2

LES LANGUES EN DANGER : UN DÉFI POUR LES TECHNOLOGIES DE LA LANGUE

Nous vivons une révolution numérique qui a un impact fort sur la communication et la société. Les développements récents des technologies de communication numérique et les réseaux sont parfois comparés à l'invention par Gutenberg de l'imprimerie. Que peut nous dire cette analogie de l'avenir de la société de l'information européenne et de nos langues en particulier ?

Nous vivons actuellement une révolution numérique comparable à l'invention par Gutenberg de l'imprimerie.

Après l'invention de Gutenberg, de réelles avancées dans la communication ont été accomplies à travers des efforts comme la traduction par Luther de la Bible dans une langue vernaculaire. Dans les siècles suivants, des techniques culturelles ont été développées pour mieux gérer le traitement des langues et l'échange des connaissances :

- la normalisation orthographique et grammaticale des langues majeures a permis la diffusion rapide des nouvelles idées scientifiques et intellectuelles ;
- le développement des langues officielles a permis aux citoyens de communiquer au sein de certaines frontières (souvent politiques) ;
- l'enseignement et la traduction des langues ont permis des échanges entre communautés linguistiques ;
- la création de directives éditoriales et bibliographiques a assuré la qualité des matériels imprimés ;
- la création de différents médias tels que journaux, radios, télévisions, livres et autres formats satisfait les différents besoins de communication.

Au cours des vingt dernières années, les technologies de l'information ont contribué à automatiser et faciliter de nombreux processus :

- les logiciels de PAO (Publication Assistée par Ordinateur) ont remplacé la dactylographie et la composition ;
- les logiciels comme Microsoft PowerPoint remplacent les transparents de rétroprojection ;
- la fonction courrier électronique (mél) permet d'envoyer et de recevoir des documents plus rapidement qu'à l'aide d'un télécopieur ;
- Skype gère à faible coût les appels téléphoniques sur Internet et permet d'organiser des réunions virtuelles ;
- les formats de codage audio et vidéo facilitent l'échange des contenus multimédias ;
- les moteurs de recherche fournissent un accès par mots-clés aux pages Web ;
- les services en ligne comme Google Translate produisent rapidement des traductions approximatives ;
- Les plates-formes de médias sociaux telles que Facebook, Twitter et Google+, facilitent la communication, la collaboration et le partage d'informations.

Bien que ces outils et applications soient utiles, ils ne permettent pas encore d'avoir une société de l'information européenne vraiment multilingue, une société moderne et inclusive où la libre circulation des individus, des marchandises et des informations n'est plus freinée par les barrières linguistiques.

2.1 LES FRONTIÈRES LINGUISTIQUES ENTRAVENT LA SOCIÉTÉ DE L'INFORMATION EUROPÉENNE

Nous ne pouvons pas prédire précisément ce à quoi la société de l'information à venir va ressembler. Il y a cependant de fortes chances pour que la révolution dans les technologies de la communication rassemble d'une nouvelle façon les personnes qui parlent des langues différentes. Cela pousse à la fois les individus à apprendre de nouvelles langues et les développeurs à créer de nouvelles applications pour assurer une entente mutuelle et accéder à des connaissances partagées.

> Dans un espace économique et informationnel global, nous sommes confrontés à des langues, des locuteurs et des contenus très nombreux.

Dans un espace économique et informationnel mondial, nous sommes confrontés à un accroissement des contacts avec des langues, des locuteurs et des contenus à travers les nouveaux types de médias. La popularité actuelle des médias sociaux et collaboratifs (Wikipédia, Facebook, Twitter, YouTube et Google+, par exemple) n'est que la face émergée de l'iceberg.

Aujourd'hui, nous pouvons transmettre des giga-octets de textes autour du monde parfois en moins de temps qu'il ne faut pour se rendre compte qu'ils sont écrits dans une langue que nous ne comprenons pas. Selon un récent rapport commandité par la Commission Européenne, 57% des utilisateurs d'Internet en Europe achètent des biens et des services dans des langues qui ne sont pas leur langue maternelle (l'anglais est la langue étrangère la plus utilisée, suivie par le français, l'allemand et l'espagnol). 55% des utilisateurs lisent des contenus dans une langue étrangère, tandis que seulement 35% utilisent une autre langue pour écrire des méls ou poster des commentaires sur le Web [2]. Il y a quelques années, la vaste majorité des contenus sur le Web étaient en anglais. Cependant, la situation a maintenant changé radicalement. La quantité de contenus en ligne dans d'autres langues (les autres langues européennes, mais aussi les langues asiatiques et l'arabe en particulier) a explosé.

