After ten years in development, Siemens’ METAL is now arguably the most technologically advanced machine translation system in commercial use. But its developer, German electronics giant Siemens, has a grander design. Metal is envisaged as just one element in a glittering array of future natural language processing products. Electric Word’s roving correspondent Andrew Joscelyne took the Nachtzug nach München to survey the scene.
1. The Big Picture

Rudolph Thiem
Products Manager at Siemens Natural Language Division

The day in 1978, a frantic telecoms product manager at Siemens AG, Munich, suddenly realized that his new US customer wouldn’t be able to read the 100,000 pages of German technical documentation he’d received for the company’s EWS public switching system.

When company executives learned how long it would take — and even worse, how much it would cost — to translate the documentation into English, they went into red alert and started contacting machine translation companies for a fast, but high-quality, solution.

Unfortunately, they were too late. One valuable lesson was learnt from this episode: If Siemens was to avoid a recurrence of such an embarrassment of unneeded translation, they would have to have the tools for the job themselves.

So Siemens entered the chancy world of MT. Its first move: to buy a system already being developed at the University of Texas, Austin. That system’s name: METAL (Machine Edited Text Aspiring to Legibility).

A decade later, as a direct result of continuous coordination and PR work by Thomas Schneider, honorary leader of the project, the fruit of the research team’s labors is now a commercial product — one of the few operational MT rigs with a price tag on the box. Today, METAL is the registered trademark of a state-of-the-art machine translation system offering complex German into quality English.

According to Siemens Natural Language Products Manager Rudolph Thiem, since the METAL configuration runs on Siemens hardware and software, the company is now using the rig to attract heavy-duty data-processing users to Siemens computer technology in general. Even the managing director of IBM West Germany brought a group of US businesspeople to visit the METAL stand at last year’s CeBIT computer fair in Hannover.

Understandably, after ten years of R&D costs, Siemens is now pulling out the stops to make the product pay its way. And for the company’s Applications Programs division, which houses Thiem’s department, there’s an extra incentive to try something harder; the company seemingly draconian rule whereby every new product must break even in its third year on the market.

Not that the DM25m that Siemens has so far sunk into METAL is exorbitant. Compex (Zürich) and Schönau and Danieli (Rendsburg) — as well as by Philips Kommunikations-Industrie in Nuremberg and Mannesmann Kienzle in Wittenberg. In addition, the University of Hildesheim has been using METAL with its translation students, reporting regularly on usage and providing dictionary support.

There has been a generally positive response to its key user features: sophisticated document-formating utilities; user-friendly coding for adding new lexical data; and general high-tech backup from the mother company.

In Compex’s preliminary evaluation of METAL, the translation company estimated that on large-scale contracts, the use of METAL resulted in savings of 40%. Compex also found that unedited output was often sufficient for clients needing translations for purely informative purposes.

Other recent entries to Rudolf Thiem’s user list include the Dresdner Bank (Frankfurt-am-Main), which has been using METAL in its Advanced Products Department and has recently recommended the system for its main translation center. And Henkel, the Düsseldorf-based chemical giant, has just started using METAL to translate patent summaries for its newly acquired US subsidiary.

Another new METAL customer is the Geschlossene Fürsamer Mathematik and Datenverarbeitung in St Augustin, near Munich, where translation chief Ursula Bernhard will soon be using METAL. Bernhard has recently tested Logos, a rival MT system developed by US-based Logos Corp. As a result of her experiences, she is setting up a product-independent machine translation user group to promote objective discussion on work in this highly specialized area.

Russian Ambitions?

Although METAL so far offers only a single one-way language pair — German-to-English — development of further language pairs is underway at a number of locations, employing a total of some 50 researchers worldwide. Of these, German-to-Spanish and English-to-German will be ready for pilot testing in spring 1990.

With METAL, the bulk of the research — on average of four person-years per language pair — goes on analyzing the source language. Once this is done, producing a synthesis module for the same language as a target is much easier.

