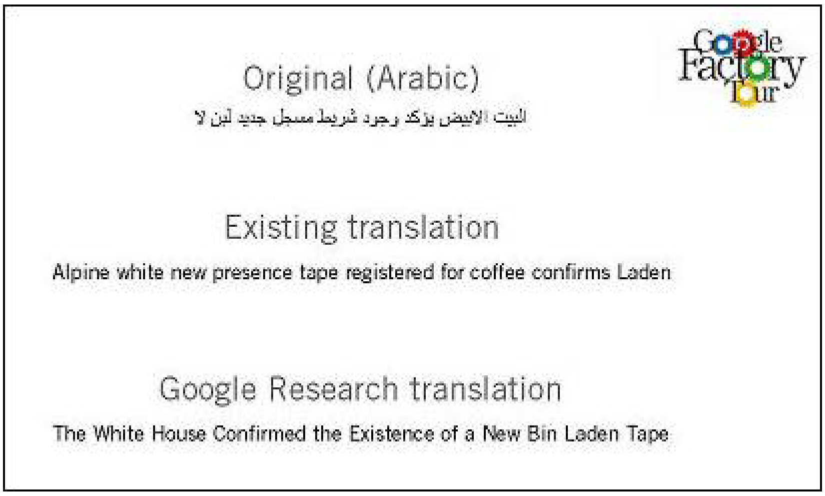
[from: *Google Factory Tour*, Sunday, May 22, 2005 <http://blog.outer-court.com/archive/2005-05-22-n83.html>]

Google Translator: The Universal Language



Original (Arabic)

البيت الابيض يؤكد وجود شريط مسجل جديد لبن لا

Existing translation

Alpine white new presence tape registered for coffee confirms Laden

Google Research translation

The White House Confirmed the Existence of a New Bin Laden Tape

At the end of the 19th century, L. L. Zamenhof proposed **Esperanto**; it was intended as a global language to be spoken and understood by everyone. The inventor was hoping that a common language could resolve global problems that lead to conflict. Esperanto as a planned language might have had some success, but today, English is much more universal. 30 countries have it as an official language, and in many other countries it is taught in school and understood fairly well. The internet can be suspected to further increase the adoption of English.

Still, many people can't speak English. The collected, shared knowledge that makes up the web is therefore only partly accessible to them. The reverse, of course, is true as well. When you surf the web, you will sometimes come across languages and characters you don't understand – like Chinese, Arabic, Korean, French, German, Italian, Spanish, or Japanese. Would you be able to fluently read these languages, those sites wouldn't be a dead end for you. You would discover a wealth of knowledge, and more importantly, opinions. If you're an US citizen, how many Arabic, German or French sources do you read to get a good understanding of how the world sees the US? How many blogs do you read in foreign languages? Probably not many, unless you're fluent in those languages.

At the recent web cast of the Google Factory Tour, researcher Franz Och presented the current state of the **Google Machine Translation Systems**. He compared translations of the current Google translator, and the status quo of the Google Research Lab's activities. The results were highly impressive. A sentence in Arabic which is now being translated to a nonsensical "Alpine white new presence tape registered for coffee confirms Laden" is now in the Research Labs being translated to "The White House Confirmed the Existence of a New Bin Laden Tape."

How do they do that? It’s certainly complex to program such a system, but the underlying principle is easy – so easy in fact that the researchers working on this enabled the system to translate from Chinese to English without any researcher being *able* to speak Chinese. To the translation system, any language is treated the same, and there is no manually created rule-set of grammar, metaphors and such. Instead, the system is learning from existing human translations. Google relies on a large corpus of texts which are available in multiple languages.

This is the *Rosetta Stone* approach of translation. Let’s take a simple example: if a book is titled “Thus Spoke Zarathustra” in English, and the German title is “Also sprach Zarathustra”, the system can begin to understand that “thus spoke” can be translated with “also sprach”. (This approach would even work for metaphors – surely, Google researchers will take the longest available phrase which has high statistical matches across different works.) All it needs is someone to feed the system the two books and to teach it the two are translations from language A to language B, and the translator can create what Franz Och called a “language model.” I suspect it’s crucial that the body of text is immensely large, or else the system in its task of translating would stumble upon too many unlearned phrases. Google used the United Nations Documents to train their machine, and all in all fed 200 billion words. This is brute force AI, if you want – it works on statistical learning theory only and has not much real “understanding” of anything but patterns.

One can suspect Google will release their new translation system soon (possibly, this or next year). The question is: what will they do with it – where will they integrate it – and what side-effects would it have? If via Google we get our universal language, would that resolve many global problems by fostering cross-cultural understanding, like Zamenhof was hoping for? Here is a speculative list of translation applications Google might implement; the key is *auto-translation*.

The Google Translation Service

This one is the most obvious: Google will still allow you to translate any document from their search results by the click of a link. What might be less obvious is that they might enable you to search foreign languages in your native language. All translating would be done behind-the-scenes, so that when you search for “thus spoke”, you might as well get results which only contain “also sprach.”

The Google Browser

If Google ever releases their own browser, they could seamlessly integrate translations of foreign languages; the user would just have to define what languages she reads fluently. It would be the Google Auto-translator (and surely it would be attacked using similar arguments than those brought forth against Google’s auto-linking.) And if it's not a Google Browser, it would be a Google Toolbar feature.

Now imagine this: you specified you speak English only. What does the Google Browser do when it encounters a Japanese page? *It will show you an English version of it.* You wouldn't even notice it's Japanese, except for text contained within graphics or Flash, and a little icon Google might show that indicates Auto-translation has been triggered. After a while, you might even forget about the Auto-translation. To you, the web would just be all-English. Your surfing behavior could drastically change because you're now reading many Japanese sources, as well as the ones in all other languages.

You are now enabled to get a better understanding of cultures outside your own country. Would there be any negative side-effects? Well, one for sure: people would have less incentive than before to learn foreign languages. And as soon as they encounter a foreign speaker in real life, they'd be just as lost as before.

The Google Instant Messenger

The GIM (Google's Instant Messenger) could be a chat application – web based, of course – which automatically translates from and to any language. You can now chat all around the world and get to know people where before you'd have run against the language barrier.

The Google Babelfish

This would be the most advanced implementation of the Google Translator. It would be a smart device you plug-in to your ear, and it would have speech recognition and Auto-translation built in. You can now visit a foreign country and understand people who talk to you in languages you never learned. For them to understand you as well, either they would also have a Google *Babelfish*, or there would be the need of a second gadget you speak into, which then translates what you said. While the needed text-to-speech and speech-to-text technologies are far from perfect at the moment, they are still realistic possibilities.

[Thanks to Alex Ksikes and Dominik Schmid for our brainstorming session.]

Google Translator: The Universal Language by Philipp Lenssen