, GSDL continues to enhance TM/2. The most recent release includes a utility for generating translation memory files from existing collections of bilingual texts. According to Lippmann, IBM customers clamored for such a tool to be able to exploit previously translated materials and therefore get up to speed with TM/2 quickly. While it sounds very promising, the effectiveness of the translation memory generator will largely depend on the correspondency of the parallel texts. Automatic alignment of bilingual pairs of sentences is a tricky business. AT&T Bell Labs' Ken Church, author of some influential research on bilingual sentence alignment, discovered while working with AT&T's translation company, Language Line, that his well-known alignment algorithms simply weren't robust enough for the kinds of materials the average translation company has at hand. On a more mundane but in no way less vital level, GSDL has now extended TM/2's file format compatibility to include such industry standards as WordPerfect, Am¡ Pro, Interleaf, and Ventura.

On the other hand, there are certain things that IBM is not planning to develop. One of them is bilingual dictionaries. "IBM is not interested in competing with publishers," says Lippmann. Neither does the company see itself as the exclusive supplier of such consulting services as building termbanks and translation memories for TM/2 users, and it is actively looking for suitable business partners in the translation world.

The Heidelberg Science Center is also developing a tool, TransLexis, to build and maintain termbases and LMT dictionaries in an integrated fashion. With a relational database at its core, TransLexis will allow multiple users to add and modify terminological and lexical entries, and will ensure that termbases and LMT dictionaries are consistent. This addresses a serious problem faced by MT users in organizations where people use terminology management software alongside an MT system. TransLexis has a data exchange interface for communication with applications and an interface to LMT, thereby making centralization and integration a reality. The GSDL developers are tackling the inevitably complex task of entering lexical information with an example-based approach. For adding the morphological and syntactic information LMT entries require, TransLexis prompts the user with examples generated by LMT itself.

Will TM/2 be a success? As always, it depends on how this is measured. If the package serves IBM's internal needs—as it apparently does—then discussion is irrelevant. If IBM moves into translation services, TM/2 and LMT could provide IBM with a formidable competitive advantage. Is IBM getting into translation services? Says Flemming Svanholm, "right now, IBM is actively seeking and exploring new business opportunities, and translation may be one of them."

Whether TM/2 will be a market success as an independent entity remains to be seen. One unfortunate obstacle to this is the fact that it runs under OS/2. As much as it must disappoint IBM's surprisingly dynamic desktop systems group, Windows is irrefutably becoming the platform of choice on today's desktops—at least for the time being. If IBM wants to achieve any kind of market penetration with its translation software, it will have to consider some kind of Windows- based client package offering a subset of TM/2 functionality. Preferably, it should be modular in nature, whereby the program's services can be called upon from other word-processors—this approach is clearly the wave of the future.

Trados, for example, is moving in this direction, forsaking the dos version's dedicated editor for data interchange facilities in the upcoming Windows version of its Translator's Workbench package.

Another obstacle facing TM/2 might well come from within: IBM's own monolithic sales staff (Gerstner is said to be sharpening his knives). It is disconcerting to hear Lippmann say of a five thousand dollar package, "our host system salespeople aren't used to selling something this low priced." Welcome to Mainframe Country. IBM's natural language interface package LanguageAccess, launched in late 1990, was withdrawn from the market after several months, owing to lack of interest. Apparently, IBM sales staff didn't know what to do with it. Let's hope the same fate does not await TM/2 and LMT.

IBM Deutschland Entwicklung GmbH, Postfach 1380, D-7030 Böblingen, Germany; Fax +49 7031 166736