Esperanto interlingua, networks, parallel processors, CD-ROM, and AI – BSO's Distributed Language Translator aspires to be

LEADER OF THE PACK

by Peter Rutten / Illustration by René van Asselt

"Translating is not just using a dictionary and applying grammar rules. It's also knowing what things are about." The mystical ring to the word "knowing" lingers fleetingly in BSO's solemn, stately conference room in Utrecht. Toon Witkam sits at the white, U-shaped table. He pauses, smiles distantly as though searching for the right words, before slowly continuing.

"It's also a form of understanding, and it's this element of understanding that we want to put into the computer.

This is the challenge, the reason for all this multi-million research race going on all over the world." He should know. Toon Witkam (42) directs one of these research projects himself. Head of research at the Dutch software house, Bureau for Systems Development (BSO) in Utrecht, his department has been working since 1979 on just one project — the Distributed Language Translation program (DLT). His current staff includes 19 specialists, both linguists and computer scientists. His 15 million guilder budget (US$7.5 million) is funded by the Dutch government and by BSO itself.

BSO's goal, says Witkam, is to stay in the forefront of language technology. Because Toon Witkam knows too that the first company to cross the tape with a successful second generation machine translator will garner not just laurels, but riches from a world hungry for a solution to the language problem.

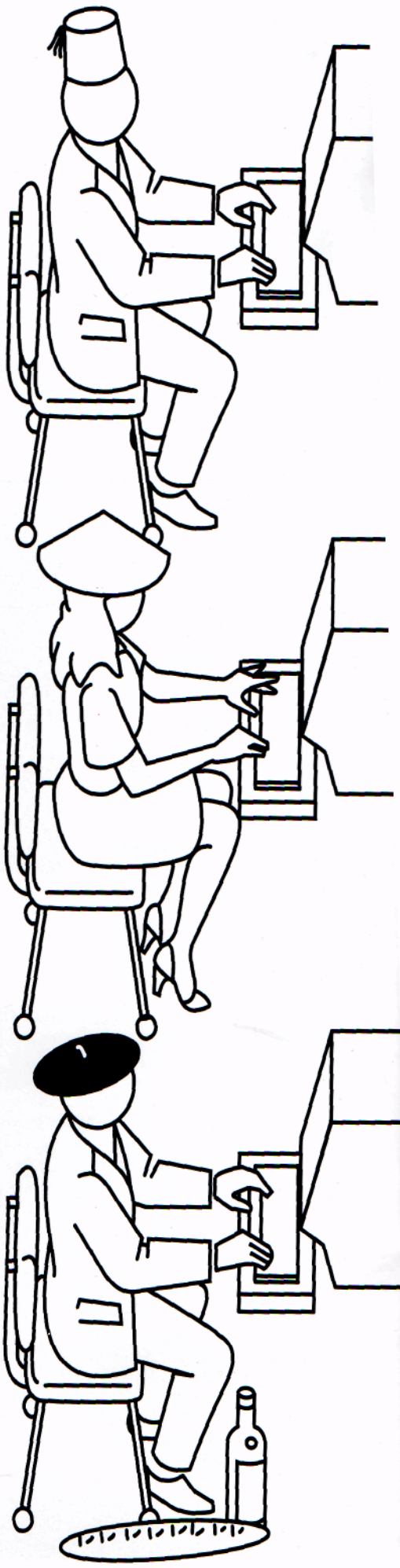
A HIGH TECH SOLUTION

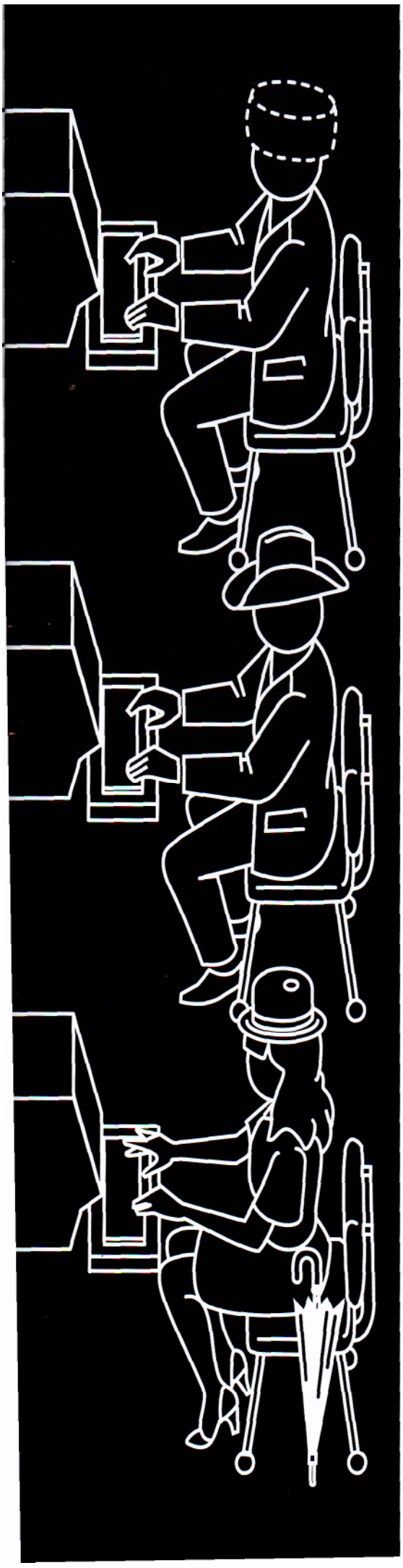
Real world knowledge— artificial intelligence —is only one way BSO intends to stay in the forefront of language technology. Other aspects of its design are equally radical.

For one thing, it is designed to run on the new parallel processing computers which are only now leaving the laboratories and entering commercial operation — and which are pegged to supply the heavy duty computing power of the next decade.

"In the 1990's, something as popular as a Macintosh," Witkam notes, "will be very powerful indeed."

Secondly, the system is a *distributed* language translator. That is, the system is a network of translators, either at one location or distributed over time and distance. Every translator at his/her workstation on the network can process source text into an interlingua semi-product, which can then be stored or used by any other node on the network to





generate target text. This contrasts starkly with the central batch processor paradigm of the old school of machine translator thinking.

"Each translator will have the program in the CD-ROM drive in their PC," Wilkam says matter-of-factly.

