Half of messages on Twitter are not in English
Japanese is the second most used language

A quantitative and semantic study of Twitter

Paris, France — February 24, 2010 — English messages account for only half of the messages on Twitter, a study of 2.8 million tweets has revealed. The analysis, carried out by Semiocast, showed that the top 5 languages used on Twitter are: English, Japanese, Portuguese, Malay and Spanish.

The study was conducted on messages gathered over a period of 48 hours, from February 8 to February 10, 2010 to establish Twitter's most used language ranking. The messages were processed with Semiocast's analysis tools which can identify the language used in short messages among 41 languages in all major writing systems (including Greek, Hebrew, Chinese, Korean, Tamil, Devanagari, …).

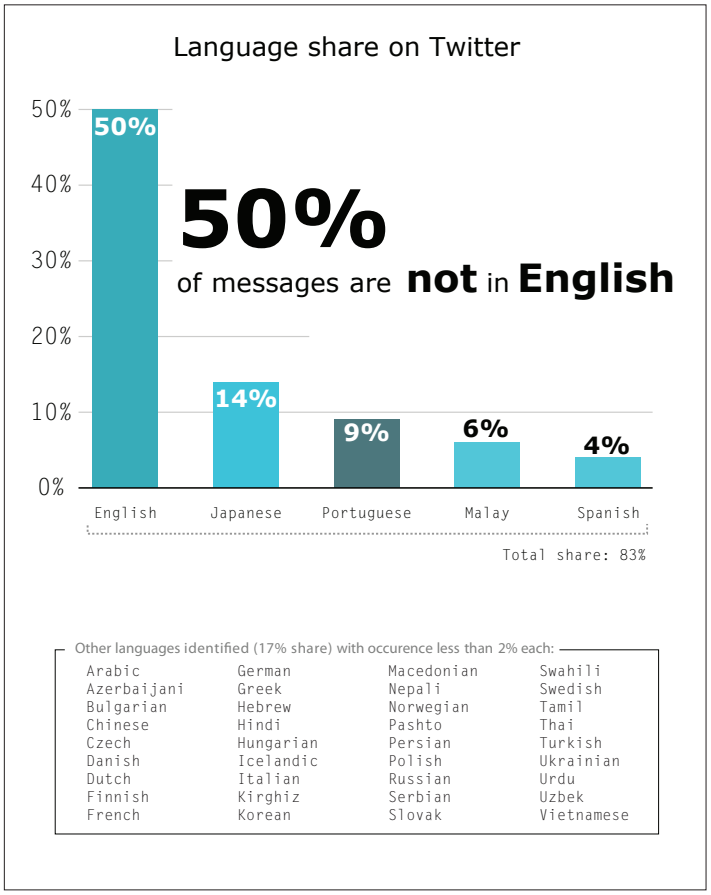
English is still the most used language on Twitter, with 50% of messages, reflecting its high penetration rate in English-speaking countries and the tendency of Twitter users that are non-native English speakers to tweet in English. This is nonetheless a sharp decline from the two-third share English was representing in the first half of 2009. In the near future, English's share should drop even further as the strongest growth for Twitter is expected to come from non-English speaking countries.

Japanese is the clear second with 14% of messages. This figure proves Twitter's continuous popularity in Japan and confirms Japan has been the first milestone of Twitter's international development.

The third most used language is, without surprise, Portuguese, mirroring the huge success of social networks in Brazil. Portuguese already makes up 9% of all messages, about 4.5 million messages per day.

The rapid adoption of Twitter in Malaysia and Indonesia, where Twitter concluded partnerships with two mobile careers, shows in their rankings. Malay languages, including Bahasa Malaysia and Bahasa Indonesia, now represent the fourth most used language on Twitter, with 6% of messages. This means that every day about 3 million messages are exchanged in Malay-related languages.

Spanish comes fifth with 4% of all messages, thanks to the large number of Spanish speakers around the world, in particular in Twitter's home country, the U.S.



The ranks six to eight are occupied by major European languages, namely Italian, Dutch and German, each accounting for about 1% to 2% of total messages. As for our home base, French represents a little less than 1% of total messages (rank 11).

Language detection is a prerequisite to semantic analysis: even the most standard indicator, such as a mood associated to a message, can only be assessed with some precision after the language of the message has been correctly identified.

About Semiocast

Semiocast is a company based in Paris, France, providing data intelligence and research on the real-time Web. Semiocast helps companies measure and evaluate reactions of consumers to a campaign or product launch, to understand what consumers are saying about their products, services and brands, and to engage in real-time Web conversations happening today mostly on Twitter and Facebook.

Semiocast's offer is based on proprietary technologies and tools to index real-time web conversations and perform semantic analysis of conversations. With these technologies and its expertise, Semiocast provides its customers with in-depth insights into thousands of conversations concerning their brands and products, solutions to monitor these conversations and tools to engage with consumers.

More information : semiocast.com
Contact : Paul Guyot (paul@semiocast.com — +33.175000290)