The system kernel is maintained at Munich along with German-to-English development and some English-to-German. The rest of the English-to-German is handled by a development team including two full-time Siemens employees at the University of Texas Language Research Center, Austin.

At Louvain (Leuven University, Belgium), a joint team from Brussels-based Siemens Data and the university is working on the Dutch-to-French and French-to-Dutch modules, of which the Belgian Ministry of the Interior is due to run a
pilot test in 1990.

Eurowatchers should note that two of the Louvain group recently migrated from ongoing work on the EC's troubled Eurotra project—despite the common prejudice among non-Belgian francophones that the poor Walloons will never get the rules right for everyone else's French.

In 1987, Barcelona became the fourth regional center for METAL development, with German-to-Spanish and English-to-Spanish modules planned for implementation in the near future.

Meanwhile, the Danish government—now Siemens—is financing a Danish analysis module at the Southern Denmark School of Business Administration and Modern Languages in Sæby, under Prof. Gert Engel. Siemens welcomes this initiative to develop large-scale electronic tools for one of Europe’s "smaller" languages.

No doubt, as the old order crumbles in Central Europe, Siemens will find new markets and hence new user languages—and the company’s recently announced major contract with the USSR for the provision of factory control systems sets METAL staff dreaming of stitching together a Russian synthesis module too.

Work such as this, however, costs millions of marks in development costs, and Thiem’s budget requests don’t always get immediate acceptance.

"Siemens is generally thought of as a cost-conscious company—mainly because it has pretty rigorous financial control at every stage of its projects. And natural language processing has an extra hurdle to clear—the suspicion about it in the minds of most data-processing people. We still get the kind of comment one of my bosses made recently on seeing some METAL output: "My son in the fifth grade could do better than this.""

**Natural Language Processing**

Though METAL has started to become highly visible, with ads suddenly appearing in the press and regular presentations at Siemens offices throughout export-conscious West Germany, the company is quick to point out that this is "just the tip of the iceberg."

Other NLP activities are underway at the vast Siemens Data and Information Systems headquarters at Neuperliche to the south of Munich.

As NLP head Gregor Thurmair explains: "Work is ongoing in two separate groups at Siemens: one doing research under the control of the centralized research division; and the other handling product development in the framework of the Office Systems unit in Applications Software."

What ties all this work together is document-processing, and Thurmair’s teams are investigating a range of technologies for handling its various phases: document preparation, or the development of pre-editing tools; document translation (the METAL system, but also the major TEAM terminology database and the new Term-PC product—see TEAM sidebar); and document-filing and retrieval (using REALIST, a powerful indexing and text analysis tool conducting semi-linguistic searches of very large full-text databases).

The most obvious area in which these other NLP products might be used in conjunction with METAL is technical writing. Thurmair looks forward to the day when technical writing will be prepared in advance for MT systems by running it through pre-editing grammar and style checking programs, thereby reducing postediting when the output arrives.

With the life cycle of consumer products—and consequently, their documentation—shrinking all the time, technical translation needs are bound to rise. Siemens is therefore working on a pricing system for MT work related to its level of difficulty.

On the pure NLP research side, with no products yet in sight, Siemens speech-tech boffins are working on a spectrum of tools including those to handle speech input, language processing and generation, speech output for database access, and voice control (see SPICOS sidebar).

Aside from theoretical work on unification grammars and research into the whole thorny problem of modelling discourse and pragmatics for NLP, Siemens’ other current NLP project is ONEIS. This "natural language explanation" program will allow an expert-system user to get his rig to explain in natural language how it reached a conclusion via its production rules.

**Multiple Applications**

"Each project is extremely expensive," explains Thurmair. "So our basic strategy is to combine the knowhow acquired in different areas into 'knowledge sources,' which can then be used in more than one application." For example, the two million terms in the Siemens TEAM terminology base offer a useful source of lexical data for the translation software, even though grammatical information has to be added to make it applicable. But this terminological material might also be used to generate other products. The METAL parser can also be recycled into a robust stylechecker. In German, too, word order is different than in English and allows for problems about an example that designed stylechecker can never solve. And any of these text-based tools could be plugged into the speech-tech applications to offer linguistic support to the text-processing stages.

