

CABRE, M. T.; DE YZAGUIRRE, LL.; GARCIA, Y. (2001) «La terminologie et l'accès multilingue à l'information». En: Actes del congrés Le plurilinguisme dans la société de l'information (Paris, 9-10 de març de 2001).

Le plurilinguisme dans la société de l'information

Commission Française pour l'UNESCO

Paris, 9-10 de mars de 2001

LA TERMINOLOGIE ET L'ACCES MULTILINGUE A L'INFORMATION

M. Teresa Cabré

Lluís De Yzaguirre

Yannick Garcia

Institut Universitari de Lingüística Aplicada

Universitat Pompeu Fabra (Barcelone)

Avant-propos

On part de l'idée que toutes les langues sont également aptes pour exprimer toute connaissance mais elles ne se trouvent pas au même niveau de développement. Pour quelques-unes, il manque ou il a manqué un corpus littéraire consolidé en leur permettant d'établir un registre formel de la langue; d'autres ont été créées sur la base de registres écrits, la plupart des cas administratifs, et elles ont des lacunes dans le domaine général ou familial; il y en a aussi qui ont profité d'une certaine «normalité» linguistique jusqu'à la dernière dizaine d'années, où elles ont été interdites aux nouvelles technologies de la réalité mondiale et elles sont devenues des langues diglossiques *strictu sensu*. Si l'on considère l'usage des langues, on peut les distinguer en termes quantitatifs (nombre d'usagers) ou en termes de diversité d'usages (usage complètement normalisé ou diglossie). Les raisons ne sont donc pas internes, mais fondées sur des circonstances externes aux langues, sociales et historiques.

De la totalité des usages, il y en a qui sont considérés prestigieux par leur idiosyncrasie et leur représentation sociale. Il y en a d'autres qui ne donnent pas de prestige aux langues. Dans les premiers, on peut situer quelques registres particuliers (cientifico-techniques, professionnels et les spécialisés en général) et les usages liés aux nouvelles technologies de l'information et de la communication (TIC). Le scénario communicatif donné par Internet est un des exemples les plus représentatifs de cette option de futur. On dit qu'une langue qui ne s'intègre pas dans les TIC (c'est-à-dire, une langue qui utilise les TIC et qui est utilisée par elles) est destinée à disparaître ou, au moins, à perdre la considération de langue capable de transférer de la connaissance plus ou moins spécialisée et de rester, en même temps, réduite à la langue parlée et familière.

Face à ce défi, les efforts de quelques communautés et de quelques pays pour entrer dans les nouveaux espaces communicatifs thématiques et technologiques ont été très grands, bien qu'il soit un fait incontestable que la présence des langues dans les domaines des nouvelles technologies n'est pas égalitaire.

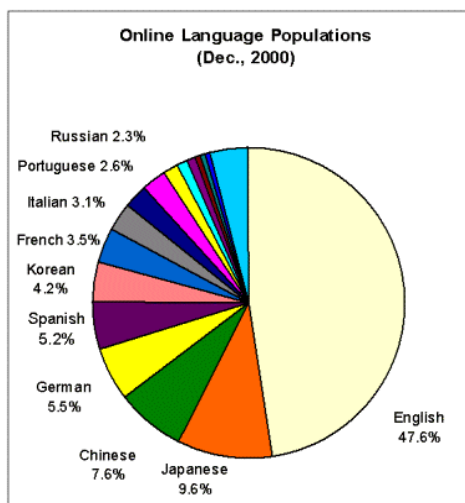
Etat de la question

Internet est une ressource universelle, l'accès à la masse d'information virtuelle disponible est garanti là où se réunissent les conditions économiques nécessaires pour en devenir bénéficiaire. Le nombre d'utilisateurs possible est, donc, très divers quant aux niveaux de formation et linguistiques. Pour atteindre un traitement égalitaire dans l'accès à l'information spécifique, une conduite bipartite est nécessaire: *a)* d'un côté, les outils de gestion d'information électronique —ceux qui se chargent des sites et des moteurs de recherche— doivent prévoir cette diversité et doivent fournir les mécanismes et les stratégies pertinentes afin que l'outil ait une couverture réelle la plus étendue possible; et *b)* les propriétaires des pages sur le web doivent être conscients du travail qu'ils peuvent éviter aux outils de gestion avec une phase efficace de pré-édition des ressources qui comporte une analyse très détaillée concernant une indexation future du document.

