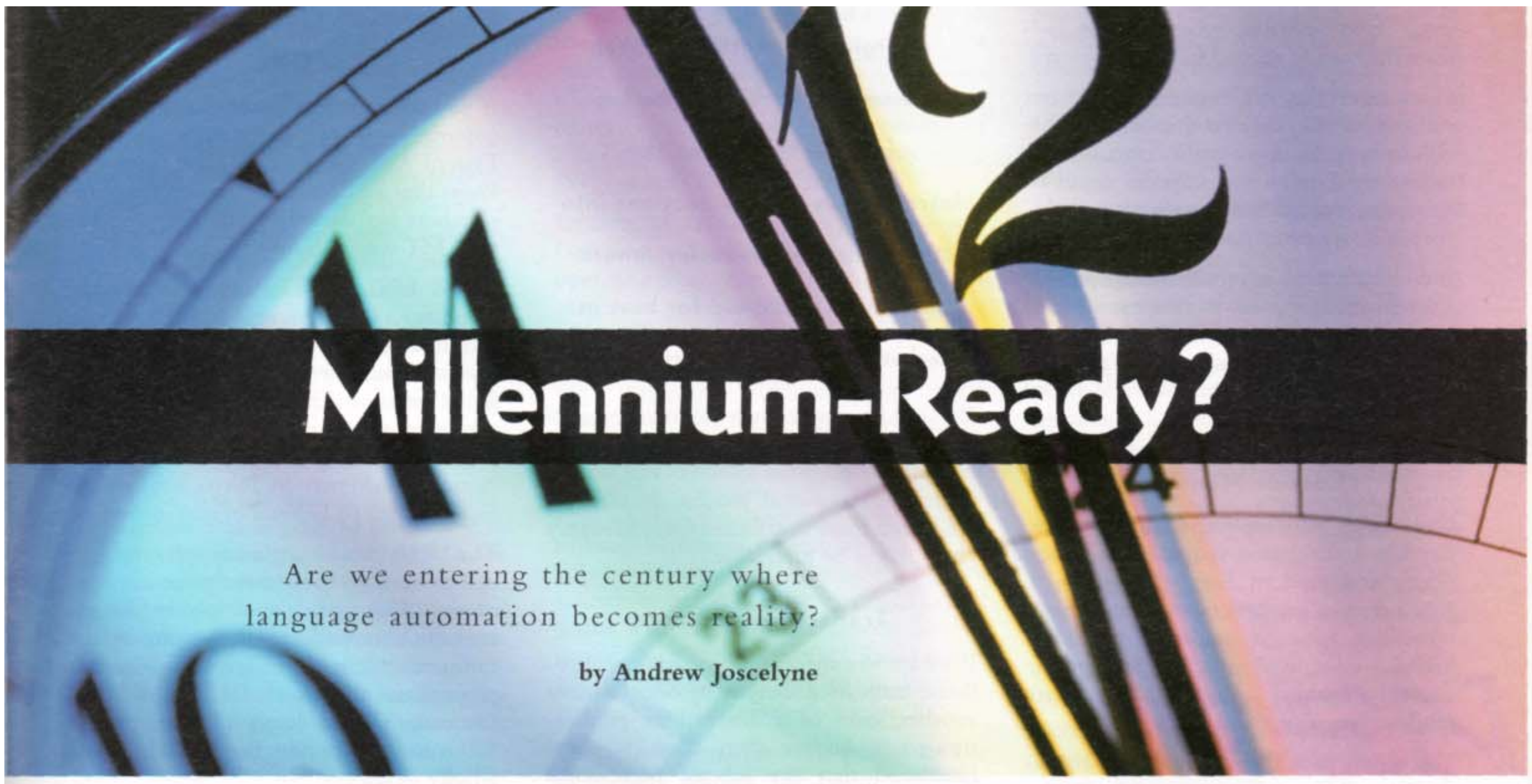

Millennium-Ready?

Are we entering the century where language automation becomes reality?

by Andrew Joscelyne

What's all this? The third millennium in a few weeks time, and still no viable multitask machine-translation system available? Ah, the old dream, and precisely, just a dream. If we try and listen to the technology and drop that old-fashioned "machine," then the answer is yes, there are indeed viable translation systems at work today.They are built of people and their machines, of brawn and brain, coevolving in a single system to meet the ever-growing demand for MTAQ—more translations anywhere quickly.

The various strands of translation-engineering development, from big-iron MT via translation memories to a bilingual word list in Excel, are all converging, blurring the once radical differences between MT systems and CAT tools. If the agenda of the recent MT Summit held in Singapore is anything to go by, it suggests that the professional MT community is happy to embrace a hybrid combination of technologies, from parsers to term-management tools, in their effort to automate translation.

