It's been called the most elegant machine translator currently in development. But can it solve the ambiguities of language without artificial intelligence and real world knowledge?

PHILIP'S ROSETTA M
A Question

by Peter Rutten

"I am not at all optimistic about fully automatic machine translation. It'll certainly be the 21st century before we are doing an acceptable job on natural language that's actually interesting to read."

Sobering words – but not cheap ones. Jan Landsbergen's considered judgement is grounded in seven years of solid research. He it is who laid the foundation of one of the most promising Machine Translation systems now being put together: Philips' Rosetta Machine Translation Project. What he is optimistic about is that semantics and not artificial intelligence will be at the core of "doing an acceptable job on natural language," as he puts it. Indeed, running through the whole Rosetta project is a scarcely concealed challenge to its scientific rivals: "Look out AI believers – the linguists are striking back!"

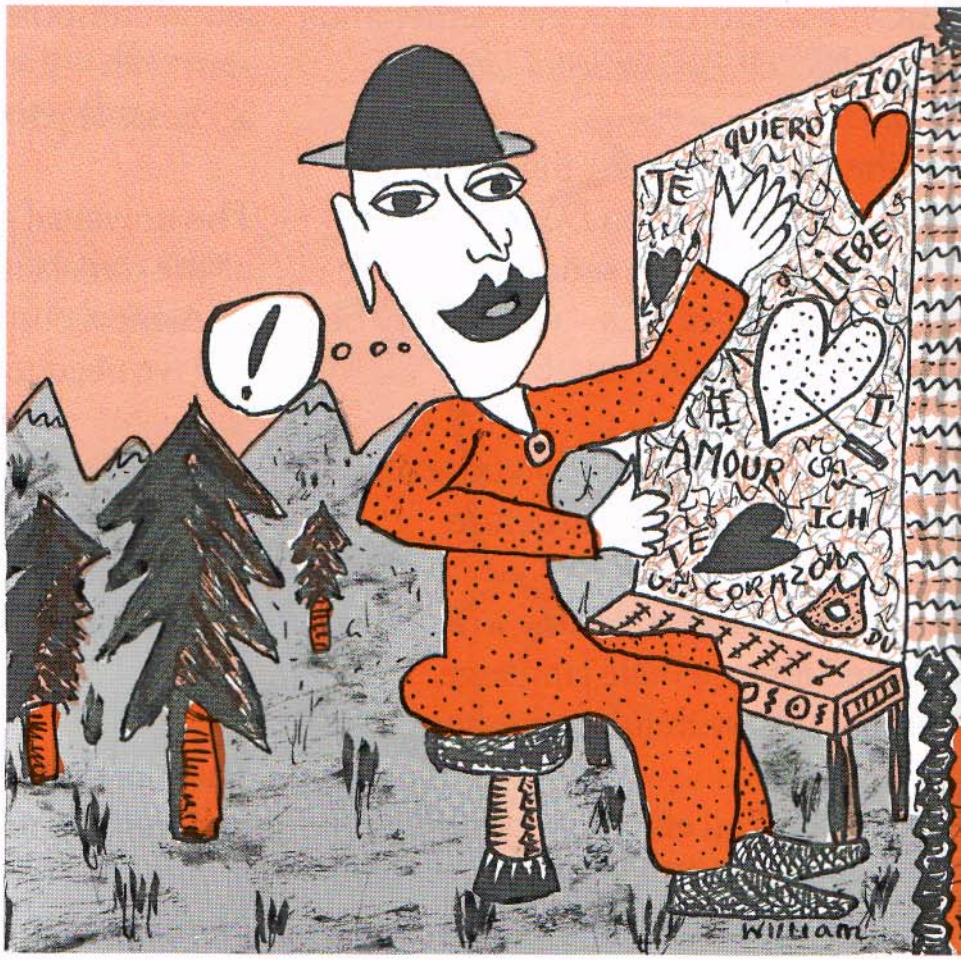
For although we are not on the eve of an outburst of rivalry in the peaceful and academic world of machine translation, the theoretical schism cannot be denied. Especially in the little, pioneering country of the Dutch, with two major Machine Translation projects – BSO's DLT and Philips' Rosetta – not 80 kilometers apart, both gaining momentum, and together lapping up megaguilders in research.

