Meta
Journal des traducteurs
Translators' Journal

META

Apport contrastif des dictionnaires généraux de la langue au problème de l'indexation automatique dans le discours technico-scientifique

Jacques Jansen

Volume 34, numéro 3, septembre 1989

1. Actes du Colloque Les terminologies spécialisées : Approches quantitative et logico-sémantique et 2. Actes du Colloque Terminologie et Industries de la langue

URI : https://id.erudit.org/iderudit/003253ar
DOI : https://doi.org/10.7202/003253ar

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

Jansen, J. (1989). Apport contrastif des dictionnaires généraux de la langue au problème de l'indexation automatique dans le discours technico-scientifique. *Meta*, *34*(3), 412–427. https://doi.org/10.7202/003253ar

Résumé de l'article

L'exposé a trait à l'analyse de textes disponibles sur support lisible par ordinateur. Il est centré sur la présentation d'un analyseur expérimental dont l'objectif est d'obtenir d'un texte une indexation thématique par des moyens automatiques.

L'analyseur exploite des bases de données extraites de dictionnaires généraux de langue (monolingues, comme LDOCE, Longman Dictionary of Contemporary English, ou bilingues, comme Le Robert & Collins, dictionnaire français - anglais/english - french dictionary).

L'indexation recherchée est obtenue par une double technique de contraste: contraste d'une part entre le profil thématique du texte à analyser, évalué par consultation des dictionnaires, et un profil thématique plus neutre évalué dans les mêmes conditions sur un texte ou corpus de textes choisi comme référence; contraste d'autre part entre le vocabulaire du texte et celui des dictionnaires consultés.

Il est fait état d'une large expérience où le corpus anglais LOB (Lancaster-Oslo/Bergen) a servi à la fois d'objet d'analyse et de corpus de référence.

L'indépendance des techniques utilisées vis-à-vis de la langue analysée est soulignée, ce qui rend leur application possible à la langue française.

APPORT CONTRASTIF DES DICTIONNAIRES GÉNÉRAUX DE LA LANGUE AU PROBLÈME DE L'INDEXATION AUTOMATIQUE DANS LE DISCOURS TECHNO-SCIENTIFIQUE

JACQUES JANSEN
Université de Liège, Liège, Belgique

SOMMAIRE

L'exposé a trait à l'analyse de textes disponibles sur support lisible par ordinateur. Il est centré sur la présentation d'un analyseur expérimental dont l'objectif est d'obtenir d'un texte une indexation thématique par des moyens automatiques.

L'analyseur exploite des bases de données extraites de dictionnaires généraux de langue (monolingues, comme LDOCE, Longman Dictionary of Contemporary English, ou bilingues, comme Le Robert & Collins, dictionnaire français — anglais/english — french dictionary).

L'indexation recherchée est obtenue par une double technique de contraste: contraste d'une part entre le profil thématique du texte à analyser, évalué par consultation des dictionnaires, et un profil thématique plus neutre évalué dans les mêmes conditions sur un texte ou corpus de textes choisi comme référence; contraste d'autre part entre le vocabulaire du texte et celui des dictionnaires consultés.

Il est fait état d'une large expérience où le corpus anglais LOB (Lancaster-Oslo/Bergen) a servi à la fois d'objet d'analyse et de corpus de référence.

L'indépendance des techniques utilisées vis-à-vis de la langue analysée est soulignée, ce qui rend leur application possible à la langue française.

0. INTRODUCTION

JUSTIFICATION DU TITRE DE LA COMMUNICATION

L'exposé a trait à l'analyse de textes disponibles sur support lisible par ordinateur. Il sera centré sur la présentation d'un analyseur expérimental dont l'objectif est d'obtenir d'un texte une INDEXATION thématique par des moyens AUTOMATIQUES, domaine où le terminologue peut à la fois apporter son concours et trouver un intérêt propre.

Cet analyseur exploite de l'information que nous avons extraite de DICTIONNAIRES GÉNÉRAUX de la langue (monolingues, comme LDOCE, ou bilingues, comme Le Robert & Collins), que nous présenterons d'ailleurs sommairement. On percevra que le caractère général de ces dictionnaires peut ouvrir des perspectives sur le traitement du DISCOURS TECHNO-SCIENTIFIQUE.

