



MultiLingual

Language | Technology | Business

MultiLingual Computing, Inc. • 319 North First Avenue, Suite 2 • Sandpoint, Idaho 83864-1495 USA • 208-263-8178 • Fax 208-263-6310

Tech

Multilingual search with PanImages

Susan M. Colowick

Finding images on the internet can sometimes be tricky. Even though a growing number of images are tagged with keywords, search engines usually look for search terms only in the text that appears near an image on a web page. That text may have little or nothing to do with the content of the image.

Another obstacle to effective searching is ambiguity. The search term entered may have multiple meanings. A search for *lock* or *locks*, for example, will retrieve images of hair, canal segments and padlocks; a search for *spring* will retrieve images of metal coils, water sources and fields of flowers.

If image searching is so hard for people who are fluent in English – today’s dominant web language – what hope is there for those who speak Guarani, Sardinian or one of the other thousands of languages that are poorly represented online? To address this issue, researchers at the University of Washington’s Turing Center (<http://turing.cs.washington.edu>) have developed an application that makes it easier for people, especially speakers of minority languages, to locate images on the web.

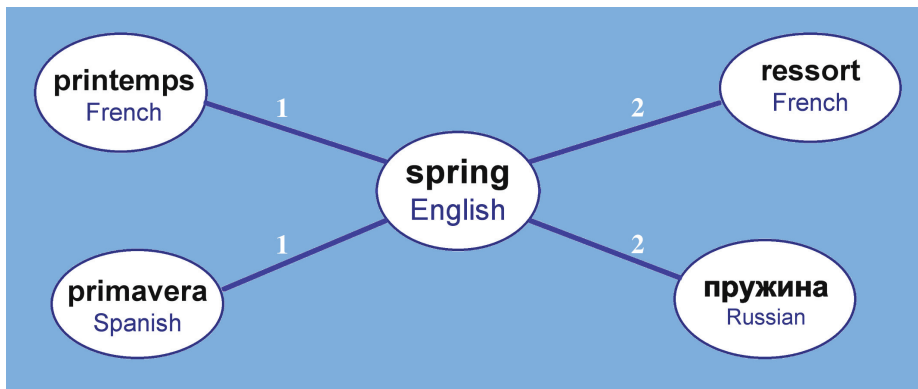


Figure 1: Nodes in the translation graph are ordered pairs (w, l), where w is a word in language l. Lines in the graph indicate translations between words. Each line is labeled with a word sense ID.

In the PanImages system (www.panimages.org), the user can enter a word or phrase in any of several hundred languages. When searching in a particular language by using the Advanced Interface, the user can take advantage of an auto-complete function to select terms that exist in the database. The system then shows the available translations of the user’s term, often giving a choice of several senses. The user chooses a sense and then selects one or more translations of the original term. The selected translations are used to perform a search in Google Images and Flickr.

Source of the data

The millions of expressions in PanImages come from TransGraph, a continually expanding database developed

Susan M. Colowick is a writer, editor and researcher who worked for 10 years as a reference librarian.



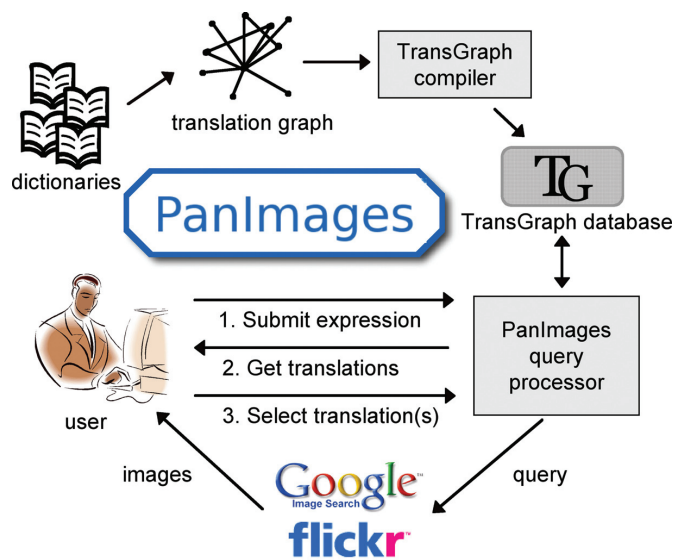


Figure 2: Data from bilingual and multilingual dictionaries make it possible to search for images in hundreds of languages.

at the Turing Center. TransGraph gets its raw data from a translation graph of expressions and translations, which are acquired from machine-readable dictionaries.

