

## Article

---

« Un système de recherche documentaire multilingue comme outil d'aide à la traduction »

Claire Gerardy et Walter Brüls

*Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 39, n° 1, 1994, p. 159-167.

Pour citer cet article, utiliser l'adresse suivante :

<http://id.erudit.org/iderudit/002881ar>

Note : les règles d'écriture des références bibliographiques peuvent varier selon les différents domaines du savoir.

---

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter à l'URI <http://www.erudit.org/apropos/utilisation.html>

---

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche. Érudit offre des services d'édition numérique de documents scientifiques depuis 1998.

Pour communiquer avec les responsables d'Érudit : [erudit@umontreal.ca](mailto:erudit@umontreal.ca)

# UN SYSTÈME DE RECHERCHE DOCUMENTAIRE MULTILINGUE COMME OUTIL D'AIDE À LA TRADUCTION

CLAIRE GERARDY ET WALTER BRÜLS  
*Université de Liège, Liège, Belgique*

## POURQUOI UN SYSTÈME DE RECHERCHE DOCUMENTAIRE POUR LE TRADUCTEUR ?

Pour effectuer une bonne traduction, la consultation de dictionnaires, glossaires et banques terminologiques n'est pas toujours suffisante. En effet, outre leur manque d'exhaustivité et leur fiabilité qui, même pour les meilleurs, est loin d'être absolue, ces références présentent le désavantage de se limiter au mot ou au terme. Or, le traducteur soucieux de rédiger un texte dont on ne pourra deviner qu'il s'agit d'une traduction souhaite trouver non seulement le terme, mais aussi la phraséologie adéquate. Pour cela, il doit généralement s'inspirer de documents écrits dans la langue cible.

L'utilisation de textes réels est un impératif absolu pour les documents stéréotypés (contrats, polices d'assurance, normes, etc.) auxquels le traducteur professionnel est souvent confronté. Ici, l'entité à traduire globalement va du multiterme à la phrase, voire au paragraphe entier. La formule de politesse que l'on trouve en fin de lettre est sans doute l'illustration la plus simple de ce phénomène. En français, les possibilités sont multiples et nuancées : *Veillez agréer, Je vous prie d'agréer... l'expression de mes sentiments les meilleurs, salutations distinguées, etc.* L'allemand et l'anglais sont plus concis puisqu'on utilisera respectivement *Mit freundlichen Grüßen, Yours sincerely* ou *Yours faithfully* en fin de lettre.

Dans les cas mentionnés ci-dessus, la traduction tient donc moins de l'acte créatif que de la reproduction à bon escient de formules consacrées. Cette approche de la traduction basée sur le texte réel peut s'apparenter aux théories de la traduction automatique basée sur l'exemple exposées notamment par Gale, Church et Yarowski (1992 : 101), Somers et Hutchins (1992) et Doi et Muraki (1992 : 525). La démarche caractérisant ces théories consiste à avoir recours à un important corpus d'exemples réels auquel on peut accéder en cas de difficulté, plutôt que de définir des règles collocationnelles, sémantiques, voire encyclopédiques à l'infini.

Un problème de traduction peut donc devenir ainsi un problème de recherche documentaire et c'est là que les choses se compliquent, car le traducteur perd souvent un temps précieux à dénicher l'extrait de texte dont il a besoin. Dans une telle situation, il peut opter pour la solution classique, qui consiste à dépouiller des piles de documents archivés sur support papier, ou opter pour une solution plus efficace qui tire profit des possibilités offertes par l'ordinateur.

## INTÉGRATION D'UN SYSTÈME DE RECHERCHE DOCUMENTAIRE DANS LE POSTE DE TRAVAIL DU TRADUCTEUR

À la fin des années 80, le PC est devenu partie intégrante de l'environnement du traducteur et le traitement de texte a remplacé avantageusement la machine à écrire clas-

sique. Mais faciliter l'encodage des données est loin d'être la seule possibilité offerte par l'ordinateur pour assister l'homme dans le processus de traduction.

Notre fin de siècle se caractérise par un flux croissant d'informations que le traducteur doit apprendre à maîtriser s'il veut être efficace et compétitif. Or, l'ordinateur s'est avéré un outil idéal, non seulement pour archiver, mais aussi pour gérer et récupérer ce flux d'informations.