Landsbergen (46) does not want to talk about it. There's been some tension between Philips and BSO in the past, arising from the smaller company's request to the multinational to become its partner. The reason Philips said no turned out to be that they themselves had plans to "do something in the language business," is all he will reveal.

And whereas BSO's researchers have a strong belief that artificial intelligence, in the form of real world knowledge, holds the key to solving machine translation's most exasperating stumbling block – "semantics," or translating real meaning – the Philips team chose to go down the linguistic road. Landsbergen and his workers – four computer scientists and seven linguists – have even designed what they claim to be the first ever translation-focused grammar, called M-grammar, an uncompromising attempt to tackle the problems of Machine Translation linguistically.

Amazingly, before Landsbergen started creating it, there was no formal grammar for translation. This still puzzles the project leader. "We were – and still are – doing things linguists apparently never felt like doing. A few years ago I was surprised at the things we had to do for the first time, but not anymore."

The Rosetta approach makes demands on various underexplored – and sometimes unpopular – fields of linguistic research, with the side-effect that Philips has also ended up contributing large funds for the furtherance of pure linguistics. As if the electronics giant, bored with lightbulbs, had turned its attention to words.



Nuts and Bolts

In Rosetta, most of the work is done in Analysis and Generation, which each contain four parallel modules, connected through interfaces. These modules are the nuts and bolts of the translation process. Transfer is a relatively short and simple phase between Analysis and Generation.

To illustrate how Rosetta works, let's type a simple sentence in English and translate it to Dutch. The sentence: "All bishops like her." In Dutch: Zij bevalt alle bisschoppen. There are a number of not-immediately apparent ambiguities that will have to be dealt with.

Translation begins with Analysis. Step One: the Morphological Component. This module's dictionary contains source language words and morphological rules. It breaks down the English sentence by assigning all possible lexical functions to all the words (noun, verb, conjunction, etc.), and all possible morphological functions to all the bound morphemes (singular, third person, etc.). For instance, the word 'bishops' is represented as:

NOUN: stem: bishop
number: plural

The word "like" in this example can be either a verb or a conjunction, an ambiguity that will be solved in Step Two: the Surface Syntax Component. This component strings the identified, loose words according to the English surface grammar, and shows that a string

ACHINE TRANSLATOR
of Semantics

Illustration by William Wilson

The Image of the Ever-Living King

The last two lines on the Rosetta Stone, which Philips' Machine Translation Project Rosetta has been named after, run as follows: "This decree shall be inscribed on a stela of hard stone in sacred and native and Greek characters and set up in each of the first, second and third temples beside the image of the ever-living king".

The Rosetta Stone is a slab of compact black basalt found in July 1799 by French soldiers in an Egyptian town called Rosetta. The stone shows three distinct inscriptions that date from the time when King Ptolemy ruled over Egypt. The inscriptions are each versions of a single text in different scripts: hieroglyphs (the sacred characters), Demotic (the native characters) and Greek. The stone enabled a frenchman called Jean-Francois Champollion (1790-1832) to decipher Hieroglyphs for the first time.

The Stone is exhibited in the British Museum in London.

where "like" is a conjunction does not produce a sentence. It also shows that there is a sentence with "like" as a verb.

At this point a less sophisticated translation machine would proceed to Transfer and translate "fast and dirty." But not Rosetta. It would be abhorrent to the linguists if "All bishops like her" came out in Dutch as something that resembles "All bishops like to eat her." Besides, what would the bishops be? Wooden chess pieces?

Analysis Step Three, a module called the Deep Syntax Component, contains Landsbergen's *pièce de résistance*: the new translation-focused grammar M-Grammar. M-grammar is based on theories developed by the semanticist Richard Montague, and Noam Chomsky. It makes Landsbergen laugh that he can still drone the basic principle of M-Grammar after getting so deeply involved with its detailed consequences for the past years: "Two sentences that are each other's translation have two things in common: they have the same meaning, and that meaning has been derived from basic expressions for both sentences in the same way." In other words: every meaning expressed in one language can always be expressed in another.

