[*Language Industry Monitor* 14, March-April 1993]

PANGLOSS: Interlingua Vivat?

PANGLOSS is one of three us government-sponsored MT projects currently under way. As ISI's Eduard Hovy explains, part of the challenge is to balance research needs with those of development.

In the field of machine translation, an interlingua-based system remains a fascinating but elusive goal. It is an intriguing concept; an intermediary, language-independent representation which, in theory, separates the analytical side of a system from the generative side, thereby suggesting a measure of modularity. Much has been written in the research literature about interlinguas. Bruce Tolin, founder of Toltrans, allegedly even obtained a patent for one. The accepted wisdom is that up until now the interlingua approach has not yet been proven to be a viable alternative to direct or transfer-based MT

"Just a minute," interjects Eduard Hovy, "I'm not so sure direct and transfer-based MT is a technology to get excited about - Systran notwithstanding!" Hovy maintains that the interlingua approach does not require a bold leap of faith, rather it is a logical extension in terms of depth of analysis beyond the "deep syntactic" analysis of a system like Logos. Hovy and his colleagues Kevin Knight and Richard Whitney at the University of Southern California's Information Sciences Institute (ISI) are currently participating in the development of one of the most ambitious interlingua MT systems to date. Called PANGLOSS, it is a three-way collaboration between three American research groups with funding coming from ARPA (yes, DARPA is now ARPA again). Initially, the researchers are working on a Spanish-English system; it will be followed eventually by a Japanese-English one.

For the analysis stage of PANGLOSS, New Mexico State University's Computing Research Laboratory (CRL) is contributing its ULTRA parser, which will work in tandem with a Spanish lexicon tagged with the Longmans LDOCE sense keys; this will be the initial bilingual lexicon for the system. At Carnegie Mellon University in Pittsburg, the Center for Machine Translation is developing the interlingua for PANGLOSS. Here, the output from CRL's parser will be mapped to a complex scheme of case frames, in which such entities as agent, theme, and experiencer are identified. These are mapped to the database of concepts - individuals, events, or states, which in turn have their own attributes and are linked causally, temporally, or spatially to other concepts. There are slots in these concepts where discourse and other kinds of pragmatic information can be tracked, such as speaker attitudes (degrees of certainty and formality), thematic structures, and aspects of time (speaking, reference, and event). Much of this is based on AI research from the past decade; the CMU MT Center worked for many years at applying AI techniques to MT. The resulting Knowledge-Based Machine Translation (KBMT) methodology is also at the heart of the ambitious fifteen-language project the Center is now undertaking with the Carnegie Group at the behest of Caterpillar.

Working in close collaboration with the Carnegie Mellon and New Mexico groups, Hovy and his colleagues at ISI are developing the module to generate the English output in PANGLOSS. ISI is regarded as a bastion of text generation, primarily on account of penman, its text generation system. penman has been very generously funded over the past decade and it now circulates freely within the research community. As Hovy explains, because generation is more straightforward than parsing, penman is reasonably accurate. When parsers encounter something that is not defined, they generally fail; generators, however, can fail gracefully by incorporating the unknown data literally and proceeding onward. A system like PANGLOSS will display the problematic data interweaved through the text. This is where the user comes in.

The PANGLOSS prototype is intended to be used interactively, with users resolving ambiguities and choosing formulations in the source language at a workstation being developed at CMU. The Caterpillar system is being designed for Caterpillar's Simplfied English. Will PANGLOSS also be able to take advantage of the constraints imposed by some form of restricted input? "No," says Hovy, "that's the distinction between dissemination and assimilation in MT. Caterpillar wants to disseminate information and can afford to impose input restrictions. Our task is to gather information. We have to be prepared for any kind of input, albeit within our assigned domain, that of finance - mergers and acquisitions in particular."

An important component of PANGLOSS is the PANGLOSS Ontology, a large conceptual network which supports the semantic processing of the other PANGLOSS modules. When it is complete, this network will contain 100,000 nodes representing commonly encountered objects, entities, qualities, and relations. It is being built partly by merging WordNet, the semantic word database based on psycholinguistic principles developed by George Miller at Princeton, and Longmans' LDOCE dictionary. "Each of these resources has something to offer a large-scale natural language system," explains ISI's Kevin Knight, "but each is missing important features present in the other." Knight and colleagues have developed a suite of algorithms which match LDOCE definitions to WordNet definitions in order to flesh out this network. Because they are aiming at broader coverage than has previously been possible in MT systems, part of the group's strategy is to develop automatic and semi-automatic methods of knowledge acquisition for the system. Since dictionaries and corpora are not always perfect sources of knowledge, initially they still check the results.

