

## Lost in translation

*Efforts to design software that can translate languages fluently have encountered a problem: how do you program common sense?*

by Stephen Budiansky

In one famous episode in the British comedy series *Monty Python* a foreign-looking tourist clad in an outmoded leather trenchcoat appears at the entrance to a London shop. He marches up to the man behind the counter, solemnly consults a phrase book, and in a thick Middle European accent declares, "My hovercraft ... is full of eels!"

Eventually the scene shifts to the Old Bailey courthouse, where the prisoner at the bar stands accused of intent to cause a breach of the peace for having published an English-Hungarian phrase book full of spurious translations. For example, the Hungarian phrase "Can you direct me to the railway station" is translated as "Please fondle my buttocks."

This episode is brought to mind by some recently available computer programs that claim to provide automatic translation between English and a number of other languages. Translation software that runs on mainframe computers has been used by government agencies for several decades, but with the advent of the Pentium chip, which packs the power of a mainframe into a desktop, such software can now easily be run on a personal computer. You can buy a program for translating in one direction between English and a major European language of your choice for as little as \$29.95. But what has really brought machine translation, or MT, into the mainstream lately is two Internet sites that offer Spanish, French, German, Italian, and Portuguese translations free. Systran, a company that three decades ago pioneered the field under contracts from the U.S. government, provides its translator at [babelfish.altavista.digital.com](http://babelfish.altavista.digital.com) as part of AltaVista's Web site. ("Babelfish" is an allusion to a diminutive piscine character in *The Hitchhiker's Guide to the Galaxy* that, when inserted in the ear, provides instant translation of any language.) Babelfish the Web site lets you type in a block of text up to fifty words long, click on a button, and watch as, seconds later, a translation appears above your words.

The second company to offer its services free (at least for now) is Globalink, a major retailer of translation software that as this article went to press had just been acquired by Lernout & Hauspie. Globalink's online translator, Comprendre, will exist in a different form under the new ownership. At <http://www.lhs.com/> you can send E-mail or participate in a live multinational "chat" with translations to and from English.

When faced with criticism of their products' translations, MT vendors tend to invoke the "talking dog" -- as in, Don't be picky; it's amazing that a dog can talk at all. And in fairness, the outright breaches of the peace that these programs cause are far fewer than one might expect. When the field was still in its infancy, in the early 1960s, an apocryphal tale went around about a computer that the CIA had built to translate between English and Russian: to test the machine, the programmers decided to have it translate a phrase into Russian and then translate the result back into English, to see if they'd get the same words they started with. The director of the CIA was invited to do the honors; the programmers all gathered expectantly around the console to watch as the director typed in the test words: "Out of sight, out of mind." The computer silently ground through its calculations. Hours passed. Then, suddenly, magnetic tapes whirred, lights blinked, and a printer clattered out the result: "Invisible insanity."

When I tried out Systran's Babelfish and Globalink's Comprendre, Babelfish handled that highly figurative phrase with aplomb, rendering it in idiomatic, even nuanced, French as "*Hors de la vue, hors de l'esprit.*" Both systems also translated "My hovercraft is full of eels" into French and back and into Italian and back without a glitch. Competent performances were turned in on "We have nothing to fear but fear itself," "Don't bank on it," "I fought the law and the law won," "Wild

thing, you make my heart sing" (*"La chose sauvage, vous faites mon coeur chanter"*), "I shot an elephant in my pajamas," "Can you recommend a good, inexpensive restaurant?," and "The komodo dragon is the world's largest living lizard" (*"Le dragon du komodo est le plus grand lézard vivant du monde"*).

But the Pythonesque possibilities were all too manifest in what Babelfish did to "I have lost my passport." After a trip into French it came back as "I have destroyed my passport" -- arguably better than "I have means pass lost," by way of German. "All's well that ends well" by way of Portuguese became "All gush out that the extremities gush out," and "Would you like to come back to my place?" returned from German as "Did you become to like my workstation to return?"

Most translations fell somewhere between impressive and nonsensical; in general they were surprisingly understandable, if odd and stilted. Particularly fetching was the tendency of both Babelfish and Comprendre to finish English-to-French round trips having picked up a diction vaguely reminiscent of Inspecteur Clouseau's: "Where is the room of the men?" "Do you like to return to my hotel?" All that was missing was an occasional substitution of "zee" for "the." And some uncannily Teutonic cadences emerged from excursions into German and back. "A penny, which becomes secured, is an acquired penny" has a stolidly Germanic pedantry about it. "Pepsi-Cola strikes the point, twelve full ounces, those is much" was perfect stage German.