Au-delà de l'accroissement des ressources lexicales fournies par les dictionnaires électroniques et les banques de données terminologiques accessibles en ligne, sur CD-ROM ou sur support magnétique classique, on a assisté, ces dernières années, au développement de plusieurs logiciels de gestion terminologique destinés à faciliter la tâche du traducteur et à remplacer les fiches en carton dont la mise à jour exigeait un travail de bénédictin.

Mais un des nombreux avantages de l'informatique est de permettre l'accès à des ressources autres que purement terminologiques. En effet, les bases de données textuelles comme les corpus multilingues, par exemple, offrent au traducteur une quantité non négligeable d'informations souvent inexploitées faute d'outils d'archivage et de gestion documentaire.

#### QUEL SYSTÈME DE RECHERCHE DOCUMENTAIRE ?

Les systèmes de gestion de bases de données relationnelles (SGBDR) classiques comme Oracle®, dBase®, Paradox®, Foxpro®, pour ne citer que quelques-uns des plus connus, ne se prêtent pas vraiment à la gestion de données textuelles. Ces systèmes ont été conçus pour accéder rapidement à des données structurées souvent modifiées et dépassant rarement un nombre restreint de caractères. Quant aux données textuelles, elles ne sont pas ou peu structurées, et ne nécessitent généralement pas de mise à jour (il s'agit en quelque sorte d'archives). Il est donc évident que l'approche adoptée pour ces deux types de données est fondamentalement différente.

Il existe différents types de systèmes de gestion documentaire plus ou moins évolués et plus ou moins efficaces. Une classification grossière permet de distinguer six types différents que nous allons brièvement passer en revue. Les critères d'évaluation pour un système de recherche documentaire portent essentiellement sur deux aspects : la recherche (rapidité, convivialité, efficacité) et l'automatisation de la création de la base textuelle.

Le problème principal en recherche documentaire est de minimiser simultanément le bruit et le silence. Cette affirmation peut paraître déroutante à première vue, mais elle s'explique facilement par l'acceptation donnée aux notions de bruit et de silence en sciences documentaires. En effet, on entend par bruit tous les documents non pertinents extraits par le système, et par silence tous les documents pertinents que le système n'a pas pu trouver. L'impératif absolu est donc d'exclure le silence, car le but étant la recherche de l'information, le silence équivaut à des informations perdues. Seulement, s'il y a trop de bruit, «on n'entend plus rien» : l'utilisateur est confronté à une telle quantité de documents inintéressants que le dépouillement des documents proposés devient une tâche trop lourde.

Passons rapidement en revue les principales méthodes utilisées en gestion documentaire :

#### Systèmes à mots clés

Le texte à indexer est d'abord entièrement lu par un éditeur humain qui choisit les mots lui paraissant pertinents et les signale comme mots clés.

Ce procédé offre l'avantage de s'intégrer facilement dans un SGBDR classique.

Mais les désavantages sont nombreux et de taille : la liste des mots clés attribués par l'éditeur est par définition incomplète et il est impossible d'effectuer une recherche

sur les mots réellement présents dans le document. Dans le cas du traducteur, cette restriction rend impossible toute recherche sur les collocations, par exemple.

Un autre inconvénient majeur de ce système est la subjectivité de l'éditeur humain. Deux éditeurs attribueront des mots clés différents à un même texte. De plus, l'attribution des mots clés peut varier en fonction des besoins de l'utilisateur. Pour un flux important de données, le coût de l'indexation peut devenir prohibitif, étant donné que l'éditeur est obligé de lire intégralement tous les documents à indexer.

### **Systèmes à opérateurs booléens**

Ce genre de système indexe automatiquement les documents qui lui sont soumis. Pour retrouver un texte donné, l'utilisateur doit entrer les différents mots qui l'intéressent, combinés par des opérateurs booléens. Le système lui propose alors en réponse tous les documents correspondant à la structure donnée.