Les langues

Une des questions de base à se poser, en ce qui concerne les langues, est la possibilité qu'un parlant trouve de l'information dans la propre langue ou dans une autre qu'il possède. Puis, on doit ajouter que dans ces communautés avec une taux d'analphabétisme élevée il faudrait éviter l'erreur de négliger la propre langue au moment de dessiner des stratégies d'accès à l'information.

En premier lieu, pour se faire une idée préliminaire de la question linguistique sur Internet, on doit jeter un coup d'œil sur les données du nombre de cybernautes.



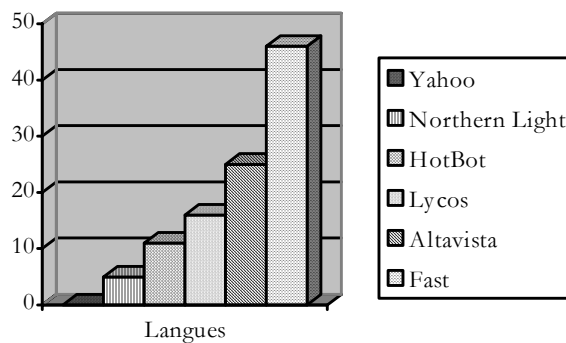
(Global Reach)

Langue	Population
Anglais	172,3 millions
Japonais	27,3 millions
Allemand	19,9 millions
Espagnol	19,5 millions
Chinois (mandarin)	18 millions
Français	13,2 millions
Coréen	11,7 millions
Italien	10 millions
Portugais	7,7 millions
Russe	6,7 millions

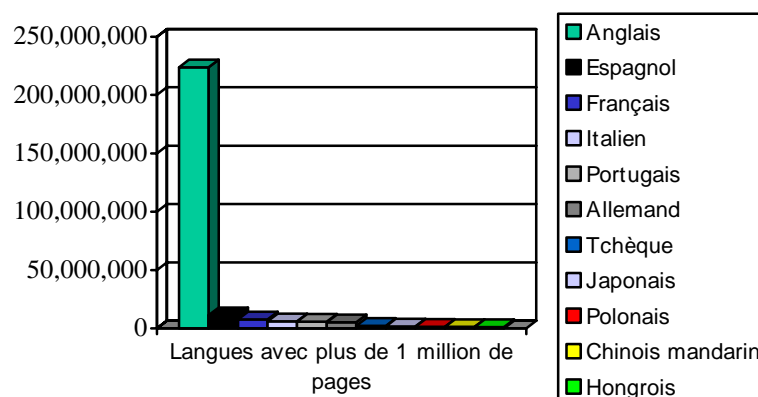
(Blue Earth)

Ces données ne font pas référence à la population totale qui parle une langue en particulier, elles indiquent les millions de parlants natifs qui ont accès à Internet ou qui sont connectés régulièrement, et on se rend compte que presque la moitié de la communauté consommatrice a l'anglais comme langue habituelle.

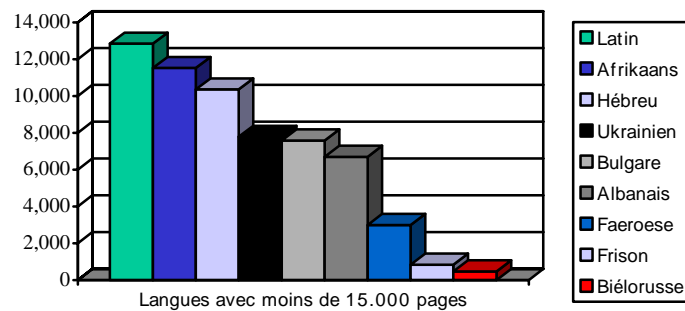
En deuxième lieu, il faut considérer quelles sont les langues envisagées et reconnues par les principaux chercheurs d'information (moteurs de recherche, directoires, etc.); le degré de conscience du plurilinguisme des outils de gestion de moteurs de recherche est bien divers. En considérant le nombre de langues incluses comme un filtrage dans l'option de recherche avancée dans les outils les plus reconnus (soit Yahoo!, Northern Light, HotBot, Lycos, AltaVista et Fast), on se rend compte de l'existence d'une grande variété:



