Some industry watchers say the Web will usher in MT's golden age. One industry watcher takes a critical look.

by Andrew Joscelyne

Is It Showtime?

My tutelary spirit is the benign ghost of Christmases ahead, not dead. It surfs the epoch, echoing the kind of fantasies that machine translation (MT) people in suits probably dream up around the fire after hours.

These musings are just ideas in progress. They are worth airing, if only to feed the pool from which the next generation of MT business plans and industry discussion will be selected.

MT is now an established, if hardly central, feature of communication across the Web. Articles in the consumer press regularly cover the topic, and there is a vast archive of informal Web lore about what works and what doesn't in the MT game.

The Web itself is growing a substantial new multitrillion-dollar economy of online information services, advertising, and B2B and consumer e-commerce, backed by a glittering constellation of Web-enabling technology and content start-ups. And the number of people online is growing daily. The spirit of the times asks: How will MT assets grow in the years ahead?

MT as we know it seems custom-designed for the Web. The technology appears to squeeze a complex set of cognitive operations into a single computing function—let's call it the language switch. It gives users a comforting illusion of seamless access to a powerful service, with few workflow hassles, and no "foreign language" work to do themselves.

In its current form, MT provides a perfect paradigm of job sharing between human and machine: the MT system does the rapid word and syntax replacement, and the human slowly figures out what it means. The alternative—having the machine work out what it means while humans tweak, adjust, input, and fine-tune terminology to tame the beast into

How much are Web-based MT systems used, and what do we know about the demographics, geographies, and content of their translation work? The answer is not much—yet. For the usual reasons of corporate secrecy, it is hard to squeeze out statistics from the suppliers.

semantic submission—is still light years away.

Wedded to the Web

This extraordinary democratization and instrumentalization of MT, beyond our wildest dreams a decade ago, has been strongly encouraged by the very texture of our online lives. It's the information age, remember?

The informational quotient of text—as used in search engines, for example—largely boils down to collections of words, proper noun collocations, and short phrases. This highly infocentric mode of expression and communication on the Web is custom-designed for MT.

We can read texts translated into our own languages through the filter of this new-found "word-'n'-phrase" idiom, where we compensate for the lack of grammatical flexibility by mobilizing our powerful human interpretative capabilities.

Since this mode of reading is becoming second nature, navigating our way through looking-glass MT output is only more of the same, not a qualitative leap into the cognitive unknown. We can easily learn to reframe the "translationese" of online MT in a way that gives it meaning in the info-texture of our lives.

In many ways, this stripped-down "word-'n'-phrase" mindset (as opposed to full grammatical context) that MT output exploits is not a new phenomenon. Predigital telegraphy, military messages from the Trojan War all the way up to Kosovo, the Tironian notes of Cicero's slave, modern court-reporter shorthand, or SMS messages tapped into GSM mobile phones—all these practices testify to our cognitive ingenuity in making meanings from words and phrases which lack the full syntactic setting.

With its capacity to reduce language items to electronic strings, digital technology largely mimics this ancient skill. By running a variety of links between keywords and phrases across vast document landscapes, it gives the illusion of semantic depth beneath the surface of what are simply word connections.

Systran and the MTs

At last count there were some 15 different MT services now on the Web. (The British MT scholar John Hutchins will be keeping a tally at www.eamt.org/archive/compendium.pdf.) Several of these services (e.g., the major AltaVista and Infoseek search portals) use the same translation engine—Systran. So it's probably a more accurate assessment to claim no more than half a dozen serious contenders in the Web-MT ring today.

Some, like e-Lingo, offer a complete Web search and translation service, while others such as www.handyserv.com seem to offer just an automatic dictionary lookup. www.yupi.com/traductor will only translate URLs, while Go Translate requires you to download a language kit to use the system.

The very latest offering in this cornucopia that I know of is the recent bundling of the Canadian firm Alis Technologies' "Gist in Time" MT system with the Netscape 6.0 browser, so that Netscape users will have a hotline to multilingual Alis when searching Web sites.

According to GlobalReach, 20 million new users came on stream between February and April this year. If true, this suggests corresponding MT growth of 40 million pages per month or another 1.3 million pages per day. Even through this very fuzzy picture, we can dimly perceive the outlines of a market.

Remember that Netscape merged with AOL, and is closely involved with the exploitation of Sun's ASP technologies. So this looks like a riposte to Lernout & Hauspie's agreement with Microsoft to install a "translate" option in Office 2000.

How much are these systems used, and what do we know about the demographics, geographies, and content of their translation work? The answer is not much—yet. For the usual reasons of corporate secrecy, it is hard to squeeze out statistics from the suppliers.

Page-count numbers used to play a key role in the MT-watching business. The question, "What is the size of the MT market?" echoes back down the years to the mid 1980s, when the whole thing started. Throughput mattered then because you could calculate the size of the MT market on the basis of pages.

