Meta
Journal des traducteurs
Translators' Journal

META

Outils linguistiques et système terminologique multilingue

Yahya Hlal

Volume 42, numéro 2, juin 1997

Lexicologie et terminologie II (1) et Traduction et post-colonialisme en Inde
Translation and Postcolonialism: India (2)

URI : https://id.erudit.org/iderudit/003073ar
DOI : https://doi.org/10.7202/003073ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Hlal, Y. (1997). Outils linguistiques et système terminologique multilingue. *Meta*, *42*(2), 321–327. https://doi.org/10.7202/003073ar

Résumé de l'article

Cet article présente un système de TAL basé sur l'emploi d'une base de données terminologique multilingue (arabe, français, anglais) et d'outils linguistiques permettant l'aide à la néologie, l'exploitation de textes et de corpus pour l'extraction de termes ainsi que la génération automatique d'éléments morphologiques utiles pour l'exploitation et l'édition de dictionnaires. Le système offre une série de fonctions dont l'aide à la traduction.

OUTILS LINGUISTIQUES ET SYSTÈME TERMINOLOGIQUE MULTILINGUE*

YAHYA HLAL
EMI LIT2A, Agdal, Rabat, Maroc

Résumé

Cet article présente un système de TAL basé sur l'emploi d'une base de données terminologique multilingue (arabe, français, anglais) et d'outils linguistiques permettant l'aide à la néologie, l'exploitation de textes et de corpus pour l'extraction de termes ainsi que la génération automatique d'éléments morphologiques utiles pour l'exploitation et l'édition de dictionnaires. Le système offre une série de fonctions dont l'aide à la traduction.

Abstract

This article presents a TAL system using a multilingual (Arabic, French, English) terminology data base and linguistic tools for assistance in creating terms, identifying terms in texts and computerised generation of morphological elements for use in dictionary writing. The system also has several functions including a translation aid.

1. INTRODUCTION

La terminologie se caractérise par un volume de données très important et par la nature des traitements qui demandent une compétence pointue dans le domaine. Cela ne pourra se faire de façon rationnelle et efficace sans l'emploi d'une technologie appropriée. Celle-ci puise ses sources, en particulier, dans la technologie base de données pour la gestion des données (acquisition, stockage, mise à jour) et leur exploitation (requêtes, état global ou partiel selon des critères). Mais cela demeure insuffisant si l'on ne fait pas appel à la technologie industrie de la langue pour apporter une aide au niveau de la compétence propre au domaine (aide au néologisme, traitement de corpus, linguistique automatique).

Nous présentons dans cet article un système basé sur l'emploi d'une base de données terminologique multilingue (arabe, français, anglais) et d'outils linguistiques permettant l'aide au néologisme, l'exploitation de textes et de corpus pour l'extraction de termes ainsi que la génération automatique d'éléments morphologiques (racine, schème) utile pour l'exploitation et l'édition de dictionnaires. Le système offre une série de fonctions dont l'aide à la traduction.

La base de données est conçue en empruntant le modèle relationnel. Les différents objets sont en particulier les entrées relatives à chaque langue. Chaque entrée est caractérisée par des attributs dont des éléments morphologiques (racine et schème : cas de l'arabe). La correspondance se fait par le biais d'une relation associant les entrées multilingues. Cette relation est caractérisée par les critères de validité (domaine d'utilisation, par exemple). D'autres relations permettent de faire état des synonymes, des commentaires, des expressions associées à une entrée, etc. Le système est ouvert pour inclure à tout moment tout type de relations ou d'attributs nouveaux.

La technologie industrie de la langue se traduit dans notre système par la mise en place de primitives linguistiques dont nous citons en particulier :

- l'analyse morphologique qui renvoie pour un mot arabe (lexical ou textuel) les éléments morphologiques premiers qui entrent dans sa composition (préfixes,

suffixes, racine, schème), ainsi que les informations du type valeurs grammaticales, genre, nombre ;

- la génération morphologique de termes théoriquement acceptables à partir d'éléments premiers (racine, schème) associés à un concept pour lequel on cherche un équivalent dans une langue (cas de l'arabe qui souffre d'un grand déficit terminologique) ;
- traitement de corpus et de textes pour constituer une base de données morphologiques et statistiques associés aux corpus traités.

