Préface

Ce numéro contient les articles retenus lors de l'appel thématique sur le multilinguisme et le traitement automatique des langues. Depuis le dernier numéro de la revue TAL sur cette thématique (numéro 40-1), il y a maintenant dix ans, le multilinguisme s'est imposé comme une réalité sociétale incontournable notamment sur l'Internet : alors que la langue anglaise dominait largement il y a dix ans, la représentation du chinois, de l'hindi ou de l'espagnol tend peu à peu à correspondre à l'importance réelle de ces langues. L'Internet reflète maintenant mieux la diversité des langues nationales et régionales. Logiquement, les applications informatiques développées – que ce soit en traduction, en recherche d'information interlangue, en aide à la communication écrite ou en apprentissage des langues… – doivent s'adapter à cette réalité en prenant en compte des langues de plus en plus diverses avec parfois une grande distance linguistique entre elles.

Le présent appel a suscité seize soumissions, parmi lesquelles quatre articles ont été retenus pour publication à l'issue du processus de relecture (taux de sélection : 25 %). Cinq soumissions provenaient de pays non francophones : deux des États-Unis, deux de l'Inde et une d'Allemagne.

La traduction automatique est probablement l'une des principales applications qui fait intervenir le traitement de plusieurs langues. Les recherches dans ce domaine remontent aux origines de l'informatique dans les années 50 : en 1954, l'expérience de Georgetown-IBM montrait un premier système de traduction automatique pour traduire soixante phrases du russe vers l'anglais, en pleine période de guerre froide, et les premiers travaux russes sont publiés dès 1957 par Panov et Ljapunov. Après quelques années de recherches intensives, le rapport ALPAC[1] a conclu, en 1966, que la traduction automatique est plus coûteuse, plus lente et moins précise qu'un traducteur humain et qu'elle n'atteindrait pas la qualité d'une traduction humaine dans un futur proche. Ceci a entraîné l'arrêt de quasiment toute activité de recherche dans le domaine aux États-Unis, avec l'exception de la création des entreprises SYSTRAN et Logos en 1968 et 1970, respectivement. Ce n'est principalement qu'en Allemagne, Canada et France que la recherche a continué. Enfin, dans les années 80 plusieurs nouvelles approches ont été proposées comme la traduction fondée sur les exemples ou l'approche statistique. Aujourd'hui, la traduction automatique est utilisée de plus en plus dans le milieu professionnel et par le grand public. Quatre articles soumis à ce

1. Disponible à `http://www.nap.edu/openbook.php?record_id=9547&page=R1`

numéro spécial concernaient directement le développement de systèmes de traduction automatique.

Le traitement automatique de documents disponibles dans plusieurs langues pose également des questions scientifiques intéressantes. Est-ce qu'une méthode particulière fonctionne bien pour différentes langues ? Quelles informations peuvent être extraites par un traitement conjoint de deux ou plusieurs langues ? Les textes dits parallèles, c'est-à-dire des ensembles de phrases dans une langue associées avec leur traduction dans une autre, sont particulièrement intéressants pour plusieurs domaines du traitement du langage. D'abord, ils constituent une ressource indispensable, mais malheureusement assez rare, pour la construction de systèmes de traduction statistiques. Ensuite, ces textes permettent également l'extraction de dictionnaires terminologiques ou l'extraction de paraphrases. Les recherches actuelles couvrent de nombreux aspects du traitement des textes parallèles : comment obtenir des textes parallèles, comment aligner les phrases ou les mots en langue source et cible, comment extraire des informations pertinentes, etc. On s'intéresse aussi beaucoup à l'utilisation de *corpus comparables*. Il s'agit de textes rédigés indépendamment dans des langues différentes, mais qui peuvent être mis en relation sur la base d'une similarité de leur contenu. Wikipédia en est un exemple bien connu. Quatre soumissions traitaient du traitement de documents multilingues.

Les techniques de recherche d'information sont aujourd'hui un outil indispensable pour consulter des informations sur Internet. Ainsi en 2010, environ 25 % des utilisateurs parlent anglais ou chinois alors que moins de 4 % utilisent le français, l'allemand ou l'arabe comme langue principale[2]. D'autre part, les pages rédigées en anglais étaient encore largement dominantes avant 2000[3], mais les langues utilisées se sont beaucoup diversifiées ces dernières années. Il est donc devenu indispensable de considérer la recherche d'information dans plusieurs langues. Les recherches dans ce domaine ont débuté en 1996 lors du premier atelier CLIR (*Cross-Lingual Information Retrieval*) à la conférence SIGIR. Ces ateliers ont lieu annuellement depuis 2000. Nous avons reçu deux propositions d'articles dans ce domaine.

Les autres articles soumis traitaient du multilinguisme dans un sens plus large dans des domaines très variés tels que l'extraction de paraphrases, la localisation de logiciels ou la reconnaissance de caractères.