L'indexation thématique recherchée est obtenue par une double technique de CONTRASTE: contraste d'une part entre le profil thématique (concept à expliquer) du texte analysé, évalué par consultation des dictionnaires, et un profil thématique plus neutre évalué dans les mêmes conditions sur un texte ou corpus de textes de référence; contraste d'autre part entre le vocabulaire du texte et celui des dictionnaires consultés.

OBJECTIFS DE CETTE COMMUNICATION

Le premier sera bien sûr de mettre en évidence les possibilités et les perspectives offertes par notre analyseur: nous ferons ainsi état d'une large expérience où le corpus anglais LOB (Lancaster-Oslo/Bergen) (réf.3) fut à la fois objet d'étude et corpus de référence pour nos travaux. Nous tâcherons par là de valoriser la conjonction d'idées et de techniques somme toute assez simples dans le recherche de résultats pertinents en indexation automatique.

Le second sera de souligner l'intérêt de ressources lexicales existantes non codifiées à cet effet mais mises en œuvre par de telles techniques, et de faire ainsi percevoir l'apport possible de travaux lexicaux qui seraient préparés dans ce but.

Ensuite, nous serions heureux que les possibilités comme les imperfections de notre analyseur, que nous ne masquerons pas, soient génératrices d'idées susceptibles d'en repousser les limitations, et que des terminologues et lexicologues spécialisés puissent nous aider à en accroître le pouvoir de résolution.

Enfin, nous souhaiterions que transparaisse bien le fait que les techniques mises en œuvre ne sont pas tributaires de la langue jusqu'à présent étudiée, à savoir l'anglais. Nous ferons d'ailleurs état de perspectives pour le traitement du français.

I. A LA BASE : UNE IDÉE ET UN OBJECTIF

L'IDÉE

L'idée était de tenter de donner à un automate des possibilités qui imiteraient celles d'un être humain essayant de saisir les sujets évoqués dans une conversation voisine dont ne lui parviennent que des mots au hasard.

Dans une telle situation, l'esprit envisage les différents domaines de signification des mots qu'il croit saisir et se forge une idée de ce dont on parle autour de lui en tâchant de leur trouver des interprétations communes ou dominantes, et des interactions possibles.

Une telle opération pourrait être qualifiée d'indexation mentale. Ainsi, dans la phrase suivante où seuls les mots en majuscules auraient été compris par l'auditeur:

> «JEAN a le BRAS tout ROUGE : il va consulter un MÉDECIN»
> l'esprit percevra bien la connotation «médecine».

Par contre, dans la phrase :

> «LOUIS a le PIED tout BLEU : il a renversé l'ENCRIER»
> l'esprit pourra exclure la connotation «médecine» et percevoir la connotation «couleurs».

L'OBJECTIF

Notre objectif était l'élaboration d'un système visant à se substituer à l'homme dans le dépouillement de gros volumes de textes, c'est-à-dire capable de produire des index thématiques, ou encore de sélectionner du texte sur base de critères thématiques, sans pour autant perdre de vue l'intérêt indéniable dans le secteur de la traduction automatique (cf. projet EUROTRA). Différents secteurs d'activité (économie, recherche, enseignement, ...) peuvent en effet exiger une analyse rapide et condensée de volumes importants (la presse par exemple) qui prime d'autres considérations .

La démarche se devait d'être simple, économique et générale. Elle aboutirait à une grande perte d'information: incapacité à saisir la fiction, la polarité, l'intérêt scientifique (actuel ou dépassé), ou tout simplement l'idée véhiculée par un texte, qui transcende toute mesure. Avec cependant des corollaires positifs: condensation extrême des textes bruts et valeur indicative des informations retenues.

II. MISE EN ŒUVRE DE L'IDÉE

A. ESQUISSE D'UN AUTOMATE

Si l'indexation véritable requiert l'exercice de l'intelligence humaine, il est quand même permis d'envisager un automate, rudimentaire certes, qui considérerait le texte comme un ensemble de mots dotés d'attributs statiques (connotations associables à chaque mot considéré indépendamment des autres) et dynamiques (connotations des mots voisins). Le ou les sujets abordés dans le texte seraient déduits de l'observation de ces attributs.

