Technologies facilitatrices pour la dissémination de la communication scientifique multilingue

Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Lyne.Da.Sylva@UMontreal.CA

Introduction

L'anglais occupe une place prédominante dans la communication scientifique. Plusieurs travaux permettent de chiffrer ceci

On peut regarder par exemple des données portant sur la proportion de la production traditionnelle d'articles scientifiques selon les pays. Selon Inönü (2003, p. 139 et ss.) et d'après des chiffres provenant du *Institute for Scientific Information*, pour l'année 1999 les États-Unis et le Royaume-Uni occupaient respectivement la première et la deuxième place en matière de nombre d'articles publiés. Sur les 95 pays recensés, cela représente environ 39 % de la production scientifique[1]. En matière de langue de publication cette fois, on peut regarder les données du même ISI disponibles par le serveur *Dialog* ; les index cumulatifs des citations des bases de données présentent des profils différents pour le secteur des sciences naturelles et exactes, celui des sciences sociales et celui des lettres et sciences humaines. L'anglais affiche respectivement 91,8 %, 90,8 % et 70,9 % des publications dans chaque domaine[2].

Cependant, prédominance ne veut pas dire exclusivité. En effet, s'il est vrai que l'anglais occupe la première position absolue en matière de nombre d'articles publiés tel que rapporté dans Inönü (2003), il reste que plus de 60 % des articles proviennent d'autres pays, où la langue d'expression n'est pas limitée à l'anglais.

Il est intéressant d'examiner également le rôle des langues autres que l'anglais dans le réseau de communication que constitue Internet. Au niveau de l'utilisation d'Internet en général, selon l'IDC (*International Data Corporation*[3]), l'utilisation d'Internet par des utilisateurs dont la langue maternelle n'est pas l'anglais a augmenté de 10 % à 50 % dans les quatre dernières années. On peut présumer qu'une part toujours croissante sera effectuée dans une langue autre que l'anglais.

Pour la communication scientifique numérique, Thelwall et al. (2003) ont analysé les sites Web des universités de 16 pays d'Europe de l'ouest, afin d'illustrer les patrons d'utilisation du Web selon la langue. Les données à l'étude sont les hyperliens entre universités (une justification de l'intérêt d'étudier les liens entre les pages des universités en tant qu'indicateur de la production scientifique de ces universités est présentée dans l'article, ainsi que dans Borgman et Furner, 2002). Une première donnée intéressante de leur étude est que l'anglais est utilisé dans 56 % des pages Web étudiées. Une deuxième, est que dans plusieurs cas (difficiles à chiffrer d'après les données dans l'article), des liens sont effectués entre des paires de langues qui n'incluent pas l'anglais.

La communication par voie numérique prend de l'essor dans les milieux de recherche. Borgman et Furner (2002) documente la croissance de l'utilisation du numérique dans la communication savante. Spécifiquement, des initiatives diverses de développement de répertoires de textes scientifiques foisonnent : périodiques en ligne, archives ouvertes pour la publication scientifique, dont le répertoire e-Print Archive[4] (anciennement connu sous le nom de LANL e-Print Archive ; Ginsparg, 1994) ; le portail de diffusion Érudit[5] ; l'initiative de Budapest[6], etc. Le forum « *Open Archives Forum*[7] » répertorie d'ailleurs un nombre important de telles archives d'accès public.

Sur support numérique, l'importance des langues autres que l'anglais est croissante.

---

[1] Ces données comprennent les sciences fondamentales et appliquées mais excluent les sciences sociales.
[2] Le cumul commence en 1972 pour le Social Sciences Citation Index (SSCI), en 1974 pour le Science Citation Index (SCI) et en 1980 pour le Arts & Humanities Citation Index (A&HCI). Les chiffres sont en date du 28 avril 2003.
[3] http://www.idcresearch.com
[4] http://www.arXiv.org
[5] http://www.erudit.org
[6] http://www.soros.org/openaccess/fr/index.shtml
[7] http://www.oaforum.org/oaf_db/list_db/list_repositories.php

Et a priori, le numérique n'a pas de langue ; toutes les informations sont transformées ultimement en code binaire, et la représentation d'une langue plutôt qu'une autre n'est, en théorie, pas plus difficile. La prédominance de l'anglais s'explique par un grand nombre de facteurs, bien sûr, économiques, politiques, sociaux, etc. Ce qui nous intéressera ici c'est d'en examiner les facteurs technologiques. Spécifiquement, quels outils technologiques doivent être en place pour faciliter la dissémination de la communication scientifique multilingue en format numérique ? En d'autres termes, si les documents sont dans plus d'une langue, qu'est-ce qui est nécessaire pour favoriser également leur dissémination ?

La chaîne documentaire et la communication scientifique

La communication scientifique est le reflet d'une communauté, nécessairement ancrée dans une culture. Le travail scientifique naît de la collaboration, de l'échange des idées afin de nourrir l'innovation. Toute barrière à la communication est une barrière au travail scientifique. Notamment, la langue est potentiellement une barrière. Dans toutes les traditions, la formation d'un chercheur implique qu'il se familiarise avec les travaux du domaine où il veut œuvrer, et lorsque nécessaire, avec la langue prédominante dans le domaine. Les langues minoritaires ont été défavorisées puisque, étant donné l'infrastructure inhérente à la publication scientifique, et la difficulté de se faire connaître à l'extérieur de sa propre communauté (linguistique), les auteurs ont privilégié une langue commune, parfois l'allemand, le russe, le français ou l'anglais.

Traditionnellement, la communication scientifique a été encadrée par les producteurs des revues savantes : les communautés de pairs qui procèdent à l'évaluation des contributions, les maisons d'édition qui s'occupent de publier les contributions reçues favorablement par les pairs. Un service documentaire dédié à la diffusion de la communication scientifique effectue une sélection parmi les travaux publiés et traite la collection en vue de l'utilisation finale. Les documents suivent la chaîne de traitement documentaire : l'identification des documents pertinents pour la collection, leur description, l'organisation de ces descriptions, le stockage adéquat des documents et finalement la mise au point d'outils permettant aux utilisateurs de repérer les documents utiles à leurs besoins. Dans l'univers des documents papier, ce traitement repose sur un ensemble de « technologies » bien connues : au préalable, mise en pages et impression des documents ; puis, lecture humaine, production humaine de descriptions des documents, etc. L'avènement du numérique apporte plusieurs changements dans la façon de produire et de traiter les documents. Spécifiquement, le réseau de diffusion qu'offre Internet apporte deux nouveautés : la facilité de publier, éventuellement soi-même, des œuvres (dans la langue de son choix), et la disponibilité d'outils variés (par exemple les outils de traduction automatique en ligne) qui peuvent accroître la visibilité de ces travaux[8].

La problématique qui nous intéresse est la suivante : étant donné un réseau de diffusion de la communication scientifique numérique exprimée dans diverses langues, quelles sont les technologies qui facilitent la gestion de ces documents ?

Éventail de technologies nécessaires

Pour structurer la discussion, cet examen se fera avec les lunettes du documentaliste : si l'on considère dans sa globalité la chaîne documentaire, ou le parcours d'un document de son auteur à son auditoire cible via un service documentaire dédié, nous examinerons ce qui est nécessaire à chaque étape pour assurer le traitement des documents dans une langue quelconque. Nous présenterons l'état actuel des technologies disponibles et, parmi ce qui manque, ce qui est envisageable ou au contraire difficile. Nous incluons dans le terme « technologies » diverses notions : outils, ressources, systèmes, normes, etc. Chacune occupe une place dans la chaîne de traitement informatique.

