Traitement informatisé des langues, langues nationales et partenariat francophone

Je participe à cette manifestation avec plusieurs casquettes à la fois : je porte celle de chercheur à l'Université de la Manouba, Tunisie, celle de membre du Groupe de recherche expérimentale sur les systèmes informatisés de la communication de l'Université Bordeaux 3 (GRESIC) et celle de secrétaire général de l'Association internationale des écoles des sciences de l'information. Chacun de ces profils m'a conduit, au fil des ans, à faire des recherches sur les questions du multilinguisme, l'un des axes fondamentaux des travaux menés dans le secteur du traitement informatisé des langues (TIL).

Je commence toujours ce genre d'intervention en faisant un recours à l'exemple mythologique de la tour de Babel. Comme vous le savez, le mythe de Babel prétend que les hommes avaient décidé un jour de construire une énorme tour pour se rapprocher du Ciel. Devant cette arrogance humaine, Dieu décida de punir leur vanité en faisant en sorte qu'ils ne puissent plus parler la même langue et, de la sorte, ne plus se comprendre. L'édifice resta inachevé. Et la Terre se transforma en un univers multilingue et multiculturel.

Quels seront les effets des technologies de l'information (TI) dans un contexte pareil ? Renforceront-elles les partenariats linguistiques et les échanges interculturels pour essayer de pallier la conséquence de ce « péché originel ? » Ou au contraire, amplifieront-elles les inégalités entre les langues des groupes majoritaires (dominantes) et celles des communautés minoritaires (soumises), comme les Basques, les Kurdes ou les Berbères ?

Je crois qu'une approche multilingue pourrait assurer que la première réponse prévaut. Je qualifie cette approche de *multilinguisme numérique du Sud.* Permettez-moi d'expliquer ce que j'entends par cette idée.

Actuellement, on remarque que des langues comme l'arabe, l'hébreu ou l'ourdou sont des langues statistiquement importantes sur le plan démographique, mais qu'elles sont *technologiquement minoritaires,* en ce sens que leur traitement informatique n'est pas très avancé par comparaison avec les acquis des langues latines.

D'abord, lorsque ces langues sont prises en compte, elles ne sont jamais les seules composantes d'une application. Par exemple, lorsqu'un outil traite l'arabe, c'est qu'il a l'anglais ou le français comme langue pivot. La situation est bien différente dans le cas de ces deux dernières langues : la nature unilingue (française ou anglaise) quasi dominante du marché mondial du logiciel atteste de ce déséquilibre linguistique et de l'autonomie de certaines langues devant la dépendance d'autres.

On note également que les développeurs de logiciels et les fabricants de matériel informatique ne tiennent pas toujours bien compte, dans leurs produits, de certains aspects culturels, sociologiques ou même cognitifs importants pour les locuteurs du Sud. Par exemple, les logiciels d'origines *latines,* adaptés pour la langue arabe, présentent souvent un nombre assez important d'incohérences aux yeux de l'utilisateur arabophone type. Ces incohérences peuvent avoir de grandes répercussions sur l'acceptabilité, l'utilisabilité et l'efficacité même de l'interface.

La source du problème d'interface émane du fait qu'il existe deux types de multilinguisme : un multilinguisme *souple*, très répandu et couvert par la recherche, et un multilinguisme *lourd*, peu documenté et moins connu. Le multilinguisme souple renvoie à la coexistence de deux ou plusieurs langues latines ou germaniques (i.e. français, anglais) dans un seul espace applicatif ou logiciel donné. Il est relativement facile à gérer par les développeurs et à appréhender par l'utilisateur.

Le multilinguisme lourd, par contre, oppose deux ou plusieurs familles de langues dans le même contexte logiciel ou applicatif sur la base des deux points majeurs : le rendu visuel de la forme graphique des caractères (i.e. graphie latine ou arabe) et la directionnalité de l'acte graphique (directionnalité Gauche-Droite ou Droite-Gauche). Ce cas de figure est nettement plus difficile à gérer, ce qui explique les difficultés toujours enregistrées au niveau de la gestion d'interface homme-machine et des systèmes internes de traitement des données codées.

Ce phénomène de bidirectionnalité, par exemple, est très manifeste dans les cultures et les langues à directionnalité opposée par rapport à ce qui est courant dans les cultures occidentales à directionnalité Gauche-Droite. Dans la culture arabe, on prononce en théorie les nombres en passant des unités aux dizaines, aux centaines, aux milliers, etc., évoluant dans une directionnalité Droite-Gauche. En principe, on écrit ces mêmes nombres également de droite à gauche. Par exemple, le chiffre 1988 se prononce « huit quatre-vingt neuf cent mille » et s'écrit dans la même orientation Droite-Gauche « 1988 », ce qui

donne un état de conformité entre l'acte d'écriture, le mode de prononciation et le modèle de calculabilité arithmétique. Cette pratique constitue encore de nos jours l'une des marques de distinction de la bonne maîtrise de la langue arabe.