Présentation des articles

Ce numéro regroupe les quatre articles sélectionnés suivants :

1) *Apprentissage non supervisé de familles morphologiques : comparaison de méthodes et aspects multilingues*, Delphine Bernhard

2) *Paradocs : l'entremetteur de documents parallèles indépendant de la langue*, Alexandre Patry et Philippe Langlais

2. Selon http ://www.internetworldstats.com/stats7.htm, juin 2010.

3. Supérieures à 70 % selon http ://www.clickz.com/clickz/stats/1697080/web-pages-language

3) *Micro-adaptation lexicale en traduction automatique statistique*, Josep Maria Crego, Gregor Leusch, Aurélien Max, Hermann Ney et François Yvon

4) *Transliteration as Alignment vs. Transliteration as Generation for Crosslingual Information Retrieval*, Anil Kumar Singh, Sethuramalingam Subramaniam, and Taraka Rama

L'article de D. Bernhard présente deux méthodes pour l'apprentissage non supervisé de familles morphologiques. La première technique forme des familles par groupements successifs, d'une manière similaire aux méthodes de classification ascendante hiérarchique, tandis que la seconde technique exploite une représentation des relations morphologiques sous forme de réseaux lexicaux. Des communautés sont ensuite détectées par des algorithmes fondés sur les graphes. Les méthodes proposées sont relativement génériques et ne dépendent pas de la langue traitée. Ainsi, des résultats expérimentaux sont présentés pour les langues allemande et anglaise, dont les différences sont soulignées.

Les textes parallèles jouent un rôle crucial dans les applications multilingues du traitement automatique de la langue, notamment en traduction statistique. MM. Patry et Langlais étudient le problème de détection et d'alignement de ces textes. Des techniques de recherche d'information sont appliquées, utilisant les nombres et les hapax comme mots-clefs, suivi par un simple classifieur. La méthode donne des résultats très intéressants sur plusieurs couples de langues : les onze langues du corpus Europarl, des textes du *Monde Diplomatique* en français et arabe, les comptes rendus des débats parlementaires du Nunavut (inuktitut/anglais) et des articles en anglais et français sur Wikipédia.

MM. Crego, Leusch, Max, Ney et Yvon décrivent des techniques pour effectuer une *micro-adaptation* du modèle de langue d'un système de traduction statistique. L'idée de base consiste à produire un modèle de langue supplémentaire à partir de traductions alternatives. Plusieurs possibilités sont envisagées pour obtenir ces traductions additionnelles : des systèmes alternatifs pour le même couple de langues, des système de traduction par pivot (un grand nombre de langues intermédiaires sont analysées) et la traduction multisource. De nombreuses expériences sur le corpus des débats du Parlement européen montrent très clairement l'intérêt de ces méthodes.

Enfin, l'article de A.K. Singh, S. Subramaniam et T. Rama aborde le multilinguisme dans le cadre de la recherche d'information. Lorsque l'on cherche des documents dans une langue, la procédure habituelle consiste à traduire les requêtes et à effectuer ensuite la recherche en langue cible. De nombreuses études ont montré que les noms propres représentent une partie importante de la requête, ce qui demande leur translittération lorsque les deux langues n'utilisent pas le même système d'écriture. L'article de A.K. Singh, S. Subramaniam et T. Rama compare plusieurs techniques de translittération de l'anglais vers le hindi et le marathi.

La préparation de ce numéro sur les recherches en multilinguisme et en traitement automatique des langues a montré qu'il s'agit d'un domaine actif. Il y a clairement deux applications clefs, la traduction automatique et la recherche d'information mul-

tilingue, mais le traitement de plusieurs langues se généralise dans d'autres domaines du TAL. Il y a de nombreuses équipes en France et à l'étranger qui travaillent sur ces problématiques. On peut donc anticiper que le traitement du multilinguisme jouera un rôle important dans les recherches en traitement automatique des langues dans les années à venir.

Emmanuel MORIN
LINA (Université de Nantes)

Holger SCHWENK
LIUM (Université du Maine)

Remerciements

Nous tenons à remercier les relecteurs sans qui cet ouvrage n'aurait pu paraître :

– Núria Bel, Universitat Pompeu Fabra, Espagne
– Laurent Besacier, LIG, université de Grenoble 1, France
– Romaric Besançon, CEA, France
– Hervé Blanchon, LIG, université de Grenoble 1, France
– Nicola Cancedda, XRCE, France
– Pascale Fung, Hong Kong University of Science and Technology, Chine
– Éric Gaussier, LIG, université de Grenoble 1, France
– Gregory Grefenstette, Exalead, France
– Tony Hartley, University of Leeds, United Kingdom
– Pierre Isabelle, NRC-CNRC, Canada
– Kyo Kageura, université de Tokyo, Japon
– Olivier Kraif, LIDILEM, université de Grenoble, France
– Marie-Claude L'Homme, université de Montréal, Canada
– Aurélien Max, LIMSI, université Paris Sud 11, France
– Bruno Pouliquen, OMPI, ONU
– Violaine Prince, LIRMM, université de Montpellier, France
– Aarne Ranta, université de Gothenburg, Suède
– Jean Senellart, Systran, France
– Michel Simard, NRC-CNRC, Canada
– Monique Slodzian, INALCO, France
– Kamel Smaïli, LORIA, université de Nancy, France
– Éric Wehrli, LATL, université de Genève, Suisse

ainsi que les membres du comité de rédaction de la revue (voir pages d'en-tête de ce numéro).