Wikipedia seeks to speak your language
Using language detection to search the right Wikipedia
by Trey Jones, Software Engineer, Wikimedia Foundation

Wikipedia readers speak many languages. Sometimes they search for phrases not in the language of the current wiki. This unfortunately leads to poor search results. A recent survey of English Wikipedia queries identified searches in 40 different languages!

To better help people find what they are looking for, the Wikimedia Discovery department has rolled out language identification software to detect and then redirect unsuccessful searches in other languages (if needed). Searches will be redirected to a Wikipedia that is more likely to provide results, and those cross-wiki results will be displayed along with the local-wiki results, if any. We’ve started by enabling the language identification and search redirection for English, French, German, Italian, and Spanish Wikipedias, with more to come.

“Foreign language” searches on a given Wikipedia can work just fine—phrases and mottos (scientia est lux lucis, ἰνεωμέτρητος μηδείς εἰσίτω), places (さいたま市, गृहस्थारप), people (גננת הבן, Екатерина Гордеева), events (حصل, होली), and other things (한글, లోగింగ్, Հայերեն)—but they don’t work quite so well in many cases. Hence the idea to try to detect the language of unsuccessful searches—i.e., those that get fewer than three results—and redirect them to another Wikipedia when appropriate.
The automatic language detection works similarly to the way you might identify a language you don’t know—by looking for bits of text that “look like” a particular language. If you see lots of *les*, *aux*, *l’*, and *-eux*, that’s probably French. *Los*, *las*, *-o*, *-a*, and an *ñ* or two seems most likely to be Spanish. Words with *sch*, *ah*, *ü*, *pf*, and *aus* might make you think of German. The automatic language detection looks at a lot of these snippets, and knows a lot more about how common they are across languages—from longer combinations down to individual letters and characters.

While the best Wikipedia articles are written in a formal, encyclopedic style, people often search using much more informal language. They use more nouns and fewer verbs because they tend to search for *things*. They ask more questions because they want to know something. They omit diacritics in languages that use them because typing is hard. For these and other, similar reasons, we’ve built our language detection models using actual user searches, so they are more attuned to these deviations from more formal writing—and it’s paid off in more accurate identification.

Language detection can be difficult—especially on short snippets of text, like many Wikipedia searches. Of course, that’s what makes the task of trying to make it work so much fun! Sometimes, though, it’s impossible to tell what language is intended. The word *cubo*, for example, might look like it’s Spanish to you—unless you know Portuguese… or Italian, or Galician, or even Latin! The shorter the search, the more likely it is to be truly ambiguous.

One way to get around the ambiguity problem—and the more general problem of confusion among similar words and languages—is to limit the options for language identification. For example, in a survey of poorly-performing searches on Italian Wikipedia, we found a small number of searches in Spanish. However, Spanish resulted in four times as many wrong identifications as correct ones (just among Italian Wikipedia searches), because the vast majority of searches on Italian Wikipedia are in fact in Italian. We disabled Spanish detection on Italian Wikipedia because it hurt more than it helped. And of course not considering languages that almost never show up on a given Wikipedia—i.e., are not in a sample of one thousand to ten thousand searches—improves accuracy and speed. We’ve done this kind of analysis for the Wikipedias we’ve deployed language detection to so far.
Old results on the English Wikipedia for a search in Russian:

Search results

You may create the page "Первым экспериментом".
There were no results matching the query.

Screenshot by Deborah Tankersley, CC BY-SA 3.0.
New results on the English Wikipedia for a search in Russian (after the addition of language detection):

We have other language detection ideas and plans in the works. The next batch of Wikipedias to have language detection enabled will probably include Indonesian, Japanese, Portuguese, and Russian. We’ll be investigating ways to more easily bring language detection to many more Wikipedias, and to projects other than Wikipedias. We’re also thinking hard about ways to improve language detection with smarter measures of confidence. Another not-quite-language form of detection we’re exploring is recognizing when someone has typed a search in one character set while using the keyboard for another character set—early experiments with English and Russian are promising!
You can find more technical details about our language detection module, TextCat, on MediaWiki. If you want to use it yourself, PHP and updated Perl libraries are available; they include the search-based language models we are using on Wikipedia, along with dozens of additional language models based on the more formal text in Wikipedia articles. You can also directly test TextCat with our online demo, which lets you try all the different language models on your own text, and includes tutorials and lots of additional information about the TextCat’s internal workings.