> Quelles sont les langues européennes qui vont prospérer ou survivre dans l'information en réseau et la société du savoir ?

Le fossé numérique omniprésent causé par les frontières linguistiques n'a étonnamment pas suscité beaucoup d'attention dans le discours public, et pourtant, il soulève une question très pressante : « Quelles langues européennes vont prospérer et persister dans la société de l'information et du savoir en réseau, et quelles sont celles qui sont susceptibles de disparaître ? »

2.2 NOS LANGUES EN DANGER

L'arrivée de l'imprimerie a contribué à un inestimable échange d'information en Europe, mais elle a aussi conduit à l'extinction de certaines langues européennes. Les langues régionales et minoritaires ont été rarement imprimées. En conséquence, des langues, comme la langue de Cornouailles ou la langue dalmate, étaient limitées par leur forme orale de transmission, ce qui a restreint leur adoption, leur diffusion et leur utilisation par

rapport aux langues imprimées. L'Internet aura-t-il le même impact sur nos langues actuelles ?

La grande variété des langues en Europe est l'un de ses plus riches et plus importants atouts culturels

Les quelques 80 langues de l'Europe sont une composante essentielle de son modèle social unique [3]. Alors que les langues largement répandues comme l'anglais ou l'espagnol vont certainement maintenir leur présence dans la société numérique émergente et sur le marché international, beaucoup de langues européennes pourraient être coupées de la communication numérique et devenir sans importance dans une société en réseau. Cela affaiblirait la réputation mondiale de l'Europe et serait en contradiction avec l'objectif stratégique d'une participation équitable pour tous les citoyens européens indépendamment de leur langue. Selon un rapport de l'UNESCO sur le multilinguisme, les langues sont un moyen essentiel pour l'exercice des droits fondamentaux, tels que l'expression politique, l'éducation et la participation dans la société [4].

2.3 LES TECHNOLOGIES DE LA LANGUE SONT DES TECHNOLOGIES-CLÉS HABILITANTES

Dans le passé, les investissements relatifs à la sauvegarde des langues ont porté essentiellement sur l'enseignement des langues et sur la traduction. Selon certaines estimations, le marché européen pour la traduction, l'interprétation, la localisation des logiciels et la globalisation des sites Web a été de 8,4 milliards d'Euros en 2008 et on s'attend à une croissance de 10% par an [5]. Pourtant, cela ne couvre qu'une petite partie des besoins actuels et futurs permettant d'assurer une communication entre les communautés linguistiques. La seule solution pour promouvoir un usage large et entier des langues en Europe est d'utiliser les technologies qui le permettent, tout comme on utilise des technologies pour couvrir nos besoins dans les domaines de l'énergie ou des transports, entre autres.

Des technologies de la langue traitant toutes les formes de textes écrits et de discours parlés peuvent aider les individus à collaborer, à entretenir des échanges commerciaux, à partager des connaissances et à participer à des débats sociaux ou politiques indépendamment des barrières linguistiques ou des compétences informatiques. Elles opèrent souvent de façon cachée dans des logiciels complexes qui nous aident déjà aujourd'hui lorsque nous :

- trouvons des informations avec un moteur de recherche sur Internet ;
- vérifions l'orthographe et la grammaire dans un traitement de texte ;
- obtenons des recommandations de produits dans un magasin en ligne ;
- entendons les instructions verbales d'un système de navigation routière ;
- traduisons des pages Web, des méls, des blogs, etc. avec un service de traduction en ligne.

Les technologies de la langue sont des technologies habilitantes dans le cadre d'applications plus larges comme les systèmes de navigation ou les moteurs de recherche. La mission de cette collection de livres blancs réalisés par META-NET sur les langues est de déterminer les capacités de ces technologies de base pour chacune des langues européennes.

L'Europe a besoin de technologies de la langue robustes et abordables pour toutes les langues européennes.

Pour maintenir sa position à la pointe de l'innovation, l'Europe aura besoin de technologies de la langue pour toutes les langues européennes, qui soient robustes et abordables, et qui puissent être étroitement intégrées au sein des environnements logiciels clés. Sans les technologies de la langue, nous ne serons pas en mesure de donner aux utilisateurs des moyens de communiquer réellement interactifs, multimédias et multilingues dans un futur proche.

