

PARTENARIAT FRANCO-ALLEMAND

Un institut pour naviguer d'une langue à l'autre

Utiliser les nouvelles technologies pour passer de l'oral à l'écrit et d'une langue à une autre, c'est l'ambition d'un nouvel institut franco-allemand créé à Orsay.

Ce n'est peut-être pas une tour de Babel où les hommes ne parleraient qu'une seule langue, mais le tout nouvel Institut franco-allemand des technologies multilingues et multimédias de l'information (IMMI), à Orsay, aspire quand même à abattre quelques frontières linguistiques. Ses outils ? Les nouvelles technologies. « *Celles développées par l'Institut et par les autres partenaires du programme Quaero (lire l'encadré), lancé en 2008, pourront transcrire un discours ou un dialogue, reconnaître la langue utilisée et la traduire, identifier le locuteur d'après son visage, ou encore résumer automatiquement des textes ou des contenus de sites* », explique Joseph Mariani, directeur de l'IMMI, qui associe l'Université technique de Rhénanie Westphalie (RWTH), à Aix-la-Chapelle, l'université de Karlsruhe et le CNRS – à travers le laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi) ¹.

Pour cela, les scientifiques de l'IMMI – qui seront à terme une centaine – poursuivront le développement de nouvelles technologies linguistiques. « *Notamment sur le traitement du langage, le traitement de la parole, la traduction automatique (que ce soit texte à texte, parole à texte, ou parole à parole), le traitement de documents multilingues et l'indexation de documents multimédias*, poursuit Joseph Mariani. *Parce que ce sont les compétences des trois laboratoires fondateurs.* » Ces derniers, qui regroupent des linguistes, des informaticiens, des sociologues et des spécialistes de l'ergonomie partagent en effet la même approche méthodologique. « *Celle de l'apprentissage, de l'évaluation et de la validation par des méthodes statistiques à partir d'ensembles de documents (des corpus)* », précise Joseph Mariani. En clair, pour mettre au point une transcription écrite automatique de données sonores (discours ou émissions radio, par exemple), on fournit au système des fichiers sonores et les transcriptions correspondantes. Il s'en nourrit, les analyse et « apprend » à associer le bon son à la bonne transcription. Plus le corpus de fichiers donnés au système est grand, plus la base statistique est importante et meilleures sont les transcriptions.

La naissance de l'IMMI a été longuement préparée. Les trois partenaires fondateurs ont créé en décembre 2007 une unité mixte internationale pour faciliter la gestion de l'Institut. Puis, en décembre dernier, ils ont fondé un laboratoire européen associé (LEA), l'IMMI-Labs, qui fédère les efforts de recherche des trois partenaires,

auxquels est aussi associée l'université Paris-Sud-XI. Outre les financements apportés par le consortium Quaero aux trois partenaires, le CNRS, le Conseil général de l'Essonne et le réseau thématique de recherche avancée Digiteo contribuent à couvrir les coûts de la construction et des équipements informatiques. D'ici à trois ans, les chercheurs de l'IMMI seront en effet installés dans un nouveau bâtiment de 3 000 m² construit à Orsay, à proximité du Limsi, dans le sud de l'Île-de-France.

À terme, l'Institut sera l'un des plus grands pôles de recherche mondiaux sur ces thématiques. Et sa place en Europe, où plus d'une vingtaine de langues cohabitent, est loin d'être anodine. « *Nous cherchons à développer des technologies qui permettent à chaque citoyen européen d'utiliser sa propre langue et de passer facilement d'une langue à l'autre* », rappelle Joseph Mariani. Le laboratoire pourrait par exemple développer des outils de traitement automatique ou de traduction des 23 langues européennes officielles et répondre aux besoins de nombreuses institutions européennes : Com-

mission ou Parlement, office européen des brevets, bibliothèque numérique ou agences de sécurité...

Virginie Lepetit

1. Laboratoire CNRS / Universités Paris-VI et XI.

LE MULTIMÉDIA AU PROGRAMME

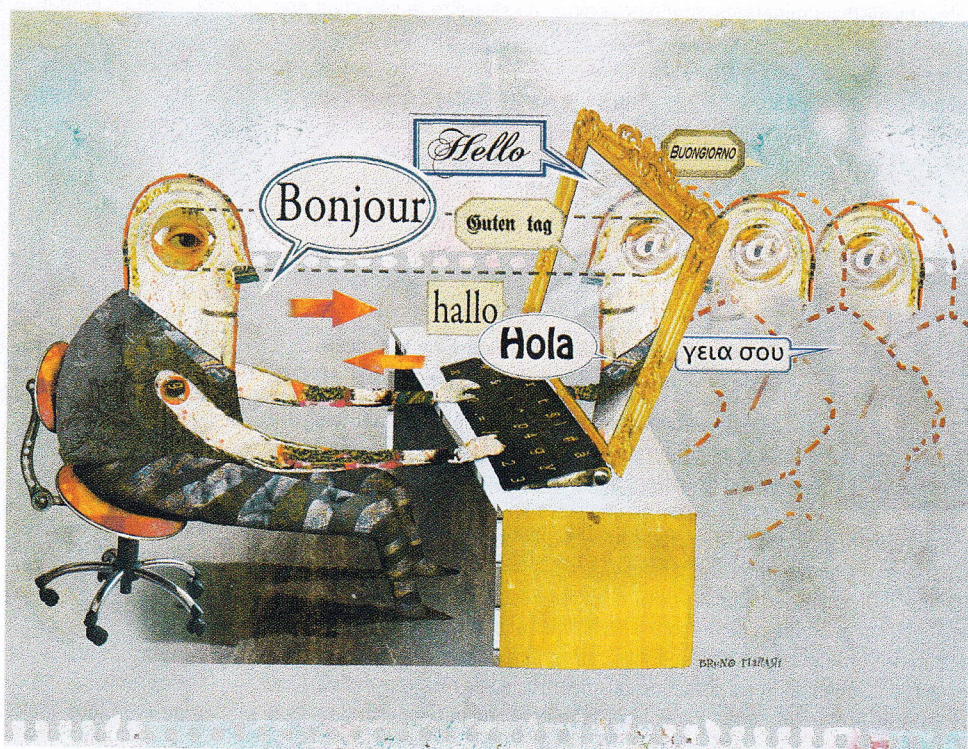
Quaero est un programme de recherche de cinq ans (2008-2013), qui associe plus d'une vingtaine de partenaires de la recherche publique et de l'industrie et bénéficie d'un budget proche de 200 millions d'euros. Son but est de produire des technologies avancées de traitement de l'information multimédia (texte, parole, musique, images, vidéos) et de développer des applications innovantes les utilisant : moteurs de recherche, portail de communication, numérisation des œuvres, navigation dans les contenus audiovisuels, gestion des droits numériques...

V.L.

CONTACT

→ Joseph Mariani

Institut des technologies multilingues et multimédias dans l'information (IMMI), Orsay
joseph.mariani@limsi.fr



© Illustration : B. Mallart pour le journal du CNRS