[*The Observer*, 19.12.10: *The New Review*, p.22-23.]

# Finally, you are speaking my language

## Accurate translation by computer is the holy grail of machine learning. Yet despite Google's best efforts it still seems a long way off, Tim Adams asks whether it will ever be possible, and where success could lead

Were you to run perhaps the most famous line in literature, the opening sentence of *Anna Karenina,* through Google Translate from Russian to English, this is what you would get: "All happy families resemble one another, each unhappy family is unhappy in its own way."

The translation, which approximates to the best "human" version of the sentence, looks like a triumph for what used to be called artificial intelligence and now is called, less ambitiously, machine learning. The computer can understand language, we are invited to think. Run the subsequent lines of *Anna Karenina* through the system, though, and the picture, along with the grammar, is not quite so clear.

"All mixed up in a house Oblonskys. Wife found out that my husband was in connection with the former in their house, a French governess, and told my husband that he could not live with him in the same house. The situation is now lasted three days and were painfully conscious of themselves and their spouses…"

It is just about explicable, if we know the original, but barely readable. The reason for this discrepancy lies in one of the nuances of Google's system that allows interested users to improve translated texts where they can. Somebody has obviously got to the first line of Tolstoy's masterpiece and put it right. What follows is more representative of what the system is capable of.

Ever since computers were a reality, the possibility of using their logistical power to break down barriers of language has been something of a holy grail in machine learning. The initial – unsuccessful – attempts were based on the principle that all languages could be distilled into two components: a lexicon of words with specific meanings, and a set of rules of grammar and syntax by which those words were linked together. The cold war prompted ambitious efforts by American intelligence agencies to understand the "code" of the Russian language on an industrial scale. It produced mostly gibberish.

The first significant breakthrough in the potential of mechanised translation came in the early 1990s when IBM produced a model that abandoned any effort to have the computer "understand" what was being fed into it and instead approached the task by installing in the computer the comparative versions of as much translated text as possible and having the system compute the probability of meanings of words and phrases based on statistical precedent. The approach was pioneered by Frederick Jelinek at IBM, who, distrusting models that grew from analogies with human learning of grammar, insisted: "Whenever I fire a linguist, the performance of our system improves."

A decade or so later, though, the statistical-based system was becoming severely limited, particularly so when it attempted translations from languages in which there was comparatively little text to "learn" as reference. It was at this point that Google entered the field in earnest. The impetus for Google's translation machine can be traced, corporate legend has it, to a particular meeting at the company's California headquarters in 2004.

One of the search engine's founders, Sergey Brin, had received a fan letter from a user in South Korea. He understood that the message was in praise of the innovative scope of his company, but when Brin ran it through the machine translation service that Google had then licensed it read: "The sliced raw fish shoes it wishes. Google green onion thing!"

Brin believed that Google ought to have the capacity and determination to improve on that particular piece of nonsense. In the years since, as its global interests have grown, the free Google Translate service has evolved to attempt instantaneous translations from 52 languages; it offers a "toolkit" for speakers of more marginal languages to establish their own services, and it is used tens of millions of times a day to translate web pages and other text.

The great improvements Google has pioneered in that time have been based almost entirely on its unique access to vast quantities of translated text, billions of sentences, trillions of words, that can be searched for likely matches in seconds. A good deal of these data come from transcripts of United Nations meetings, which are routinely translated by humans into six languages and those of the European Parliament, which are translated into 23.

Google has incorporated text from its comprehensive book-scanning project and other internet sources to add still further to that syntactical database. (In this it has the edge over its chief translation rivals, Microsoft's Bing and Yahoo's Babel Fish, which are based on broadly the same principles.) As a company, it is in the habit of making great claims for the possibilities of this effort. It announced earlier this year, for example, that the translation tool was being combined with an image analysis application that would allow a person to take a mobile phone picture of a menu in Chinese and get an instant English translation. In the summer, it suggested that it would use speech recognition technology to generate captions for English-language YouTube videos, which could then be immediately dubbed into 50 other languages.

"This technology can make the language barrier go away," Franz Och, who leads Google's machine translation team argued. "It will allow anyone to communicate with anyone else."