The computer talks this way for very much the same reason that Inspecteur Clouseau does -- both use literal renderings of foreign idiom. Corny imitations of a French accent abound with the likes of "And now, would madame care for the dinner?" and "The car, she is magnificent" because they more or less follow the syntactical peculiarities of the original. In French one eats "the dinner," not "dinner," and all nouns are assigned to either the masculine or the feminine gender.

The earliest computer translators were even more literal; they were "direct systems," which means they looked up each word or phrase in a lexicon and substituted an equivalent word or phrase in the target language. It ought to have been obvious that this approach had serious shortcomings. But such was the "naive optimism" -- as Eduard Hovy, a researcher at the University of Southern California and the president of the Association for Machine Translation in the Americas, puts it -- that it took a surprisingly long time for practitioners to realize that they had a lot of hard work ahead of them to produce even passable translations. In the 1950s the surging Cold War demand for translations of thousands of pages of Russian technical articles figured into the exciting belief, held by many computer scientists, that the new, programmable computers could duplicate the human mind through "artificial intelligence." Andrew Booth, an early MT researcher, recalls that right after the Second World War he tried to get the Rockefeller Foundation to fund the development of a computer that would perform the tedious calculations required to deduce a chemical compound's three-dimensional structure from the pattern of x-rays it diffracted. That was a task perfectly suited to a computer, but the foundation was not interested. It wanted a machine that would explore how people think. Booth obligingly switched gears, proposed to build a language translator, and got his funding.

Compounding the naiveté was a simplistic analogy advanced by Norbert Wiener, a brilliant and eccentric mathematician at the Massachusetts Institute of Technology who was at the forefront of computer theory: computers were used during the war to help break enemy codes; decoding is a matter of transforming a set of symbols; language translation could be the same.

Of course, it isn't. One huge snag is word order. Forty to 50 percent of the words in a typical English sentence end up in a different position in the corresponding French sentence. In French, adjectives often appear after the nouns they modify; objects of a verb often appear before the verb. In going from English to Japanese, the rearrangement rate hits almost 100 percent, partly because in Japanese -- as in most languages, with the notable exceptions of English and the Romance languages -- verbs regularly come at the end of the sentence. Getting the word order wrong not only makes for horrible-sounding sentences but also can change meaning, often in comic ways.

Another problem with the direct approach is the sheer amount of computational resources required. To carry out direct substitution effectively, a lexicon must include every inflected form of every verb and every plural of every noun: separate translations have to be provided for “walk,” “walks,” “walking,” “walked,” and so on. Idioms can be handled as chunks (for instance, “got out,” “got by,” and “got even,” each of which is expressed by a different verb in French). But if other words interrupt those phrases (“I got the eels out of my hovercraft”), a direct system is helpless -- there is just no way to anticipate every possible combination of words.

Like all modern translation systems, the latest products from Globalink and Systran seek to overcome these limitations by incorporating at least some grammatical rules for figuring out what words are performing what functions in a sentence. The programs literally construct a parse tree, just as students used to do in school. The computers’ lexicons specify the parts of speech that might apply to each word; the parser then looks up grammatical rules that describe how different parts of speech can be combined, and tries to identify noun phrases, modifiers, auxiliary verbs, and so on. Once the sentence is parsed, the resulting syntax tree is “transferred” to the target language with the aid of a second set of rules governing grammatical combinations in that language.

Much of what is impressive about these programs lies in such “transfer” algorithms. The komodo-dragon sentence, for example, required a substantial amount of word rearrangement to come out in even passable French. To say “the world’s largest living lizard” in French, one says literally “the most big lizard living in the world”: the sequence ABCD becomes BDCA -- a 75 percent rearrangement.