Each node in the translation graph represents a word in a particular language, and each edge represents a translation with a particular word sense (Figure 1). The probability that two expressions share a sense is determined by a TransGraph algorithm that examines the paths in the graph between one expression and the other. The expressions, senses and sense probabilities are then stored in TransGraph. When a user searches for a word or phrase, PanImages displays the terms that have the greatest probability of being accurate translations (Figure 2).

Another source of errors is the incorrect parsing of dictionary entries, which can happen when there are inconsistencies in the formatting of dictionaries. A parsing error may cause a usage note, cross-reference or other text to be mistakenly identified as a translation of a term. This leads to such unhelpful translations as the English *see Día de la Ascensión* for the Spanish word *ascensión*.

TransGraph contains data from more than 350 machine-readable bilingual and multilingual dictionaries, including 12 Wiktionaries. Because the scale of the project requires automatic processing of the dictionaries, the data can end up with a few quirks. One problem is that the dictionaries often do not distinguish among the various senses of a word. For example, the French Wiktionary lists 12 senses for the word *caisse*, but, when translating the word into other languages, it treats all the senses as one, causing the list of English translations to include *box*, *car*, *motor*, *money-box* and six other terms.

An in-depth description of the research behind PanImages appears in the paper “Lexical Translation with Application to Image Search on the Web” by Oren Etzioni, Kobi Reiter, Stephen Soderland and Marcus Sammer, which was presented at the Machine Translation Summit XI in September 2007.

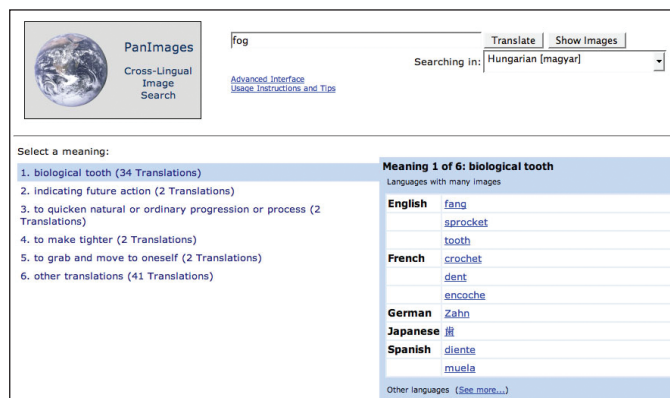


Figure 3: By choosing translations of the Hungarian word *fog*, the user gets a few images of teeth.

A collaborative resource

As TransGraph’s processing methods improve, errors will occur less often. However, as with most fully automatic systems, some human intervention will always be required. Fortunately, errors in TransGraph can be fixed as soon as they are discovered because PanImages is a collaborative tool. Any user can add or delete expressions and translations, thus improving and growing the database.

Volunteers have also assisted by translating the user interface. PanImages can currently be used in about 50 languages, including Asturian, Estonian, Tamil and Thai. Anyone who wishes to add an interface language can do so by submitting a form with translations of the phrases used in the interface.

For now, PanImages is an open system where anyone can make changes. In the future it may become preferable to have people register and log in if they want to help build the database. The addition of a user forum is also being considered.

What PanImages can and cannot do

The difficulty of searching in an under-represented language can be compounded by the fact that a common word in that language could mean something completely different in one of the better-represented languages. For example, the word *fog* in Hungarian means *tooth*, but a search in Google Images turns up mainly photos of misty landscapes. With PanImages, a Hungarian user can select one or more translations of *fog* in languages that have a substantial number of images of teeth (Figure 3).

With its sense-distinguished translations, PanImages makes it possible for speakers of any language to identify terms that are less ambiguous than those in the user’s native language. After getting translations of *spring* or *lock*, for example, the user can choose the preferred meaning and find a less ambiguous word, such as the French *printemps* or the Spanish *esclusa*.

PanImages can also be used to compare visual depictions of customs from around the world. One easy way to do this is to search for terms such as *breakfast*, *wedding* or *market* in one’s own language and then click on each of the translations provided (Figure 4).