But now that the service is free on the Web, presumably financed with other revenue streams, concern for the page-counts is less important. Or perhaps the figures are so massive that we lose any perspective on what the rows of zeros mean.

Creative Counting

So let's do some creative counting of our own: Systran was recently reported as translating two million pages per day on the Web. This makes for a total of 7.3 billion pages per year. If this is a plausible figure for what the leading machine can do with a well-positioned "Translate" button on some major search engine portals, let's presume that the rest of the pack did as much again last year.

After all, some only got started during the year, others are not easily visible to average Web surfers, and others again might simply not be used much. This makes a conservative online total of 14 billion pages MT'd in a year. Is this a lot or a little?

Inktomi, the search-engine outfit, recently crawled around a total of 1.2 billion distinct Web documents. Our MT figure is the equivalent of saying that one page of every single document on the Web was translated 12 to 13 times in a year. But if we look at the figures through the prism of Web surfers, we get a more interesting take on MT usage.

The latest figures suggest a total Web population of nearly 300 million. If we accept our probably low 14 billion pages a year of MT, this is like saying that each of the planet's Web surfers had 24 odd pages translated a year—about two per month. That sounds faintly likely, doesn't it?

Now, the Web population is an expanding universe. Judging by the growth registered by GlobalReach, 20 million new users came on stream between February and April this year. If true, this suggests corresponding MT growth of 40 million pages per month or another 1.3 million pages per day.

Even through this very fuzzy picture, we can dimly perceive the outlines of a market.

Does It Mean Business?

There is, of course, no evidence that translation figures rise in direct proportion to a rising Web population. For all we know, the Web's entire MT-user population

(continued on page 40)

Showtime (continued from page 19)

might be a small gang of French or Japanese university students laboriously translating technical material from English, page by page in those little MT windows.

But it is more likely that MT usage is spread more uniformly across the Web, and will tend to spread further in the coming years.

In terms of other growth drivers, we know that some six billion emails zip their way across the ether every day. Only a tiny proportion of emails are translation candidates today, but the figure will almost certainly rise.

Then there is the potential megamarket of mobile phone users, who will need either to extract information from the multilingual Web or exchange messages with other language speakers. No one knows how to evaluate the MT dimension in these markets, but they could add plenty of zeros to the equation.

On the other hand, there are contextual brakes on MT. For example, it looks as though most sustainable third-generation Web sites—those built for industrial-strength e-business rather than mere Web presence—are tending to localize their sites in some way, at least for most major language communities. And a localized Web site—even in just one other language—is in some sense a loss of business for MT engines.

One way or another, then, the real MT figures will start to matter soon. Not simply to compare performance, but more strategically to understand the market as something to be grown, not given. If the emergence of the net economy has taught us anything, it is that knowledge of and care for your clientele is likely to be critical to success, just as in any other economy.

This explains the range of new software coming on the e-commerce market, designed to help virtual boutiques learn about and manage their customers through the Web-site interface.

There's no reason why the same logic will not apply to MT. With a rack of systems lined up to go, any serious MT promoter (or their financial backers) will want to know who their customers are, what they translate, and how often they return. This sort of data must be easy to store, structure, and mine. Indeed, one wonders what use is being made of it already.

So once the performance figures, customer-segmentation statistics, and other bits of basic market research data have been collected, MT will be ready to do some businesslike word strumming over the Web. Here are three fairly obvious scenarios for the future of the MT business:

The M&A Scenario

If you take the market approach seriously, you can imagine that as the usage numbers start rising into the trillions of pages, online MT suppliers will try to differentiate themselves from their competitors to snap up more market share.

This could take a number of forms—offers of more post-editing services for serious translation jobs, email specialization, language-pair expansion, deals with anyone from e-commerce sites to operating-systems manufacturers—all to ensure that your system gets pride of place in the Web-visibility wars.

We know that some six billion emails zip their way across the ether every day. Only a tiny proportion are translation candidates today, but the figure will almost certainly rise. [...] Then there is the potential megamarket of mobile phone users, who will need either to extract information from the multilingual Web or exchange messages with other language speakers.

Competition among players seeking MT users on the Web will heat up as globalization gives the planet a new spin. The user base will grow as more MT language pairs come on stream. Word will eventually get around that one or two of the MT engines are better than the rest, and the market will react with a stream of acquisitions, mergers, and other forms of consolidation. In this mid-term scenario, we'll end up with a couple of mega MT companies running the rest.

Is this plausible? The obvious argument against it is that any player who tries to stack MT systems together will not necessarily boost his speed or quality of service. The lucky owner who aggregates a bunch of smaller players would still be selling four competing brands of washing powder, not one super-cleaner. But at least you own them all.

The usual reason for gobbling up your competitor is because it owns part of a technology you need. If the competitor's MT system has a language pair you lack, buy it up, strip out what you need (the language software, some word lists, translation memories, and the like) and plug it into your best workflow system.