And DLT is an interactive system. Instead of using post-editing to clean up text processed by a batch system, DLT relies on translators to resolve the ambiguities the program can't handle, while the system is translating. And in the process, teach the system the real world knowledge it needs to progressively reduce ambiguities in the future.

ESPERANTO

As radical as these ideas are – this is the only MT system under development anywhere that incorporates them all – BSO's system has drawn most attention, and skepticism, from another radical departure: its use of Esperanto, the artificial language invented some 100 years ago by the Polish eye doctor Zamenhof, as its interlingua.

Witkam's inspiration for Esperanto came from his reading about the 1977 European conference, "Overcoming the Language Barrier," held in Luxemburg. At the conference, a group of Esperanto enthusiasts advocated computerized Esperanto as the way to eliminate language problems in Europe. Witkam was already aware of some British experiments using Esperanto as a sort of interlingua in international telex traffic.

Linguist Klaus Schubert (32) is head of the DLT project's grammar group. He was recruited by BSO because of his knowledge of Esperanto. Schubert says, "It's unusual to use a human language as an interlingua. It's supposed to be twice the work. But that's not true, because Esperanto makes a text less ambiguous. Other systems often use symbolic interlinguas. But symbols are based on human language anyway. You're better off using a real, self-sufficient language."

Witkam admits that a great many eyebrows have been raised at the use of Esperanto. "It's true there's a lot of prejudice. Esperanto has this kind of amateuristic, starry-eyed, beardy-weirdy image. It doesn't seem scientific. But as a linguist you can't be too scared of being laughed at."

He himself sees Esperanto as "a perfect compromise to meet design demands." He points to its use of existing roots from Romance, Germanic and Slavonic languages and its legibility, ease of pronunciation and widespread use. He plans a public demonstration of the working parts of DLT this year, partly, he admits, "to remove the last vestiges of doubt as to whether Esperanto is a suitable central interlingua."

NUTS AND BOLTS

Klaus Schubert speaks perfect Dutch with a slight German accent, giving him the aura of a professor in an old American film. He says, "The networks we're aiming for will process information rapidly. Typing everything in first and then translating it is a waste of time. DLT begins immediately with the first word." As the typist types, the growing sentence is syntactically analyzed by the part of the program called the parser.