Parsing carries with it a certain amount of clarification as a bonus. Words that in the source language can be both verbs and nouns (“Book him, Danno” versus “Give me the book, Danno”; “Fight the good fight”) are distinguished automatically by a computer that recognizes valid sentence structures, and a certain number of semantically absurd translations are weeded out in this manner too. For example, Babelfish botched Groucho Marx’s punch line “What an elephant was doing in my pajamas, I’ll never know” by translating the first three words as “*Quel éléphant,*” a construction bearing the meaning embodied in a phrase like “What a pity” (“*Quel dommage*”); but Comprendre recognized that such a rendering produced a grammatically unfeasible sequence in the sentence as a whole, and instead came up with “*Ce qu’un éléphant*” – “That which an elephant.”

One thing demonstrated by all of this effort is that language is far more complex than even linguists ever imagined. Yorick Wilks, of Sheffield University, in England, is one of a number of MT researchers who have been developing automated programs that comb through a corpus of text to derive grammatical rules empirically. The best known of these corpora is the Penn Tree Bank, which contains millions of words of text to which parse trees have been attached; the computer sifts through thousands upon thousands of sentences and “reads” the rules implied by those trees. This whole approach flies in the face of the tradition of Noam Chomsky, which, Wilks says, led linguists to believe that “they know language through intuition and introspection and don’t need ‘evidence.’” Wilks’s lab has developed an English grammar that so far contains 18,000 rules -- many times the number that linguists ever dreamed would be necessary.

The entire exercise of machine translation has been sobering -- even “depressing,” Hovy says -- because the best MT systems so far appear to have less to do with linguistic breakthroughs at a conceptual level than with simple doggedness. “It doesn’t matter how brain-damaged the approach,” Hovy says. “The older the system, the better” -- because the programs need repeated and extensive fine-tuning, lexicon building, and tweaking of the rules.

Even more galling is the impressive performance turned in a few years ago by a system that made a total mockery of the theoretical excursions of linguists. This system was designed by a group of physicists at IBM who got the idea of treating translation as a problem of simple probability. Rather than working out grammatical rules themselves, they created a program to exploit the expert knowledge built into actual translated texts. The idea was brilliant stupidity: a huge corpus of

bilingual text -- a chunk of the proceedings of the Canadian Parliament, which by law must be published in both French and English -- was fed into the machine. The computer then began tallying coincidences in the crudest fashion possible: if the word "dog" appeared in an English sentence, what words appeared in the corresponding French sentence? After tallying the French words in thousands of sentences that corresponded to "dog"-bearing English sentences, the computer had in hand a table of probabilities: "*chien*" might appear, say, 99 percent of the time; much less frequent would be "*veinard*" ("lucky dog!") or "*talonner*" ("to dog one's footsteps"); least frequent of all would be every other word that had happened to appear in one "dog" sentence or another -- everything from "*anticonstitutionnel*" to "*biscuiterie*" to "*pamplemousse*." Whenever the machine subsequently encountered "dog" in an English text it was given to translate, it substituted the most probable translation -- "*chien*." With enough computing power it's possible to create probability tables for combinations of two or even three words, which can distinguish, for instance, "lucky dog" from "bird dog." (Beyond about three words the number of permutations grows too enormous for any existing computer to handle.)

The IBM system then tossed all the words in its translated sentence into a bag and let the laws of probability decide what order to put them in. Let's say the words in the bag were "*chien*," "*maison*," "*est*," "*la*," "*le*," and "*dans*." Although it is unlikely that any Canadian parliamentarians ever uttered the words "*Le chien est dans la maison*," the millions of sentences they have uttered provide plenty of statistics on how often in a French sentence each of these words directly follows another of them. The words "*la*" and "*le*" almost never follow each other; "*chien*" almost always follows "*le*" (the masculine) rather than "*la*" (the feminine); "*dans*" almost never precedes "*est*" but often follows it; and so on. The computer would simply string the words together in different combinations until it came up with the one that contained the most statistically probable sequence of word pairs.

The embarrassing thing is that the IBM system performed almost as well as systems that incorporated the latest in highfalutin linguistic theory. With no programmed intelligence, no rules about meaning or grammar or word order, statistical systems got as many as 50 percent of their translated sentences correct, whereas the rate is about 65 percent for systems like Babelfish. Hovy says that developers of statistical systems were going around bragging, "Every time I fire a linguist, my system gets better."