Pourtant, dans toutes les solutions multilingues proposées par les grandes instances internationales de normalisation, les chiffres sont exclusivement traités en partant du plus gros au plus petit, de gauche à droite. En français (et autres langues latines), le chiffre « 1988 » est prononcé « mille neuf cent quatre-vingt-huit » et rédigé dans une logique directionnelle Gauche-Droite « 1988 ». Ceci engendre un état de non-conformité entre l'acte graphique et le mode de prononciation d'une part (G-D), et le mode de calculabilité arithmétique de l'autre (D-G). Il est certain que les personnes d'origine arabe éduquées dans les écoles occidentales s'adaptent relativement bien à ce système directionnel Gauche-Droite ; sauf que ces personnes ne représentent qu'une infime portion de la population arabophone. Il faudrait, pour généraliser l'utilisation de l'ordinateur et des technologies numériques par cette dernière, s'adapter à ses attentes linguistiques et cognitives.

En somme, un certain nombre de questionnements se pose actuellement devant un déséquilibre évident entre les langues universelles :

- comment gérer l'aspect linguistique intensif du contenu informationnel sur le réseau (codage des langues) ?
- comment ont été abordés les deux principes de l'internationalisation (I18n) et de la localisation (L10n) dans les systèmes et les ressources d'information multilingues (langues complexes) ?
- quels sont les apports de l'I18n pour les langues « complexes » (c'est-à-dire arabe) au sein des systèmes d'information multilingues ?
- quelle forme de partenariat peut exister entre le français comme langue pivot et l'arabe comme sa première langue partenaire du Sud ?

Il est certes évident que le traitement automatique des langues du Sud n'est pas au même seuil d'importance que les langues du Nord. Il suffit de constater l'assise linguistique des produits logiciels dans le marché pour conclure que les langues dites « complexes » n'ont pas une présence massive et, si elles sont prises en compte, elles ne le sont pas toujours adéquatement. La Francophonie devrait s'efforcer de corriger ces lacunes. Quelle devrait être sa stratégie pour ce faire ?

Pour favoriser l'informatisation des langues du Sud, il faudra d'abord régler certains problèmes terminologiques. En effet, les langues partenaires sont habituellement plus pauvres à ce chapitre que celles du Nord. Bien sûr, il y a des causes endogènes à cela, mais il y a aussi des causes extérieures.

Il conviendrait aussi à la Francophonie, pour combattre l'hégémonie des grandes compagnies de logiciels, d'appuyer les initiatives de développement de solutions locales par le renforcement des initiatives d'internationalisation (I18n) et localisation (L10n). Ces initiatives devraient viser le traitement des langues médianes du Sud. Les langues médianes représentent un assemblage entre les langues officielles et les dialectes. Ces langues transversales s'adressent à tout le monde ; elles sont de plus en plus utilisées dans les journaux, dans les revues, etc.

Cependant, je suis réaliste et je me rends compte que la plupart des logiciels continueront de provenir des pays du Nord, en particulier des États-Unis. Je me permets cependant d'espérer que la localisation de ces outils sera faite, à l'avenir, par des personnes-ressources vivant dans le marché linguistique et culturel visé, plutôt que par des experts n'ayant qu'une connaissance théorique de la langue cible.