Deux attitudes, pas véritablement exclusives et pâlement inspirées d'ailleurs du comportement de l'esprit, sont déjà possibles à ce stade :

1. tenter de trouver des connotations «communes», du moins entre mots voisins,

2. s'intéresser aux connotations les plus «fréquentes» . Le tout dans le cadre d'une modélisation statistique destinée à corriger tout possible biais.

En ce qui nous concerne, et du moins pour l'instant, nous avons choisi la deuxième attitude, qui nous évitait le délicat problème de la disambiguïsation entre mots voisins, au détriment bien sûr du bénéfice de celle-ci !

Mais revenons aux deux petites phrases précédentes, pour montrer qu'aucune de deux attitudes possibles ne parvient dans tous les cas à résoudre la question du sujet abordé. Imaginons que seuls les mots suivants soient affectés de connotations, en l'occurrence :

BRAS	→ médecine
ROUGE	→ couleurs, médecine
MÉDECIN	→ médecine

et

PIED	→ médecine
BLEU	→ couleurs, médecine
ENCRIER	→ couleurs

Il apparaît bien que, tant sur base de considérations «communes» que «plus fréquentes» , l'automate trouverait la connotation du sujet abordé dans la première phrase, mais pas dans la seconde !

B. UNE HYPOTHESE NÉCESSAIRE

Notre idée de départ s'assortissait donc d'une hypothèse simple qu'il nous appartenait de confirmer par l'expérience, à savoir que plus un texte traite d'un domaine déterminé du savoir, plus il se trouve dans le texte de mots pouvant se voir attacher une connotation évocatrice du domaine en question. Si cette hypothèse est vérifiée, alors un relevé quantitatif des connotations associables aux mots d'un texte, qui peut être envisagé de manière automatique, peut aider à déterminer les sujets marquants.

Précisons néanmoins deux choses :

1. Le texte analysé devra être d'une taille «suffisante» : les deux petites phrases citées ci-avant le montrent à suffisance.

2. Un équilibre est à trouver entre taille du texte et variabilité du sujet.

III. COMPOSANTS D'UN ANALYSEUR AUTOMATIQUE

A. DES RÉSERVOIRS DE CONNOTATIONS

Nous avons cherché à exploiter les possibilités que nous prétendions pouvoir trouver dans des dictionnaires généraux de la langue qui soient «lisibles sur ordinateur».

Pourquoi ?

1. D'abord parce que les dictionnaires, tant monolingues que bilingues, associent souvent aux définitions des mentions assez standardisées diversement appelées «field codes», «codes matière», «field labels», ou encore «champs sémantiques», mais que nous désignerons simplement par les mots «connotation» ou «étiquette», et dans le registre desquelles nous pensions pouvoir trouver certaines vertus d'un thésaurus,

2. ensuite parce que ces «étiquettes» y sont associés à des mots, ce qui nous permet de faire le pont avec le texte par des voies automatisables,

3. enfin, parce que l'information qu'ils détiennent peut à la fois être récupérée (pas toujours sans peine) et exploitée par ordinateur.

Disposant donc par convention de recherche des fichiers magnétiques du LDOCE et du Robert & Collins, ce sont ces dictionnaires que nous avons tenté d'utiliser. Présentons d'abord de brefs extraits de leur registre d'étiquettes :

LDOCE

NB: les étiquettes de LDOCE comportent en réalité une subdivision additionnelle que nous n'avons pas encore exploitée

			fréquence dans corpus LOB
AE	Aeronautics	|***	2724
AO	Aerospace and astronautics	|*	929
AS	Astronomy	|****	3730
AU	Automotive	|**********	9621
CM	Communications	|*****	4984
DP	Data processing and computer technology	|*	552
EC	Economics	|********************************	31938
IS	Information science	|**	1735
MF	Manufacturing	|****	4351
ML	Meteorology and climatology	|**********	9954
MT	Metallurgy	|***	3006
PT	Printing and publishing	|**********************	22710
TN	Transport	|****************	16177
VH	Vehicles	|***	2893