The object of the principle is to impose the target language's deep syntactic structure onto that of the source. The Deep Syntax Component therefore contains a set of syntactic rules which express the deep structure of English in terms of the deep structure of Dutch.
One of these rules states that in an English sentence containing the verb "to like," the surface syntactic object will become the deep syntactic subject ("her" will become "she"), and the surface syntactic subject ("all bishops") will become the deep syntactic object. The verb "like" therefore will come to resemble the verb "to please" or "to make glad." The Deep Syntax Component's output is expressed in a derivation tree, representing the words' deep, "Dutchified" syntactic functions.

Men of the Cloth

"OK, let's translate," you may be saying, but there is still one small ambiguity to be resolved, concerning the word "bishops." Are these "bishops" men of the cloth or are they wooden chess-pieces? To any human being the answer is obvious, but to a machine – from which real world knowledge, the dreaded AI, has been withheld by its designers – it's not.

Eliminating this ambiguity is the task of Analysis Step Four: the Semantic Component. The "deep" structure from the previous component will, in general, contain a so-called predicate-argument: something is being said about the subject. The Analytical Semantic Component performs "checks" to find out whether the predicate and the arguments fit together.

In "All bishops like her" it will consider the verb "like" a two-place predicate of which the agent argument should be "living being." The computer is thus informed that though people or things can't please chess pieces, they can please men of the cloth.

For another example of the way Landsbergen and his group handle semantics, consider the sentence: "The boys are sleeping." In this sentence the meaning of "boy" is "a property of individuals," namely the property "being a boy." The meaning of "sleep" is also "a property of individuals." And attached to those basic meanings are two semantic rules: one rule yields the properties all boys have, and the

other yields "true" if the property of "sleeping" is a property that all boys have; otherwise "false." In logical terms, that would be:

FORALL (x) [boy' (x)->sleep'(x)]

Back to "All bishops like her." Analysis is now complete. The English sentence's final ambiguity has been ironed out, and its representation – in the form of a semantic derivation tree – is at last passed on to Transfer. This is a short and simple process. Landsbergen calls it "trivial" – a word he loves and uses whenever he gets the chance. Transfer is not where the real translation takes place. Actually, the transfer of this "tree" to the generative part, where the Dutch translation will be generated, shows that Rosetta is an interlingual system, with the semantic derivation tree as the interlingua.

After Transfer, the generation of the Dutch sentence follows exactly the same steps as the English Analysis, but takes place in reverse order: Semantic, Deep Syntax, then Surface Syntax Derivation Trees. Eventually, Morphological i's are dotted and t's crossed. And our formal output is the short, sweet and perfectly formed Dutch sentence: Zij bevalt alle bisschoppen.

Philips Research Laboratories presented a paper at the International Conference on Machine and Machine-Aided Translation at Aston University in Birmingham, U.K., that stated: "The preservation of meaning can be guaranteed without an analysis in terms of logical formulae and without semantic representations of sentence contents in the traditional AI sense. This point is one of the central tenets of the Rosetta method." And they are trying hard to prove it.

Not as Straightforward

Yet, not all Rosetta translations are as straightforward as the one mentioned above. For some sentences, if you take a close look at Landbergen's monitor, you'll notice that Rosetta did not produce just one Dutch sentence, but many. When ambiguities are still present by the end of Analysis, Generation will produce all possible target language translations. Rosetta Two, the working prototype completed two years ago, is not yet capable of making contextual judgments. It will translate the sentence "Wij zagen haar," for example, as both "We saw her" and "We saw hair," because "haar" means both "her" and "hair."

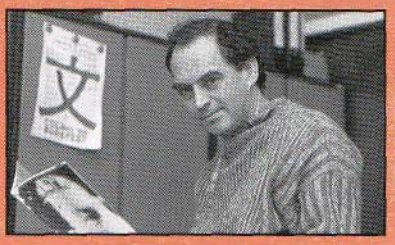
Rosetta has so far been developed for English, Dutch ("If we don't include Dutch, who in the world will?") and Spanish, and in all possible combinations. Auxiliary linguistic research is being conducted for Philips at the University of Utrecht, and the dictionaries are being provided by Kluwer N.V., publishers of the prestigious Van Dale dictionaries. The next stage in the project – Rosetta Three – is expected to be ready in 1988, and the final result – Rosetta Four – is projected for 1991.