L'éventail que nous présentons contient, nous le reconnaissons, certains éléments qui sont fortement souhaitables mais pas toujours nécessaires ; d'autres éléments en revanche sont absolument essentiels.

Notons enfin qu'on exclura presque totalement, pour la discussion, l'option de traduction humaine ou automatique systématique des documents originaux vers une langue commune quelconque. Mis à part le fait que l'état de la technologie ne permette pas des traductions de qualité « humaine », cette possibilité est écartée d'emblée puisque l'on s'intéresse ici à la coexistence de documents en différentes langues.

Sélection des documents

Pour décider d'ajouter un document à une collection, il faut d'abord pouvoir lire le document, et, avant de pouvoir le lire, il faut s'assurer de savoir lire la langue dans laquelle il est écrit. Non seulement la langue, mais aussi, pour les fichiers numériques, l'encodage utilisé. En effet, plusieurs langues (dont les langues asiatiques, par exemple) peuvent être

[8] Il faut noter bien sûr que de prendre Internet dans son ensemble comme exemple de collection soulèverait de nombreux problèmes, dont l'authenticité et la fiabilité des informations. Et qu'effectivement, un service documentaire incorpore une étape de sélection des documents, absente du Web.

encodées différemment. Des logiciels d'identification automatique de la langue (par exemple, le logiciel SILC[9]) peuvent déterminer, par une analyse de séquences de caractères, de lettres ou de mots courts discriminants, la langue d'un document et l'encodage de celle-ci.

Un logiciel d'identification de la langue permet de gagner du temps, de choisir les outils de traitement appropriés, et bien sûr, dès lors que l'on envisage une procédure automatique de sélection des documents, de pouvoir sélectionner ou refuser des documents déjà sur la base de la langue dans laquelle ils sont rédigés.

Par ailleurs, on peut imaginer des logiciels d'analyse globale des documents qui serviraient à décider d'ajouter (ou non) un document à une collection, sur la base de son contenu. Nous couvrirons ceci ci-dessous, dans la discussion de l'analyse documentaire.

Création et représentation des documents

Normalement, la création de documents est une opération qui n'appartient pas à la chaîne documentaire ; toutefois, son importance est incontournable dans le contexte numérique puisqu'elle touche au format de représentation même des documents en question.

La façon de représenter les informations dans les documents doit être applicable n'importe quelle langue. Cela déterminera la possibilité d'effectuer subséquemment leur transfert, leur affichage, et tout autre traitement ultérieur. D'abord, il doit être possible de représenter adéquatement tout jeu de caractères utilisé par une langue : caractères romains, cyrilliques, asiatiques, etc. Ensuite, les langages de balisage (comme SGML, XML, HTML) représentent un moyen extrêmement puissant de structurer les documents de manière à faciliter la préhension (par des outils automatiques) de la structure logique du document ; voir Marcoux et Sévigny (1997) pour une justification de l'utilisation de cette technologie dans les milieux documentaires. En contexte multilingue (ou par exemple un même document ou une partie de celui-ci, comme le résumé, existeraient en plus d'une langue), cela prend encore plus d'ampleur. En effet, il est raisonnable de penser que la structure logique d'un document resterait la même si celui-ci était traduit dans une autre langue[10]. Le balisage XML peut donc représenter une charpente pour guider l'interprétation du document, indépendamment de la langue. On peut en imaginer deux applications utiles : d'une part, la structuration identique de versions multilingues du même document (ce qui facilite la mise en correspondance entre les deux) ; d'autre part, la description normalisée de documents d'un certain type afin d'en faciliter l'utilisation par une communauté élargie (par exemple, multilingue) qui partage la norme. Faut-il ajouter que les langages de balisage comme SGML et XML ne présupposent pas de jeux d'étiquettes prédéfinis et sont par conséquent tout aussi aptes à des utilisations dans une langue quelconque ? Ainsi, ils permettent une représentation dans n'importe quelle langue de documents éventuellement multilingues.

Enfin, parmi les outils création de documents, on reconnaît l'utilité d'avoir des logiciels de traitement de texte, d'éditique et de conception de pages Web sensibles à la langue, dotés d'outils linguistiques comme des correcteurs d'orthographe et de grammaires, des dictionnaires, etc. dans la langue du document ou effectuant des ajustements typographiques selon la langue.

Il convient peut-être d'ajouter ici les logiciels d'affichage des documents, par exemple les navigateurs. Ils doivent incorporer les normes portant sur la représentation des documents, soit les jeux de caractères et les langages de balisage. Ici aussi, une détection automatique de la langue est nécessaire pour afficher correctement le document du premier coup.

Acquisition des documents

Du point de vue technologique, si les jeux de caractères universels sont tolérés par le fournisseur et le client, le téléchargement ou la copie de documents numériques sont insensibles à la langue. Toutefois, dans le cas de l'acquisition d'un document papier qui doit être numérisé, il peut être nécessaire d'utiliser des logiciels de numérisation qui sont, eux, dépendant des langues. Les logiciels de reconnaissance optique des caractères peuvent utiliser des informations sur les propriétés de la langue reconnue pour guider le processus de reconnaissance, dans le cas d'ambiguïtés ou bien dans la correction d'erreurs.

Description des documents

Le traitement des documents passe d'abord par une première étape de description des propriétés dites externes au document : le titre, l'auteur, la maison d'édition, la taille, etc. Ces propriétés définissent ce que l'on pourrait appeler un premier ensemble de métadonnées sur le document (ce qu'ont fait les bibliothécaires et documentalistes depuis des

[9] http://www-rali.iro.umontreal.ca/ProjetSILC.fr.html
[10] Si l'on exclue du moins les traductions littéraires.

siècles a trouvé récemment le nouveau nom de métadonnées). Différents schèmes de métadonnées peuvent être utilisés. Ce qu'on entend par schèmes, ce sont des ensembles déterminés d'éléments jugés essentiels pour décrire adéquatement une ressource. Le schème peut être considéré comme l'équivalent d'une notice bibliographique traditionnelle.

Dans l'univers numérique, certaines métadonnées s'ajoutent : le format du fichier, la version, l'adresse URL de la ressource, etc. L'initiative du *Dublin Core*[11] et celle du *Resource Description Format*[12] sont des exemples de schèmes de métadonnées spécifiques aux documents numériques.

Dans un souci d'universalité linguistique, il serait important de se doter de schèmes de métadonnées dont les étiquettes mêmes sont multilingues.

Notons que l'utilisation de logiciels d'identification de la langue et de l'encodage peut permettre de déterminer quel schème de métadonnées utiliser, c'est-à-dire, quelle version linguistique.

Analyse documentaire : indexation, classification, condensation des documents

Cette étape de la chaîne documentaire vise à produire, pour chaque document, une représentation succincte qui sera utilisée pour le repérage. Dans le cas de l'indexation, on produira des termes qui saisissent l'essentiel des thématiques abordées ; les termes seront vraisemblablement tirés d'un langage documentaire comme un thésaurus. La condensation produira un résumé qui capte l'essentiel de la discussion. Pour la classification, on identifiera, dans un plan de classification choisi, dans quelle classe on peut situer le document ; le résultat sera alors l'énoncé de la classe en question.

Que le mode de production de ce deuxième type de métadonnées (qui portent sur le contenu intellectuel des documents) soit humain ou automatique, des ressources ou outils documentaires seront nécessaires, principalement des langages documentaires comme les thésaurus et les plans de classification. En contexte multilingue, ces outils devront être disponibles pour chaque langue, avec des correspondances entre les versions.