"Purely university research is not constrained by product development," concludes Thurmair. "We’re always trying to see how to integrate research into products and to use proven tools in new product environments.

"Actually, we find it difficult to convince people that natural language products can do much more—and there aren’t too many really successful ones around. As a result, we have a vast background in available techniques, lexicons and so on, we feel we’re in a position to really produce these things."

With experience of project management in both speech-tech research and NLP projects departments, Thurmair feels he knows what product potential exists in the field.

Moreover, Siemens provides a very full set of procedures to follow when developing a product, offering strong support to the team. "If something goes wrong, corrective action can always be taken early on. You have to come up with good reasons not to follow the tested procedures, and that makes you think hard about what you’re doing."

What Siemens is doing is amassing the biggest and most momentous concentration of natural language processing development in Europe—rivalled in the rest of the world only by IBM and the Japanese megacorporations.

**Pedal To The Metal**

Gregor Thurmair, head of Natural Language Processing at Siemens

We still get the kind of comment one of my bosses makes recently on seeing some METAL output: "My son in the fifth grade could do better than this."
2. Metal: How It Works

Metal is more than just a translation engine. It is an integrated solution for large-scale, repetitive technical translation.

The system package requires two central computers to run the software: a standard Siemens workstation (the SINIX, Siemens' Unix-based mid-range computer) used for text preparation, deformatting, reformatting, and postediting; and a Symbolics LISP machine, containing the grammar and lexicons that perform the translation. The two machines are connected via Ethernet.

The user is connected to the SINIX workstation via a terminal; a Siemens PC, on which original source-language text is input either via disk or OCR. An office configuration for Metal can support up to 30 such terminals.

The text is sent from the terminal to the workstation, where it is prepared for translation. It is then sent in batch mode to the Symbolics, where it is machine-translated and then returned to the workstation for postediting. The classic MT drudgery of entering new lexical material and finetuning the grammar is all carried out on the Symbolics screen.

Text Preparation

Metal is designed to respond as much to document and formatting problems as it is to mere language transfer. As Siemens NLP head Gregor Thurmair says: "Integration into office environments is one of the key points

TEAM TERMINOLOGY

Coming to your screens soon from Siemens is a new version of Term-PC, a terminology database management tool designed to work on large, multilingual termbases.

Siemens became interested in terminology after many years of running TEAM, the huge inhouse termbase used both by the company's translation center and by online subscribers such as Philips and Bayer.

TEAM is now 15 years old and consists of over two million variegated terms. The original idea behind Term-PC was to offer remote users a way of exploiting this vast multilingual collection of technical words.

TERM-PC

Running under Microsoft Windows, Term-PC has handy screens to access its various functions. You can use the standard record entry mask to generate a Siemens-style term record (fully compatible with a TEAM record), or you can build your own, using anything up to 800 possible fields.

"Users always tell you they want freedom to choose fields when handling terminology, and Term-PC certainly gives it to them," explains Otto Vollnhals, head of Computerized Lexicography.

The system allows the full and economical storage and indexing of synonyms. And a dozen entries can be generated from a single record, thereby saving space in your system.

In developing Term-PC, Siemens had in mind a server-gles-satellite configuration. This was meant to have a constellation of PC terminals emulating the handler interface to the mainframe TEAM termbase. Terminological material would be accessed, updated, downloaded, and generally managed on TEAM.

Smaller users, however, would only ever need a subset of all the fields on TEAM. So Term-PC has been designed to run as a fully autonomous application, able to access data from files gleaned from TEAM or from other user-based sources. No doubt, interfaces with CD-ROM termbases will eventually become necessary.

Term-PC is also due to be made accessible from inside a standard wordprocessor, which will allow terms to be pasted directly into a text.

As in Siemens' other NLP applications, in Term-PC the company has paid special attention to print output. A printout function allows the user to quickly build a customized layout file for a paper glossary. This can then be stored for re-use.