Robert & Collins

			fréquence dans corpus LOB
Anat	anatomy	|****	8641
Astron	astronomy	|*	2665
Aut	automobiles	|**********	41116
Aviat	aviation	|******	11949
Bio	biology	|*	2261
Chem	chemistry	|****	8063
Computers	computers	|*	2355
Econ	economics	|***	5602
Elec	electricity, electronics	|*******	14952
Fin	finance	|**********	32790
Phys	physics	|****	8680
Pol	politics	|**********	72600
St Ex	Stock Exchange	|*****	9107
Telec	telecom.	|********	15964

OBJECTIONS POSSIBLES :

1. La qualité des étiquettes (précision, exhaustivité, hiérarchisation,...). Les extraits précédents, et l'usage le confirme, montrent qu'elles se situent à la frontière du domaine techno-scientifique spécialisé.

2. L'absence de coefficients traduisant la probabilité relative d'usage dans l'association d'étiquettes à un mot.

3. L'omission, même justifiée, d'étiquettes possibles, ou même de toute étiquette.

4. Les limites des dictionnaires en tant que lexiques. Le caractère général des dictionnaires utilisés fait qu'un certain nombre de mots ne s'y trouvent pas : noms propres, acronymes, mots trop techniques ou simplement mots trop nouveaux. Mais cet inconvénient peut se retourner en avantage si l'on veut bien considérer que la liste de ces mots peut constituer un élément d'appréciation indéniable, tant à elle seule que comme complément à l'indexation thématique a priori recherchée (première facette de l'utilisation contrastive des dictionnaires).

B. DES FONCTIONS D'ASSOCIATION «TEXTE — > THÉSAURUS»

Deux types de fonctions s'avèrent nécessaires :

1. Des outils d'analyse morphologique pour assurer le redressement de formes infléchies en formes canoniques possibles

2. Des outils de consultation des dictionnaires, qui présupposent une organisation de l'information en «bases de données» .

C. UNE NÉCESSAIRE MODÉLISATION

Revenons un moment aux étiquettes présentées ci-dessus et jetons un coup d'œil sur les fréquences comptabilisées par examen complet du corpus LOB. Nous avons sous les yeux des tables (partielles) de fréquence. Nous les appellerions d'ailleurs «profils thématiques» si les fréquences étaient exprimées sous forme de pourcentage.

La première constatation est une grande disparité de fréquence entre étiquettes d'une même table, et il est douteux que les sujets abordés dans le corpus LOB le soient dans les mêmes rapports. La seconde constatation qui résulterait d'un examen plus minutieux, est que des étiquettes analogues semblent proportionnellement moins représentées dans une table que dans l'autre.

Il y a donc un biais probable dû à la notion même de discours, mais aussi un biais probable dû au dictionnaire utilisé pour constituer ces tables. Les fréquences observées ne peuvent donc pas être interprétées sans correction : l'analyseur devra éliminer un double biais pour distinguer les étiquettes caractéristiques et les traduire en index.

La technique que nous avons adoptée pour nous débarrasser de ce double biais est la suivante :

1. On fait, choix d'un dictionnaire.

2. On fait, choix d'un texte ou d'un corpus de textes dont on connaît le contenu et que l'on choisit comme référence.

3. On évalue, par consultation du dictionnaire choisi, un profil thématique de référence en comptabilisant toutes les étiquettes qu'il est possible d'associer à chaque mot du texte ou corpus de textes de référence.

4. On évalue de la même manière le profil thématique du texte à indexer.

5. La mise en évidence des étiquettes caractéristiques du texte analysé est alors basée sur les différences observées entre le profil thématique du texte analysé et le profil thématique du texte de référence (deuxième facette de l'utilisation contrastive des dictionnaires).

D. L'INDISPENSABLE RÉFÉRENCE

La nécessité d'un profil de référence et la souplesse de cette technique nous permettent de viser plusieurs objectifs :

1. Soit mettre en évidence les connotations majeures d'un texte (constater par exemple qu'il parle de médecine). Le profil de référence nous semble alors devoir être issu d'une grande collection de textes qui parlerait «de tout et de rien», comme le corpus LOB pris dans son ensemble. C'est le choix que nous avons fait dans l'analyse de ce corpus.