Witkam squeezes himself into a huddle of researchers working at three screens which have been pushed together. He types in English, "Make sure that the bolts are removed." On the lefthand screen a histogram appears with branches, each indicating a point of analysis. On the middle screen is a tree diagram depicting the same analysis in words, and their grammatical relations. The words hang like ripe apples from the branches.

"Ah," muses Witkam, "it's responding with two interpretations." In fact, "removed" was being interpreted both as predicative adjective and a passive past participle: the English could either mean "Make sure the bolts have already been removed" or "Make sure you remove the bolts." Meanwhile, the "Metataxor," the actual translation component, has applied various rules and translated both interpretations. On the righthand screen are two tree-diagrams — the parsing for English and Esperanto.

DLT often supplies more than one correct translation, only one of which will be appropriate in a given context. This is why not only parsing, but semantic knowledge is necessary too. The system has to be able to fathom several possible meanings and pick out the most meaningful.

"It has to understand the sentence, which is why we use AI and a knowledge bank," says Schubert, who immediately qualifies what he means by "understand." "Human translators don't always understand everything either, but they always understand some of what a sentence is about. This system tries to include some of that meaning."

REAL WORLD KNOWLEDGE

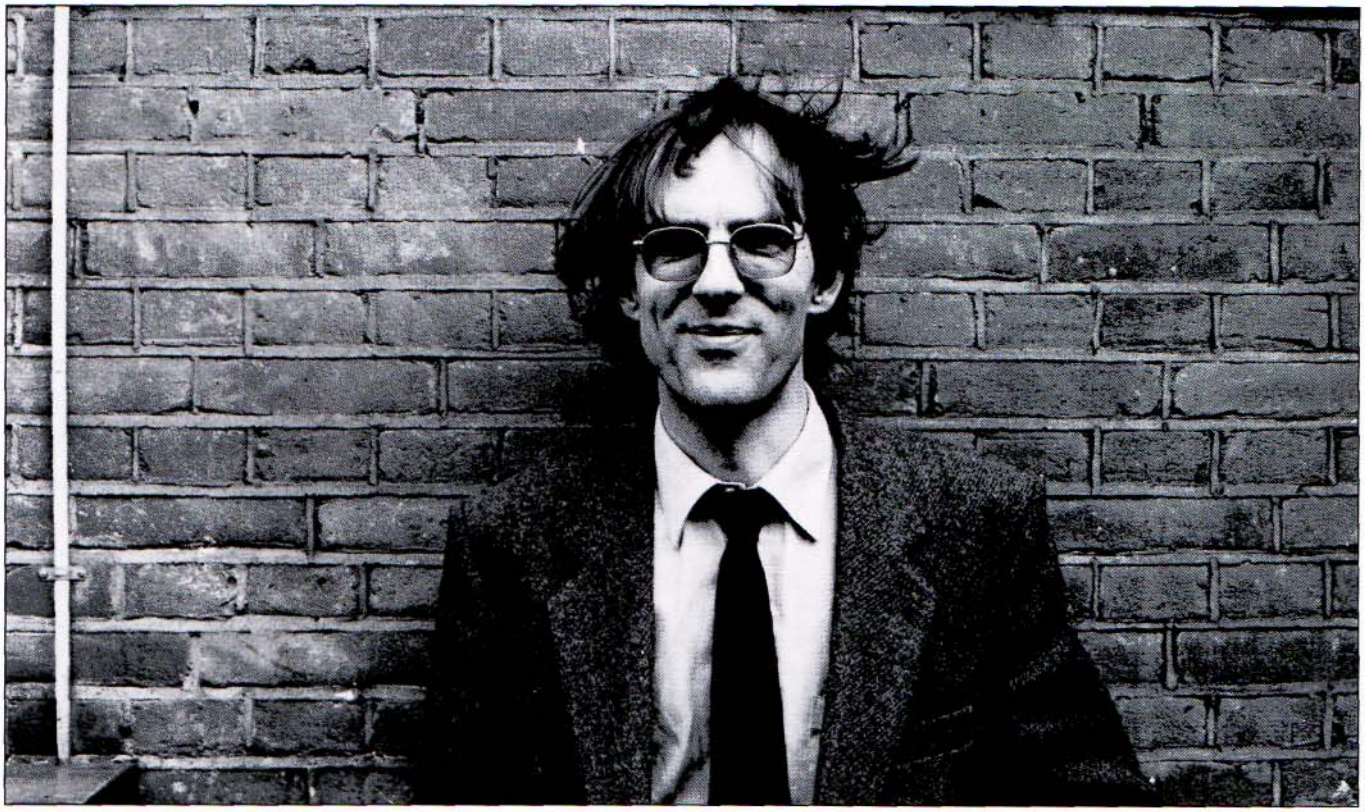
The "understanding" component of the DLT system consists of an expert system called SWESIL, the "Semantic Word Expert System for the Intermediate Language." It's the component with "real world knowledge." SWESIL splits the Esperanto tree up into word pairs and compares them with word pairs in a knowledge bank.