"Rosetta Three will be a far more powerful package than Two," says Landsbergen. "Number Three will contain some important subgrammars, the dictionaries will be expanded dramatically and the morphology fine-tuned. However, the program will still not be capable of disambiguation based on discourse analysis. For this reason, in many cases it will still generate more than one translation." Still, Landsbergen says he will be "satisfied" anyway.

Modular testing of the system has been successful so far, though Landsbergen admits he is intimidated by the amount of work. "Some of the grammar's Pascal programs contain no less than one million lines," he says, astonished at the growth the project has been through. "The project can no longer be comprehended with just intuition. It needs discipline and organization now."

Rosetta Four is expected to have additional capabilities – one of which is the ability to process specialized text for customers in particular domains, such as insurance and tourism. Landsbergen's main ambition for 1991 is to increase Rosetta's power of disambiguation, incorporating domain knowledge into the Semantic Components. But Landsbergen remains pessimistic about full automatic translation, and has therefore decided to continue working on the system's power to operate "interactively."

This means that whenever an ambiguity arises, the system will put questions to the user about the meaning of a source language word or structure. Landsbergen: "This user should usually be the author or someone else who understands the text, mainly because the questions will not always be trivial, but also because these questions might be quite unendurable for a professional translator. The user will not need to know the target language. Therefore the system must be quite reliable after all the ambiguities have been solved in interaction."



Jan Landsbergen: Everything is Under Control

Jan Landsbergen (46) studied Mathematics at the Technische Hogeschool Delft in Holland. He graduated with a degree in Formal Language Theories in a decade when computer science was not yet being taught. He wrote an extended essay on Transformational Generative Grammar, "as far as possible, considering the extent of that area in those days," he notes.

After his studies he started working with Philips' "Electrologica" division in Apeldoorn and designed a Question and Answer system for them. At this point he became interested in M-Grammar, and decided to create a new version of it that would suit the purposes of his Q&A project. The outcome was a parser for automatic translation.

"And I also discovered the possibility of using a derivation tree as an interlingua," he remembers.

The Q&A system became the foundation for Landsbergen's next idea, called Speakos. It was a Q&A System based on voice input and output, developed with the cooperation of West Germany's Siemens.

In 1981 he made a proposal to Philips' staff to build a translation computer, was granted the necessary funds and started building Rosetta together with two coworkers.

"We put eight modules together at a speed of two per month, and demonstrated a working system the very same year," he beams with pride.

The only thing that bothers him today is the seemingly uncontrolled, exponential growth of the project. "Everybody underestimates the consequences of transforming a cute little parser into a 'heavy duty' big one. I certainly did!"

Yet, he claims everything is under control.

Weird Little Translation Calculator

"Philip's strategic goal is, of course, language processing in general. The Rosetta project is a means of gaining experience in the field," says Landsbergen. What the eventual product — as manufactured by Philips' HIS group (Home Interactive Systems) or Telecommunication & Data Systems Division — will look like, no one knows. "There isn't a master plan," says Landsbergen, and he mutters about "Spelling and Grammar Checkers." He doesn't really seem to care. His enthusiasm is fired by the research. Its marketability is other people's business.

When the first parallel processing computer arrives, Rosetta will be ready to run on it. Algorithms are currently being rewritten in a parallel language by programmers from the DOOM project (Decentralized Object Oriented Machine), a branch of the European ESPRIT project.

It will certainly be possible to adapt Rosetta for transmission across computer, telephone and television networks – or by satellite. The span of applications could vary from office networks to continent-wide Minitel or Teletext systems, within which journalists would type their stories in one country while Rosetta translated them into local languages. Rosetta could also be used to transmit tourist and traffic information: a driver could pick up the news about routes or roadblocks and then have it transformed into a synthesized voice speaking his or her own language.

Philips at least seems convinced that the market will be there and the research is worth every cent of the 20 million guilders it will eventually cost. A position Landsbergen himself finds pleasantly puzzling: "They just figure out how many people we need for the job and hire them."

A far cry from Rosetta's modest beginnings seven years ago, when the Dutch economy was in the doldrums and Landsbergen was brooding over what to do next. "In those days you could buy these weird little translation calculators that could translate 2500 words. They were a pretty useless novelty and weren't much of a success. Sharp, Texas Instruments, even Philips were in the business. One day I just looked at one and said:'Come on, we can do better than this!'"

Peter Rutten is LT's Associate Editor.