Ce type de système présente l'avantage d'une indexation entièrement automatique, donc nettement moins onéreuse et plus rapide que l'indexation manuelle. En outre, cette méthode permet d'effectuer les recherches sur les mots réellement présents dans la base.

Le langage d'interrogation est néanmoins très peu convivial, la liste de documents proposés n'est pas triée par ordre de pertinence et le traducteur peut perdre un temps précieux s'il n'a pas bien formulé sa question.

### **Systèmes statistiques**

Ces systèmes se basent sur l'hypothèse selon laquelle des mots proches l'un de l'autre dans un texte traitent probablement du même sujet. Le système calcule les relations statistiques entre les mots du texte. À l'interrogation, l'utilisateur entre une liste de mots et le système fournit en réponse les documents qu'il juge pertinents.

Dans ce genre de système, l'indexation est entièrement automatisée et l'interrogation est généralement conviviale.

Malheureusement, l'absence totale de composante linguistique signifie que la qualité des résultats obtenus n'est pas toujours satisfaisante.

### **Systèmes hypertexte**

Ce genre de système permet de truffier les documents de liens entre les différentes parties afin de permettre à l'utilisateur de se déplacer dans le document en sautant de nœud en nœud.

L'indexeur peut établir des liens intéressants et l'interface est généralement très conviviale.

Mais, comme dans le cas des systèmes à mots clés, l'éditeur est obligé de lire l'intégralité du texte, tâche extrêmement lourde, et, de plus, le nombre de liens est généralement limité.

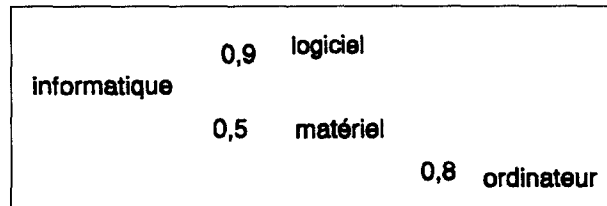
### **Systèmes basés sur les graphes de concepts**

L'éditeur crée des graphes de concepts pondérés, c'est-à-dire des réseaux de mots avec un poids informationnel attribué arbitrairement par le créateur du réseau à chaque arc. À l'interrogation, l'utilisateur sélectionne un concept et la recherche est lancée sur tous les mots inférés par le(s) mot(s) demandé(s) dans le(s) graphe(s).

Ce genre de système présente de nombreux avantages : la classification des documents trouvés, la convivialité et l'indexation entièrement automatique.

La création des graphes de concepts reste néanmoins une tâche extrêmement lourde et surtout très subjective, notamment en ce qui concerne l'attribution d'un poids informationnel à un arc donné.

Voici comment se présente un graphe de concept à un niveau extrêmement simple :



Il est évident que la complexité des graphes varie en fonction du degré de hiérarchisation du domaine étudié. Les graphes doivent donc être créés par des éditeurs hautement spécialisés.

#### LE SYSTÈME SPIRIT®

L'approche du système SPIRIT<sup>1</sup> repose sur une analyse linguistique du document et une pondération statistique en fonction de la distribution des mots dans le texte. Les deux principales composantes du système sont la création de la base et l'interrogation.

#### Création de la base

Le système accepte en entrée des documents au format ASCII (standard ou étendu) ou émanant d'un traitement de texte.

La première étape consiste à découper la chaîne de caractères (c'est-à-dire le document) en mots et en phrases. Ce premier découpage est pris en charge par un automate à nombre fini d'états. Cette étape est plus complexe qu'il n'y paraît à première vue, car de nombreux caractères considérés communément comme des séparateurs de phrases peuvent avoir un statut ambigu (le point dans un sigle par exemple).

Lorsque les mots du document ont été identifiés, le système procède à leur analyse morphologique et réduit tous les unitermes au lemme. Dans le cas d'ambiguïtés, dues à un cas d'homographie, par exemple, les différentes possibilités de normalisation sont gardées afin d'être résolues lors d'une étape ultérieure. La normalisation repose sur un dictionnaire *full-form* plutôt que sur des règles flexionnelles, notamment pour des raisons d'efficacité du système.

