Building a Universal Translation Service

Limits of machine translation

Statistical methods rely on large collections of parallel texts. These exist in only a few sets of languages, almost none of which are languages of developing countries.

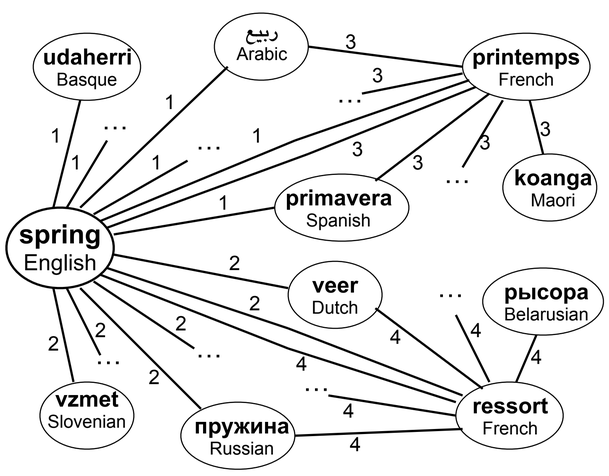
Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему.

≅

Les familles heureuses se ressemblent toutes; les familles malheureuses sont malheureuses chacune à leur façon.

An alternative: lexical translation via a translation graph

A database is constructed from hundreds of machine-readable bilingual, multilingual, and monolingual dictionaries. Inference algorithms consolidate senses and discover indirect translations.

