Statistical analysis of multilingual dictionary queries - a case study of Lingro.com

Łukasz Olejnik (Adam Mickiewicz University, Poznań), Artur Janc (Worcester Polytechnic Institute)

Dictionaries are the primary linguistic reference tool for the majority of the population; understanding their usage patterns can provide major insight into their educational applications and provide lexicographers with detailed understanding of the most relevant entries to help them identify vocabulary with missing or inadequate translations.

Until recently, little such data was available to researchers. However, in the last few years the advent of on-line dictionaries made it possible for maintainers of dictionary websites to gather detailed usage information.

In my talk, I will present detailed statistics based on the data collected using the Lingro on-line dictionary (http://lingro.com). Lingro is a dictionary website with a global reach, providing free multilingual translations and language study tools. Throughout the last 18 months, Lingro has provided over 2 million translations in 11 languages to Internet users in more than 50 countries.

In particular, I will give insight into the most often used source and target languages, along with frequency data to determine the most often looked-up words and phrases. I will also present the relative frequencies of different query categories (such as single words, phrases, and entire sentences), and the availability of results for each category. In addition, I will present other statistical information, including the average number of dictionary queries per user per day, percentage of users adding translations to the system, and other similar information.

For each of the described quantities I will attempt to highlight relative similarities and differences between users in different countries and regions, based on GeoIP data.

References:

-