Si le but visé est de produire automatiquement ces métadonnées, il faut des systèmes de traitement automatiques, basés sur des ressources linguistiques, des analyseurs linguistiques automatiques génériques et des systèmes d'analyse documentaire automatique (indexation, condensation, classification automatiques) ; chacun est commenté ci-dessous. Également, il faudra des ressources documentaires numériques (les langages documentaires décrits ci-dessus, lisibles par ordinateur).

Ressources linguistiques

Diverses ressources dites linguistiques sont utilisées par les outils de traitement mentionnés ci-dessous. Elles encodent des connaissances sur les propriétés de chaque langue et guident le travail des analyseurs. Notons les dictionnaires (unilingues, et peut-être multilingues) ; les terminologies, ontologies ou taxonomies liées aux unités lexicales présentes dans une langue (qui pourront être différentes des thésaurus documentaires) ; les listes de fréquences de mots, éventuellement par domaine ; les répertoires de mots-outils utiles pour isoler le vocabulaire significatif ; les corpus, associés à chaque langue, nécessaires à la détermination de ces éléments. Ce sont autant de ressources qui doivent être développées pour chaque langue à traiter.

Analyseurs linguistiques automatiques

Pour soutenir les analyseurs spécialisés dans l'analyse documentaire, il faudrait en fait plusieurs analyseurs unilingues, dont les principaux types suivants : outils de segmentation des paragraphes en phrases (la segmentation dépend des règles de typographie, de majuscules, d'abréviation propres à chaque langue) ; lemmatiseurs (*POS taggers*), ou encore règles de morphologie, afin de reconnaître les mots d'une langue, leur racine, leurs suffixes et préfixes, leur catégorie grammaticale ; outils d'extraction ou de repérage de termes (expressions nominales, multitermes, unités nominales complexes), dans le but de fournir une indexation automatique basée sur les termes principaux par exemple ; identification de noms propres, afin de fournir une indexation par nom propre ou bien de classifier les documents sur la base des entités (personnes, organismes) mentionnées dans le document.

Divers autres types d'outils plus sophistiqués seraient éventuellement utiles : analyseurs syntaxiques qui identifient les composantes des phrases et les liens qui les unissent ; analyseurs sémantiques (lexicaux, phrastiques ou textuels) qui rendent compte du contenu conceptuel ou propositionnel des textes. Comme nous le verrons ci-dessous cependant, ils demeurent accessoires dans le traitement documentaire.

---

[11] http://dublincore.org
[12] http://www.w3.org/RDF/

Logiciels d'analyse documentaire automatique

Des logiciels spécialisés dans l'indexation, la classification ou la condensation de documents s'appuient sur les analyseurs linguistiques (et aussi éventuellement sur des techniques davantage statistiques mais dépendantes néanmoins des langues) pour produire automatiquement les métadonnées attendues. Ces derniers ne couvrent pas toute la dimension linguistique de la tâche d'analyse documentaire, se limitant généralement aux interactions intraphrastiques. Une sensibilité à la langue doit être incluse au niveau supérieur d'analyse documentaire pour traiter le texte en entier.

Stockage des documents

Pour le stockage des documents, la problématique linguistique peut être circonscrite à la possibilité de représenter adéquatement les jeux de caractères et d'utiliser des formats de fichiers normalisés, énoncées précédemment.

Diffusion des documents

Un système documentaire multilingue doit être capable de diffuser dans la langue du destinataire. On voit réapparaître ici l'exigence minimale de permettre différents jeux de caractères. On comprend que la diffusion bénéficiera de l'existence de métadonnées en plusieurs langues. Cela peut prendre la forme de la traduction des représentations des documents, de l'utilisation de thésaurus multilingues ou des pointeurs vers des ressources utiles (par exemple, un hyperlien vers un système de traduction automatique en ligne).

En d'autres termes, les ressources multilingues peuvent être introduites à différents endroits dans la chaîne. Dans l'éventualité où toute la chaîne est assurée par un même organisme, le multilinguisme peut être implanté à différents endroits, en quelque sorte indifféremment. Mais si plus d'un joueur se trouve à intervenir dans la chaîne, chacun doit fournir des fonctionnalités multilingues si l'on veut s'assurer que celles-ci soient présentes du début à la fin.

La traduction des pages du site hôte vers une ou plusieurs autres langues peut bien sûr davantage favoriser la visibilité des documents ; une gestion efficace de ces traductions reposera, encore une fois, sur la possibilité de représenter en parallèle (par le langage de balisage, par exemple), les différentes versions linguistiques.

Notons finalement que notre discussion n'aborde pas les problèmes, importants, de sécurité des documents, de droits d'accès ou de droits d'auteur. La dimension linguistique ou multilingue ne semble pas pertinente de prime abord pour les questions d'encryptage ou de signature numérique, par exemple.

Repérage des documents

Un utilisateur peut choisir de faire une recherche de documents de différentes façons. L'aspect linguistique prendra alors des formes différentes.

Outils de recherche par requêtes

La problématique de la recherche d'information est bien documentée (voir notamment Baeza-Yates et Ribeiro-Neto, 1999). Pour la dimension linguistique, cependant, les références se font plus rares.

Du côté monolingue, on pense à des langages de requêtes qui, d'une part, peuvent être utilisés pour toutes les langues (par exemple, tolérants aux accents et autres diacritiques), et d'autre part sont sensibles à une langue particulière ; par exemple, des langages qui soient capables de faire une expansion de requête à l'aide d'un thésaurus dans la langue désirée, ou qui utilisent un algorithme de recherche capable de lemmatisation selon la langue (reconnaître les pluriels et les différentes formes du verbe, notamment).

Dans une optique multilingue, les exigences sont plus grandes encore. On cherche des capacités de rechercher dans une langue des documents qui pourraient être dans des langues diverses ; les résultats des requêtes pourraient, eux, être monolingues ou multilingues, selon les préférences de l'utilisateur. Également, on voudrait guider une recherche monolingue dans un environnement multilingue, en spécifiant par exemple au moteur que l'on recherche « file » (en français) et non « *file* » (en anglais) ; ceci implique ici aussi d'avoir des systèmes d'identification de la langue pour ne rechercher que dans les documents écrits dans la langue désirée.

Outils de butinage

Dans l'optique d'une recherche d'information par butinage ou furetage de la collection (et non basée sur une requête), ce qui est présenté à l'utilisateur s'apparente à des plans de classification ou d'autres structures hiérarchiques. Comme pour les thésaurus ci-dessus, dans un contexte multilingue, les hiérarchies seraient idéalement multilingues (développées indépendamment ou bien traduites), laissant à l'utilisateur le choix de la langue.

Il est à noter que, pour les navigateurs Web grand public, les préférences linguistiques fixées par chaque utilisateur sont utilisées afin de déterminer quelle page d'accueil utiliser. Même en l'absence de préférences linguistiques, le choix peut être dicté par l'adresse IP du client. Bien sûr, la différence entre ces pages d'accueil dépasse le purement linguistique et relève des stratégies de marketing. On peut se demander quels seraient les autres facteurs déterminants dans la communication scientifique.

Préservation des documents

Les documents numériques (ou plus précisément l'information numérique) ont un statut temporel paradoxal, à la fois intemporel (puisque leur contenu n'est pas forcément matériel, une simple suite de bits) et limité dans le temps, par le matériel de support et les environnements logiciels nécessaires à leur consultation. Ces derniers évoluent rapidement, ce qui met en péril notre capacité à consulter leur contenu à l'avenir.

