

LES DÉFIS POUR LES LANGUES À L'ÈRE DU NUMÉRIQUE (JOSÉ ANTONIO MILLÁN)

Dans une société déjà très développée numériquement, le langage naturel sert de plus en plus de moyen de communication entre les systèmes et les personnes et entre les personnes de langues différentes, et ce pour une raison évidente : la langue est un système de communication utilisé quotidiennement. Il s'agit de l'interface non seulement la plus commune mais aussi la plus intelligente : aucun menu déroulant, aucun bouton ne pourra jamais remplacer toutes les subtilités d'une phrase simple. Et si l'on considère la langue orale, même les analphabètes et les personnes peu habituées au clavier et à la souris, sont capables de dicter ce qu'ils souhaitent à un système automatique bien conçu.

À l'ère du numérique, la langue –un bien commun de création collective, gratuit et à usage illimité– est devenue une marchandise. Pour pouvoir communiquer avec les machines, elles doivent utiliser des programmes dont le développement est long et coûteux et qui exigent des ensembles de données structurés (corpus, dictionnaires). Même si ces programmes existaient et que nous voulions –et pouvions– les acheter, ils ne répondraient certainement pas à l'ensemble des besoins de nos sociétés.

Quel type de système les langues vont-elles utiliser comme interface ? Des systèmes de saisie de données en général : de l'agenda personnel aux systèmes professionnels ; de commerce électronique : systèmes capables de rechercher des produits en fonction de caractéristiques données, de les décrire et de les comparer ; de loisirs : location de spectacles, réservation de restaurants, informations touristiques, etc. ; d'éducation et de formation : systèmes automatiques d'apprentissage et d'évaluation ; ou encore de recherche : localisation de matériel, accès intelligent à des bases de données.

Nous utiliserons de plus en plus ces programmes, et parfois sans le savoir. Ils seront multilingues et capables de deviner notre intérêt envers telle ou telle information, de la traduire –à plusieurs degrés de fiabilité– et de la résumer. Ils deviendront nos outils de travail intellectuel et professionnel¹.

Un secteur économique d'envergure naîtra de l'environnement de ces systèmes. Néanmoins, dans le cas d'une langue répandue comme l'espagnol, une dépendance technologique croissante des pays hispanophones et un déséquilibre de la balance commerciale² en résulteront. Si les locuteurs d'une langue répandue doivent payer pour utiliser ces programmes, ceux des langues ou variantes minoritaires n'auront pas cette possibilité, de par l'absence de tels systèmes pour leur langue. Ainsi, nous payerons pour un dictionnaire électronique des synonymes en français de France ou en espagnol d'Espagne –même s'ils font implicitement partie des logiciels de traitement de texte–, mais il sera impossible de se procurer un tel dictionnaire en français du Sénégal ou en espagnol de Bolivie, même moyennant finances ...

1 Voir mon livre *Internet y el español* (Madrid, Fundación Retevisión, 2001), partie III.

2 José Antonio Millán, "La lengua que era un tesoro", 28 mars 2001 : <http://www.jamillan.com/tesoro.htm> et sa version résumée en anglais "How much is a language worth. A Quantification of the Digital Industry for the Spanish Language": <http://www.jamillan.com/worth.htm>.

Pourquoi n'existe-t-il pas, comme c'est souvent le cas pour nombre de langues, des études et des recherches –presque toujours financées par le public– qui pourraient servir de base au développement de logiciels linguistiques à divers niveaux généraux et locaux ? Pourquoi ce secteur industriel, important et stratégique, sera-t-il presque totalement colonisé ?

