[*Language Indsutry Monitor* 19, January-February 1994]

The Spirit of Eurotra

Eurotra may have never replaced Systran, but its ghost lives on in Denmark, where a brand new MT system based partly on Eurotra work recently went into operation.

This past December saw the installation of a new English-to-Danish machine translation system at Lingtech, a translation company based in Copenhagen specializing in translating patents. The new system, called PaTrans, was developed by the Center for Sprogteknologi (CST) previously Eurotra Denmark. The deployment of PaTrans marks a new era for post-Eurotra MT in Europe and is a splendid achievement for the CST. Under the leadership of director Bente Maegaard, the CST, predominantly researchoriented in the past, has delivered a functional MT system to specification, on time, and within budget.

To understand the precise nature of this success, it should be noted exactly what PaTrans is. This English-to-Danish system has been designed for the domain of patents, initially petrochemical patents. It is decidedly not a general purpose system, and consequently lacks the kind of facilities and resources that a shrink-wrapped package would be expected to have; its users are also prepared to both pre-edit the input and post-edit the output of this Unix-based system. But the CST's customer is getting what it wants. Lingtech director Viggo Hansen says that PaTrans satisfies all of the design and performance criteria that were established at the beginning of the project. Previous to installation at Lingtech, the software ran for several months at the CST to iron out any remaining wrinkles.

When the system was delivered in December, there were simply no surprises. "You're a brave man," we remarked recently to Mr Hansen, who replied with a smile, "many people say that." Hansen's confidence has clearly been rewarded; Lingtech's MT gamble has paid off.

The inevitable question arises: How much of PaTrans is based on Eurotra? "It depends on who is asking," quips project manager Annelise Bech. Commission technology boosters would obviously love to hear that the system is a direct spinoff of the Eurotra program, while Eurotra insiders would shudder at the thought of production software based too closely on that unwieldy collective effort. In the final tally, the question is virtually impossible to answer; moreover, it is not terribly relevant. While you might be able to count lines of code or lexicon entries, how do you quantify the collective experiences of a group of researchers over ten year's time? Maybe Eurotra demonstrated how not to do certain things. The CST feat, then, is not that it somehow recycled Eurotra technology, thereby putting something back in the communal kitty, but that it has been able to capitalize on the human resources cultivated during the Eurotra years. And that, in hindsight, makes the Eurotra exercise seem at least partly worthwhile.

Inevitably, the CST has drawn on past work, benefitting, points out Bech, from the formal grammatical descriptions the group previously developed for Eurotra. But whereas Eurotra was by definition designed to be as general as possible to accommodate all potential language pairs, PaTrans was expressly designed and heavily optimized for just one language pair, going in only one direction. Among other things, that means parsing times can now be measured in seconds, not minutes, as was the case with Eurotra.

Speed is one consideration in a production system; robustness is another. PaTrans includes what Bech calls a *failsoft* strategy. This encompasses not only morphology-based heuristics for unknown words, but also a robust parser that does not plead "no parse," even when dealing with ungrammatical input. "The system does not require the user to pre-edit input texts to conform to a controlled language," says Bech. "There are no restrictions on the linguistic constructs

it will accept. However, its coverage definition does specify what the parser can deal with before it resorts to the failsoft strategy."

Additional robustness is provided by the PaTrans pre-parser, largely the work of the CST's Brad Music. Explains Music, "we can do some useful things in the pre-parsing stage, such as delineate sentence boundaries and lists and tag some constituent parts. We also identify some coordinate structures and prepositional phrases. The more you can trap at this stage, the better." The fewer the ambiguities, the less likelihood of the parser overspecifiying, and this means fewer parse trees generated by the system's chart parser. And that, in turn, means better performance.

Reflecting the nature of the input material, PaTrans is largely syntax-based, performing, in the words of Bech, "deep syntactic analysis, like the Eurotra model." However, she adds that the PaTrans lexicons do include a few attributes to facilitate some translation ambiguities, eg, there is a feature called *measure*. PaTrans also draws on semantic distinctions in the partitioning of lexical items in domain-specific term dictionaries, such as petrochemical, chemical, and mechanical engineering. "These partitions are physically distinct," explains Bech. "The items are actually labelled for domain." Accordingly, a PaTrans term, dictionary will be assigned a basic semantic category, e.g. composition (chemical), composition (legal), etc. While PaTrans's grammatical coverage may be limited to patent texts, control of its lexical coverage is squarely in the hands of the user.

Naturally, a lot more goes into a production MT system than just linguistics; you also need a document-handling component and end-user facilities, both of which are non-trivial development tasks. While the interface to the translation module was rather spartan when the CST was demonstrating the system in November, it will eventually be integrated into an "administrative" front-end. This will provide a menu,based interface for both the translation module and the pre- and post-editing facilities. It will also offer some document-handling tools, such as an archiving system for the patent texts. However, in terms of the user-interface, the highest priority has been the term- and dictionary-entry tool, called PaTerm. Jacketed in an elegant X/Windows interface, PaTerm offers two levels of support to end-users.