2.4 DES OPPORTUNITÉS POUR LES TECHNOLOGIES DE LA LANGUE

Dans le monde de l'imprimerie, l'avancée technologique a été la possibilité de reproduire rapidement l'image d'un texte en utilisant une presse d'imprimerie suffisamment puissante. Il revenait aux humains de faire le dur travail de rassembler, d'évaluer, de traduire et de résumer la connaissance. Il a fallu attendre Edison pour enregistrer la voix humaine – et cette fois encore, la technologie n'était capable que de produire des copies analogiques.

Les technologies de la langue sont une solution pour traiter le handicap lié à la diversité linguistique.

Les technologies de la langue peuvent à présent simplifier et automatiser les processus de traduction, de production de contenus, de traitement de l'information et de gestion des connaissances pour toutes les langues européennes. Les technologies de la langue peuvent également favoriser le développement d'interfaces vocales pour les appareils électroniques domestiques, les machines, les véhicules, les ordinateurs, les téléphones ou les robots. Les applications commerciales et industrielles sont encore dans les premiers stades de développement, même si les récentes réalisations en R&D ont créé un éventail d'opportunités. Par exemple, la traduction automatique atteint déjà une qualité raisonnable dans des domaines spécifiques et des applications expérimentales permettent d'effectuer la fourniture d'information multilingue, la gestion des connaissances ainsi que la production de contenus dans de nombreuses langues européennes.

Les technologies de la langue représentent une formidable opportunité pour l'Union européenne. Elles peuvent aider à traiter la délicate question du multilinguisme en Europe – le fait que plusieurs langues coexistent naturellement dans les entreprises européennes, les organisations et les écoles. Les citoyens veulent cependant pouvoir communiquer en franchissant les frontières linguistiques qui existent encore dans le Marché Commun européen et les technologies de la langue peuvent aider à surmonter ce dernier obstacle, tout en soutenant une utilisation libre et ouverte de chacune des langues. Si l'on se projette encore plus loin, les technologies de la langue multilingues innovantes pour l'Europe peuvent aussi servir d'exemple à nos partenaires mondiaux lorsqu'ils s'adresseront à leurs propres communautés multilingues. Les technologies de la langue peuvent être vues comme des sortes de technologies « d'assistance » qui aident à résoudre le « handicap » de la diversité des langues et rendent les communautés linguistiques plus accessibles les unes aux autres. Finalement, un champ de recherche actif concerne l'utilisation des technologies de la langue pour les opérations de sauvetage dans les zones sinistrées [6]. Dans un tel environnement à haut risque, la précision de la traduction peut être une question de vie ou de mort : dans le futur, des robots intelligents ayant des capacités de communication interlangues auront la capacité de sauver des vies humaines.

2.5 LES DÉFIS DES TECHNOLOGIES DE LA LANGUE

Bien que les technologies de la langue aient fait des progrès considérables au cours des dernières années, le rythme actuel des progrès technologiques et des produits innovants est trop lent. Les technologies de la langue largement utilisées, comme les correcteurs orthographiques ou grammaticaux dans les traitements de texte, sont généralement monolingues, et ne sont disponibles que pour une poignée de langues. La traduction automatique en ligne, même si elle est utile pour produire rapidement une bonne approximation du contenu d'un document, génère beaucoup trop d'erreurs pour être utilisée lorsque des traductions précises et complètes sont nécessaires.

> Le rythme actuel du progrès technologique est trop lent.

Du fait de la complexité du langage humain, la modélisation informatique de la langue et son expérimentation en milieu réel est une tâche longue et coûteuse qui nécessite des engagements financiers durables. L'Europe doit donc maintenir son rôle de pionnier en affrontant le défi technologique posé par une communauté multilingue en inventant de nouvelles approches pour accélérer le développement sur toute sa superficie. Cela peut faire appel à des avancées informatiques ou à des techniques comme le *crowdsourcing*.

2.6 ACQUISITION DE LA LANGUE PAR LES HUMAINS ET LES MACHINES

Pour illustrer comment les ordinateurs manipulent le langage et pourquoi il est difficile de les programmer pour traiter différentes langues, nous porterons un bref regard sur la façon dont les humains acquièrent leurs premières et deuxièmes langues, puis nous regarderons comment les technologies de la langue fonctionnent.