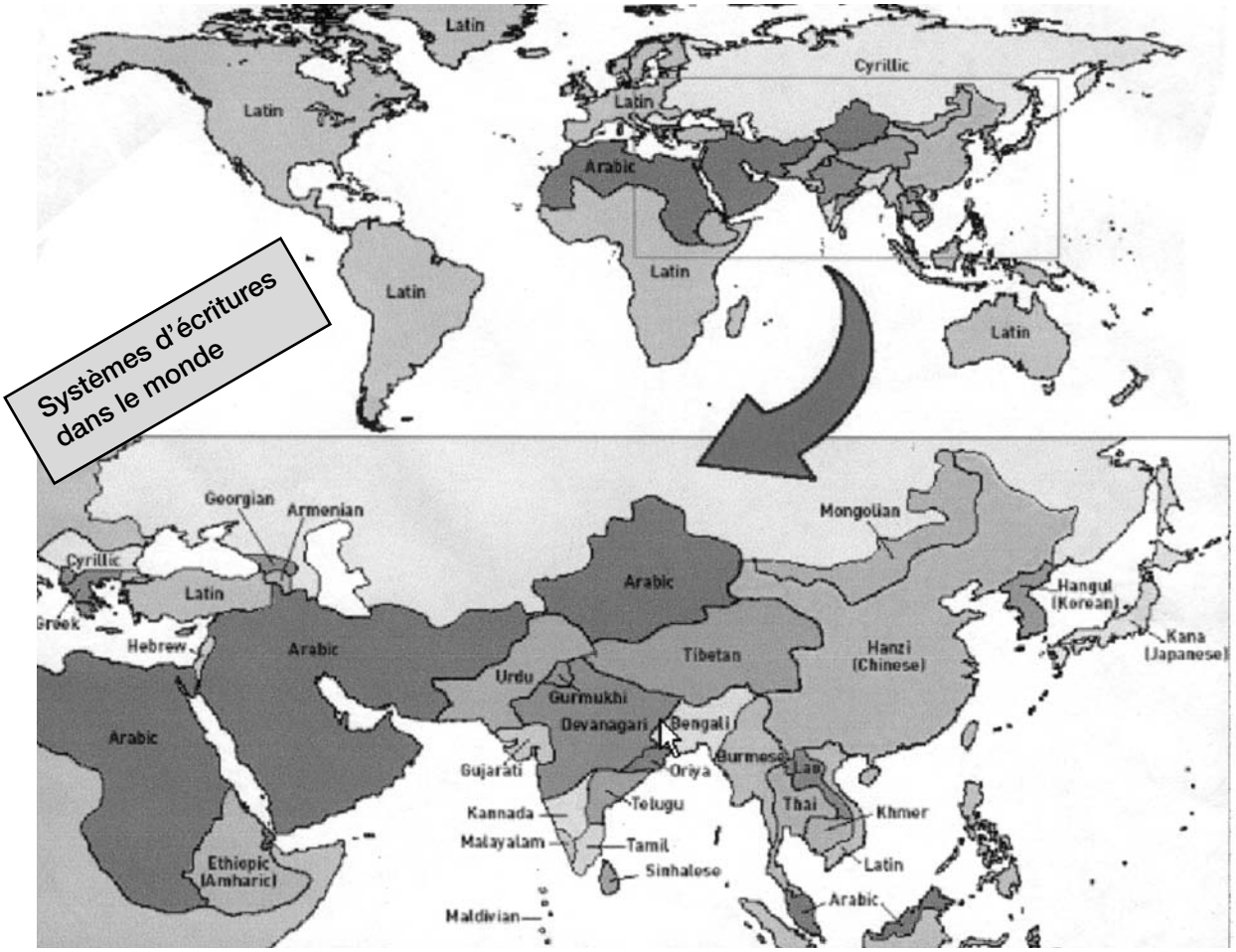
Enfin, le français devra chercher à occuper le terrain comme langue pivot, sinon l'anglais le fera à sa place. Par exemple, pendant plusieurs années, les arabophones ont disposé de versions arabe/français des systèmes d'exploitation de Microsoft, tel *Windows 95.* Cependant, *Windows 98* fut seulement offert en version anglais/arabe. Résultat : rapidement, on a remarqué que les étudiants se sont mis à utiliser une terminologie anglophone. Avec la pénétration d'Unicode, les choses se sont corrigées un peu, mais il reste des choses à faire.

Dans une veine un peu différente, il arrive souvent qu'il faille d'abord, pour traduire un texte depuis l'arabe, le traduire en anglais, parce qu'aucune passerelle directe n'existe entre l'arabe et le français.

Rien de cela ne sera possible sans l'intensification de la collaboration internationale entre les équipes de chercheurs et les équipes de développeurs, pour assurer que l'on tient bien compte des attentes linguistiques et culturelles des

locuteurs du Sud, au niveau phonétique, syntaxique, sémantique, et ainsi de suite.

Illustration 1
le contexte linguistique mondial



Je termine cette intervention par la citation d'un projet conduit dans le contexte de la Francophonie dans lequel on essaye de renforcer le partenariat entre le français et la langue arabe. Il s'agit du projet Cyberprof, développé sous forme d'une application pédagogique pour les enseignants en information et en communication dans les écoles des sciences de l'information francophone du Sud. Cette application a pour but d'enseigner la matière de la recherche de l'information sur intranet en langue française. Il s'agit d'une application qui consiste à développer un corpus textuel francophone sur lequel est greffée une couche logicielle permettant d'indexer et de rechercher des données en utilisant les techniques des stratégies de recherche booléenne. C'est un outil qui permet aux enseignants de l'informatique documentaire de disposer d'un outil pédagogique pour conduire des travaux pratiques de

recherche d'information. Dans une deuxième phase, il est question d'incorporer des textes dans les langues nationales, de les indexer pour que ce soit un support pédagogique multilingue conforme aux règles de l'internationalisation et de la localisation. En effet, il est actuellement question de collaborer avec les concepteurs français de l'outil d'indexation pour couvrir une plus grande diversité de langues par le robot indexeur. Le partenariat francophone avec les diverses langues partenaires aura ainsi abouti à quelque chose de concret au profit de la communauté scientifique francophone du Sud.

Mokhtar Ben Henda,
CEM-GRESIC, Université de Bordeaux 3, France, et
ISD, Université La Manouba-Tunisie.

Illustration 2
le schéma fonctionnel de Cyberprof