Full font choices avoid the usual dependence on ASCII file printouts, which most similar products offer. Future versions of Term-PC may even offer interfaces for high-end dtp packages such as PageMaker or FrameMaker, allowing professional dictionary production straight from the PC workstation.

This emphasis on output quality derives from a powerful facility available in TEAM, whereby full page-typsetting commands can be introduced into a glossary selection file. The list can then be sent straight to a publisher.

In this way, Siemens has so far provided camera-ready texts for almost 60 dictionaries, including the famous Ernst Science and Technology dictionaries and the recent Routledge French/English Information Technology dictionary.

TEAM also supports Russian using an iso-transliteration character set, which allows input in Roman script but prints out in Cyrillic. It now looks as if these TEAM features will gradually be made available in Term-PC.

NEOPHYTE-FRIENDLY

Languages accessible on Term-PC include most West European ones, including Catalan — as a result of software developments at Siemens' Barcelona subsidiary.

"The Catalans are obviously very proud of having an editing facility on Term-PC," says Vollnhals.

"It was especially interesting to discover that the accented upper-case letters that appear in Catalan — which we thought unique — are also found in Gaelic. We now have plans with the Irish standards institute to use Term-PC to produce Gaelic glossaries. Not a vast market, of course, but it's nice to be useful."

The onscreen environment for Term-PC is designed to be friendly and powerful. Sorting can be run on headwords (not just character strings) or numerical data, as well as on characteristics of different languages — Spanish "h" and "ñ," German "ö" (do you listnames as "Mueller," "Muller," or "Mueller"?)

For functions that Term-PC cannot handle itself, Siemens offers a full-dictionary production system using the TEAM database. Users who wish their floppy-based material to be converted to the ISO Standard for Terminological Data Exchange, can get Siemens to transfer their material to magnetic tape.

This year, Siemens also hopes to start a Directory of Term-PC Users willing to exchange term data with each other.

Vollnhals plans a newsletter including the latest termbases available from different users. And for the neophyte Terminologist who purchases an empty Term-PC, Siemens plans to publish start-up in the form of 5000-termbilingual glossaries on floppy disk.

If, on the other hand, you already have all your records in dBase II or III, then Siemens wants to lure you to Term-PC by offering a free program to convert dBase files to Term-PC format.

Further developments in 1990 will include the extension of Term-PC to Russian and then to Arabic character input, where applications are seriously lacking.

"Despite the many congresses and symposia which tell us about the need for terminology software for Arabic," says Vollnhals, "the years go by and nothing emerges. Finally we have a product to handle the language."
of the whole system. If you only concentrate on the translation part of
document preparation, you save time later on formatting problems."

But if a system can easily be interfaced with other text editors, users can
afford a certain loss in translation quality — since they’ll have more time for
postediting. Documentation preparation alone can account for up to 50% of
total translation costs.

METAL developers argue that what is needed is a balanced system with
both a friendly text environment and acceptable translation quality. And that,
they claim, is where METAL shines bright. For one of its key features is the
preservation of input layout information such as graphics, tables, columns,
bold, and italics.

However, though the aim — to produce a visually identical document in
the target language — is simple enough, it is not a simple task, since both word
length and word order change during translation.

To skirt round this problem, METAL first deformats the document, separating
layout from text and marking up all fonts and format conventions for eventual
output. The system accepts source text prepared on a variety of
wordprocessors — including Wordstar, WordPerfect, and Siemens editors like
the powerful ViewPoint dtp program.

Users find that there is always a certain amount of layout that has to be re-
input in the target text, but this is far better than having to reformat the whole
text, as with most current systems.

**Pre-Translation**

Once deformatted, the text is separated into phrases and short sentences
("translation units") and sent to the Symbolics for translation.

The Symbolics first pre-analyzes the text by running a dictionary check for
unknown words. Once these have been listed, the user codes them linguistically
via the METAL Intercoder, a user-friendly windows-based lexical coding
environment.

Pre-analysis shows the number of occurrences of new words plus the
locations of the translation units they occur in. This gives the user a quick idea of
their use. It is also a convenient stage at which to find typographical errors and
hence improve source-text quality.