Pour certaines langues, cela peut s'expliquer par une absence de recherches, et ce pour des raisons historiques ou en raison d'un manque de moyens et de personnel universitaire. Cependant, pour l'espagnol, le portugais, le français ou l'italien, c'est leur Gouvernement respectif qui manque cruellement de politique linguistique numérique. Il s'agit en effet d'un terrain particulièrement glissant, sur lequel deux domaines s'affrontent : la politique linguistique et la politique numérique. En général, les Gouvernements s'y connaissent peu et ne cherchent pas à en savoir plus. Souvent, ils ne savent même pas qu'une politique linguistique est possible, sauf peut être dans les pays francophones du Nord et dans les communautés où une langue minoritaire sert d'instrument politique ; quant à la portée sociale de la question numérique, elle est encore loin d'être comprise.

Quels objectifs doit se fixer une politique linguistique numérique ?

- S'assurer que les ressources –corpus et programmes de développement par exemple– et les ensembles de données structurés qui alimentent les systèmes automatiques comme les dictionnaires, soient disponibles pour de nouveaux développements.
- Augmenter le nombre de développeurs de logiciels linguistiques afin d'élargir la qualité et la quantité des options.
- Faciliter l'incorporation des langues minoritaires ou des variantes locales dans les logiciels linguistiques.

En réalité, ces trois points peuvent se résumer en un seul : permettre l'utilisation des ressources et des données grâce à une licence garantissant l'ouverture et la réutilisation des produits dérivés. Ce sont des points de base : dans le cas de l'espagnol et sans doute d'autres langues, il est difficile d'utiliser, pour les développer, certaines ressources des institutions publiques et historiques. Certaines d'entre elles sont dites « ouvertes » sur l'Internet, mais cela signifie simplement qu'il est possible de les consulter : pour connaître par exemple l'occurrence d'un mot ou son analyse grammaticale, mais il est impossible de les utiliser à des fins de développement. Pour être clair, ouvrir signifie remettre à qui le demande³ un DVD ou tout autre système de stockage contenant l'intégralité de la ressource. À la fin de cet article, vous trouverez les objections que cette manière d'agir peut soulever. Quant à la réutilisation, elle peut être assurée en donnant la ressource via une licence type GPL⁴ ou Creative Commons⁵.

La situation actuelle, tout du moins pour l'espagnol, est telle que les ressources linguistiques des centres de recherche publics ne parviennent pas à toutes les entreprises qui pourraient les utiliser sous leur forme transparente, mais sous une forme très réduite qui correspond aux programmes destinés aux utilisateurs finaux. Une politique efficace

3 Une récente initiative du catalan, langue minoritaire dont l'activité numérique est foisonnante, a ouvert la banque de données terminologique TERMCAT (<http://www.termcat.net/>) sous la licence Creative Commons. Voir article sur SoftCatalà (<http://www.softcatala.org/noticies/13042005271.htm>)

4 <http://www.gnu.org/copyleft/gpl.html>.

5 <http://creativecommons.org/>

serait d'autoriser la diffusion de toutes les ressources de développement d'outils linguistiques –qu'elles viennent d'institutions publiques ou privées– auprès de toute personne souhaitant développer un logiciel linguistique. La pensée dominante est que ce travail est réservé aux grandes entreprises, et par-dessus tout aux sociétés nord-américaines, mais, en réalité, chacun peut développer à sa façon, tant au niveau des données que des programmes de développement ou des petits logiciels d'utilisateurs finaux : en utilisant par exemple des dictionnaires spécialisés écrits et oraux pour compléter des lexiques de logiciels de reconnaissance vocale.

Aussi, les Gouvernements devraient facilement adopter une proposition aussi simple, bon marché et source de bénéfices certains pour la Société ; une solution encourageant les capacités des entreprises et des groupes d'utilisateurs au détriment des oligopoles ; une proposition qui permettrait de contrôler un secteur stratégique et de passage obligé pour les entreprises et les citoyens ...

José Antonio Millán, un madrilène qui vit en Catalogne, est un professionnel freelance des TIC, éditeur électronique, linguiste et spécialiste du thème langue et Internet avec de nombreuses publications déterminantes sur ce sujet. Son site Internet vaut le détour (<http://jamillan.com/>) !