Convergence

The result is a virtual translation toolbox, composed of various techniques designed to speed up this bit or that of the translation process. Depending on the purpose of the exercise, there will usually be some degree of human intervention included in the overall package.

This continuum of translation technologies has largely been enabled by the advent of the Web, which provides an ideal platform for delivering the sort of seamless, transparent translation services that people seem to want. It has also provided new business opportunities, by putting different bits of the toolkit to work as revenue makers in portals or in other ways.

Translation Factories?

Various Web sites now offer an automatic translation service for online browsing or document gisting, while supplier sites may soon be proposing online translation factories, using translation-memory-type technologies to serve the needs of networked clients.

Technologically, the way is now open for comprehensive language outsourcing—i.e., servicing all the language-based content needs of a company right across its various data streams, including the Web, the intranet, marketing materials, competitor monitoring, and so on. As always, the right mix of translation strategies and technologies will depend on the source material and end use.

The action, though, is not just out there in the marketplace, adapting off-the-shelf software for in-your-face new Web applications. Research and development in the translation field also appears to be thriving, and R&D teams seem happy to keep pushing the translation technology envelope.

Long haunted by the specter of no funding, at least in Europe and the US, translation processing is now a booming field again, for both fundamental and applied research. People realize that there are a number of crucial engineering challenges that can be met without the need for full-fledged language understanding systems, only small fragments of which still exist today for a few languages. This is largely due to greater and cheaper computing muscle, and far more generous computer memories, but also because there is now a huge stock of multilingual versions of electronic text for experimentation.

Setting an Example

There seems to be a trend for R&D to be focusing on example-based machine translation (EBMT). This approach applies the logic of translation memory to truly massive parallel corpora of already translated material, and is not in itself a radical new idea. Using special software, the translation system can be activated to use statistical analysis to derive translation equivalents of some kind—phrases, words, sentences—from the mass of aligned, bilingual material.

One interesting experiment, carried out earlier this year, showed that a reasonably efficient—i.e., low-quality—translation system for English and Chinese could be generated using EBMT methods on a bi-corpus in approximately 24 person hours.

If confirmed, this approach would provide a rough and ready way to create some kind of a rapid translation system for language pairs where adequate bilingual dictionaries don't exist, but where there were plenty of electronic versions of texts and translations.

Corpus Manager

One rather odd application might be to develop a translation system for Classical Greek and modern European languages. Kilometers of printer's tapes from some of our most eminent academic publishers probably contain all you need to align the ancient authors electronically with modern language translations.

You simply pack them all into your corpus manager and, given enough time, you ought to be able to get the system to learn-by-example and generate a plausible automatic English or other language translation of any long passage of classical Greek.

All of which raises interesting questions about the potential market value of these parallel corpora. For years, technologists have been telling translators and large organizations who use their services that they are sitting on a virtual gold mine of validated versions of source and multiple language targets—if only someone knew how to exploit the wealth they contain.

Habeas Corpora

Translation-memory software as we know it now lets people leverage these dormant riches, but the actual process of deploying translation memory is still not as streamlined as many would wish. When effective EBMT engines start delivering decent results, then some of your bi-corpora could turn into handy assets in a wired world.

One suspects that the translation-automation suppliers who have been gradually amassing data warehouses full of parallel corpora—many of them offering the added advantage of being recorded in an advanced mark-up language such as SGML—are hoping to do a little key provisioning for what may become an interesting new language-resources market.

If we are avid believers in the emerging institutions of Web commerce, it may not be too fanciful to imagine the introduction of Web auctions into the translation industry, where employers could bid for best in market text editors or translators.

We shall have to see whether ownership of the translation portal will also mean owning the language resources that underwrite the translation process on offer, or whether a more distributed form of supply chain might be preferable.

Translation Editors

If we are to expect a greater array of automatic translation output in one form or another, some of it destined for publication-ready quality, then there is every likelihood that the job of professional translation editing is set to take on value.

Translation editors will simply be professional editors with language competency who can rapidly craft deliverable final versions from translation-system output, however it was derived. The Web itself is now riddled with content editors, capable of writing, framing, integrating text and images, and repurposing textual information in various ways. These skills could well be beneficial to the translation millennium as well.

For translation suppliers, the availability of good cross-language text editors would be a boon. Using the reach of the Web, any competent translation editor would be just a mouse click away, potentially able to integrate with a networked team and do their job online from anywhere in the globe.

Auction, Anybody?