A more likely M&A narrative is vertical integration. In this case, it will be the big Web-development companies who buy out a likely MT candidate, and then hard-wire the translation engine into their own platforms for e-commerce, customer-relationship management, etc. They will then pump some money into fine-tuning MT development on the basis of their customer data.

The market is already here. We will almost certainly be able to compare notes on some chunks of this scenario in the near future.

The MT Portal Scenario

This is a special case of the M&A scenario. Instead of aggregating MT systems into a virtual network of unrelated points of service all owned by a large player, someone might decide to squeeze revenue streams from the perceived user need for multilinguality by building a single point of service—the MT portal.

You might find the MT portal concept leveraged across language/national markets or even across business sectors (e.g., "the banking MT portal"), but it is essentially modeled on the entrance-to-the-pipeline principle. A "dedicated" portal like this would have to add value to the translation experience (the "Translate" buttons scattered across other people's portals do not). This would mean fighting hard against the accepted wisdom that most users use MT for fast access to a Web page: MT as a language switch, not a translation service.

We suggested above that aggregating MT systems has obvious drawbacks. What would be some of the benefits? Perhaps a high-profile MT site could attract users who wish to choose the most readable translation of a document or Web site.

Now that the cost of MIPS has dropped to almost nothing, you could afford the luxury of firing all your barrels—all six MT systems, say—and letting your customer decide which translation is the best for them.

This sort of approach would also reveal which systems are best suited to which language pairs, and to which text types. But experience suggests that, like other language information or skills that can be more or less successfully modeled as computer functions, the less visible these functions are, the more appeal they will have for users.

For example, all sorts of people need glossaries and dictionaries, but no one really wants to switch tasks midstream and start looking for online dictionaries. What they want is a dictionary-type function planted directly on their browser or desktop GUI.

There seems to be a general law here about language technology—the more visible it is to users, the greater their disappointment. The more synonymous it is with the digital interface itself, the more effective it becomes.

If MT portals do in fact emerge, they could offer some unexpected—and hopefully more playful—services. MT has coasted along on the old joke about spirit-and-flesh, vodka-and-meat for far too long. It needs to turn its own technology inside out and attract customers with more machine-stimulated pleasure.

And not just examples of idiotic translations. Let's hope that some language surrealist creates a function that takes your text and chains it through all the 15 odd systems online, looped together into a Formula-1 MT circuit.

Roll up for a round of lingual *cadavres exquis*, the game the real-life Surrealists liked to play where the last line of one person's verse formed the input to the next player's line, and so on round the group.

In the MT version, you could run bets on the chances of having your text return to the wording of the source text after N runs through Z languages and back. And since MT has so far been perceived as a boring black-and-white text function, why shouldn't some creative design types come along and brighten up the whole service? How about some sexy graphics displaying the translation process in full sound and vision in a special window, a brain scan of the linguistic pathways zigzagging across the computer's cortex?

The OS Scenario

Which segues to our final "technology"-driven scenario. Not MT technology—we take it for granted that MT systems will somehow improve. Here we plunge beneath the visible surface of the Web, and try to imagine what would happen if a powerful software publisher decided to build some sort of MT enablement right into an operating system.

This would have the effect of pushing our language switch down deeper into the computing infrastructure, allowing it to draw seamlessly on the best possible resources, translation memories, and example pools, by sending word- or syntax bots scuttling around the network in search of the right stuff for good translations.

The big Web-development companies might buy out a likely MT candidate, and then hardwire the translation engine into their own platforms for e-commerce, customer-relationship management, etc. They might then pump money into fine-tuning MT development on the basis of their customer data.

Like all techie dreams, this ignores the Darwinian battlefield of real markets, and dangerously presumes that what is rational is inevitable. But there are signs that players in such a scenario are waiting in the wings.

We have already seen Lernout & Hauspie and Alis Technologies attempt to corner the user market with an MT button judiciously positioned on desktop software instead of on the Web site. The technology approach would go much further and attempt to render our whole computing experience totally "language sensitive" to the machinery and fully "language transparent" to the user.

This would ultimately depend on a revolution in the operating-system market. But the current spread of the Linux freeware "version" of Unix, for example, would make a wonderful test bed—perhaps it already is—for such radical customizability. If designed with full multilinguality in mind, such a system would enable users to tune their relationship to digital knowledge linguistically, so that all textual or spoken material would ultimately appear to be written in various dialects of one language—the user's own. The Babelian barrier just fades away.

Multilingual search engines and embedded MT systems would do all the hard work down in the basement. Our old friend the language switch would be programmed right into your information appliance the day you buy it. If it were nonreversible, it would be as strange as having an MT chip implanted into your brain. Never again would you need to know what a foreign language was.

And while you've been reading this, another few hundred thousand pages have been MT'd from a Web site near you. It's okay to dream, isn't it?

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