In addition, pre-analyses generates a glossary file, which provides information
on how any given term will be translated by different special-purpose
modular lexicons (see below). The user can specify which lexicon s/he wants
to see.

There is also a compounds file (a must when translating from German!) which
offers a default translation of unrecognized compounds based on
analytical processing. For example, it will offer an acceptable "final
inspection" for Ausgangsnrung based on the sum of its parts, as well as a logical
but unwanted "salt mountain" for Saltzwelg.

This utility saves the user spending time coding every compound separately —
its predicts correct compounding in 70% of cases. Siemens has found that the
more specialized the technical usage, the more regular the interpretation of
compounds will be.

**Lexical Resources**

For the actual process of language transfer, METAL uses three basic lexicons.
There are two 50,000-entry monolingual lexicons (one German, one English)
and a transfer lexicon which specifies the parallels between words in the two
languages. The words in the lexicons are organized in a hierarchy of modules.
At the top, there are the function words, then general vocab, and finally
common technical vocab ("cable," "machine," etc.).

In addition, there is a set of special-purpose modular lexicons for subject
areas such as data-processing, telecommunications, and medicine. The focus of these
modules can be refined further and further according to user-specifiable needs.

In translating any term, the system follows the strategy of first looking for
it in the most specialized lexicon. If it doesn’t find it there, it will progressively
work up to the general. The order of this hierarchical search can itself be
reconfigured by the user. External lexical material from TEAM or other sources
can also be imported or converted into METAL format.

METAL’s lexicons obviously need more than merely lexical data. Morpho-
logical and syntactic information must also be included in each term’s entry.
The transfer lexicon is organized as a set of rewrite rules, which state
that source expression X is realized as target expression Y under particular
conditions. The user can look at these rules and, if necessary, modify them via
the Intercoder.

Incidentally, using the Symbolics high-resolution 17-inch bitmap display
clone to work on the Intercoder turns coding and grammatical rule-checking
into MT fun.

The Intercoder offers full interactive help by proposing default coding based
on its own interpretation of given words. For example, if it hits on a word ending
-inzation, it will know immediately that it must be an abstract noun of a certain
type — and so yield the corresponding coding.

According to Siemens, many users are so satisfied with system-generated codes
that they don’t even bother to check them. All the same, the company claims
that it takes just two minutes to code a German-to-English noun, as against up
to thirty minutes for other comparable systems.

**Parsing**

Translation itself is handled by the language rule database, which provides
language-pair-specific parsings of sentences.

It analyses each sentence by looking for the most deeply embedded phrases
first. It then gradually grinds up to the surface, assigning constituent structure
rules at each level. Once it has reached the surface-level nodes, it generates
a tree structure for the whole sentence.

This is performed by a technique of weighing up the probabilities for various
alternative rules until the system reaches a highest probability reading.
Generating a tree structure in this way is a greedy business memory-wise —
METAL needs an address space of over 120 Mf.

The sentence is then transformed by the Symbolics into a deep-structural
 Predicate and Argument construct similar to Case Grammar representations.
From this "deepest" level of analysis, the system builds an output tree in the target language. The tree structure for any given sentence can be called up on the Symbolics intercoder screen. The user can use a mouse to click on any given node and get a listing of coded information on its corresponding word or phrase. If needs be, coding information can be modified.

Sentence by sentence, at lightning speed, the system works through the text, offering a translation that is then collected and sent back to the SINIX workstation as a METAL Output File, where it is prepared for postediting.

Back at the terminal, the user can then either work directly on the reformatted file or call up source and target text together on a horizontally split screen. The two texts can even be presented interlinearly.

**PRACTICAL MATTERS**

A METAL translation run is ideally carried out at night, so that the output file will be ready for postediting the following day. Siemens boasts 200 pages of translation in an eight-hour run (40,000 words a day), though obviously this depends on the quality of the input text.