2. Soit mettre de évidence des connotations spécialisées dans un domaine prédéfini. Par exemple, pour analyser plus finement le discours médical, on évaluerait le profil de référence sur un ensemble de textes divers parlant de médecine. Nous tenterons de donner un exemple de cette technique.

3. On pourrait aussi viser la détection de l'évolution du sujet traité au fil du texte, en utilisant comme profil de référence une notion de «profil courant».

IV. LES DICTIONNAIRES UTILISÉS

Présentons maintenant, sommairement bien sûr, les dictionnaires auxquels nous avons eu recours, ainsi que les bases de données que nous en avons extraites.

A. LDOCE

1. Le fichier d'origine :

La version imprimée du LDOCE, dictionnaire unilingue explicatif de taille moyenne publié pour la première fois en 1978, est assez pauvre en étiquettes. Mais il en est tout autrement du fichier informatisé où chaque définition est dotée, entre autres codes, d'une étiquette mnémonique à quatre caractères puisée dans un répertoire de plusieurs centaines

de descripteurs et empruntée au système de codage MERRIAM-WEBSTER, partenaires de LONGMAN aux USA : les deux premiers caractères désignent un domaine d'intérêt de l'activité humaine auquel a trait la définition concernée, et les deux derniers, lorsqu'ils sont présents, soit amènent un complément d'information ou une mention relative à l'usage, annoncés respectivement par un «Z» ou un «U» en troisième position, soit désignent un domaine d'intérêt alternatif.

exemple :

cell	→ MD	(médecine et biologie)
enzyme	→ MDZC	(biochimie)
bailiff	→ LWUB	(législation, usage britannique)
abortion	→ MDLW	(médecine et biologie, ou législation)

NB: Dans une première étape, nous nous sommes volontairement limités à l'exploitation des deux premiers caractères de l'étiquette.

2. Bases de données utilisées par l'analyseur :

a. fichier MOTS
pour chaque mot ou variante orthographique :
- une clé d'entrée, attribuée de manière incrémentale et destinée à faire la liaison avec le fichier MATIÈRE
- le partie du discours (verbe, nom, ...), codée
- un drapeau de redoublement consonantique (ex: put — > putting)
- le mot (ou sa variante), en clair (= clé d'accès)

b. fichier MATIÈRE
pour chaque étiquette différente associée au mot :
- la clé d'entrée du mot
- l'étiquette en question (= clé d'accès)

B. LE ROBERT & COLLINS

1. Le fichier d'origine :

La version imprimée du Robert & Collins, dictionnaire traductif de taille moyenne publié pour la première fois en 1978, correspond exactement, informations typographiques mises à part, au fichier informatisé que nous exploitons. Les étiquettes utilisées, nombreuses également, tendent à préciser le ou les domaines du savoir ou de l'activité humaine dans lequels la traduction «terme source — > terme cible» est licite. Parfois, l'étiquette n'est pas citée s'il y a correspondance biunivoque entre le terme source et le terme cible :

exemple :
parabola
—E/F → parabole (Math)

mais par contre :
wine
—E/F → vin

Dans le discours techno-scientifique cependant, l'usage spécialisé de beaucoup de mots s'accompagnent davantage d'étiquettes :

exemple :
antenna
—E/F → (Rad, Telec, TV, Zool) antenne

antenne
—F/E → (Zool) antenna,feeler
(Rad) aerial, antenna
(TV) aerial
(Naut) lateen yard
(Mil) outpost ...

2. Bases de données utilisées par l'analyseur :

a. Fichier MOTS
pour chaque mot (les variantes orthographiques n'ont pas étésaisies) :
- une clé d'entrée, attribuée de manière incrémentale et destinée à faire la liaison avec le fichier MATIÈRE
- la partie du discours en clair
- le mot, en clair (= clé d'accès)

b. Fichier MATIÈRE
pour chaque étiquette différente associée au mot :
- la clé d'entrée du mot
- l'étiquette en question (= clé d'accès)

V. LE CORPUS LOB

Le corpus de textes LOB (Lancaster-Oslo/Bergen) a été notre champ d'expérience favori. Il s'agit d'un recueil de 500 textes authentiques d'anglais britannique, de quelque 2000 mots chacun, répartis en 15 catégories selon leur genre et leur thème et qui offre un échantillonnage représentatif de la langue écrite comme le montre la figure ci-dessous :

