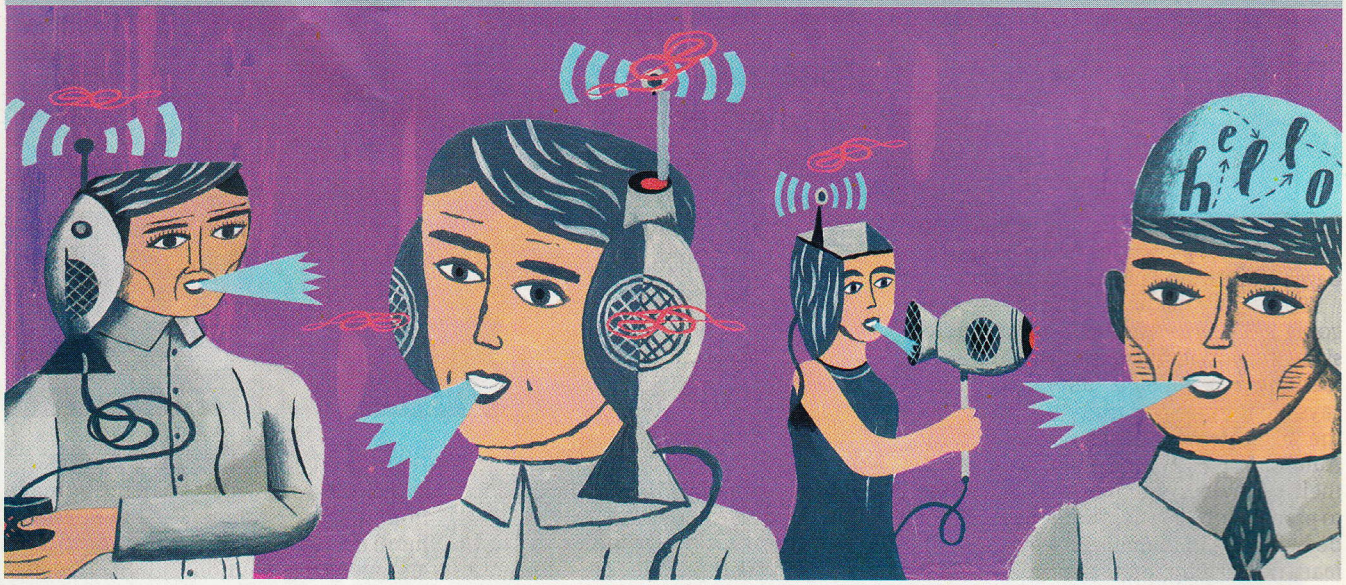

How to build a Babel fish

Translation software: The science-fiction dream of a machine that understands any language is getting slowly closer

IT IS arguably the most useful gadget in the space-farer's toolkit. In "The Hitchhiker's Guide to the Galaxy", Douglas Adams depicted it as a "small, yellow and leech-like" fish, called a Babel fish, that you stick in your ear. In "Star Trek", meanwhile, it is known simply as the Universal Language Translator. But whatever you call it, there is no doubting the practical value of a device that is capable of translating any language into another.

Remarkably, however, such devices are now on the verge of becoming a reality, thanks to new "statistical machine translation" software. Unlike previous approaches to machine translation, which relied upon rules identified by linguists which then had to be tediously hand-coded into software, this new method requires absolutely no linguistic knowledge or expert understanding of a language in order to translate it. And last month researchers at Carnegie Mellon University (CMU) in Pittsburgh began work on a machine that they hope will be able to learn a new language simply getting foreign speakers to talk into it and perhaps, eventually, by watching television.

Within the next few years there will be an explosion in translation technologies, says Alex Waibel, director of the International Centre for Advanced Communication Technology, which is based jointly at the University of Karlsruhe in Germany and at CMU. He predicts there will be real-time automatic dubbing, which will let people watch foreign films or television programmes in their native languages, and search engines that will enable users to trawl through multilingual archives of documents, videos and audio files. And, eventually, there may even be electronic devices that work like Babel fish, whispering translations in your ear as someone speaks to you in a foreign tongue.

This may sound fanciful, but already a system has been developed that can translate speeches or lectures from one language into another, in real time and regardless of the subject matter. The system required no programming of grammatical rules or syntax. Instead it was given a vast number of speeches, and their accurate translations (performed by humans) into a second language, for statistical analysis. One of the reasons it works so well is that these speeches came from the United Nations and the European Parliament, where a broad range of topics are discussed. "The linguistic knowledge is automatically extracted from these huge data resources," says Dr Waibel.

Statistical translation encompasses a range of techniques, but what they all have in common is the use of statistical analysis, rather than rigid rules, to convert text from one language into another. Most systems start with a large bilingual corpus of text. By analysing the frequency with which clusters of words appear in close proximity in the two languages, it is possible to work out which words correspond to each other in the two languages. This approach offers much greater flexibility than rule-based systems, since it translates languages based on how they are actually used, rather than relying on rigid grammatical rules which may not always be observed, and often have exceptions.

