A FRANCO-GERMAN INSTITUTE

Breaking Language Barriers

The Institute for Multilingual and Multimedia Information (IMMI), a new Franco-German institute founded in Orsay, will use new technology to link spoken words with their written counterparts and navigate from one language to another.

It's not quite the tower of Babel, but a brand new Franco-German institute has high hopes of breaking down language barriers like never before. This Orsay-based institute called IMMI,[1] dedicated to multimedia language technologies, brings together the Rhenish-Westphalian Technical University Aachen (RTWH) and the University of Karlsruhe (UKA), both in Germany, with CNRS' LIMSI,[2] a computer science laboratory specialized in language and speech processing. The joint institute will house powerful tools for its researchers, based on the newest available technologies. "Those developed by both the institute members and the other partners of the Quaero program, started in 2008 (*see box*), will be able to transcribe a speech or conversation, recognize the language and translate it, identify the speaker by his face or voice, and make automatic summaries of texts or website contents," explains IMMI director Joseph Mariani. To do this, the institute's scientists–who will eventually be a hundred strong–will work on the development of new language technologies. "Notably on speech and language processing, machine translation (whether this be text to text, speech to text, or even speech to speech), processing of multilingual documents or indexing of multimedia documents," continues Mariani, "because these are the skills of the three founding laboratories."

The research teams bring together computer scientists, linguists, sociologists, and specialists in ergonomics, with a common methodological approach: "statistical learning and evaluation based on corpora," explains Mariani. In simple terms, to improve the automatic textual transcription of sound data (speeches or radio broadcasts, for example), sound files and their corresponding transcriptions are fed into the system, which analyses them and "learns" to associate a given sound with its transcription. The larger the sound and text corpora provided for training, the better the statistical model–and the better the system's transcriptions.

Preparations for IMMI have been underway for some time. As early as December 2007, the three founding partners created an International Joint Unit (UMI) to facilitate the institute's management. Then, last December, to bring their research under a common umbrella, they formed a European Associated Laboratory (LEA IMMI-Labs), to which the Paris-Sud University was associated. Funding is provided by the Quaero program, CNRS, RWTH, UKA, Paris-Sud University, the Essonne departmental council, and the Digiteo Advanced Thematic Research Network, which jointly contribute to covering the costs of construction, computing equipment and functioning. Within three years, it is hoped that IMMI researchers will be working in their new 3000 m² building, built close to LIMSI, in Orsay.

On completion, the institute will be one of the world's largest centers dedicated to this field of research. Its role will also be very important in Europe, where more than 20 languages are spoken. "We seek to develop technologies that will allow European citizens to use their own tongue and easily switch from one language to another," says Joseph Mariani. The laboratory could, for example, create natural language processing or translation tools for the 23 official European languages, a much needed service also for numerous European institutions like the European Commission, Parliament, Patent Office, Digital Library, and Security Agency.

Virginie Lepetit

1. Institute for Multilingual and Multimedia Information.
2. Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (CNRS / Universités Paris-VI and XI).

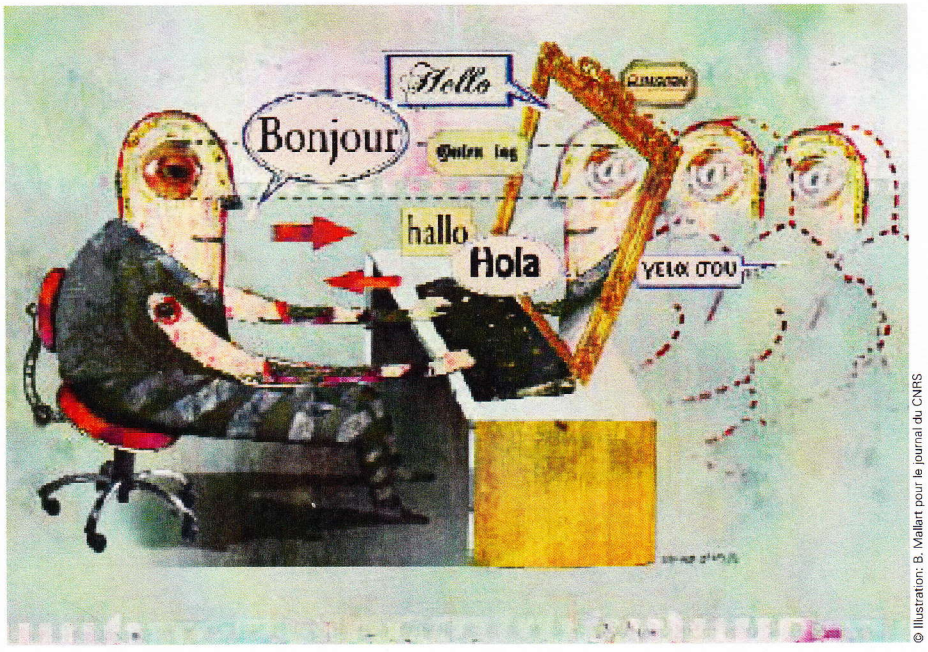
A MULTIMEDIA PROGRAM

The Quaero program aims to produce advanced information technologies to process multimedia data (text, speech, music, images, and video) and use them to develop innovative applications: search engines, communication portals, content digitization, personalized video, digital media assets management, etc. Quaero is a five-year research program (2008-2013) involving more than 20 partners from public research and industry. It has a budget close to €200 million.

V.L.

> http://www.quaero.org

CONTACT INFORMATION
→ Joseph Mariani
IMMI, Orsay.
joseph.mariani@limsi.fr



© Illustration: B. Mallart pour le journal du CNRS