Catégorie	Nombre de textes	Thèmes majeurs annoncés
A	44	Press : reportage
B	27	Press : editorial
C	17	Press : reviews
D	17	Religion
E	38	Skills, trades and hobbies
F	44	Popular lore
G	77	Belles-lettres, biography, essays
H	30	Miscellaneous (government documents, foundation & industry reports, college catalogue, industry house organ)
J	80	Learned & scientific writings
K	29	General fiction
L	24	Mystery and detective fiction
M	06	Science fiction
N	29	Adventure and western fiction
P	29	Romance and love story
R	09	Humour

VI. EN PRATIQUE ...

Dans la pratique, l'opération d'analyse comprend les étapes suivantes :

1. Choix d'un dictionnaire

2. Découpe du texte en mots

Le texte est découpé en formes brutes élémentaires sur base de conventions typographiques. Au stade actuel, la découpe est sévère, en ce sens que tous les caractères non purement alphabétiques (signes de ponctuation, chiffres, etc) sont utilisés comme séparateurs. Dans la version non interactive de l'analyseur dont il est question ici, le résultat de la découpe est transformé en une table de fréquence de mots. On comprendra ici que nous renonçons actuellement à la détection de mots composés ou de collocations.

3. Analyse morphologique (réalisée par Archie Michiels)

Par inversion des règles grammaticales d'inflexion, assorties de leurs exceptions, les formes brutes ainsi obtenues sont alors redressées en formes canoniques possibles (accompagnées d'exigences diverses : partie du discours, comportement grammatical, etc.), qui seront confrontées au dictionnaire choisi. À ce stade, le mot est analysé isolément, en dehors de tout contexte syntaxique ou sémantique.

4. Consultation du dictionnaire

À ce niveau, les formes en majuscules (partiellement ou entièrement) sont translatées en minuscules (la forme majuscule étant arbitrairement imputée à la présence de signes de ponctuation, que nous ne vérifions pas, ou à une mise en évidence volontaire). Ce choix est destiné à mieux rejeter dans un «contre-index» les mots en majuscules par nature.

5. Le contre-index

Les limitations des dictionnaires en tant que lexiques permettent d'obtenir un sous-produit intéressant que nous appellerons «contre-index», où se retrouvent les mots du texte dont aucune forme canonique possible n'a été trouvée dans le dictionnaire.

6. Obtention d'un profil thématique

Chaque forme canonique possible trouvée dans le dictionnaire donne lieu à la comptabilisation d'une occurrence pour chaque étiquette qui lui est associée. Nous sommes alors en possession d'une table de fréquence d'étiquettes rapidement transformée en profil thématique par expression des fréquences sous forme de pourcentage.

7. Comparaison à un profil de référence

Nous l'avons déjà évoquée. Précisons que la démarche suivie dans la mise au point du processus de comparaison a un caractère heuristique : l'arbre est jugé à ses fruits. Les indexations produites par la machine sont évaluées par l'homme et les observations qui en ressortent servent à affiner le modèle. Nous avouons donc notre porte-à-faux vis-à-vis d'une démarche statistique régulière qui s'assurerait a priori des conditions d'application des modèles étudiés. Néanmoins, nous allons faire part de quelques considérations prises en compte dans ce processus de modélisation.

VII. LA MODÉLISATION STATISTIQUE

Nous pouvons considérer le résultat de l'indexation d'un texte comme le tableau affichant le classement des étiquettes selon les valeurs significatives décroissantes d'une fonction numérique évaluée pour chacune par comparaison du profil thématique du texte et d'un profil de référence.

Soient T(i) et R(i) les taux de citations, exprimés en pourcents, de l'étiquette E(i) respectivement dans le profil du texte et dans celui de référence. Soient encore N(i) le

nombre de formes canoniques différentes auxquelles est associée l'étiquette E(i) dans le texte. Soit S(i) la fonction de modélisation qui régit le classement.