Fast (All The Web) est le chercheur qui présente plus de langues parmi ceux que nous avons consulté: 46. Si nous considérons que *Ethnologue* compile une liste avec à peu près 6.700 langues parlées actuellement dans le monde, les données sont assez significatives. Le chercheur même nous donne la frontière pour délimiter les options linguistiques entre *langues de basse diffusion* (LBD) et *langues de grande diffusion* (LGD). Sur la page de présentation (<http://www.alltheweb.com>), le serveur contient 575 millions de pages. Le système ajoute le nombre de pages trouvées dans les directions des pages. Si nous faisons donc une recherche assez simple —la lettre «a», par exemple— dans toutes les langues, nous aurons des données quantitatives pour diviser les résultats. On a pris ceux avec plus de 1 millions de pages et ceux avec moins de 15.000 pages:



Évidemment, toutes ces langues ne peuvent pas être considérées LGD (d'après les autres moteurs de recherche analysés, seuls l'anglais, le français, l'allemand, l'espagnol, le japonais et le chinois sont des LGD). Cependant, ces données démontrent l'existence de matériaux dans des langues très diverses. Par contre, il y a des langues avec une représentation assez pauvre sur la Toile:



Bref, on peut seulement considérer comme des LGD ceux qui ont beaucoup de parlants avec accès à Internet et un grand nombre de pages accessibles. Pour ces communautés, la recherche monolingue d'information peut devenir insatisfaisante dans la plupart des cas, indépendamment du fait que peut-être dans une autre langue ils auraient pu trouver une information plus adéquate.

Pour la grande majorité de parlants des autres langues, la probabilité de trouver de l'information dans la propre langue est infime.

Comment est-ce qu'on peut agir face à cette situation?

- Accepter le déséquilibre et laisser le marché décider la sélection des langues.
- Donner des subventions officielles à la traduction dans toutes les langues.
- Chercher des solutions intermédiaires avec le support des nouvelles technologies.

Si l'on laisse le choix à la liberté du marché, on présuppose la conscience et la militance des parlants dans un scénario très souvent vu comme un moyen d'atteindre une finalité, et non comme une finalité en elle-même. Les usagers veulent des réponses rapides à Internet, et face à l'absence de ressources dans leur langue, ils n'ont que deux options: *a*) faire l'effort d'accéder à l'information dans une autre langue en utilisant des traductions externes, bien qu'ils soient conscients que l'information vraiment utile peut être écrite dans une autre langue que le moteur de recherche ne catalogue pas; ou *b*) laisser tomber, et, dans ce cas, le but principal d'Internet considéré une ressource universelle n'est pas atteint.

La traduction officielle subventionnée est, pourtant, une solution partielle et à courte échéance. Partielle, parce que seules les communautés économiquement fortes et avec un système d'aménagement linguistique consolidé peuvent en profiter. À courte échéance, parce que la conception monolingue ne se perd pas et la subvention reste soumise à un cercle vicieux: plus d'information, plus de traduction, et en conséquence, plus de traduction, plus de subvention. Cette dépendance augmente de plus en plus, étant donné que la croissance de pages sur Internet se multiplie mois après mois et, en plus, cet augmentation oublie de trouver des méthodes pour optimiser les efforts.

Comme démonstration de ce qu'on vient de dire, regardons les premières chiffres consignés de la croissance de pages sur la Toile (1993-1997). Évidemment, l'augmentation est exponentielle et, quand on a dépassé un certain seuil, les chiffres deviennent simples estimations. Actuellement, ces calculs sont encore plus difficiles.