Le logiciel procède ensuite à la reconnaissance de locutions et d'expressions idiomatiques (recensées dans un dictionnaire séparé). Ceci permet d'éviter d'effectuer d'inutiles et dangereuses recherches sur les composants d'un multiterme dont le sens n'est pas déductible de la somme des sens de ses composants (chemin de fer, par exemple).

Pour résoudre les cas d'homographie et établir des liens de dépendance entre mots adjacents ou uniquement séparés par un mot vide qui n'est pas considéré comme un séparateur fort, SPIRIT dispose d'un module d'analyse syntaxique. Cet analyseur repose sur des matrices binaires et ternaires qui précisent les séquences de catégories grammaticales que l'on peut trouver dans une langue donnée. Dans cette approche dite markovienne, les matrices sont constituées par apprentissage : un texte est soumis au système afin d'être analysé. Les résultats de l'analyse morphologique sont alors présentés à un «professeur» humain pour validation et désambiguïsation, et l'ordinateur mémorise les résultats validés dans les matrices.

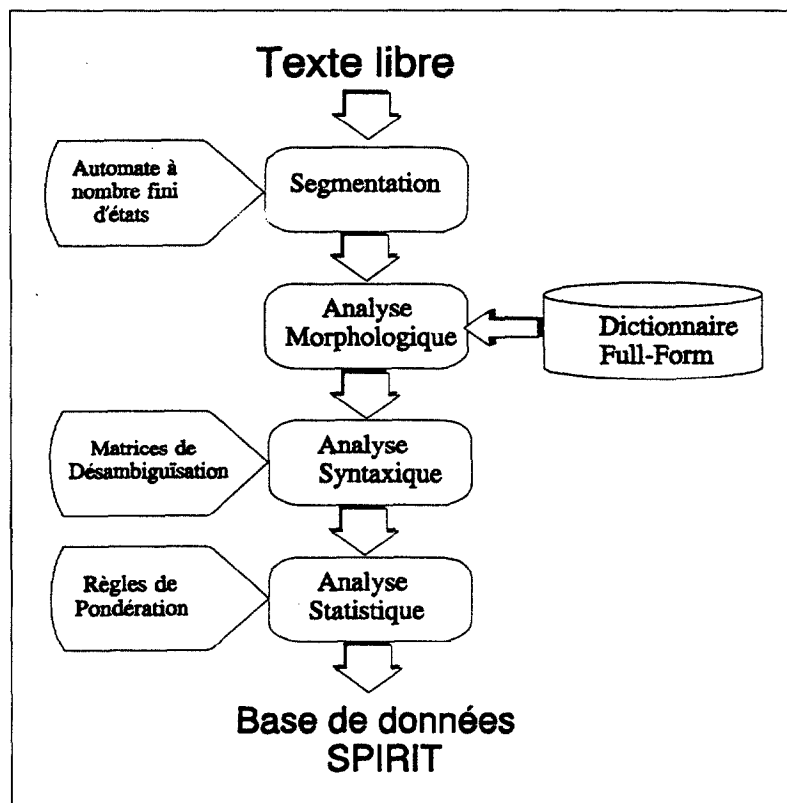
Après avoir complété toutes les étapes de l'analyse morphologique et syntaxique, le système procède au calcul du poids informationnel de chaque mot plein en fonction de deux critères : la fréquence du mot dans la base et sa distribution dans la base. La pondé-

ration repose donc sur le principe suivant : plus un mot est fréquent dans la base, moins il est intéressant, mais plus un mot est fréquent dans un même document, plus il est intéressant. Imaginons, par exemple, que l'on s'intéresse à la TVA sur le caviar. Dans une base de données traitant de l'imposition, le mot TVA apparaîtra très souvent, alors que le mot caviar sera relativement rare. Le poids informationnel du mot TVA sera donc faible. En outre, il est probable que dans les documents traitant réellement de caviar, le mot caviar soit répété plusieurs fois. Le système attribuera donc à ce mot un poids informationnel élevé.

La dernière étape de la constitution de la base est la création du fichier inverse, c'est-à-dire de la liste de tous les mots (tous les lemmes) accompagnés des renseignements qui leur sont associés : poids informationnel, catégorie grammaticale, etc.