"This knowledge bank is actually a large one-language dictionary," according to Schubert. "The meanings of interlingua words are explained in the interlingua. The knowledge bank contains contexts in which words are typically found."

SWESIL compares words and calculates a score based on probability. Artificial Intelligence specialist Bert Kessels says, "The farther apart two words are from one another the lower the score, and vice versa. 'Tree' and 'forest' get a high score, for example. So do 'tree' and 'man' because they're both living organisms."

The idea is that the knowledge bank will eventually fill up with this sort of information. In the future, the DLT will also learn from daily experience. In the last phase of translating into the interlingua, the computer submits any problems to the operator, and over the years builds up knowledge by this means of question and answer. This is DLT's spectacular learning aspect.

In the future, it will be able to study independently to gain further knowledge of the world. "Day in, day out, selected 'books' will be put through the computer's laser scanner," says Witkam, "and DLT will automatically compile its own body of knowledge."



Fabiana van de Paverd

In a couple of decades, the best-read translator will be a translation machine.

BSO doesn't intend to wait a couple of decades, of course. Once the system has achieved a 20% "intuitive capacity" for translation, BSO intends to put it on the market.

Bunkum and the Money Chase

BSO began work on DLT in 1980. Two years later, pleas to the Dutch government for financial assistance fell on deaf ears.

"Artificial intelligence? Automatic translation? They called it bunkum," says Witkam.

A year later, however, the European Community came across with 250,000 guilders ($125,000) for a design study. "The result was a nice thick book and a lot of interest," says Witkam.

In 1984, with the help of this study and four independent advisers – including Brigham Young's Alan Melby – The Hague was finally convinced and shelled out seven million guilders ($3.5 million). "We told them that in six years we could have an English-Esperanto-French translation machine ready for a trial run," Witkam points out.

In 1984, well into the DLT program, BSO began looking around for new partners to shoulder the burden. It cast hopeful glances at nearby Philips, trying to interest the electronics giant in helping to launch DLT in 1992.

Witkam explains: "BSO invents things, but we have no marketing apparatus. For that we need to go in with a computer manufacturer, electronics giant or large publisher, any concern that can steer the system in the most promising market direction. Without a large company that can add translation as an extra service to its network, we don't have a chance."

But even though Philips seemed the most obvious choice, they are developing their own translation system *[Rosetta – Ed.]* and so rejected BSO's offer.

BSO is still looking for a partner. Witkam will not name names but admits that it might be one of a number of large Dutch publishers.

Toon Witkam: "There's a lot of prejudice. Esperanto has this kind of amateuristic, starry-eyed, beardy-weirdy image. It doesn't seem scientific. But as a linguist you can't be too scared of being laughed at."

Milestone

Early this year, the DLT achieved its most important development milestone so far: it passed the "Melby test," created by and named after the American linguist, Prof. Alan Melby. The test is designed to measure how well a machine translator can determine the right meaning in the right context. In the Melby test, a 480-word French-English DLT translation was compared with the work of human translators. For one-third of the English words the computer chose the correct French equivalents.

"We did well," says Witkam laconically, "though there were flaws, all of which were just points of detail. For example, we still had no provision for idiomatic expressions or for expressions such as 'to take into account.'"

Nevertheless, in January, BSO announced that it had achieved a major advance over its competitors. Witkam is now prepared to qualify this claim:

"The Japanese have much more power, thanks to their eight-company joint venture with the Japanese government. We have less money and less manpower, but we do have a good interlingua! We're excited about who'll get there first."

Witkam says BSO is on target to get DLT out there by 1993. ●

Peter Rutten is a journalist and technical writer.