Mois	Nombre de pages
6/93	130
12/93	623
6/94	2.738
12/94	10.022
6/95	23.500
1/96	100.000
6/96	230.000 (est)
1/97	650.000 (est)

Une solution efficace devrait prévoir le besoin de traduction pour n'importe quel usager dans toutes ses recherches. Si l'on ne trouve pas des voies améliorées dans le domaine de la traduction automatique, il faudrait au moins lui offrir des «substituts» de l'information qui raccourciraient le chemin jusqu'aux données. Si on nous permet une métaphore biblique, une traduction subventionnée serait comparable à un poisson donné à l'utilisateur pour manger un jour, et une réponse efficace d'information à n'importe quel type de consultation dans la propre langue de l'utilisateur serait comme lui donner une canne à pêche pour pêcher et manger toute la vie.

Proposition

Notre option se situe dans la troisième voie, plus possibiliste, qui ouvre le chemin à plus de langues (sans arriver à la totalité) pour avancer vers l'égalité linguistique. Actuellement, la plupart des moteurs de recherche sont satisfaisants seulement si nous les interrogeons dans la langue du document que nous voulons obtenir:

The automated ones such as AltaVista will work perfectly well in any language. To find American hotels, type *hotel*. To find Italian hotels, type *albergo*, which is Italian for hotel.

AltaVista will search for titles, text and keywords one character at a time, without «realizing» what language. (Hopkins 1996)

Ce processus est appliqué aussi bien aux chercheurs qui utilisent descripteurs ou mots clé —la plupart d'entre eux—, comme aux chercheurs qui transforment le langage naturel au langage contrôlé, avec l'aide de thesaurus, par exemple AskJeeves.

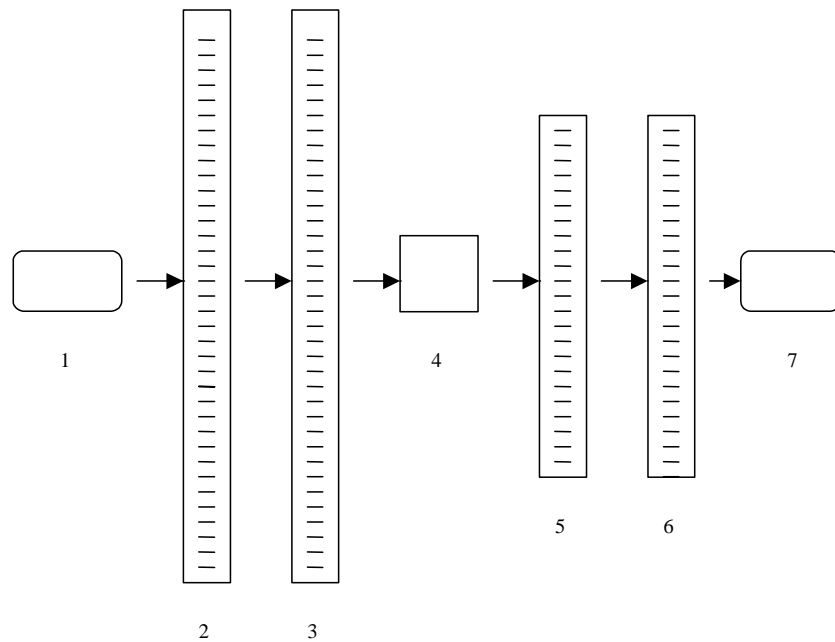
Pour réduire cette diglossie informative, l'utilisateur doit pouvoir interroger le chercheur dans sa propre langue et recevoir des données écrites dans n'importe quelle langue. Ce type d'opérations est possible seulement si l'on conçoit les ressources du point de vue plurilingue et s'il est possible d'accéder sur la Toile unilatéralement dans une langue, sans l'existence d'aucun domaine déficitaire. Pour atteindre ce but, il faut avoir des ontologies, qui permettent de standardiser les matériaux disponibles. Voyons un exemple.