2. DES OUTILS LINGUISTIQUES

2.1. Généralités

La linguistique automatique a pour objet la mise en évidence d'outils d'analyse et de génération permettant d'une façon générale le dialogue «homme-machine». Il est convenu dans ce cadre de considérer quatre étapes :

- ***morphologie*** : se caractérise par la considération du mot hors contexte pour mettre en évidence les éléments autonomes entrant dans sa composition. Il peut en résulter, hors contexte, des ambiguïtés d'ordre morphologique (plusieurs décompositions possibles) et grammatical (au niveau d'une décomposition, on associe plusieurs valeurs grammaticales) ;
- ***syntaxe*** : cette étape suppose acquis les résultats de la morphologie et met à profit le contexte positionnel pour lever les ambiguïtés (autant que possible). Le but ultime de cette étape est la mise en évidence des structures syntaxiques (arbres syntaxiques) associées aux discours. Cela se fait, en principe, sans considération sémantique ;
- ***sémantique*** : cette étape a pour objet de lever les ambiguïtés syntaxiques (multiplicité d'arbres) en recourant à des considérations sémantiques (relations lexicales sémantiques, par exemple). L'objet, généralement, est la mise en évidence d'une représentation interne de l'information véhiculée par le discours en vue d'une application déterminée ;
- ***pragmatique*** : cette étape a pour objet la levée d'ambiguïtés issues de la sémantique en recourant à des considérations extra-linguistiques (monde environnant).

2.2. Analyse morphologique[1]

L'objet de l'analyse morphologique est la décomposition du mot textuel en éléments morphologiques pertinents qui entrent dans sa composition. Plus précisément, nous avons les décompositions suivantes :

Mott —> EEP + Motlex + EES
Motlex —-> Racine + Schème

Où

Mott	:	mot textuel
Motlex	:	mot lexical
EEP	:	Éléments en état de préfixation
EES	:	Éléments en état de suffixation

On demande également à l'analyseur d'associer les valeurs grammaticales hors contexte au mot textuel traité.

REMARQUE

- La décomposition en racine-schème est propre à l'arabe (langue sémitique).

L'analyse morphologique est basée sur les éléments suivants :

- Une base de données du type linguistique où l'on trouve les préfixes, les suffixes, les racines, les schèmes, etc.

■ Une base de règles morphologiques où chaque règle est associée à une classe de mots représentée par un triplet «préfixe, suffixe, longueur d'un reste».

Le principe d'analyse est le suivant : à partir du mot textuel, on détermine le préfixe et le suffixe le plus long reconnu dans le mot ; ce qui permet de procéder à une décomposition mettant en évidence un reste (ce qui reste du mot en écartant le préfixe et le suffixe). Cela conduit au triplet d'accès à la règle à utiliser pour traiter le mot. La règle est constituée d'actions à appliquer pour parvenir au résultat souhaité.

Ce qui caractérise cette approche est le fait que l'essentiel de l'analyseur (90 %) se présente sous forme de données externes ; ce qui confère à ce produit une grande portabilité.

Le produit est exploité par le biais de primitives que l'on peut solliciter depuis les applicatifs pour demander le type de résultat voulu (racine, schème, mot lexical, valeur grammaticale, cas).

2.3. Génération morphologique[2]

Le principe de la génération est le suivant :

à partir d'une forme première (canonique) et par emploi d'attributs associés au contexte de cette forme, on génère la forme finale :

forme première + attributs —> mot lexical.

◆ Dans la génération lexicale, il s'agit, à partir d'une racine et d'un schème, de produire un mot lexical potentiellement existant :

racine + schème —> mot lexical.

◆ Dans la génération textuelle, il s'agit, à partir d'un mot lexical (obtenu par transfert lexical en traduction) et d'un ensemble d'attributs caractérisant le contexte (valeur grammaticale, détermination, cas, personne, genre, nombre, etc.), de produire le mot textuel :

mot lexical + attributs —> mot textuel.

3. SYSTÈME D'AIDE AU NÉOLOGISME

3.1. Généralités

Le problème, dans le domaine du néologisme, se présente comme suit :

à partir d'un terme nouveau (représentant un concept nouveau) existant dans une langue (français, anglais) trouver un terme équivalent dans la langue que l'on traite (l'arabe en l'occurrence). Ce terme à trouver doit respecter un certain nombre de contraintes dont :

1. le respect du système morphologique ;
2. l'acceptation du terme par les différentes communautés arabes (un terme acceptable au Maroc peut avoir une consonance grossière en Syrie, par exemple).

Cette dernière contrainte ne pourra être satisfaite que dans le cadre de réunions pluripartites pour valider l'unification du terme ; ce qui se fait, en principe, au niveau des différents congrès pour l'unification de la terminologie sous le guide de la Ligue arabe.