Avant de passer à la description du processus d'interrogation de la base de données, il faut noter la possibilité qu'offre SPIRIT à l'utilisateur de définir lui-même les mots qu'ils souhaite considérer comme vides afin que ceux-ci soient ignorés par le système au moment de l'indexation et de l'interrogation. La classification en tant que mot vide peut se faire de deux façons : par catégorie grammaticale et par mots isolés. Certaines parties du discours comme les prépositions, articles, etc., sont généralement considérées comme mots vides en recherche documentaire. Cependant, l'utilisateur peut souhaiter considérer comme vides certains mots appartenant à d'autres classes grammaticales en fonction de la base à indexer.

#### Processus d'indexation d'un document par le système SPIRIT



### Interrogation

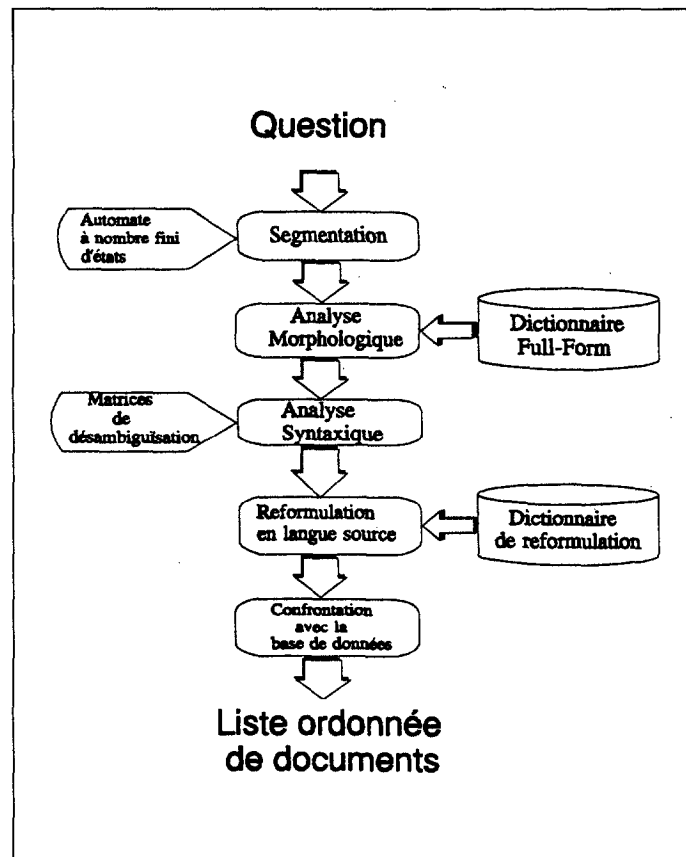
L'interrogation d'une base de données par le système SPIRIT offre le grand avantage de combiner convivialité et puissance.

En effet, l'utilisateur n'est pas obligé d'avoir recours à des opérateurs booléens ou d'apprendre un langage d'interrogation complexe, puisqu'il peut poser sa question en langage naturel, par l'intermédiaire du clavier.

Une fois transmise au système, la requête est soumise aux mêmes algorithmes de lemmatisation et d'analyse syntaxique que ceux auxquels a été soumis le texte de la base lors de l'indexation. La liste de mots ainsi obtenue est alors confrontée au fichier inverse, et le système fournit une liste des documents présentant une intersection avec la requête. Ces documents sont présentés à l'utilisateur, classés par ordre décroissant de pertinence, ce qui permet un gain de temps précieux.

Il est également possible de procéder à une reformulation des mots de la question, c'est-à-dire à une extension de la question aux synonymes, hyperonymes et mots dérivés de la même famille (Londres/Londonien, foie/hépatique, etc.). Cette reformulation permet de découvrir des documents qui traitent du même sujet que la question mais en termes différents.

### Processus d'interrogation d'une base par le système SPIRIT



**Interrogation multilingue : EMIR**

Jusqu'à présent, nous avons exclusivement parlé de SPIRIT, la version monolingue du système de recherche documentaire qui est commercialisée actuellement en français, anglais, allemand et arabe. Cependant, le but du projet EMIR<sup>2</sup> (European Multilingual Information Retrieval), auquel participe notre équipe de recherche, est de rendre ce système multilingue, c'est-à-dire de permettre l'interrogation dans une langue différente de celle de la base. Au cours du projet, un système français-anglais / anglais-français a déjà été élaboré et un système allemand-français / français-allemand est prévu pour la fin de l'année 1993.