L'aspect matériel de la préservation des documents n'est pas sensible à la langue. L'aspect logiciel, lui, rejoint la problématique des formats normalisés de documents. Ceci prend une grande ampleur dans le cas de périodiques électroniques. La communauté scientifique veut être assurée d'avoir accès à ces périodiques dans un avenir plus ou moins lointain, ce qui rend primordiale la définition de formats normalisés, indépendants des plates-formes actuelles. Et c'est justement à ce moment que les problématiques multilingues devraient être prises en compte, si l'on veut assurer l'existence d'environnements qui permettent des documents dans des langues quelconques.

Dans l'optique d'assurer l'authenticité et l'intégrité de l'information contenue dans les documents, il est utile de mentionner certains outils linguistiques qui permettent des vérifications sur le contenu des documents ; encore ici, des logiciels de détection de la langue peuvent aider à signaler une corruption des fichiers (si, par exemple, la langue détectée ne correspond pas à la langue affichée dans les métadonnées, ou si aucune langue ne peut être identifiée). Également, des logiciels peuvent comparer deux fichiers pour en identifier les similarités et les différences (sur la base du contenu textuel et lexical), ce qui permet entre autres de détecter une détérioration ou encore un cas de plagiat. Ces outils sont basés sur l'exploitation de propriétés linguistiques propres à chaque langue.

État actuel des technologies

À la lumière de cet éventail, nous brosserons un tableau de l'état actuel des technologies qui facilitent la communication scientifique multilingue. Le lecteur ne s'étonnera pas de constater plusieurs lacunes. Pour avoir une idée de la distance à parcourir, nous examinerons du même coup, pour ce qui n'est pas disponible, le niveau de difficulté associé.

Nous aborderons d'abord les ressources (données statiques) puis les outils de traitement. Pour chacun, nous effectuerons une distinction binaire assez simpliste, nous en convenons, entre ce qui est plutôt disponible et ce qui pose davantage de problèmes.

Ressources assez largement disponibles

Les technologies actuelles en place permettent assez bien la première étape de représentation de documents multilingues.

Jeux de caractères

Avec l'adoption d'Unicode (recommandation ISO/IEC 10646 en 1993) et ses divers codages (UTF-1 à UTF-8) on s'assure de pouvoir représenter et traiter adéquatement toutes les langues du monde. Tout système qui ne s'y conformerait pas accuserait un sérieux retard technologique.

Formats de documents

Le format ASCII et le jeu de caractères Unicode forment le dénominateur commun assurant l'interopérabilité des systèmes et la portabilité des fichiers. C'est vers cette dernière solution que l'on peut se tourner. Par contre, les formats propriétaires (liés aux logiciels de traitement de texte grand public, par exemple) n'étant ni publics ni normalisés, ils présentent des problèmes lorsque l'on veut les intégrer à une chaîne de traitement générique, en l'occurrence ici multilingue.

Langages de balisage

En revanche, un langage comme XML (recommandation du W3C datant d'octobre 2000[13]) permet de baliser les documents à l'aide d'étiquettes qui explicitent le contenu sémantique de chaque élément du document. Spécifiquement, des normes de description existent pour les textes littéraires (suite aux travaux du *Text Encoding Initiative* ou TEI, Sperberg-McQueen et Burnard, 1995[14]), les textes documentaires (DOCBOOK[15]), les documents d'archives (*Encoded Archival Description*[16]) et les bases de données en sciences sociales (*Data Documentation Initiative*[17]). C'est déjà là un excellent début.

XML est multilingue par définition, c'est-à-dire que le jeu d'étiquettes de balisage est défini par son créateur dans la langue de son choix. XML est défini sur le jeu de caractères Unicode UTF-8 et permet donc réellement des étiquettes dans une langue quelconque. Par contre, les DTD (*Document Type Definition*[18]) existantes qui servent à décrire les documents, elles, sont monolingues dans les faits. Les jeux d'étiquettes définis dans une DTD sont fixes et on ne permet pas, par exemple, un affichage d'étiquettes traduites pour une même DTD. Les DTD ne sont pas non plus nécessairement adaptées aux différences culturelles qui apparaissent dans la structure même des documents (prenons comme exemple bien connu la position de la table des matières dans les ouvrages de tradition française vs. anglo-saxonne)

Dictionnaires unilingues et multilingues

Il existe un certain nombre de producteurs de ressources linguistiques, dont ELRA (*Evaluations and Language Resources Distribution Agency*[19]) et le LDC (*Linguistic Data Consortium*[20]) qui distribuent (gratuitement ou non, selon le cas) des dictionnaires pour un bon nombre de langues que l'on pourrait appeler « majeures ». ELRA offre notamment les dictionnaires suivants (en date du 30 avril 2003) :

- unilingues : pour la plupart des langues européennes majeures (allemand, anglais, danois, espagnol, français, grec, italien, néerlandais, portugais, suédois, etc.), ainsi que pour le bulgare, le catalan, le polonais, le portugais brésilien, le turc ;
- bilingues ou multilingues : allemand, anglais, bosniaque, coréen, croate, danois, espagnol, estonien, finnois, français, grec, hongrois, islandais, italien, japonais, néerlandais, polonais, portugais, portugais brésilien, roumain, russe, serbe, suédois, tchèque (toutes les paires ne sont pas représentées ; le plus grand nombre de paires contient l'allemand, l'anglais, l'espagnol et le français).

Par contre, pour plusieurs langues « également majeures » (des langues asiatiques notamment) et un très grand nombre d'autres langues, les ressources sont rares, voire inexistantes.

Ressources plus problématiques

Les ressources décrites ci-dessous ont une accessibilité beaucoup plus limitée. Ceci est souvent dû à l'effort considérable nécessaire à leur développement.

Thésaurus multilingues

Il existe un certain nombre de thésaurus multilingues, dont certains sont assez connus : AGROVOC[21], EUROVOC[22], le thésaurus de l'UNESCO[23], etc. De plus, les besoins qui les justifient ainsi que des règles d'élaboration sont documentés, entre autres, dans Association français de normalisation (1990) et dans Hudon (1997).

Mais l'élaboration de thésaurus multilingue soulève des problèmes importants. Il est faux, en règle générale, que la traduction d'un thésaurus unilingue produira un thésaurus multilingue que l'on pourra utiliser de la même façon. On se heurte à des problèmes (bien connus en traduction) de non-correspondance de certains termes d'une langue à l'autre, de

[13] http://www.w3.org/TR/2000/REC-xml-20001006
[14] http://www.tei-c.org/
[15] http://www.docbook.org/
[16] http://www.loc.gov/ead/
[17] http://www.icpsr.umich.edu/DDI/index.html
[18] En quelque sorte, une déclaration de la façon dont les étiquettes sémantiques sont combinées pour former la description d'un document de ce type.
[19] http://www.elda.fr
[20] http://www.ldc.upenn.edu
[21] http://www.fao.org/agrovoc/
[22] http://www.psp.cz/cgi-bin/eng/kps/knih/ee_geninfo.htm
[23] http://www.ulcc.ac.uk/unesco/

traductions multiples selon le sens, de l'absence de termes dans une langue et de découpage différent d'un champ sémantique selon les langues. Fietzer (2002, p. 89) reprend la gradation de relations entre des paires de termes issus de thésaurus multilingues, présentée dans les recommandations de l'AFNOR : équivalence exacte, inexacte, partielle, d'une à plusieurs, ou non-équivalence. Cette gradation donne une idée des difficultés auxquelles on peut s'attendre dans l'élaboration de thésaurus multilingues. Les divers travaux en traduction automatique sont autant de témoignages de la difficulté inhérente à la traduction (voir notamment Kay, 1997). Et comme les thésaurus sont des structures hiérarchiques, l'absence de correspondance entre les nœuds de deux hiérarchies pourra voir des répercussions importantes sur les deux sous-structures correspondantes.