Level one is for the user who requires lots of guidance. He or she is taken through the process step by step, prompted a question at a time by on-screen examples. Level two is for the experienced user; there is only one screen and no examples. The user fills in the information in short form by clicking on boxes or using the keyboard. Level two also offers a template function for adding a batch of similar terms which differ only in lexical values. The coding tool is "intelligent," says Bech; it computes itself a number of values based on the input of the user. It is also very fast. The CST is quite proud of PaTerm; it designed and implemented the tool from scratch.

"It was absolutely essential that the PaTerm be easy to use," explains Viggo Hansen. "We wanted people with only a general knowledge of grammar to be able to use it." Lingtech had a person coding the PaTrans dictionary fulltime before the system was introduced at Lingtech; she spent several months coding terms in the chemical and petrochemical domains. PaTerm reflects the fact that not only linguistic expertise also software engineering skills were required for the system; Bech says that programming the X/Windows interface posed more difficulties at given moments than developing and implementing the linguistic engine.

Lingtech was established several years ago by two large Danish patent attorney firms, Lehmann & Ree and Hofman-Bang & Boutard, which found themselves spending so much time on translation that they decided to join forces and establish a separate translation company. As in most countries, patents are valid in Denmark only when a Danish-language translation is registered locally. While patent translation may be one aspect of the job, a patent attorney's primary role is to assist inventors and researchers obtain patent protection and other kinds of industrial property rights, such as trademark and design protection. Viggo Hansen was originally hired as a consultant by the two firms to ascertain the feasibility of automating the patent translation process. This resulted in the newly reorganized CST

performing a feasibility study in 1992 that led to the signing of a contract to develop a translation system.

Patents are legal documents and, as such, call for meticulous, literal translations. Is MT suitable for such an application? Hansen acknowledges that a PaTrans translation requires both pre-editing and careful reviewing by an experienced translator, but the effort is worth it, although the actual benefits in terms of increased efficiency and accuracy can only be measured after the system has been in use for some time. “In a patent, each word has one and only one meaning. And the terminology is fixed. Hansen adds that these are fairly long documents, four thousand words on average — although patents of American origin tend to be even longer. “It would not be worthwhile running a one to two page text through such a system,” he points out. The company already translates between one and two thousand patents a year, and the volume is growing, exceeding even projections supplied by the European Patent Office in Munich.

A stumbling block remains the fact that most of the English patent texts are supplied in hardcopy form; only one of Lingtech’s customers regularly supplies electronic files. This implies the extra step of scanning in a text, not always a faultless process. Hansen acknowledges the problem, saying, “we know that each of these texts must be stored on a diskette somewhere in the world. It is just a matter of tracking it down.” In the meantime, Lingtech provides an economic incentive — a discount — for patents supplied in electronic form.

Lingtech selected the petrochemical domain because it seemed a suitable subject for an initial foray into MT — and because its two parent companies do a lot of business in this field. But if everything goes well, petrochemicals could be just the start; Hansen suggests that the system could also be extended to cover pharmaceuticals, electronic equipment and other domains. The goal, says Hansen, is “full coverage” of all the technical areas in which the patent offices are active. Further down the line, other kinds of technical documentation for third parties could also be a possibility. Whatever direction the Lingtech takes, it faces coding lots and lots of terminology. Says Hansen, “PaTerm is the key to the future expansion of the system. That’s why we put so much effort into it.”

Looking back on the two and half year trajectory of the project, Hansen has nothing but praise for the CST. “It is a highly competent group.” he says “There were no mishaps.” But Hansen, a computer industry veteran, no doubt contributed substantially to the success of the endeavor himself by bringing to bear his previous experiences in large office automation projects. He notes that a thorough system analysis was performed before a line of code was written.

With the introduction of new MT systems a rare occurrence, PaTrans seems like a very promising paradigm for the present age. General-purpose MT is still an oxymoron. Systems like Systran, Logos, and Metal are largely used in narrow domains — and only after extensive lexicon work. Maybe it would be better to market MT systems in the form of applications like PaTrans, for at least then the substantial cost of “customization “ would be up front.

In any event, PaTrans is more evidence — if you still need it — that the most interesting work in the field of MT applications is taking place in Europe, CMU notwithstanding. But why Denmark in particular? MT competence and patents can be found in many European countries. PaTrans seems to be the result of a happy confluence of two factors: an experienced researcher who had the formidable skills needed to pilot her team through the transition from research to development; and a customer who could formulate exactly what he wanted.

Center for Sprogteknologi, Njalsgade 80, DK-2300 Copenhagen S, Denmark; Tel: +45 31 54 22 11, Fax: +4531546197

Lingtech, Vesterbrogade 24, DK-1620 Copenhagen V, Denmark; Tel: +45 33 25 71 71, Fax: +45 33 25 61 71