L'intérêt d'un système multilingue pour le traducteur semble évident, d'autant qu'EMIR ne permet pas seulement d'entrer la requête au clavier, mais aussi de considérer un texte ASCII comme requête. Le traducteur peut donc soumettre le texte à traduire (ou du moins une partie de ce texte) au système et récupérer en sortie les documents en langue cible qui sont les plus proches du texte en langue source. On imagine aisément l'intérêt que peut avoir une telle démarche dans la traduction de textes stéréotypés comme des contrats, par exemple.

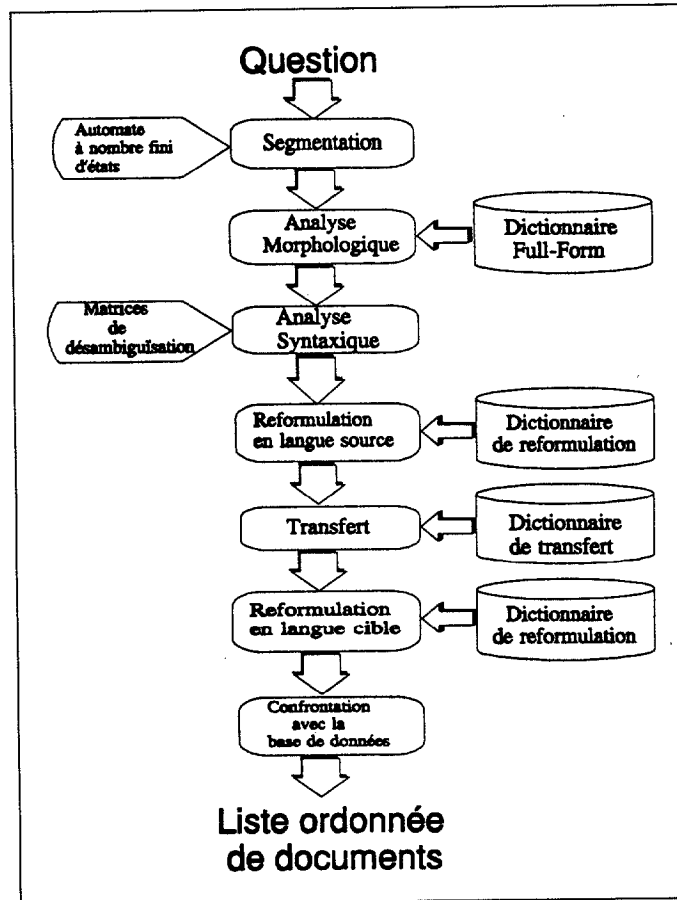
Le système multilingue procède de la même façon que le système monolingue : la question est d'abord découpée en mots, ensuite, on procède à la réduction morphologique, et enfin les ambiguïtés sont levées avec l'aide de l'analyseur syntaxique. Les mots de la question sont alors traduits avant d'être confrontés au fichier inverse de la base consultée.

C'est évidemment dans la bonne interprétation et donc la traduction correcte des mots de la question que réside toute la difficulté du système multilingue, car la correspondance mot langue source / mot langue cible est rarement univoque. Heureusement, les mots les plus polysémiques (ayant donc, par définition, de nombreuses traductions) sont également les plus généraux et les plus fréquents. Ils ont donc un poids informationnel faible. Ceci explique pourquoi les performances du système ne sont que très peu affectées par la présence éventuelle d'une traduction erronée.

Si la possibilité de gérer des multitermes est intéressante pour le système monolingue, elle s'avère indispensable pour le système bilingue, car bon nombre d'expressions et de locutions ne peuvent se traduire mot à mot. Un des problèmes les plus difficiles à résoudre est celui des verbes à particules anglais, qu'il faut être capable d'identifier, puisque le sens du verbe à particule est parfois très éloigné du sens du verbe sans particule.