METAL doesn't claim to offer the fastest run in MT land, but it does pride itself on its "acceptable" output quality. And users who want to double their METAL output can always invest in the new Symbolics XL400, which is twice as fast as -- much though dearer than -- the standard 36 machine.

However, more rapid machine translation times can end up bottle-necking human postediting. METAL users report an average 40 pages a day per experienced posteditor. At that estimate, you'd need five full-time posteditors to polish up a night's run in a single day.

Latest prices on the configuration offer the SINIX MX300 plus associated add-ons, laser printer and peripherals at DM 30,900 with a DM 347 monthly maintenance fee (including new versions, updates and hotline to Munich). SINIX software costs an additional DM 3,200.

The Symbolics USP Machine weighs in at DM 99,840, and the METAL translation software at a cool DM 90,000, with an additional DM 900 a month maintenance. This brings your total METAL rig bill to about DM 250,000 DM, or US$125,000.

---

**3. Spicos and Beyond:**

If their hopes are well-founded, Siemens speech technologists are on the brink of a breakthrough. If all goes well, their SPICOS II speech-recognition-synthesis system will be up-and-listening in 1990. In a pilot scheme, it will run as a speech-entry interface to automatic bank-telling machines.

But SPICOS is not the last word in Siemens' speech-tech efforts. With a little help from JESSI, a knowledge-processing consortium of several large European companies, Siemens hopes to beat the Japanese in its bid to automate speech-to-speech translation.

**EINS, DREI, ZWEI, LET'S GO**

Ambitious SPICOS II delights speech-tech freaks. The continuous-speech, speaker-adaptive system recognizes up to 1,000 words.

The system's speech-understanding module first analyzes input speech's formant properties and signal energy features, which the language-modelling module then turns into linguistically meaningful text by using APS grammar and predicate-logic representations.

The input then enters the relevant application database as a query. Once the database has thrown up a reply, this in turn is sent to the speech-synthesis module and is "spoken" to the customer by means of diphone procedure.

At present, the test system speaks in decidedly metallic German, and a lot of finetuning remains to be done. But the overall architecture looks sound.

The problems of a fully speaker-independent system are formidable. To solve them, the SPICOS project has mobilized help from Philips in Hamburg and the Philips-run Eindhoven Institute for Phonetic Technology, as well as from a number of German university researchers. There is also some federal government money behind the project.

SPICOS's future is being planned even as the current module slips up on whether you say "eins" or "dreie." And Harald Hög, the project's general manager, is already thinking about SPICOS III and its proposed 10,000-word recognition capacity.

**BIG JESSI**

However, Siemens' grand design for speech-tech in the 1990s goes far beyond the restricted domain of dialoging with bank-telling machines.

Many EU readers will have seen a recent Siemens ad showing a Japanese talking on the phone to a German in some future natling paradise where, thanks to your favorite German computer company, language barriers have miraculously disappeared.

Apparently, the ad was nothing more than an outburst of creativity in the advertising department -- and in no way related to products or projects. But the company's research strategists liked the idea, and life went on to imitate art.

What Siemens wants is to combine SPICOS and METAL in a speech-to-speech translation system reminiscent of the Japanese ATH automatic telephone interpreter.

As Hög says, research will be needed in many linguistic fields, including dialectal variation in speakers, the precise features of spoken language (as opposed to written) grammatical structure, handling pragmatic information such as anaphora and presuppositions, and more general human-behavior modelling for the dialog scenario.

In order to realize this, Siemens has enlisted the help of JESSI, a European knowledge-processing consortium including Alcatel, Daimler-Benz, Philips, and LIMSI. These companies will hook up with Thomson's transputer technology and a new Prolong processor to provide a whole range of Euro-techniques for the speech-to-speech translator.

"We want to be able to put a speech recognition system on a single chip with 64-Kb-RAM technology and associated logic; JESSI should provide us with the basic tools for this," says Hög.

And with the eventual production of cheap dedicated Euro-chips doing massive parallel processing, the multilingual voice-activated typewriter that speech technologists dream about may just be the first real product to emerge.

Andrew Josceline is Electric Word's 1992 Fulfillment Coordinator.