Plans de classification

Au sujet des plans de classification, une première source de difficulté réside dans le fait que la communauté informatique et la communauté documentaire ne s'entendent généralement pas sur la nature des plans de classification utiles pour organiser les documents.

Dans les milieux documentaires, peu de plans de classification « traditionnels » sont utilisés pour classer les ressources numériques (cf. McKiernan, page Web). Parmi ceux-là, les plans multilingues sont rares. Par ailleurs, on peut voir une certaine équivalence formelle entre les thésaurus et les plans de classification (bien que cette équivalence soit démentie par les praticiens). On peut envisager, par conséquent, de voir apparaître dans l'élaboration de plans de classification multilingues des problèmes équivalents à ceux pour les thésaurus multilingues. Cependant, les plans de classification étant davantage conceptuels, et moins dépendants des mots que le sont les thésaurus, on peut prétendre pouvoir élaborer plus facilement des plans de classification multilingues ; et d'ailleurs, Svenonius (1983) avait remarqué le potentiel du plan de classification Dewey pour le repérage en contexte multilingue.

Dans les milieux informatiques, les systèmes de classification automatique de documents utilisent souvent des plans construits a posteriori, par des méthodes statistico-linguistiques qui se basent sur les occurrences des mots ou concepts dans les documents (par exemple, Aas et Eikvil, 1999 ; Steinbach et al., 2000). Ces plans de classification seront forcément déterminés par la langue des documents. Dès lors, ces approches excluent (ou du moins le font dans une interprétation simple) la gestion simultanée de documents dans des langues différentes.

Schèmes de métadonnées

Les schèmes de métadonnées comme RDF (basé sur XML) ou les balises META de HTML permettent heureusement d'inscrire des métadonnées dans n'importe quelle langue. Il subsiste toutefois trois problèmes dignes de mention. D'abord, les étiquettes mêmes sont en anglais ; ceci peut biaiser la description selon un cadre de référence unique, anglo-américain ou anglo-saxon en l'occurrence. Ensuite, si les langages permettent un encodage quelconque, les logiciels utilisés pour traiter, afficher ou imprimer ces données doivent eux aussi utiliser le même jeu de caractères universel. Enfin, des valeurs exprimées dans une autre langue pourront être inutiles à un utilisateur qui ne comprend pas cette langue ; une traduction de ces valeurs peut être souhaitable (ou bien à la source, à l'aide d'un langage documentaire multilingue, ou bien à la cible, à l'aide d'outils de traduction automatique).

Et, tel que l'on aurait pu prévoir, les schèmes existants sont en grande partie restreints à l'anglais.

Terminologies, ontologies, taxonomies

L'initiative du Web sémantique (Berners-Lee et al., 2001) vise une description conceptuelle du contenu des ressources du Web ; celle-ci utiliserait un vocabulaire déclaré dans des « ontologies » publiques. Il devient alors impérieux de développer lesdites ontologies et de les rendre largement accessibles. L'approche prévoit de multiples ontologies ; le développement de chacune de celles-ci dans différentes langues posera bien sûr la même problématique, évoquée précédemment, que celle des thésaurus multilingues.

L'utilisation du latin dans les taxonomies des sciences naturelles est intéressante. Elle date d'un temps où une langue, le latin, jouissait d'une hégémonie pour la communication scientifique (comme peut-être l'anglais aujourd'hui). Cette même langue présente aujourd'hui l'avantage d'être une *lingua franca* universelle ; elle profite de la force de l'inertie de la normalisation répandue et de l'atout de n'être associée à aucun groupe linguistique à l'heure actuelle.

Le « thésaurus » WordNet (en fait pas un thésaurus dans la terminologie documentaire) est largement utilisé dans les contextes de traitement automatique de la langue[24]. Il s'agit d'une ressource unilingue anglaise qui s'avère d'une utilité qui n'est pas sans controverse pour la recherche d'information. Une version multilingue[25] (allemand, espagnol, estonien, français, italien, néerlandais, tchèque) est disponible depuis plusieurs 1999.

[24] Wordnet : http://www.cogsci.princeton.edu/~wn/.
[25] http://www.illc.uva.nl/EuroWordNet/

Corpus associés, par langue

Les organismes ELRA et LDC amassent, en plus des dictionnaires mentionnés ci-dessus, des corpus unilingues ou multilingues. ELRA offre des ressources dans la plupart des langues européennes majeures (anglais, français, espagnol, allemand, italien, suédois, grec, néerlandais, portugais etc.), et d'autres, en quantité plus limitée (turc, japonais, russe, chinois, malais, irlandais, arabe, coréen, chinois). Pour le LDC, on trouve les ressources suivantes :

- corpus : allemand, anglais, arabe, chinois, espagnol, japonais, portugais ; de façon plus limitée : albanais, bulgare, coréen, danois, estonien, français, gaélique, grec, italien, latin, lithuanien, malais, néerlandais, norvégien, russe, serbe, tchèque, tibétain, turc, uzbek, suédois
- lexiques : arabe, anglais, allemand, chinois, espagnol, japonais, néerlandais

Les ressources linguistiques habituellement dérivées de ces corpus (fréquences de mots, listes de mots-outils, etc.) seront bien sûr limitées par la disponibilité du corpus dans chaque langue.

Nous examinerons maintenant la disponibilité d'outils et de systèmes capables de traiter plusieurs langues ; ces derniers se font rares, souvent à cause des difficultés associées au développement.

Outils assez performants

Pour un certain nombre de tâches, il existe déjà un bon éventail de technologies multilingues.

Outils de création de documents

Les producteurs d'outils de traitement de texte, d'éditique et de conception de pages Web offrent aujourd'hui leurs produits dans de nombreuses langues, avec quantité d'outils linguistiques associés[26]. Aucune difficulté technologique ne freine ici les développements.

Identification de la langue

On possède aujourd'hui de plus d'un système d'identification automatique de la langue (et de l'encodage). Une liste est présentée à l'adresse http://odur.let.rug.nl/~vannoord/TextCat/competitors.html ; le nombre de langues traitées y varie entre 2 et 69, avec des taux de succès variables. L'identification de la langue serait efficace à plus de 97 % à partir de phrases de 11 à 15 mots pour 9 langues européennes (Grefenstette, 2001) (pour le français, la technique par mots courts serait efficace à 96 % pour des suites de 6 à 10 mots). Cependant, il est important de préciser que cette tâche se complique dès que le nombre de langues (et d'encodages) à reconnaître croît ; ces statistiques prometteuses devront être révisées selon les développements futurs.

Reconnaissance optique des caractères

La tâche de reconnaissance optique des caractères en est une assez bien réussie par les logiciels courants, pour les langues européennes majeures et imprimées (c'est-à-dire, non manuscrites). Il faut souligner par contre que, lors de toute numérisation, un certain nombre d'erreurs subsistent et que des outils linguistiques tels des dictionnaires ou des correcteurs d'orthographe peuvent servir à améliorer les résultats ; ceux-ci sont, bien sûr, dépendants de la langue traitée, et leur disponibilité varie tel que décrit ci-dessus.