Un usager cherche des bourses pour faire un stage en Chine. La question naturelle est: «Où est-ce-que je peux trouver de l'information sur des bourses pour étudier en Chine?». Un chercheur courant n'est pas capable de traiter la question dans n'importe quelle langue parce qu'il fragmente la séquence et il la traite comme une chaîne de conditions de recherche, en ce qui concerne les éléments pleins —les descripteurs potentiels: *bourses, étudier, Chine*—, comme les vides —*où, est-ce-que, je, peux, trouver, de, l', information, sur, des, pour, en*—. Un chercheur qui prend le langage naturel, comme AskJeeves, ne peut que le traiter et en sortir les descripteurs si la question a été posée dans une des langues qu'il connaît.

Regardons maintenant en détail chaque étape d'une recherche: l'interrogation assistée, le raffinement et la visualisation des résultats.

Interrogation assistée

L'étape de formulation de la recherche doit s'encadrer dans un environnement d'interface assistée et doit permettre l'expression de la langue de l'utilisateur. Cependant, il faut être réalistes et ne pas rêver à un système qui reconnaisse n'importe quelle langue, il ne faut pas imposer la responsabilité, sans mesure et de façon utopique, d'obliger le créateur d'une page personnelle à indexer sa ressource selon cette diversité. La solution peut s'orienter à l'usage d'un système linguistique pont dans le but d'interconnecter la question dans la langue X avec les résultats, écrits dans une langue X ou Y, en qualité de *lingua franca*. Étant donné que la plupart des documents ont été écrits, ou au moins indexés, dans une ou plusieurs LGD, ce sont ces langues qui doivent devenir la base de l'interface de recherche. Le schéma suivant illustre notre proposition:



La phase de formulation suit ces étages initiaux:

1. L'utilisateur fait la consultation en langage naturel vernaculaire.
2. Les mots lexicaux sont extraits (les grammèmes sont éliminés).
3. Ils sont traduits dans une LGD (avec tous les équivalents, dans le cas d'ambiguïté).
4. Un Système de sélection de descripteurs (SSD) choisit les descripteurs d'une ontologie, d'un thésaurus ou d'une base de données terminologiques qui améliorent la couverture de la pétition de l'utilisateur.
- 5/6. Les descripteurs en LGD (5) sont traduits par les équivalents en LBD (6).
7. L'utilisateur choisit dans sa propre langue quels sont les descripteurs qui véhiculent le plus adéquatement sa pétition. Dans le cas où les résultats ne sont pas assez satisfaisants, l'utilisateur peut recommencer le processus à nouveau. Contrairement, l'utilisateur doit définir deux paramètres, quel chercheur et quelles langues, à partir desquels on commence effectivement la recherche.