Par convention d'échelle, S(i) doit être supérieur à 1 pour que l'étiquette E(i) puisse figurer dans l'index. Par convention également, S(i) > S(j) si et seulement si E(i) domine E(j) dans le texte. S(i), différemment définissable, est en tout cas fonction de T(i), R(i).

Les facteurs suivants, utilisés de manière différentielle ou non, jouent évidemment un rÈle primordial :

$\frac{T(i)}{R(i)}$, traduisant le gain relatif de l'étiquette E(i) dans le texte par rapport à la référence utilisée

mais aussi

$\frac{100.-R(i)}{100.-T(i)}$, traduisant le mérite à avoir obtenu ce gain

Passer en effet par exemple de 2% à 3% représente évidemment un «effort» moindre que de passer de 10% à 15%.

Des relations d'un autre genre cherchent à ne retenir que des valeurs réellement significatives et non fantaisistes, comme par exemple :

$\frac{T(i)}{R(i)} > =$ seuil à définir (légèrement supérieur à 1, par exemple 1.15) traduisant le fait que le gain de l'étiquette E(i) doit être marquant

$T(i) > =$ seuil à définir (par exemple 1%) pour éviter des fluctuations accidentelles sur des étiquettes peu représentées

$N(i) > =$ seuil à définir (par exemple 12) pour assurer la représentativité de l'étiquette E(i)

Nous n'en dirons pas plus, quoiqu'il n'y ait rien de confidentiel, pour deux raisons bien précises :

1. Le principe de notre démarche étant d'affiner le modèle en fonction du jugement porté sur ses résultats, nous ne pouvons prétendre avoir trouvé la modélisation idéale. Elle a d'ailleurs souvent évolué au cours de nos essais.

2. Des modélisations légèrement différentes peuvent être jugées de qualité identique, du moins lorsqu'on prend en compte les résultats obtenus sur plusieurs textes.

VIII. DES RÉSULTATS

Que l'on veuille bien nous croire quand nous disons que les quelques résultats que nous allons vous présenter maintenant sont authentiques et ont été obtenus avec la même version du modèle.

A. CORPUS LOB : ÉTUDE DE CORRÉLATIONS ENTRE PROFILS THÉMATIQUES

Soucieux de conforter notre démarche, nous avons préalablement voulu savoir s'il pouvait exister une affinité entre la notion quantifiable de corrélation entre profils thématiques et la parenté possible des thèmes caractérisant les textes dont ils provenaient.

Pour mémoire, un coefficient de corrélation est un nombre compris entre + 1 et — 1, censé traduire de manière quantitative une notion d'affinité entre deux listes de nombres (càd, dans le cas qui nous occupe, entre profils thématiques). Les valeurs proches de + 1 traduisent une affinité étroite, les valeurs proches de — 1 une opposition certaine, et les valeurs voisines de 0 l'absence de liaison d'un type ou de l'autre. Plusieurs méthodes d'évaluation existent, dont la méthode de Kendall utilisée ici.

Rien ne nous permet de comparer les chiffres obtenus avec un étalonnage quelconque, aussi les corrélations présentées ici ne doivent évidemment pas être interprétées de manière absolue, mais relative.

Corrélations entre profils thématiques des catégories B, H, K, N, P (30 textes chacune environ) dans le corpus LOB.

■ selon LDOCE

	B	H	K	N	P	
B	—	0.86	0.81	0.77	0.78	B Press : editorial
H	0.86	—	0.77	0.73	0.74	H Misc. documents or reports
K	0.81	0.77	—	0.89	0.91	K General fiction
N	0.77	0.73	0.89	—	0.89	N Adventure & western fiction
P	0.78	0.74	0.91	0.89	—	P Romance and love story

■ selon Robert & Collins :

	B	H	K	N	P	
B	—	0.82	0.73	0.69	0.72	B Press : editorial
H	0.82	—	0.68	0.63	0.65	H Misc. documents or reports
K	0.73	0.68	—	0.88	0.88	K General fiction
N	0.69	0.63	0.88	—	0.87	N Adventure & western fiction
P	0.72	0.65	0.88	0.87	—	P Romance and love story

Les corrélations les plus élevées sont observées à l'intérieur du groupe K/N/P (de 0.89 à 0.91 selon LDOCE, de 0.87 à 0.88 selon Robert & Collins), ou à l'intérieur du groupe B/H (0.86 selon LDOCE, 0.82 selon Robert & Collins).