Analyseurs linguistiques : segmentation, morphologie, lemmatisation

La segmentation en phrases peut être effectuée plus ou moins facilement pour des langues où il est facile de déterminer les frontières de mots et de phrases ; elle est plus problématique pour les langues comme le chinois où aucun espace ne sépare les mots. En fait, tout système de traitement de la langue ou du texte doit comporter une certaine version d'un segmenteur, préalable à tout autre traitement.

Le cas de l'analyse morphologique est « assez » facile. De nombreux algorithmes plus ou moins puissants sont disponibles, plusieurs gratuitement pour les fins de recherche. Les lemmatiseurs sont à peu près du même ordre (voir notamment Brill, 1992). On rapporte des taux de précision au-delà de 96 % dans les meilleurs cas (par exemple, Leech et Smith, 2000). Mais pour tous ces cas, les langues traitées sont les langues européennes majeures, avec quelques exceptions (notamment, le finnois ou les langues sémitiques).

---

[26] Il est intéressant de remarquer toutefois que les fonctionnalités ne sont pas toutes aussi performantes selon les langues, comme une utilisation de la fonction de Synthèse automatique de Word nous a permis de constater ; elles reposent sur l'élaboration de ressources linguistiques internes qui peuvent être à des niveaux différents de développement.

Outils plus limités

Pour des traitements plus sophistiqués (ceux qui tentent de capter jusqu'à un certain point le sens contenu dans les documents), la gamme d'outils disponibles est plus limitée.

Analyseurs linguistiques : analyse syntaxique ou sémantique

Des analyseurs puissants pour détecter les structures syntaxiques et le sens des textes seraient indéniablement utiles. Malheureusement, les performances des outils existants sont plutôt limitées et par conséquent ils sont rarement mis à contribution dans le traitement des textes. Notons certaines exceptions pour l'utilisation d'analyse syntaxique locale pour identifier des groupes nominaux (Jacquemin, 2001, par exemple) ou l'analyse textuelle par la structure rhétorique pour effectuer des résumés automatiques (Marcu, 1996, 1999). Leur applicabilité est limitée ; les langues traitées ne dépassent habituellement pas les langues européennes majeures (spécifiquement, français, anglais, espagnol).

Analyse documentaire

Parmi les outils linguistiques nécessaires en tant que blocs élémentaires des outils d'analyse documentaire, seuls certains ont atteint un niveau de performance acceptable.

L'extraction ou le repérage de termes est un domaine qui a fait l'objet de beaucoup de travaux de recherche ces dernières années ; le programme du colloque Terminologie et intelligence artificielle 2001 (TIA-2001[27]) donne un aperçu de méthodes utilisées. Même s'il subsiste des difficultés, les algorithmes actuels atteignent des performances tout à fait acceptables. Cependant, les résultats encourageants portent sur l'anglais, le français, l'espagnol, l'italien, et quelques autres. Les langues à cas (le finnois par exemple) ou à morphologie nominale complexe (l'allemand par exemple) posent davantage de problèmes. Dans le même ordre d'idées, l'identification de noms propres (en anglais, du moins) atteint des niveaux de précision intéressants. Les participants aux conférences TREC (*Text Retrieval Conference*[28]), notamment, ont développé des systèmes capables non seulement de relever les noms propres, mais également de déterminer si le nom propre désigne une personne, un organisme, un produit ou un lieu géographique. Ces algorithmes ont été peu développés pour d'autres langues, comme le français (Kossein et Lapalme, 1998, entre autres).

Pour l'indexation automatique, ces deux types de travaux sont extrêmement utiles. L'objectif d'indexation automatique de qualité humaine n'est toutefois pas encore atteint, mis à part quelques implémentations circonscrites, comme par exemple l'indexation automatique dans le domaine de l'aérospatiale, au *Center for Aerospace Information* de la NASA (Silvester et al., 1994), pour n'en citer qu'un exemple.

Il y a peu de travaux en classification automatique « traditionnelle », c'est-à-dire à l'aide d'un plan de classification élaboré a priori (exceptions notables, les travaux du OCLC avec la classification décimale Dewey[29]). Par contre, un grand nombre de travaux en informatique portent sur la catégorisation, la classification ou le « clustering » de documents par des méthodes variées (voir le paragraphe ci-dessus sur les plans de classification). Les résultats divergent considérablement de ceux de la classification documentaire. Les algorithmes, généralement mathématiques et statistiques, sont indépendants des langues et pourraient être appliqués à n'importe laquelle. Dans la pratique cependant, les recherches portent essentiellement sur l'anglais. Quand des ressources linguistiques sont utilisées, le traitement de l'anglais est prédominant.

Enfin, la recherche en condensation automatique est en plein essor actuellement. Cela se traduit non seulement par un grand nombre de projets de recherche, de colloques et de subventions[30], mais aussi par un bon nombre de produits commerciaux connaissant des succès variables. Les langues traités peuvent inclure le français, l'espagnol ou le suédois, mais encore ici le traitement de l'anglais est plus répandu.

Langages de requêtes sensibles à la langue

Au minimum, pour accepter des requêtes dans n'importe quelle langue, le langage de requêtes d'un moteur de recherche doit utiliser des jeux de caractères normalisés. Un langage de requêtes « intelligent » saurait de plus reconnaître les plueirls ou autres marques de flexions, afin de rechercher toute forme dans les documents. Les langages existants le font pour un nombre de langues limité. On peut noter l'anglais, le français, l'allemand, l'italien, l'espagnol pour certains logiciels (le japonais, le finnois s'additionnent parfois). Quelques exemples de systèmes : TransSearch (du RALI[31]), DioMillennium (de Delphes[32]), TANGO (de Alis[33]). On ne tient pas compte ici des langages de requêtes qui permettent la troncature (ex.

[27] http://conferences.atala.org/conferences/fiches/tia2001.html
[28] http://trec.nist.gov/
[29] http://orc.rsch.oclc.org:6109/
[30] L'adresse suivante contient un grand nombre de références pertinentes : http://perun.si.umich.edu/~radev//summarization/.
[31] www-rali.iro.umontreal.ca/ProjetTransSearch.fr.html
[32] http://www.delphes.com

chercher cheva* dans le but de repérer à la fois « cheval » et « chevaux »), puisqu'alors c'est l'utilisateur qui fait preuve d'une sensibilité à la langue.

Il importe cependant de noter que la technologie existante permet la sensibilité à la langue dès lors qu'elle dispose d'un lemmatiseur ou d'un analyseur morphologique performant.

Recherche d'information translinguistique (*CLIR*)

Parmi les approches de recherche d'information (*information retrieval*), un nouveau courant a vu le jour ces dernières années, celui de la recherche d'information translinguistique (*CLIR* ou *cross-linguistic information retrieval*). Une synthèse en est présentée dans Oard et Diekema (1998) ; pour un portrait plus récent du domaine, voir les actes de l'atelier inscrit à la conférence *SIGIR-2002*[34]. Ces travaux visent à permettre la recherche de documents dans des langues variées à partir de requêtes dans des langues variées elles aussi, potentiellement différentes de la langue des documents pertinents. Les systèmes utilisent pour la recherche une combinaison de méthodes statistiques de repérage et de méthodes (linguistiques ou statistiques) de traduction. Les résultats finaux peuvent être affichés soit dans la langue de la requête, soit dans celle du document.