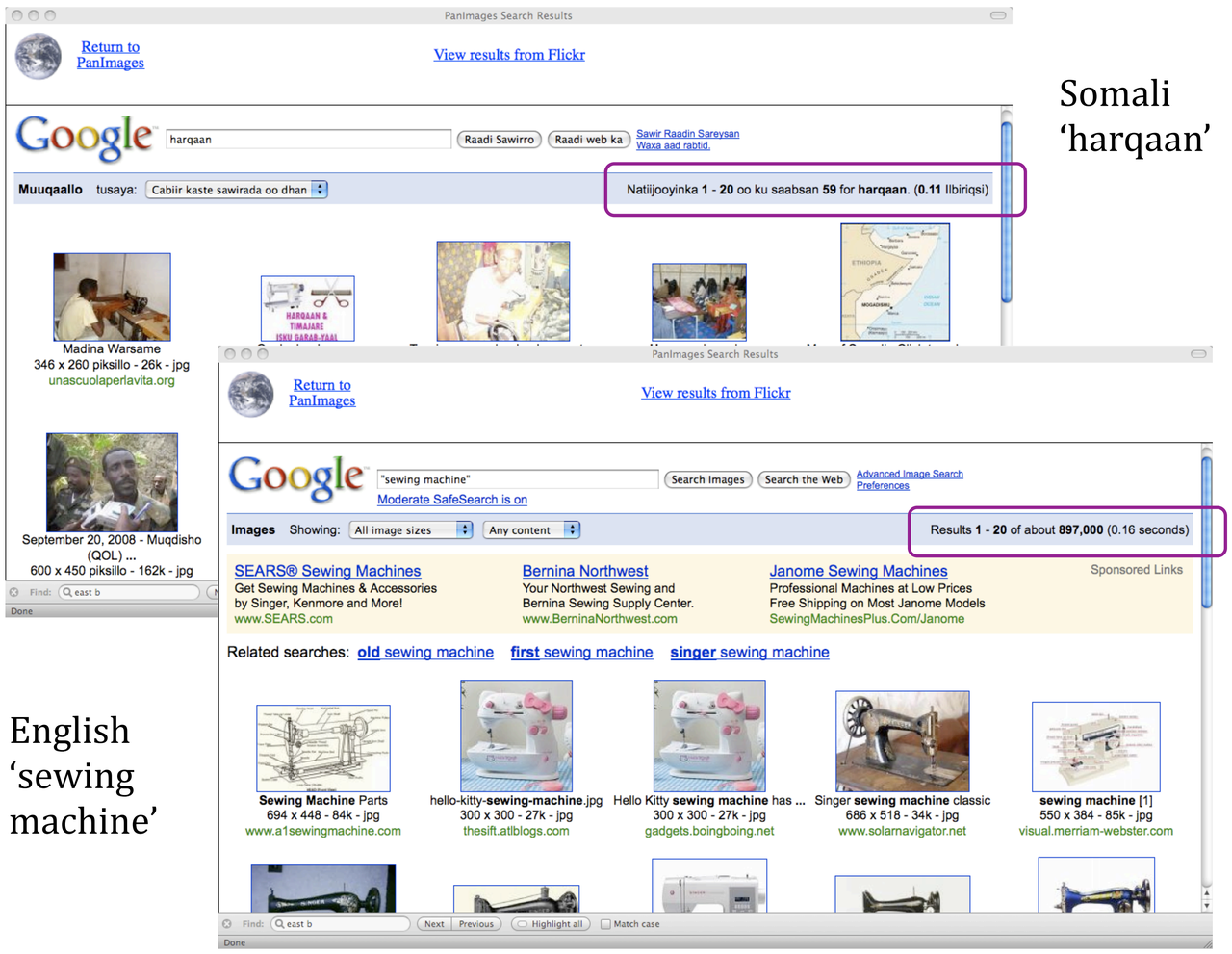


In this example, an algorithm infers that there is a high probability that word senses 1 and 3, coming from different dictionaries, are equivalent. This permits translating Basque ‘udaherri’ into Maori ‘koanga’.

An application: PanImages

At panimages.org, the user enters a word or phrase in one of several hundred languages. PanImages submits translations to Google Images and Flickr.

- Expands the results available to speakers of under-resourced languages.
- Finds alternatives to cross-lingual homonyms (e.g, English ‘bread’ instead of French ‘pain’).
- Finds alternatives to ambiguous words (e.g., Gujurati ‘વસંતઋતુ’ instead of English ‘spring’).
- Allows for cross-cultural discovery: A Thai ‘อาหารเช้า’ looks different from an English ‘breakfast’.

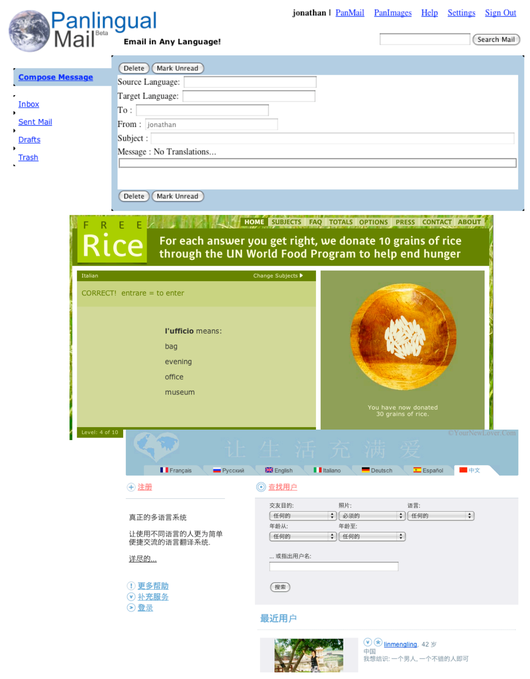


What next?

Expand and improve the resource

The database contains over 12 million words in more than 1,200 languages. But several thousand languages are not represented yet. A user interface (http://panlex.org/u) allows uploading dictionaries and editing translations.

Work is ongoing to improve the structure and performance of the database and to develop an API.



Develop more applications

A Web application under development will allow users who don’t share a language to exchange messages composed entirely of dictionary words. Future uses could include automatic UI localization, multilingual games, and social-networking applications.

Volunteers and collaborators are welcome!

Utilika Foundation • Turing Center, University of Washington

http://utilika.org • http://turing.cs.washington.edu