La première contrainte est intrinsèque à la langue arabe. Dans cette langue, il est stipulé que tout mot ayant un sens en soi est composé d'une racine (véhiculant un sens premier [générique]) et d'un schème dont le rôle est de jouer le mécanisme de spécification pour obtenir le sens souhaité. Il est à noter que les schèmes sont spécialisés pour donner une spécification déterminée. C'est ainsi que l'on parle des schèmes de lieux pour obtenir

des noms comme «école» : où l'on étudie ; «boulangerie» : où l'on fait du pain ; «bibliothèque» : où l'on trouve des livres. Tous ces mots utilisent en arabe le même schème (*mafäla*) à partir des racines associées à «étude», «pain», «livres».

3.2. Système proposé

Le système proposé est basé sur les éléments suivants :

1. base de données de schèmes où l'on associe, en particulier, à chaque schème ses classes d'utilisation (nom du lieu, nom d'outils) ;
2. base de données de racines où l'on a mis en place des relations du type thésaurus : équivalence, association, générécité, spécificité ;
3. outils de génération morphologique lexicale qui permettent de générer un mot lexical (forme canonique) à partir d'une racine et d'un schème ;
4. gestion de l'interface avec l'utilisation (terminologue) qui effectue une requête pour solliciter des mots candidats, à partir de :

- une racine, un schème,
- une racine, une classe de schèmes,
- *idem* mais demande d'emploi d'une ou de plusieurs relations au niveau des racines.

3.3. Idées d'utilisation

1. À partir du terme nouveau à traiter (français, anglais), on essaie de déterminer un niveau de champs sémantiques (un sens générique), ce qui n'est pas forcément délicat. Ce sens premier permet de lui associer une racine.

2. Ce terme traité appartient à une classe de mots : c'est un outil, c'est un endroit..., ce qui permet de mettre en évidence une ou plusieurs classes de schèmes à solliciter.

3. À partir des racines et schèmes obtenus précédemment, on sollicite le système pour obtenir des mots candidats qui serviront pour étude et validation.

Ce qui est important, dans cette démarche, c'est l'aspect exhaustivité des mots candidats (la base des schèmes est exhaustive). Autrement dit, si les mots proposés ne conviennent pas, il ne faut pas chercher à en trouver par le biais de ce mécanisme. Il faudra alors recourir aux mots composés ; puis, en dernier lieu, à l'assimilation phonétique du terme étranger.

4. TRAITEMENT DES CORPUS

L'idée, dans le cas de l'arabe, qui souffre d'un déficit important en terminologie, est de chercher un réservoir de termes candidats. Cela pourra s'envisager par le traitement de textes scientifiques anciens, écrits par des savants arabes éminents dans divers domaines de la connaissance (mathématiques, médecine, pharmacie, astrologie...) tels Avicenne ou Averroès.

Le traitement est du type suivant :

- saisie des textes (constitution des corpus) ;
- traitement linguistique (morphologie en particulier) et statistique ;
- constitution d'une base de données terminologiques et statistique associée aux corpus ;
- confrontation de cette base avec la base de données terminologiques dont on dispose actuellement ; ce qui conduit à mettre en évidence deux cas de figures théoriques :

  - mots déjà existants dans la base (une étude contextuelle permettra de savoir s'il n'y a pas eu glissement de sens, par exemple) ;

- termes existants dans les corpus mais non dans la base terminologique. Ces termes constitueront un réservoir de mots dont les études permettront de voir dans quelle mesure ils ne pourront pas être mis en circulation.

REMARQUE

Le traitement de ces corpus permet d'envisager la constitution d'un dictionnaire historique (les textes doivent couvrir les domaines et les différentes périodes), où les termes sont définis, dans le temps, en termes de contexte d'utilisation.

5. FONCTIONNALITÉ DU SYSTÈME

5.1. Nature des données de la base

Entrée arabe

- Code d'entrée (CEA)
- Entrée
- Racine
- Schème
- Classification sémantique
- Valeur grammaticale
- Domaine d'utilisation
- Commentaire
- Synonymes
- Expressions associées
- Relation type thésaurus
- Information relative à la saisie
- Information relative à la coordination

Entrée française et anglaise

- Code d'entrée (CEF ou CEE)
- Entrée
- Entrée similaire à l'arabe (sauf la racine et le schème)

Relation et correspondance

- Clé : CEE, CEF, CEA
- Domaine d'utilisation
- Condition de correspondance (ouvert)
- Information relative à la saisie

5.2. Saisie et constitution de la base de données terminologiques

Le système est conçu de façon à saisir dans une base de données de saisie (intermédiaire) qui sera constituée dans le cadre d'une organisation humaine permettant de garantir la qualité des données. Ces dernières sont utilisées pour alimenter, via un filtre, la (ou les) base(s) de données terminologiques. De sorte que l'on pourra, par le biais du filtre, constituer une base de données générale (concernant tout) et en parallèle constituer toutes les bases spécialisées que l'on voudra (informatique, mécanique, chimie...).