Bien que ce domaine tout nouveau n'ait pas atteint la maturité et qu'on puisse espérer des améliorations notables, les taux de succès laissent considérablement à désirer. À la conférence CLEF 2003 (*Cross Linguistic Evaluation Forum*[35]), on rapporte des taux de précision (en moyenne sur différents points de mesure) de l'ordre de 30 % (voir la documentation sur les résultats pour des explications sur l'interprétation à donner à ce chiffre). Les langues assez bien couvertes sont les langues européennes majeures, quelques langues asiatiques (chinois, japonais et coréen) et la langue arabe.

Traduction automatique

Il convient de dire, finalement, quelques mots sur la traduction automatique. Ses performances sont sous-optimales, malgré des décennies de travaux : un certain succès dans des domaines limités, quelques offres commerciales gratuites (Reverso en ligne, BabelFish en ligne, de SYSTRAN) ou payantes. Quiconque a tenté de se servir des systèmes commerciaux de traduction automatique a vite pu en constater les limites. Pour des mots isolés comme des mots-clés ou des termes d'indexation, de surcroît, le problème de polysémie est grand et les traductions proposées seront souvent déplacées.

Pour revenir à notre préoccupation d'assurer le traitement égal de toutes les langues, il faut souligner que les systèmes de traduction automatique existants n'existent que pour certaines paires de langues (français-anglais, anglais-russe, etc.) et que les performances peuvent s'avérer assez pauvres pour d'autres paires, lorsqu'elles sont disponibles.

Bien sûr, dans la mesure où l'on voudrait traduire une fois pour toutes des corpus limités (par exemple, les thésaurus ou les ontologies pour le Web sémantique), il est plus approprié d'avoir recours à la traduction humaine.

Avant d'émettre nos conclusions sur l'état actuel et les perspectives futures, nous trouvons intéressant d'identifier quelques-uns des joueurs dans cette industrie de diffusion de documents multilingues. La gamme de communautés impliquées témoigne de l'intérêt posé par cette problématique et laisse entrevoir les enjeux et les leaders de demain.

Joueurs et projets

En plus des sciences de l'information, plusieurs communautés sont touchées par la problématique de la diffusion de la communication multilingue.

Le monde du TAL : Le traitement automatique de la langue représente un domaine fertile de développement d'outils, ressources et systèmes de traitement. Quelques projets portent spécifiquement sur la dimension multilingue. Notamment, le rapport *Multilingual Information Management*[36], issu d'un atelier sur le traitement multilingue de l'information, a été présenté conjointement aux gouvernements américain et européen. Les linguistes informaticiens impliqués dans cet ouvrage contribuent notamment aux travaux de recherche en recherche d'information translinguistique. Également, un chapitre d'un rapport (Cole et al., 1998) sur les technologies linguistiques, mandaté par la *National Science Foundation* ainsi que par la Commission européenne, porte sur l'aspect multilingue. On note également l'important projet de recherche sur le repérage et la condensation de l'information TIDES, de DARPA, qui regroupe des chercheurs tant en TAL qu'en sciences de l'information ; voir notamment Gey et al. (2002). Ainsi, les gouvernements nationaux et fédéraux ont un intérêt marqué pour cette problématique. De plus, des groupes de recherche comme celui de Xerox, à Grenoble

[33] http://www.alis.com
[34] http://ucdata.berkeley.edu/sigir-2002/
[35] http://clef.iei.pi.cnr.it:2002/
[36] http://www-2.cs.cmu.edu/~ref/mlim/

(*MultiLingual Theory and Technology Group*) témoignent également de l'opportunité commerciale des travaux en TAL multilingue.

Le monde de la localisation : L'industrie de la localisation se préoccupe de la traduction et de l'adaptation de logiciels ou de sites Web pour une communauté linguistique et culturelle donnée. Le développement d'outils qui permettent le développement simultané de versions dans des langues différentes touche un grand nombre des problématiques présentées ici ; cela soulève par ailleurs des questions de vérification d'intégrité des versions dans chaque langue. Ici, les joueurs sont à la fois les producteurs de logiciels et de sites Web, et les fournisseurs de services en localisation et en traduction.

Le monde de l'édition : Les créateurs de documents ont intérêt à produire au départ des documents selon des formats normalisés, qui pourront profiter des outils de traitement subséquents dans la chaîne. Il y va de la visibilité de leurs produits.

Le monde de la normalisation : On voit la création de groupes de travail amenés à réfléchir sur les normes de représentations des documents numériques, ainsi que sur les normes de schèmes et de représentations de métadonnées. Une préoccupation est la définition de versions qui permettent plusieurs langues, spécialement pour ces métalangages comme XML où l'utilisateur peut définir lui-même son jeu d'étiquettes et de valeurs. Divers secteurs d'activités sont représentés : producteurs de logiciels, producteurs de documents, mais aussi gouvernements et autres collectivités.

Le monde de la « société de l'information » : Ici, les enjeux incluent l'accessibilité de l'information à tous, y compris les communautés linguistiques autres qu'anglophones. Les initiatives européennes sont plus avancées qu'en Amérique du Nord, ce qui s'explique facilement par les contextes géolinguistiques respectifs des deux territoires. Notons l'impact des programmes de recherche (FP1 à FP6) de la Commission européenne, qui suscitent la présentation de projets de recherche divers correspondant aux priorités définies annuellement. Le programme 2002-2006 (FP6) retient notamment comme priorité les technologies de l'information (*IST – Information Society Technologies*), pour assurer le leadership européen en matière de technologies génériques et appliquées qui sont au cœur de l'économie du savoir. Un champ d'action plus proprement linguistique est défini sous la rubrique *IST-4.2 Knowledge technologies and digital content*, sous-rubrique *IST-2002-2.3.1.7 - Semantic-based knowledge systems* ; ses objectifs sont de développer des systèmes sémantiques pour acquérir, organiser, traiter, partager et utiliser les connaissances du contenu multimédia. Les organismes gouvernementaux, para-gouvernementaux et les organisations non gouvernementales (ONG) y sont aussi actifs que les partenaires privés.

Toute cette présentation n'a porté que sur le texte écrit. Des joueurs additionnels sont à prévoir pour d'autres types de contenus (images, vidéo, son, parole), et des techniques additionnelles aussi.

Conclusion

Un service documentaire scientifique numérique multilingue entreposerait les documents qu'on lui fournit à l'aide d'un jeu de caractère normalisé (tous les documents pourraient y être convertis, s'ils ne le sont pas déjà). Il pourrait veiller à normaliser la structure de balisage des documents (lorsque c'est possible) afin de pouvoir reconnaître celle-ci indépendamment de la langue du document. Son interface de requête avec opérateurs linguistiques permettrait de rechercher un mot sous toutes ses formes dans la langue spécifiée ; le moteur de recherche pourrait avoir recours à un thésaurus unilingue ou multilingue pour faire des expansions de la requête, si désiré. La recherche pourrait se limiter aux métadonnées associées aux documents (traduits en amont dans la chaîne de traitement ou sur le vif) ou bien tenter de satisfaire à la requête à l'aide d'algorithmes de repérage qui examinent le plein texte du document. L'affichage des résultats tiendrait compte de la langue d'affichage préférée de l'utilisateur et pourrait inclure une notice bibliographique préétablie (éventuellement traduite) et inclure un résumé produit automatiquement.

Dans cet univers, les documents seraient accessibles avec autant d'aisance, peu importe leur langue d'expression. Il serait intéressant, alors, de voir se dessiner le profil de la communication scientifique par langue et par région du monde.