Examples abound of the ridiculous results produced by rule-based systems, which are unable to cope in the face of similes, ambiguities or bad grammar. In one example, a sentence written in Arabic meaning "The White House confirmed the existence of a new bin Laden tape" was translated using a standard rule-based translator and became "Alpine white new presence tape registered for coffee confirms Laden." So it is hardly surprising that researchers in the field have migrated towards statistical translation in the past few years, says Dr Waibel.

Now you're speaking my language

The statistical approach, which starts off without any linguistic knowledge of a language, might seem a strange way of doing things, but it is actually remarkably similar to the way humans attempt to translate languages, says Shou-de Lin, a machine-translation expert who was until recently a researcher at the University of Southern California's Information Sciences Institute (ISI). "It looks at the script and bunches symbols together," he explains, much as a human mind might try to solve the problem. But in order for this approach to work, the voracious translation systems must be fed with huge numbers of training texts. This prompted Franz Och, Google's machine-translation expert, to boast recently that the search-en-

"Most of the time, the languages that translation researchers deal with in their laboratories are so unfamiliar that they may as well be alien."

gine giant would probably have a key role in the future of machine translation, since it has such a huge repository of text.

Translation systems are of limited use if they cannot be used by people on the move, such as tourists looking for a restaurant or soldiers talking to local people in a war zone. So what is on the cards to replace the good old-fashioned phrasebook? In the past couple of years the Defence Advanced Research Projects Agency (DARPA), an American military research body, has been testing a number of projects that cram a combination of speech-recognition, machine-translation and voice-synthesis software into a hand-held device. One of these projects, developed at CMU and called Babylon, can now perform two-way translations between spoken English and Iraqi Arabic.

From Babylon to Babel fish

This is a huge improvement on the earlier one-way text-based translators used by American soldiers, says Alan Black, one of the researchers involved in the development of Babylon. For one thing, Iraqis can respond in their native language, rather than communicating through nods and shakes of the head, he says. Better still, Babylon is capable of translating completely novel sentences, rather than being limited to only a couple of hundred set phrases, as with the earlier systems.

It is still far from perfect, says Dr Black. But that is hardly surprising given the limited processing power of a hand-held computer. By comparison, the hardware used to run the lecture translator looks almost like a supercomputer, he says. The trade-off is that these hand-held systems tend to be "domain specific"—that is, they work well as long as the conversation is limited to a particular topic.

The next phase of the project, says Dr Black, will be to allow portable translation devices to be trained in the field. The idea is that when a traveller encounters people speaking a new language that is unknown by the translation device, it can be trained by exposing the software to lots of chatter. In theory, once a language model has been acquired, you could just leave the device in training mode in front of the television, although it would probably be preferable to find some bilingual people and ask them to repeat set phrases containing a lot of linguistic information, says Dr Black.

Learning a new language from scratch, as humans can, is far more difficult than statistical translation using parallel texts. But since the number of high-quality parallel texts is limited, particularly for more obscure languages, a lot of effort is now being put into the development of statistical translation systems that can manage without them. Instead, the aim is to use statistical techniques to divine the language's inherent structure, and then to work out what particular words mean. If this could be done, of course, it would open the way to a universal translator.

How far can machine translators be taken? "There is no reason why they should not become as good, if not better, than humans," says Dr Waibel. Indeed, Dr Lin and his colleague Kevin Knight at ISI have been applying statistical translation techniques to try to make sense of ancient hieroglyphics and scriptures that have baffled scholars for centuries. One example is a 15th-century work known as the Voynich manuscript, which is written in an unknown and mysterious language. Its length, of around 20,000 words, and the regular patterns in its syntax, mean it is unlikely to be a hoax, says Dr Knight. One theory is that it was written in a known language but using a novel alphabet. Some people have suggested that it is actually written in a form of ancient Ukrainian in which vowels are omitted.

Dr Knight has used a statistical-translation program to debunk this theory by showing that the order and frequencies of symbols do not match those in Ukrainian. This was not particularly surprising, says Dr Knight, because most scholars now reject the Ukrainian theory. But it was a small victory for him, because it let him test his translation software on the closest thing he could get to an alien language. "We wanted to translate documents that had never been seen before," he says.

Provided there is some common frame reference in the subject matter, there is no reason why translating an alien language should not eventually be possible, says Dr Waibel. Most of the time, the languages that machine-translation researchers deal with in their laboratories are so unfamiliar that they may as well be alien, he says. "As a joke, one of the students built a Klingon translator," he says, referring to the fictional alien language in "Star Trek".

But perhaps the best way to practice translating an alien language would be to try to communicate with dolphins, says Dr Black. By using statistical translation programs to analyse the chirps, clicks and whistles of wild dolphins off the coast of the Bahamas, he and his colleagues believe it may be possible to make sense of what the dolphins are saying. The challenge here lies in both capturing good samples and also identifying "words". Only then can the structure and frequency be analysed, he says.

So far, Dr Black and his team have managed to identify only signature whistles, the calls that dolphins use to identify themselves. But Douglas Adams's suggestion that fish-like creatures might provide the key to understanding alien languages might turn out to be true after all. ■