5.3. Confection de dictionnaires

Le système permet de confectionner des dictionnaires selon différents critères, dont entre autres :

- monolinguisme : dictionnaire français, arabe ou anglais
- bilinguisme : dictionnaires impliquant arabe-anglais-français

- trilinguisme : dictionnaires : arabe- français- anglais ; arabe ; anglais - français - arabe.

REMARQUE

La langue citée en tête constitue l'entrée principale dont l'ordre de tri pourra être selon la langue :

- français ou anglais : ordre alphabétique ; ordre alphabétique sur l'inverse de l'entrée (ordre miroir) ;
- arabe : ordre alphabétique et inversé en plus de l'ordre sur les racines (cas normal des dictionnaires du marché) et sur les schèmes.

Dans le cas des dictionnaires bilingues et trilingues, il est fourni un index sur les autres langues (ne constituant pas l'entrée principale).

- Le contenu pourra être défini par une requête d'expression «profile » du type SQL qui peut porter sur n'importe quelle rubrique de la base (sélection) et spécifier les données à faire paraître (projection).
- Le support pourra être au choix : l'imprimante (peu recommandée) ou un fichier disque récupérable pour procéder à la mise en forme au sens de l'édition. Cela pourra se faire au niveau d'un traitement de textes ou au niveau de l'imprimeur.

5.4. Consultation de la base

Cela se fait de façon classique soit par emploi du langage SQL, soit par emploi de formes permettant l'utilisation de la technique QBE (interrogation par l'exemple).

5.5. Extraction de sous-bases

Le système permet d'extraire une partie de la base (profil désiré par un abonné) spécifiée via une requête SQL pour l'exporter via le réseau ou sur support magnétique en vue d'être utilisée, en autonome, sur des postes éloignés (micro-ordinateurs).

5.6. Outils d'aide à la traduction

Autour de la base terminologique, on dispose de certaines fonctionnalités permettant l'aide à la traduction : parmi celles-ci, nous citerons :

- choix d'une base spécialisée du travail ;
- constitution d'une base de travail répondant à un profil et extraite au moyen d'une requête SQL ;
- consultation pour l'obtention de correspondants à une entrée ;
- traduction mot à mot d'un texte d'une langue source en une langue cible ;
- traduction mot à mot avec réarrangement syntaxique utilisant le contexte immédiat d'ordre 2 ou 3 ;
- traduction utilisant des algorithmes plus sophistiqués du type «analyse, transfert et génération». Ce dernier aspect n'est pas encore disponible. Des études sont en cours pour le cas de l'arabe (analyse et génération).

6. CONCLUSION

Les outils linguistiques présentés dans ce travail sont quasi opérationnels, ainsi que le système d'aide au néologisme. La base de données terminologique est en train d'être expérimentée dans le cas de l'informatique (français, arabe, anglais) avec quelque 5 000 entrées. Le traitement des corpus présenté ici, dont l'idée a été présentée par le professeur

Haj Salah, de l'Université d'Alger, constitue les prémices d'un projet qui ne pourra voir le jour que si des bonnes volontés, en particulier dans le monde arabe, se concrétisent. L'impact pourra être grand au niveau terminologique d'une part, et au niveau de la confection d'un dictionnaire historique d'autre part.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV[es] Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. Voir Hlal (1988).
2. Voir Hlal (1987b).

RÉFÉRENCES

HLAL, Y. (1985) : «Analyse morphologique de l'arabe», colloque Traitement et transmission de l'arabe, Koyeït.

HLAL, Y. (1987a) : «Génération à partir de la racine et du schème», Colloque Régional sur l'avancement de la linguistique dans le monde arabe, Université Mohamed V, Rabat.

HLAL, Y. (1987b) : «Génération morphologique de l'arabe», colloque Informatique et linguistique, Tunis.

HLAL, Y. (1987c) : «Outils linguistiques et applications», colloque organisé par l'Institut du monde arabe, Paris.

HLAL, Y. (1988) : «Un langage pour l'analyse morphologique de l'arabe», 2[e] Colloque international de la société de linguistique, Université Mohamed V, Rabat.

HLAL, Y. (1992) : «Système de gestion et d'exploitation de bases de données terminologiques», 7[e] conférence sur l'arabisation, Ligue Arabe, Khartoum.