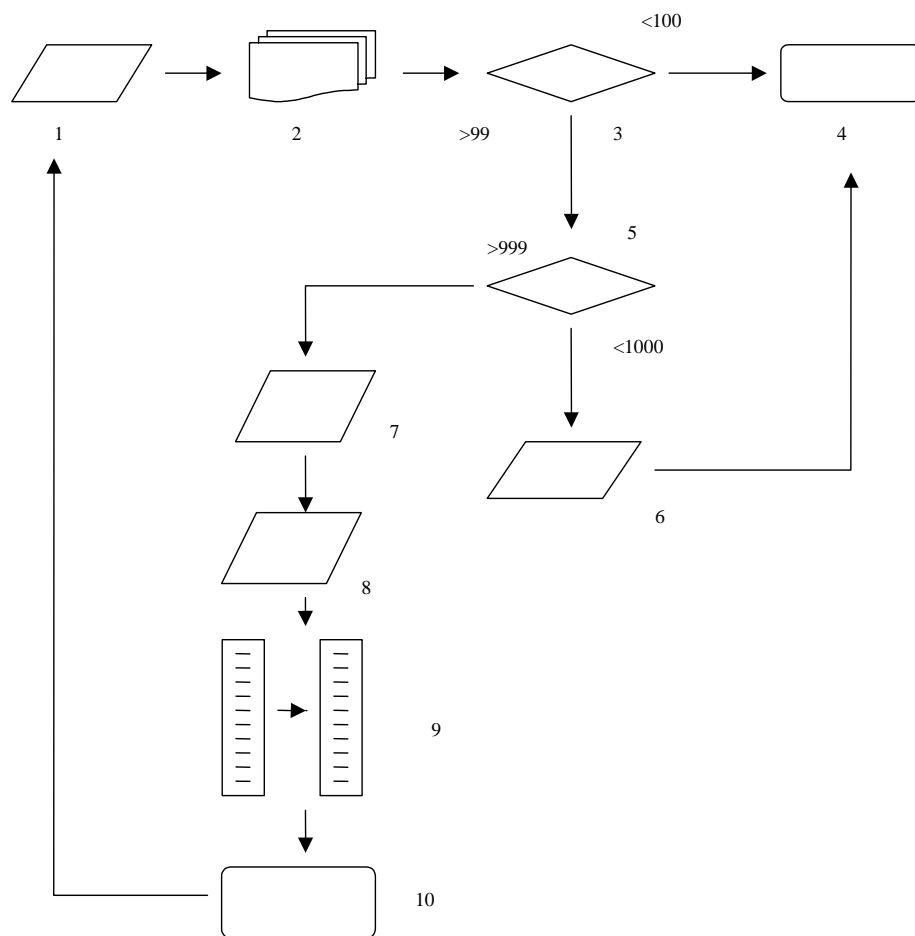
Les outils nécessaires pour le fonctionnement de l'interface sont donc:

- un dictionnaire général avec des correspondances LBD – LGD (par exemple, swahili – anglais, faeroese – allemand, dalmate – italien, etc.);
- un thésaurus, une ontologie ou une terminologie avec des descripteurs et des équivalents croisés dans la même combinaison.
- un SSD pour chaque LGD considérée; il s'agit d'un outil très similaire à un extracteur de terminologie qui sache associer le lexique général de la langue avec les descripteurs pertinents (avec WordNet ou en sortant de l'information des définitions lexicographiques, par exemple).

L'avantage de commencer par les LGD réduit en grand mesure les combinaisons et optimise les efforts, parce que la plupart de ces outils ont déjà été créés et implantés.

Raffinement des résultats

Quand le moteur de recherche a sélectionné les descripteurs, il faut faire une recherche standard dans toutes les langues possibles, fait qui doit générer en principe un excès de résultats, la plupart desquels l'utilisateur ne peut pas discriminer. Notre stratégie de raffinement est la suivante:



- 1/2. On cherche dans X moteurs dans Y langues et on obtient Z pages (2).
- 3/4. Si $Z < 100$, on peut déjà présenter les résultats (4).
5. Si $99 < Z < 1000$, on passe à 6; dans le cas contraire, à 7.
6. On applique les méthodes actuelles de préférence fondées sur des paramètres numériques et on commence l'étage de présentation des résultats (4).
7. On utilise un extracteur de terminologie pour détecter les termes les plus saillants de chaque page et des techniques de catalogation automatique pour choisir les termes discriminatoires parmi la totalité des termes relevants dans toutes les pages retenues (ceux qui permettent de distinguer entre les pages intéressantes et les pages repoussées).
8. On applique le SSD pour trouver les descripteurs des termes discriminatoires.
9. Le thesaurus multilingue (ou onthologie ou système terminologique) permet la formulation des descripteurs discriminatoires dans la langue de l'utilisateur.
10. Le système dialogue avec l'utilisateur. Le résultat de ce dialogue ne devient pas une nouvelle recherche, sinon un raffinement de la recherche antérieure, pour trouver un sous-groupe de documents obtenus dans la première recherche qui soient plus proches des intérêts de l'utilisateur.

Les ressources nécessaires dans cet étage se sont aussi groupées dans les LGD et elles sont comme celles qui sont en train de se développer pour l'extraction de terminologie, pour la classification automatique de documents et pour le repérage d'information.