The real failings of MT have less to do with the state of linguistic theory than with the fact that computers don't have any common sense. Language is full of ambiguity and multiple meanings that a correct reading of syntax goes only a short way toward sorting out. For example, Babelfish and Comprendre missed both the meanings in Groucho's line "We took some pictures of the native girls, but they're not developed yet," translating "they" into the default masculine form -- even though both girls and pictures are feminine in French. Is a "bank" a place to put money, the edge of a riverbed, or the side-to-side slope of a racetrack? A five-year-old child can grasp the differences in meaning, but getting a computer to do so is another matter. Thanks to the grammatical structure of the sentence, Babelfish correctly established that "book" in "Book him, Danno" was a verb rather than a noun. But lacking common sense, it went on to translate the phrase as "*Réservez-le, Danno*" -- in the sense of "Book a table" or "Book a hotel room."

In other words, semantics is the key to MT, and semantics is a matter of a lot more than linguistics -- it requires real-world knowledge. Indeed, one can get near-perfect translations with just about any system by limiting its lexicon to a narrow, specialized area in which there is no semantic ambiguity. Canadian radio stations translate weather bulletins from English to French every fifteen minutes. In 1974, because the professional translators were getting bored (how many times can one translate "fog this morning?"), Richard Kittredge, at the University of Montreal's machine-translation unit, was asked to develop a program to handle the task. With a lexicon of just a few hundred words, his program achieves an accuracy rate of more than 90 percent.

A major effort now under way in MT circles is to come up with elaborate taxonomies of meaning that will in effect duplicate the real-world knowledge that allows human beings effortlessly to

know which words relate to which in a sentence and which of various meanings the speaker intends. This, of course, is virtually equivalent to the challenge of artificial intelligence in toto -- creating a computer that thinks, or at least comes close enough to thinking that we can't tell the difference. Hovy and colleagues at the University of Southern California have developed an "ontology" (available on the Web at a site dubbed Ontosaurus), which sorts 90,000 concepts into a branching hierarchy according to their fundamental meaning and the way they are treated in language. The universe is divided into qualities, processes, objects, and interpersonal things; an object can in turn be a social object, a physical object, or a conscious being; and so on through multiple hierarchical layers.

The holy grail for many people in machine translation is to use a tree like this to reduce any sentence to a pure description of meaning in an "interlingua." This can then be reassembled into a sentence in any language using a grammar-driven generator peculiar to that target language. But no one seems to be holding his breath.

Except, perhaps, professional translators. "I have people here who still feel threatened, definitely," says Dale Bostad, who works in the translation branch of the National Air Intelligence Center, at Wright-Patterson Air Force Base, in Ohio. Translators are, not unexpectedly, scornful of MT, but they are also undeniably nervous about it. Babelfish's 65 percent accuracy rate on general text is terrible, but not quite so terrible that it doesn't pay to use it in many situations. A professional translator gets about ten to fifteen cents a word, and one who works in a tough language like Chinese can charge three times as much. So if a person's aim is to take a quick look at what a document is about (the bread and butter of intelligence work and, for that matter, commerce), MT can be invaluable. Brian Garr, Globalink's chief technology officer, told me, "Our biggest challenge is setting people's expectations properly. I wouldn't use [our products] to write a manual on nuclear devices." But few applications are so demanding. About 20 percent of communication on the Internet takes place in languages other than English, and to companies that sell their wares on the Web, translation -- even bad translation -- can make the difference between getting an order and not. Globalink offered a gizmo for \$100 a month that allowed whatever material was posted on a company's Web site to be read in five languages; its new owner will offer something similar.

Another major growth area for MT is the European Union, which employs about 2,000 translators to handle eleven languages. Only about 10 percent of its translations are done with Systran systems, but the figure is growing rapidly, as nontranslator bureaucrats realize that they can use the program directly on the EU computer network when they want a quick translation.

Dale Bostad thinks the biggest payoff may come from combining computers and people in a way that exploits the natural talents of each. A document can be given a once-over by MT, and a professional translator can then clean up the glitches. Bostad is developing software that will automatically flag the iffy sentences generated by machine translators' output (such as those containing acronyms, unfamiliar words, or improbable word sequences), so that the human beings can focus effort where it is needed most. Working in tandem with an MT system, Bostad finds, some people can produce eighteen pages an hour, whereas the standard quota for a professional government translator is five to six pages a day. Younger translators in particular have embraced this approach, Bostad says. Their attitude toward computers is, If you can't beat them, join them.

Or, in the words of Babelfish: "If you cannot strike it, connect them."