That utopian promise is a seductive one. In his recent book, *The Last Lingua Franca*, Nicholas Ostler, chairman of the Foundation for Endangered Languages, argues that translation engines such as Google's will eventually liberate the world from the necessity of learning dominant languages, such as English, and will reinforce linguistic diversity. When I speak to Ostler he is convinced that these changes are inevitable: "The future is easy to predict, though you don't know when it will happen."

Despite a degree of fluency in 26 languages, Ostler says he is often on the Google Translate site and believes it represents this future. "Even if you don't like what it says, you can immediately make sense of what it gives you or compare it with what you know. It still needs constructive intelligence from the user. But the fact is that it is much better than it used to be and no doubt it will continue to improve."

One consequence of its wider acceptance, presumably, will be to make people more lazy about acquiring languages?

"There is," Ostler says, "a sort of irony in that; though we may see a more multilingual future, as English starts to wane, you will see less multilingualism in individuals." The fastest-growing languages online, he points out in his book, are Arabic, Mandarin Chinese, Portuguese, Spanish and French, in that order. "The main story of growth in the net," he suggests, "is of linguistic diversity, not concentration."

Given the garbled state of much current machine translation, though, won't a shared language be as far away as ever?

Ostler argues that "mass production always gives you lower-quality stuff than artisan craftsmanship ever did. It is the same sort of consideration with Google Translate. Even so, there doubt that the more data that come in, the more languages that are assimilated, the better it is going to be."

Those who are working at the sharper end of the translation models tend to be slightly more cautionary about that future. Phil Blunsom, who lectures in machine learning and linguistics at Oxford, and has been involved in creating next-generation translation tools, suggests: "Most of the difficulties we face are what we call 'tractability'. Even in the simplest word combinations, we are searching through a massive space of possible

options. For a computer to understand how a sentence works, it basically has to iterate over all possible options of a syntactic structure between different words and then work out which is the most likely. It is an exponential computational problem, particularly as sentences get longer and more complex."

Andreas Zollmann, who has been researching in the field for many years and working at Google Translate for the last year, suggests, along with Blunsom, that the idea that more and more data can be introduced to make the system better and better is probably a false premise. "Each doubling of the amount of translated data input led to about a 0.5% improvement in the quality of the output," he suggests, but the doublings are not infinite. "We are now at this limit where there isn't that much more data in the world that we can use," he admits, "So now it is much more important again to add on different approaches and rules-based models."

That is where the old problems start. Does Zollmann see a way in which those models can eventually learn languages as well as human beings can?

"No researcher would expect it ever to become perfect," he says. "Pronouns, say, are very difficult in some languages where the masculine and feminine don't correspond to each other. If you ever solve machine translation perfectly, then you have something that is properly artificially intelligent. Language is not separate from who we are."

There are those that believe, as a result, that far from liberating us from our linguistic barriers, the translation tools will in fact serve to reinforce them. Douglas Hofstadter, author of the seminal book on consciousness and machine intelligence, *Gödel, Escher, Bach: An Eternal Golden Braid,* as well as several books on the theory and practice of translation, has been among the most trenchant critics of the hype around Google Translate. He argues that the ability to exist within language and move between languages, to understand tone and cultural resonance, and jokes and wordplay and idiom are the things that makes us most human, and most individual (one of his books was based on asking 80 people to translate the same poem and delighting in the 80 discrete versions that were produced).

The statistical models, he says, start from the wrong place. "There is no attempt at creating understanding and therefore Google Translate is doomed to the same kind of failure for ever. Of course they get occasional good results, but essentially it is mindless. They are rendering a very low-level service that will always produce something not far above the level of nonsense. I suppose that we will all bow to the pressures to use it at some level, but it will never get the flavour of phrases."

Hofstadter suggests that just as, perversely, we seem to like the idea of the world getting smaller, so we like to think that understanding language is somehow mechanical, another problem we can outsource to our screens. "Understanding the world is what humans are good at and what machines are no good at, at all," he says. We may well all be Google Translators soon, but we may also find that, more than ever, we are lost in translation.