Présentation des résultats

La dernière étape du processus —même si l'utilisateur a fait une seule recherche ou si une deuxième interrogation a été nécessaire pour le raffinement— consiste à visualiser le matériel adéquat. Dans ce cas là, l'utilisateur fait face effectivement à la langue réelle des textes. S'il ne la connaît pas, il a plusieurs options: l'extraction de la terminologie la plus remarquable des documents, la construction automatique de résumés des contenus ou d'organisations conceptuelles, ou la traduction peu raffinée —type *BabelFish*— de morceaux de texte. C'est à lui, donc, de décider s'il doit se faire traduire les morceaux des documents qu'il trouve intéressants ou le document complet qui accompli tous les critères initiaux de sa recherche.

Les ressources dont on peut avoir besoin dans ce cas sont indépendantes de celles propres des étages précédents. Évidemment, toutes les activités orientées au repérage et à la construction d'inventaires lexicaux des LBD pour les étages d'interrogation assistée et pour le raffinement des résultats deviendront une aide dans le troisième étage.

À notre avis, une stratégie efficace doit intégrer les avantages des thesaurus multilingues préexistants dans un système d'interrogation assistée qui doit guider l'utilisateur dans sa recherche. En plus, il devrait réduire le bruit du langage naturel et le convertir à un langage contrôlé pour raffiner les résultats.

Quelle est la raison principale pour considérer que le repérage de terminologie et l'élaboration de glossaires multilingues sont des voies adéquates pour favoriser le multilinguisme sur Internet?

- Les caractéristiques des textes spécialisés en comparaison aux textes généraux (la connaissance est plus dense et compactée et les termes sont les unités qui retiennent plus prototypiquement cette connaissance).
- Les textes spécialisés sont moins variés stylistiquement.
- Ils montrent plus de systematicité dénomminative.

Pour résumer, un chercheur efficace pour naturaliser l'accès à Internet devrait intégrer les étapes suivantes:

- l'interrogation originelle dans toutes les langues;
- l'extraction des descripteurs qui caractérisent la formulation;
- la localisation de ces descripteurs dans un thesaurus multilingue;
- la recherche de ressources à partir de ces thesaurus;
- la sélection des résultats selon le filtrage de l'adéquation informative;
- la visualisation des résultats les plus saillants ou «substituts», par exemple, résumés, repérage de terminologie des documents ou traductions.

Avantages de la proposition

Une naturalisation de l'accès aux données dans la propre langue en utilisant des thesaurus multilingues et des extracteurs de terminologie représente une solution:

- a) **faisable**, parce que les applications nécessaires sont déjà disponibles;
- b) **soutenable**, grâce à la charge économique limitée pour son entretien;
- c) **actualisable**, parce que l'adéquation des thesaurus et des onthologies, de même qu'une indexation efficace, ont des effets multiplicateurs dans les chercheurs;
- d) **adéquate**, parce qu'il est possible d'y entrer par chaque langue et arriver aussi à chaque langue;
- e) **intégrable** dans une chaîne technologique de développement de la traduction automatique et de l'élaboration de résumés; et
- f) **rentable**, parce que l'information résultante à partir des requêtes des usagers est beaucoup plus raffinée.

Bibliographie

- AskJeeves (1996-2001). *http://www.askjeeves.com* [Dernière consultation: 19/02/01].
- Blue Earth Language Solutions (2000). *Statistics by Language*. [Inver Grove Heights:] Blue Earth. *http://www.blueearth.net/en/resources/statistics3.html* [Dernière consultation: 19/02/01].
- Global Reach (2000). *Global Internet Statistics (by Language)*. Global Reach. *http://www.greach.com/globstats* [Dernière consultation: 19/02/01].
- Hopkins, R. (1996). *Website Translation: A Primer for Webmasters, Authors and Owners*. Global Reach. *http://greach.com/eng/ed/art/trans.html* [Dernière consultation: 14/2/01].