If we are avid believers in the emerging institutions of Web commerce, it may not be too fanciful to imagine the introduction of Web auctions into the translation industry, where employers could bid for best in market text editors or translators.

However, we can expect strong resistance from most quarters to the prospect of Web auctions of actual translation jobs. The huge number of potential candidates would presumably drive prices right down, causing mayhem on the market as we know it. But don't rule it out completely.

UNL

If you need further proof that translation automation is alive and well and living in Eastern Asia, then you will be pleased to know that a truly international band of researchers has started to build yet another total MT system from scratch.

Called UNL for Universal Networking Language, this system includes a new architecture, new software, language modules, and so on—the whole caboodle. As far as we know, UNL is not another one-way deposit for your tax money. The project is headquartered in the United Nations University in Tokyo, but not directly related to Unesco or similar bodies.

The UNL idea is simple enough on paper, and its roots go back to the pivot language model of translation that was architected into such now-dormant systems as the European Commission's massive Eurotra project, and the Dutch DLT (Distributed Language Technology) project of the 1980s and early 1990s. But UNL has been wholly developed with the Web in mind, both as communication medium and as technology platform.

UNL is a neutral conceptual language that stands like an intelligent spider at the heart of the Web. Using special software, natural-language sentences (Hindi, French, Thai, etc.) are first converted automatically into UNL conceptual code, and equally automatically "deconverted" into a target language as appropriate.

Don't Shoot the Writer

One interesting design feature of this system exemplifies yet another form of convergence within the translation matrix—the integration of the text author with the translation process. With UNL, the authors of documents or the writers of messages will be guided interactively by UNL to eliminate ambiguities and other syntactic/semantiç difficulties, so that the source version of their text is easier to translate automatically. The machine, in other words, alerts the writer about potential translation problems before the process starts—something we can call translation for writers, rather than for readers.

UNL was primarily designed for academics sharing documents and messages across the Net, and a number of national research centers around the world are participating in the development of conversion and deconversion modules for all languages in the set. It remains to be seen whether a system such as UNL will be able to scale up to become a more global resource.

Message Translation

An online message-translating resource does make sense in the essentially interactive world of the Web. Once you've had a document auto-translated from a Web site, your natural networking instinct is to get in touch with the author and open a dialog. After all, this is what the connected society of the Web is all about. In such cases an author-based translation system might offer an improvement over flying-blind translation software that makes you sound or look linguistically asinine.

The UNL is a wholly distributed translation system that appears to plug some of the translation functionalities deep down inside network software, rather than high-profile them in a large, well-branded translation portal. But the inconvenience of using language-editing software each time you want to write a multilingual message might prove greater than one thinks. There's editing, and then there's *editing*.

Speech Translation on the Way

When it comes to messaging and contacting, there are already alternatives on the horizon. If, in this fast-shrinking world, you really need to send a message to someone quickly with whom you do not share a language, then pray, be silent ladies and gentlemen for the incredible all-translating speech telephonical system—or telephone translator, or speech translator (ST)—what's in a name indeed.

Decades of research, especially in Japan, Germany, and in the US, have finally led to operating prototypes for this ultimate convergence of speech and language technology, telecommunications, and deep pockets. Most ST teams have now reached the stage of demonstrating what has turned out to be promising milestone results.

The main contenders are a Japanese concern ASR, the German government financedVerbmobil project, and C-Star, recently tested by Carnegie Mellon University in the US. But we might expect actual ST products from anyone able to deploy the technology. NEC, for example, has promised some sort of an English-Japanese translation telephone for 2000.

Certainly the market looks poised for the technology roll-out. The boom in mobile telephony has now opened up the data dimension to wireless communications, using new protocols for Web browsing and for IP telephony. People meeting recently at Telecom '99, the quadrennial industry binge in Geneva, certainly forecast a mega market for speech-enabled functions in the world of wireless telephony for anyone who can break the language barrier. All of which suggest that ST telephony's time has surely come. But once again, no one knows much about the psychology and sociology of this particular interface, so predictions about how many people will adopt it are rather unwise.

From this brief glance, one thing seems sure: if these translation projects deliver at least some of their promise, you can guarantee that there won't be much in it for plain old translation professionals. Except of course the unique privilege of knowing that it was originally you translators who produced most of those wonderful terabytes of parallel corpora now being feverishly leveraged by translation systems the world over.

Andrew Joscelyne has been closely involved in promoting the emerging language industry for the past decade. As an associate of the UK-based Equipe Consultancy, he is currently working on a number of European Commission projects for the multilingual information society. Email him at ajoscelyne@bootstrap.fr