[*Language Industry Monitor* 4 (1991), 1-3]

What Do Translators Want?

Early Computer Aided Translation (CAT) tools tended to remain close in concept to their paper-based equivalents: fast but simpl e look-up tools based on traditional notions of dictionaries, thesauri, and index card files. As developers rethink the paradigms that govern the triangular relationship between man, machine, and data, this state of affairs may change. Language Industry Monitor takes a look at two fresh new perspectives on CAT. In the Netherlands, Leo Konst has a bold new approach to multilingual online reference works. And Quebecker Claude Bédard believes he has discovered a way of not throwing the translation baby out with the MT bath water.

polyglot

The nicest surprise of the otherwise run-of-the-mill Europe Software '91 (Utrecht, May 22–24) was to discover Leo Konst, founder and director of Linguistic Systems, demonstrating his company's novel multilingual thesaurus, tentatively called Polyglot. At the core of this DOS package, he explained to onlookers, is a 70 Kb memory-resident lookup program, compatible with most common wordprocessors. Once the program is loaded, you can pop into any one of the optional lexica that you have installed, with English, French, German, Dutch, Spanish, and Italian currently available. You can then browse through terms, skipping from one language to another at will.

What sets Polyglot apart from any other electronic thesaurus is that it prompts you to select the semantic context you require for a given term when an entry displays more than one meaning. This is because Polyglot's underlying architecture is structured around hierarchies of concepts rather than lists of words. Each wordlist might have up to 100,000 lemmas (dictionary entries) categorized in sets of up to 10,000 concepts. To illustrate, Konst picks the word "teach" which is linked to conceptually associated terms such as lecture, edify, educate, coach, tutor, instruct, preach, etc. Such a concept is then grouped together under a more general or abstract concept, and this higher level is itself categorized by domain into hyper classes such as general, technical, legal, and so on.

Since Polyglot's lexicons are linked on the level of concepts not words, when you indicate the source concept you want, you get only the relevant equivalents in the target language instead of the complete set of all possible words linked to the original entry. The general sense of a common word like tender, for example, will have certain particular equivalents in another language. In the legal domain, however, tender may have equivalents with no formal (as opposed to semantic) relation to the original.

sub-lexicons

What does Konst see as the advantages of this concept-based approach? "Because Polyglot's lexicons all have the same structure, you can easily move from one target language to another, maintaining the contextual sense of the source term in question. It's a more targeted and therefore less time-consuming way of looking up lexical information. By choosing to remain within one contextual domain, you have what we call sub-lexicons (eg, medical, legal, aeronautics). If you limit yourself to a given domain, you can avoid some rather unpredicatable associations."

Other advantages of this concept-based approach include modularity (you can add lexicons as they become available) and smaller storage requirements: traditional two-way electronic dictionaries contain both languages twice, as target and as source. With Konst's system, each language is coded as source only. The six lexicons which are currently available represent over sixty man-years of dictionary coding. Konst's group did not purchase existing thesauri, preferring to start with a clean slate. "Electronic thesauri offered by traditional

reference-work publishers are obsolete," Konst maintains, "since they are organized in terms of the original printed versions using alphabetical rather than semantic arrangements. Because of this, they cannot provide a sound basis for further development."

Linguistic System's permanent staff of six is assisted in its monkish task of compiling lexicons on semantic principles by native-speaker translators and interpreters. Other lexicons planned for Polyglot include Danish, Turkish, Japanese, and Arabic. And, says Konst, an Indonesian dictionary is also on the books, to be done in collaboration with a Dutch professor at the University of Jakarta.

In addition to its vast interlinked compendium of semantic information, Polyglot's lexicons also contain morphological and syntactic information ('crucial for verbs," according to Konst) which the retrieval program uses for both stem derivation and generation as well as indicating, for example, those verbs that take a reflexive form. If you want, says Konst, you can generate all the inflections of a French verb, turning Polyglot into an online reference tool for grammar as well.

A computational linguist by training, Konst left the Catholic University of Nijmegen to start up Linguistic Systems. He received a loan from the Dutch government for the purpose of building multilingual thesauri — "something which simply did not exist at the time," Konst explains. While perhaps the most ambitious project in the company's history, Polyglot is not the first of Linguistic System's products. The company's first contract was supplying spelling checkers for Philips' electronic typewriters.