Les êtres humains acquièrent des compétences linguistiques de deux manières différentes. Les bébés apprennent une langue en écoutant leurs parents, leurs frères et sœurs et les autres membres de la famille communiquer entre eux. À partir de l'âge d'environ deux ans, les enfants produisent leurs premiers mots et des phrases courtes. Cela n'est possible que parce que les humains ont une prédisposition génétique particulière pour imiter, puis rationaliser ce qu'ils entendent.

Apprendre une langue seconde à un âge plus avancé requiert plus d'efforts cognitifs car l'enfant n'est pas immergé dans une communauté linguistique de locuteurs natifs. À l'âge scolaire, les langues étrangères sont généralement acquises par l'apprentissage de leur structure grammaticale, de leur vocabulaire et de leurs règles de prononciation à partir de matériel éducatif qui décrit les connaissances linguistiques en termes de règles abstraites, de tableaux et d'exemples.

> Les êtres humains acquièrent les compétences linguistiques de deux manières différentes : en apprenant à partir d'exemples ou en apprenant les règles linguistiques qui les sous-tendent.

Si l'on considère à présent les technologies de la langue, on voit que les deux principaux types de systèmes « acquièrent » leurs capacités linguistiques d'une manière semblable à celles des humains. Les approches statistiques (ou « fondées sur les données ») acquièrent les connaissances linguistiques à partir de vastes collections d'exemples concrets de texte dans une seule langue ou à partir de ce qu'on appelle des *textes parallèles*. Alors qu'il est suffisant d'utiliser un texte dans une seule langue pour entraîner par exemple un correcteur orthographique, des textes parallèles dans deux

(ou plusieurs) langues sont nécessaires pour entraîner un système de traduction automatique. Les algorithmes d'apprentissage modélisent la façon dont les mots, portions de phrases et phrases complètes sont traduits d'une langue à une autre.

L'approche statistique nécessite habituellement des millions de phrases pour produire des résultats de qualité raisonnable. C'est une des raisons pour lesquelles les fournisseurs de moteurs de recherche sont impatients de recueillir autant de documents écrits que possible. Certains correcteurs orthographiques dans les systèmes de traitement de textes, et les services comme *Google Search* et *Google Translate* s'appuient sur des approches statistiques.

La seconde approche pour les technologies de la langue, et la traduction automatique en particulier, est de construire des systèmes à partir de règles. Des experts de la linguistique, de la linguistique computationnelle et de l'informatique doivent d'abord coder l'analyse grammaticale (ou les règles de traduction) et compiler des listes de vocabulaire (lexiques). Cela est très fastidieux, demande beaucoup de travail et n'offre aucune garantie de couverture suffisante des phénomènes de la langue. Certains des meilleurs systèmes de traduction automatique à base de règles sont en constant développement depuis plus de vingt ans. Le grand avantage des systèmes à base de règles est que les experts peuvent contrôler de manière plus détaillée le traitement du langage. Cela rend possible de corriger systématiquement les erreurs dans le logiciel et de formuler un retour détaillé à l'utilisateur, surtout lorsque les systèmes à base de règles sont utilisés pour l'apprentissage des langues. Cependant, en raison du coût élevé de cette tâche, les technologies de la langue à base de règles n'ont été développées que pour les principales langues.

Les deux principaux types de systèmes de traitement automatique de la langue apprennent le langage d'une manière semblable à celle des humains.

Étant donné que les avantages et inconvénients des systèmes statistiques ou à base de règles tendent à être complémentaires, les recherches actuelles s'orientent vers des approches hybrides qui combinent les deux méthodologies. Cependant, ces approches ont jusqu'à présent eu moins de succès dans les applications industrielles que dans les laboratoires de recherche.

Comme nous l'avons vu dans ce chapitre, beaucoup d'applications largement utilisées dans la société de l'information d'aujourd'hui dépendent fortement des technologies de la langue, en particulier dans l'espace économique et informationnel européen. Bien que ces technologies aient fait des progrès considérables ces dernières années, il y a encore une énorme marge pour améliorer leur qualité. Dans le prochains chapitres, nous décrirons le rôle du français dans la société de l'information européenne, et évaluerons l'état des technologies de la langue pour le français.

Springer

http://www.springer.com/978-3-642-30760-7

The French Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)
2012, VI, 95 p. 37 illus., 33 illus. in color., Softcover
ISBN: 978-3-642-30760-7