À l'heure actuelle, les performances du système multilingue n'ont pas encore atteint le niveau de celles du système monolingue. Cependant, le développement d'un analyseur syntaxique plus poussé et l'intégration de dictionnaires de transfert spécifiques au domaine de la base devraient permettre de réduire sensiblement cette différence de performance entre les versions monolingues et multilingues.

### Illustration de l'interrogation d'une base par le système EMIR multilingue



#### CONCLUSION

Comme on peut le voir dans le cas de SPIRIT et d'EMIR, les systèmes de recherche documentaire s'éloignent de plus en plus des SGBDR classiques en raison des besoins différents auxquels ils répondent.

Les possibilités offertes par l'analyse du langage naturel, même si la linguistique computationnelle est loin d'avoir résolu tous les problèmes posés par l'analyse sémantique, ont permis d'améliorer sensiblement les performances des systèmes documentaires qui furent longtemps basés uniquement sur des règles statistiques.

Un des apports essentiels de SPIRIT et d'EMIR réside certainement dans l'utilisation facultative de la reformulation. Dans les versions monolingues du système, la reformulation permet de récupérer une quantité importante d'informations qui ne seraient pas détectables par les moyens traditionnels. Sans reformulation, il serait impossible de retrouver un document traitant des déficiences cardiaques si l'on a utilisé la formulation «maladies du cœur» dans la question, par exemple.

Dans la version multilingue, la traduction des mots significatifs de la question peut être considérée comme une étape supplémentaire du processus de reformulation. La cohérence et la qualité des dictionnaires utilisés par le système est donc cruciale.

Même si, à l'heure actuelle, les systèmes de traduction automatiques sont loin de pouvoir se substituer au traducteur, les recherches effectuées dans le domaine du traitement du langage naturel ont permis de développer de nombreux outils d'aide à la traduction qui s'intègrent parfaitement au poste de travail du traducteur, lui permettant ainsi d'améliorer considérablement ses performances.

#### Notes

1. SPIRIT of SYSTEX® est un produit développé et commercialisé par la firme française SYSTEX, F-91195 Saint-Aubin.
2. EMIR est un projet qui s'inscrit dans le cadre du programme ESPRIT II. Les partenaires participant au projet sont : INSTN/CEA Saclay (coordinateur), SYSTEX Saint-Aubin, TRANSMODUL GmbH, Saarbrücken et le Département de langue anglaise de l'Université de Liège.

#### RÉFÉRENCES

- ARZ, Johannes, FLASSIG, Ralph et Erwin STEGENTRITT (1992) : «Natural Language Products in Restricted and Unrestricted Domains», *Transmodul Papers*, Nr1, Saarbrücken.
- DOI, Shinichi et Kazunori MURAKI (1992) : «Translation Ambiguity Resolution Based on Text Corpora of Source and Target Languages», *Proceedings Coling 92*, vol. II, pp. 525-531.
- FLUHR, Christian (1991) : «Projet EMIR. European Multilingual Information Retrieval», *Proceedings Linguistic Engineering 91*, Versailles, vol. I.
- FLUHR, Christian (1990) : «Multilingual Access to Full-text Databases», *International A.I. Symposium 90 Nagoya*.
- GALE, W., CHURCH, K. et D. YAROWSKI (1992) : «Using Bilingual Materials to Develop Word Sense Disambiguation Methods», *Actes du colloque TMI-92, Montréal*, pp. 101-112.
- GOB, André (1992) : *Le marché belge des systèmes documentaires. Un état de l'art*, Centre informatique de philosophie et lettres, Liège, Université de Liège.
- GROSS, Maurice (1989) : *Les logiciels de traitement linguistique. Développements et perspectives d'industrialisation*, rapport présenté à la Communauté économique européenne, DG XIII / 13 / 3, BY02, pp. 21-23.
- HUTCHINS, W. John et Harold L. SOMERS (1992) : *An Introduction to Machine Translation*, London, Academic Press.
- VICKERY, Brian (1989) : *Intelligent Interfaces for User-friendly Access to Databases and Electronic Information Services. A State-of-the-art Survey*, rapport final, London, Tome Associates Ltd.