The PANGLOSS project is one of three MT projects currently being funded by ARPA. Like every good punter, this Defence Department agency is spreading its bets. Within the three-year program, it is supporting the resolutely statistics-based prototype of Peter Brown's group at IBM, a hybrid approach by Dragon Systems, which has no experience with translation but has booked impressive success in the speech arena, and the CMU/NMSU/ISI triad with its knowledge-base orientation. While Dragon is starting with a clean slate, the IBM and PANGLOSS teams are building on research which has been around for awhile. ARPA would clearly like to see some of this well-cloistered technology get out into the world. ARPA's carrot is its substantial funding of concrete, short-term goals; its stick is its annual evaluations.

The number-crunchers at IBM have vociferously promoted their approach, championing the fact that they have been able to achieve wide coverage with statistics. As a result, they are slightly ahead in the game. They are now working on obtaining higher quality. The PANGLOSS group, on the other hand, has achieved relatively high quality but in limited domains. It now needs to try to broaden its coverage. "We are all trying to get to the same place-wide coverage with robustness - but taking different paths," says Hovy. But within the short term time-scale of the annual evaluations, it is difficult for the PANGLOSS team to develop the kind of lexical resources that it needs to be able to achieve breadth in coverage. The statistics-based group, meanwhile, uses a bilingual, aligned corpus based on the Canadian Hansards. Because these texts offer such broad coverage, the IBMers were able to achieve impressive results early on using their computationally intensive algorithms that produce translations based on previously translated materials in the corpus. The game, however, is not completely over, if only because of the paucity of bilingual corpora such as the Hansards.

"We are funded to do development, not to perform research," Hovy continues, "but we're not dealing with a small, clearly-defined engineering problem - language is as large and complex as the mind. Unfortunately, we really don't understand enough about it. In the huge realm of pragmatics - theme, discourse, style, intention, connotation, and interpersonal aspects - very little is known. We need a lot more research." If MT is ever to become viable outside of narrow technical domains, understanding these aspects of language will be vital. But who will be prepared to fund the research so keenly needed in this area? "At least sixty people worked for three years on LILOG, an NLP-KR project at IBM's German labs. And it disappeared without a trace. What we could have done in three years with all of those warm bodies," he sighs despairingly, "Will that kind of money ever be available again?"

Hovy believes that too little money is being put into places where it can best be used: research into hard problems guided by a solid sense of what is going to be useful in practice (rather than what is theoretically attractive) and developing well understood solutions. "The us puts too little money into research, for too short a time, in comparison with Europe and Japan. The National Science Foundation doesn't have the budget. ARPA, or a commercial version of it, should have more money and more freedom to get away from short-term military-oriented research applications. Meanwhile, industry, with the exception of the very largest corporations like IBM and GE, can't afford real research projects because of corporate raiders," says Hovy.

If there is any certainty in Eduard Hovy's life, it is probably the next annual ARPA evaluation, commencing May 15. There, the three MT systems will be tested on a collection of twenty-two newspaper articles, mostly in the domain of mergers and acquisitions. A team of evaluators will give marks for adequacy, style, and comprehension, ranking the MT output against machine-aided and fully manual translations not identified as such.

When the program is completed after three years, and Hovy et al get their ARPA report cards, what will be next? "Hopefully, there'll be a follow up. PANGLOSS II," replies Hovy. Thereafter, parts of PANGLOSS might then be ripe for commercial exploitation in collaboration with an industrial partner - a good five years down the line. Hovy would like to see something of practical import eventually result from the beehive of activity surrounding language processing, but he warns us that we shouldn't expect radical breakthroughs. Hovy believes that with incremental progress over the next few years in such areas as speech recognition and machine(-aided) translation, these technologies should become economically feasible and they will begin to have an effect on our daily lives in some small way. "We may not all end up talking to elevators and using email translators every day," says Hovy," but enough of us will to make AI and NLP deliver a small measure of real commercial success."