Le multilinguisme est plus qu'un souhait, c'est une réalité. L'éventail dressé ici suggère des pistes de recherche, plusieurs portant sur le développement d'outils de traitement linguistique.

Car où le bât blesse, c'est lorsqu'on examine la possibilité de traiter des langues autres que les langues européennes majeures. Souvent, on constate qu'aucun obstacle technologique n'empêche le traitement d'une langue donnée, et pourtant les efforts de développement se font attendre. Ceux-ci sont limités par les ressources financières, matérielles et humaines nécessaires. De plus, certaines langues posent des problèmes particuliers plus difficiles à résoudre avec les techniques actuelles. Ces techniques de base ont été déterminées par le choix des langues sur lesquelles les travaux initiaux ont été entamés.

En l'attente de systèmes de traitement automatiques multilingues performants, il y a lieu de mettre des efforts dans les projets de normalisation des métadonnées, le développement de ressources multilingues telles que les ontologies du Web sémantique et le soutien aux efforts de normalisation portant sur la représentation et la structuration des documents.

Références

Les références aux ressources sur le Web ont été données dans le texte principal, sous la forme de notes de bas de page.

Aas, Kjersti ; Eikvil, Line. *Text Categorisation: A Survey*. Technical report, Norwegian Computing Center, June 1999. Disponible en ligne. Page consultée le 5 mai 2003. Adresse URL : http://citeseer.nj.nec.com/aas99text.html.

Association française de normalisation. *Principes directeurs pour l'établissement des thésaurus multilingues, Z47-101*. Paris : AFNOR, 1990.

Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier. *Modern Information Retrieval*. Reading, Mass. ; Don Mills, Ont. : Addison-Wesley, 1999.

Berners-Lee, Tim ; Hendler, James ; Lassila, Ora. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, avril 2001. Disponible en ligne. Page consultée le 28 avril 2003. Adresse URL : http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Borgman, Christine ; Furner, Jonathan. Scholarly Communication and Bibliomtetrics. In : *Annual Review of Information Science and Technology*, vol. 36, 2002, pp. 3-72

Brill, Eric. A simple rule-based part of speech tagger. In : *Proceedings of the Third Annual Conference on Applied Natural Language Processing*, ACL, 1992, pp. 152-155.

Cole, Ron ; Mariani, Joseph ; Uszkoreit, Hans ; Varile, Giovanni Battista ; Zaenen, Annie ; Zampolli, Antonio. *Survey of the State of the Art in Human Language Technology*. Cambridge : Cambridge University Press, 1998. Disponible en ligne. Page consultée le 2 mai 2003. Adresse URL : http://cslu.cse.ogi.edu/HLTsurvey/.

ECOTEC Research & Consulting Ltd. *Final Evaluation of the Multi-lingual Information Society Programme*, 2000. Disponible en ligne. Page consultée le 2 mai 2003. Adresse URL : http://europa.eu.int/comm/information_society/evaluation/pdf/report1mlis_en.pdf.

Fietzer, William. Integrating Metadata Frameworks into Library Description In : *Libraries, the Internet, and Scholarship. Tools and Trends Converging*. New York ; Bâle : Marcel Dekker, 2002.

Gey, F.C.; Chen, A.; Buckland, M.; Larson, R. Translingual vocabulary mappings for multilingual information access. In : *SIGIR '02: 25th International Conference on Research and Development in Information Retrieval*, 2002, pp. 455-456.

Ginsparg, P. First steps toward electronic research communication. *Computers in Physics*, vol. 8, 1994, pp. 390-396.

Hovy, E.H. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. In : *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain, 1998.

Hudon, Michèle. Multilingual thesaurus construction: integrating the views of different cultures in one gateway to knowledge and concepts. *Knowledge organization*, vol. 24, no 2, 1997, pp. 84-91.

Inönü, Erdal. The influence of cultural factors on scientific production, *Scientometrics*, vol. 56, no 1, 2003, pp. 137-146.

Jacquemin, Christian. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, Mass : MIT Press, 2001.

Kay, Martin. The Proper Place of Men and Machines in Language Translation. *Machine Translation*, vol. 12, pp. 3-23, 1997.

Kosseim, Leila ; Lapalme, Guy. Exibum: Un système expérimental d'extraction d'information bilingue. In : *Actes de la Rencontre Internationale sur l'extraction, le filtrage et le résumé automatiques (RIFRA-98)*, 1998, pp. 129-140. Sfax, Tunisia. Disponible en ligne. Page consultée le 2 mai 2003. Adresse URL : http://www.cs.concordia.ca/~faculty/kosseim/Publications/rifra98.pdf.

Leech, Geoffrey ; Smith, Nicholas. *The British National Corpus (Version 2) with Improved Word-class Tagging*. Disponible en ligne. Page consultée le 5 mai 2003. Adresse URL : http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2error.htm.

Marcoux, Yves ; Sévigny, Martin. Why SGML? Why now? *Journal of the American Society for Information Science*, vol. 48, no 7, 1997, Special Topic Issue: Structured Information/Standards for Document Architectures, pp. 584-592.

Marcu, Daniel. Building up rhetorical structure trees. In : *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996, pp. 1069-1074.

Marcu, Daniel. Discourse Trees are Good Indicators of Importance in Text. In : Mani, Inderjeet ; Maybury, Mark. *Advances in Automatic Text Summarization*, Cambridge, Mass. : MIT Press, 1999, pp. 123-136.

McKiernan, Gerry. *Beyond Bookmarks: Schemes for Organizing the* Web. Page consultée le 5 mai 2003. Adresse URL : http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm.

Oard, Doug ; Diekema, Anne. Cross-Language Information Retrieval. In : *Annual Review of Information Science and Technology*, vol. 33, pp. 223-256, 1998.

Peters, Carol. *Multilingual Information Access. Tutorial.* DELOS Summer School on DL Technologies (ISDL2001). Disponible en ligne. Page consultée le 1er mai 2003. Adresse URL : http://www.iei.pi.cnr.it/DELOS/delos2/SummerSchool/handouts1/ISDL1.ppt.

Silvester, J.P. et al. Machine-aided indexing at NASA. *Information Processing & Management*, 30, 1994, pp. 631-645.

Soergel, Dagobert . Multilingual thesauri in cross-language text and speech retrieval. In : *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*, March 1997. Disponible en ligne. Page consultée le 5 mai 2003. Adresse URL : http://www.ee.umd.edu/medlab/filter/sss/papers/soergel.ps.

Sperberg-McQueen, C.M. ; Burnard, Lou. The Design of the TEI Encoding Scheme. *Computers and the Humanities* vol. 29, no 1, 1995, pp. 17-39. Reprinted in : Ide, Nancy ; Veronis, Jean. *The Text Encoding Initiative: Background and Contexts*, Boston ; Dordrecht : Kluwer Academic Publishers, 1995.

Steinbach, Michael ; Karypis, George ; Kumar, Vipin . A Comparison of Document Clustering Techniques. In : *KDD Workshop on Text Mining*, 2000. Disponible en ligne. Page consultée le 5 mai 2003. Adresse URL : http://www-users.cs.umn.edu/~karypis/publications/Papers/PDF/doccluster.pdf.

Svenonius, Elaine. Use of classification in online retrieval. *Library Resources and Technical Services*, vol. 27, no 1, janv./mars 1983, pp. 76-80.

Thelwall, Mike; Tang, Rong ; Price, Liz. Linguistic patterns of academic Web use in Western Europe. *Scientometrics*, vol. 56, no 3, 2003, pp. 417-432.