
AXIOM PHOTOGRAPHIC/DESIGN PICS/CORBIS

Any idea what he's saying?

Language mash-ups

Inventing new languages makes it easier to translate sentences into the world's more obscure tongues

Jacob Aron

MACHINE translation can make French, Spanish or even Japanese accessible to English speakers. But it requires a wealth of documents with copies in each relevant language to learn how to translate. This works for widely spoken languages, but it can be a tall order for some of the world's 7000 or so tongues. Software that invents languages by mashing up words from existing languages could help translate foreign texts for speakers of little-used tongues.

Luis Leiva and Vicent Alabau at the Polytechnic University of Valencia, Spain, were inspired by the film *Blade Runner*, in which characters use a street language called "Cityspeak", a mix of languages including Japanese, Spanish and German.

Their system, Culturally Influenced Interlanguage (CI^2), exploits the similarity in words and grammar found in language families like the Romance family of French, Spanish and other southern European tongues.

CI^2 constructs translations for minority languages by borrowing words from languages in the same family. "The idea is to pick words from languages for which there are machine translators available," says Leiva. A resulting phrase is unlikely to be grammatically correct, and may contain unusual spellings, but it should be understandable to a minority language speaker.

For example, imagine that Spanish is a minority language, and a native speaker wants to read the English sentence "Another label with the same name already exists". Machine translations between English and Italian, French and Portuguese are fairly accurate and Spanish shares many characteristics with those languages. By choosing the appropriate words from each language, it is possible to automatically construct the sentence "*Un'altra étiquette con mesmo nome existe déjà*", which is not Spanish but should be reasonably comprehensible to a Spanish speaker.

Leiva and Alabau choose words by calculating the probability that a word from the major languages appears in the minority vocabulary by looking at the number of letter changes required to turn one word into another – so "*con*", which occurs in both Italian and Spanish, has a probability of 1,

"The idea is to pick words from languages for which there are machine translators available"

whereas the Italian "*nome*" has a probability of 0.79 because the Spanish word is "*nombre*". Choosing the words with the highest probability means that the translated sentence has the best chance of being understood.

The pair tested CI^2 by asking 17 native Spanish speakers to read a selection of Swedish sentences translated into CI^2 "Spanish" via Italian, French and Portuguese, along with direct translations in all three languages. The volunteers found the CI^2 text easier to understand than the Italian and French translations, but the Portuguese was more familiar still, perhaps because of its closeness to Spanish. Leiva and Alabau plan to test the system with a minority language, even though it is often hard to find native speakers to take part in experiments. CI^2 will be presented at the Conference on Human Factors in Computing Systems (CHI 2012) in Austin, Texas, in May.

"It is a technical solution to a cultural need, which is very exciting," says David Yarowsky, a machine translation researcher at Johns Hopkins University in Baltimore, Maryland. He says the rise of the internet means that languages with less than a million speakers will struggle to survive. "The teenagers of this planet will decide where this goes. They will decide what they want to speak and how important culture is to them." ■

DIY translation to take on Google

Google Translate focuses on the most-spoken languages, but machine translation techniques can work with any pair of languages. All that is required is a collection of documents translated into both tongues. That is why Tilde, a technology firm in Riga, Latvia, has created LetsMT!, an online system that lets users upload documents and create custom translation systems. LetsMT! can even handle technical jargon. Tilde co-founder Andrejs Vasiljevs claims that LetsMT! has been used to create a Latvian-to-English translator for IT terminology that "significantly surpasses" Google Translate. He will present LetsMT! at the World Wide Web Conference in Lyon, France, next month.