Of course, searching with just one word – in any language – can make the task of finding images very frustrating. TransGraph includes many multi-word phrases, but not nearly enough to express every concept that one might want to translate. Without the ability to combine lexemes (words and phrases), such as *purple + dress* or *boy + jumping*, the user must wade through many images to find ones that are relevant. Future work may include developing a system for translating sets of lexemes as phrases or even sentences.

The Turing Center

The Turing Center has the broad mission of enabling communication and collaboration among humans and computers. The staff and affiliated faculty include researchers in computer science, linguistics and electrical engineering. Professor Oren Etzioni of the UW Department of Computer Science and Engineering is the center's director.

The Turing Center was established in 2005 with support from the Seattle-based Utilika Foundation (<http://utilika.org>). The center also receives funding from Google, the National Science Foundation, the Office of Naval Research and the Defense Advanced Research Projects Agency.

Another Turing Center project, closely related to TransGraph and PanImages, is PanLexicon. This project involves taking expressions from machine-readable dictionaries and then analyzing their occurrences in the text in order to determine their meanings. The center's work also includes the development of methods for extracting knowledge from the web, such as the KnowItAll project, and for semi-automatically producing grammatical models of languages for use in translation among multiple languages (the LinGO Grammar Matrix).

The future of collaborative lexical resources

The PanImages website was made public in September 2007. In the first two months after its debut, the site received 84,000 visits. Users came from more than 150 countries and submitted queries to Google Images and Flickr in more than 1,000 languages.

Since November 2007, visitors to the PanImages site have also been able to try a multilingual version of Google's Image Labeler. As with the monolingual (English-only) game, two online players are randomly paired, and each tries to guess the label that the other will apply to a particular image. One user

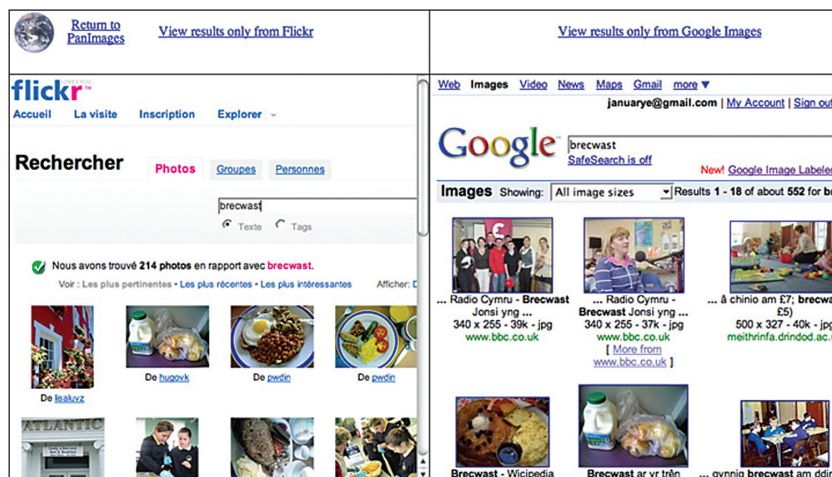


Figure 4: *Brecwast* (Welsh) may look very different from *breakfast*.

can be typing words in Thai while the other uses Turkish; the system will compare the translations to see if there's a match.

The lexical translation provided by TransGraph has many other potential uses. One possibility is the translation of tags applied by users to online content at sites such as del.icio.us, Flickr and Technorati. Ideally, TransGraph could become a resource for professional translators. Before it can be truly useful for that purpose, the database needs further enhancement and expansion. The developers will continue to refine the algorithms and parsing methods, but they hope that the users of PanImages will provide much of the necessary enrichment of the data.

Researchers at the Turing Center invite contributions of additional bilingual and multilingual dictionaries, particularly dictionaries for languages that are now under-represented in the TransGraph database. They also welcome collaborators on the development of multilingual lexical resources. They see collaboration – with web users as well as with other researchers – as the key to making progress toward the goal of universal interactivity. **M**

References:

"Lexical Translation with Application to Image Search on the Web" by Oren Etzioni, Kobi Reiter, Stephen Soderland and Marcus Sammer, which was presented at the Machine Translation Summit XI, 2007. (Proceedings, pp. 175-82, <http://turing.cs.washington.edu/papers/EtzioniMTSummit07.pdf> or www.mt-archive.info/MTS-2007-Etzioni.pdf)