Linguistic Systems has also supplied bilingual lexica to the U.S. linguistic software company Microlytics, which has since implemented these lexica in its OEM package MultiTrans, a multilingual "translation utility" toolkit for both DOS and Macintosh developers. Microlytics in turn supplied Linguistic System lexica to Casio for its range of handheld dictionaries. These handhelds are becoming increasingly more popular and sophisticated, says Konst. "It's amazing how much highly compressed data you can fit on a two megabyte chip".

atelier traductique

A well-known translator in Montreal with a strong interest in MT, Claude Bédard has been thinking long and hard about attitudes toward and expectations of MT and the level of assistance these systems afford the translator. He is now starting to implement some of his ideas. Bédard told the Monitor that this fall selected sites will begin using Atelier Traductique (ATAO), CAT software he developed using John Chandioux's high-level programming language GramR. ATAO is based on Bédard's novel concept of Machine Pre-Translation (MPT), a project which he has been nurturing for the past three years. MPT is the result of a no-nonsense look at the translation process and the level of assistance that machine translation systems have so far been able to offer. According to Bédard, the only indisputable help we get from these systems is "vocabulary retrieval." Their sentence-building capabilities are far from perfect, he points out, and introduce a high proportion of sheer noise, forcing the post-editor-dash-translator to shuttle back and forth between source and target texts. "You end up spending a lot of time wondering how your system produced what it did and even more time fixing it," he says. "Working this way interrupts the natural flow of the translator's mental processes, generating both fatigue and frustration." In the final tally, concludes Bédard, you realize that you could have done the same job as quickly and easily on your own. MT systems only do the easy stuff, while the many difficult aspects of translation remain your problem. In fact, you pay through the nose for it: in licensing costs, linguistic tuning, and substandard output."

Bédard's solution is to turn the tables on MT systems. Instead of having the translator serve the system, why not have the system serve the translator? ATAO makes no attempt at sentence building; it just "searches and replaces" about two-thirds of your text with target language (the remaining words are "wisely" left in their source form). As a result, not a single word is actually deleted, added, or moved. Bédard maintains that such a text is actually quite usable by a professional translator. "A pre-translated text can be more translator-friendly than typical MT output," maintains Bédard. "Of course there is still a fair bit of keyboard work left, but this fits in better with your train of thought because you can still "see" the source text

underneath. You work faster and feel like a translator, not a sentence-fixer."

For the moment, ATAO handles only English-to-French — the most widespread language pair in Canada. ATAO currently runs under DOS but Bédard says a Mac version is planned. Other target language versions may also be forthcoming, depending on demand. The ATAO package consists of a pre-translation engine (estimated rate: "several thousand" words per hour) and a basic 10,000-word dictionary (Bédard says about a third of the entries are labelled "DO NOT PRETRANSLATE"). Also included are a set of keyboard macros for editing pretranslated output and an array of text exploration tools to help set up user dictionaries. ATAO will be marketed progressively, with an emphasis on user-customization. At present, Bédard says he does not believe in "shrink-wrap" technology for translators: "ATAO is quite flexible — and that's an important advantage. More generally, I think CAT skills need to be disseminated among translators. It will allow them to be more creative in the organization of their work — and more demanding of the technology they use.

marketing issues

Both Konst and Bédard will have to face one crucial question: just how big is the market for translation tools? Many translators are content with "state-of-the-art" WordPerfect 4.2 and a printer that supports accents. On the whole, translators have — most fittingly — a keener interest in language than in technology. Will they be prepared to experiment with new ways of using computers for their work? Leo Konst: "Any even half-way professional translator has a bookshelf full of dictionaries." He believes translators like having lots of different reference materials at hand; why should that not include electronic ones too? Polyglot's price, however, may put off casual buyers; with one language pair, Polyglot costs about six times the price of a good hardcover dictionary. "But it's a hundred times faster," is Konst's optimistic response. He acknowledges, though, that pricing could be an issue.

Claude Bédard believes the market varies from country to country. In general, he acknowledges that for most translators automation of the translation process seldom extends beyond a wordprocessor, a spell-checker, and perhaps a simple terminology management program such as Termex. He calls it something of a "brick wall." However, by offering a product which has a modicum of linguistic sensitivity, is flexible and user-customizable, and is not too expensive, Bédard thinks he might have found a crack in this wall. One purely practical stumbling block for the independent translator using Bédard's system will be the need to obtain source texts in binary form. Bédard agrees that this requirement calls for a serious bout of consciousness-raising among translation clients.

For this and other reasons, INK TextTools author Charles Hugo may be right in his belief that big companies are the real market. Independent translators and translation agencies are not interested because they simply do not have the expertise to select a suitable system from the rather fragmented range of offerings and maintain it. Besides, he points out, "you don't need a terminology management system to translate a business letter. On the other hand, large organizations are interested in automating the translation process; take, for example, those companies still running the ALPS software on their mainframes." In other words, the degree to which translation tools can be implemented collectively among groups of networked translators may ultitmately dictate their success or failure.

Claude Bédard, 2705 Boulevard Edouard-Montpetit #1, Montreal, Quebec H3T 1J6, Canada; +1 514 738 8861, Fax +1 514 871 1269
Linguistic Systems, Postbus 1186, 6501 BD, Nijmegen, the Netherlands; +31 80 226302, Fax +31 80 242116