À l'inverse, corrélations plus faibles entre le groupe K/N/P et le groupe B/H (de 0.73 à 0.81 selon LDOCE, de 0.63 à 0.73 selon Robert & Collins).

Il serait tentant d'attribuer celà à des similitudes ou oppositions de thèmes évoquées d'ailleurs par les titres donnés à ces catégories, mais ceci doit rester bien sûr une interprétation plus qu'une explication.

Acceptons simplement, et sous réserve de contradiction, que l'affinité intuitive entre textes n'est pas contredite par les corrélations, quel que soit le dictionnaire utilisé.

B. CORPUS LOB : CONFIRMATION PAR INDEXATION AUTOMATIQUE DES CORRÉLATIONS PRÉCÉDENTES

Intéressons-nous maintenant aux indexations obtenues, à la fois selon LDOCE et selon Robert & Collins, sur les groupes de textes B, H, K, N et P dont nous venons d'examiner les corrélations.

Category B — Press : editorial

LDOCE : Political science and government, Economics, Business, Sociology, Military, Law
R.& C. : Pol, Fin, Univ

Category H — Miscellaneous

LDOCE : Education, Economics, Business, Political science and government, Sociology, Military, Engineering
R.& C. : Ind, Gram, Univ, Fin, Admin, Comm, Jur

Category K — General fiction

LDOCE : Animal names, Games, Equestrian and manège, Clothing, Medicine and biology
R.& C. : Cards, Rel, Rel, French, Brit, Naut

Category N — Adventure and western fiction

LDOCE : Meteorology and climatology, Equestrian and manège, Games, Nautical (not navy), Transport, Clothing, Medicine and biology
R.& C. : Naut, Prov, Sport, Aut, Cards

Category P — Romance and love story

LDOCE : Animal names, Equestrian and manège, Games, Clothing, Medicine and biology
R.& C. : Prov, Cards, French, Naut, Scol, Rel, Sport, Brit

Les éditeurs du corpus ne proposant pas, pour les catégories précitées, de sous-classifications qui puissent être considérées comme des indexations de référence, le lecteur pourra difficilement juger de la valeur absolue de celles fournies par notre modèle.

On constate néanmoins ceci, tant avec LDOCE que Robert & Collins même si ce dernier le montre peut-être de manière moins nette :

1. Similitude thématique entre les catégories B et H, partageant d'ailleurs les étiquettes suivantes qui collent bien au caractère a priori sérieux et officiel de ces catégories :

LDOCE : Business, Economics, Military, Political science and government, Sociology
R.& C. : Fin, Univ.

2. Appréciation identique pour les catégories K,N et P, partageant également des étiquettes communes évoquant un probable climat de fiction, de fantaisie et de loisirs :

LDOCE : Clothing, Equestrian and manège, Games, Medicine and biology
R.& C. : Cards, Naut.

3. Dissimilarité complète des groupes B/H et K/N/P : aucune étiquette de B ou de H ne se retrouve dans K, N ou P, tant selon LDOCE que selon Robert & Collins.

Ces constatations corroborent l'évaluation des corrélations rapportées précédemment entre ces groupes, mais cette fois par une formulation qualitative d'autant plus suggestive qu'elle ne fait usage que des étiquettes qui lui apparaissent réellement significatives. Et ces résultats nous semblent d'autant plus probants que, bien que cataloguées par les éditeurs, ces catégories regroupent chacune plusieurs dizaines de textes qui abordent des domaines particuliers bien différents (problème de la variabilité du sujet).

C. CORPUS LOB : COMPARAISON D'INDEXATIONS PRODUITES PAR LE MODELE AVEC DES INDEXATIONS «MANUELLES»

Les catégories de textes étudiées ici étaient les seules catégories du corpus LOB dont la présentation par les éditeurs (annoncée par «LOB» ci-dessous) nous semblait mériter le rang d'index.

Category E — Skills, trades and hobbies

LOB : Homecraft, Handiman, Hobbies, Music, Dance, Pets, Sport, Food, Wine, Travel, Miscellaneous, Trade, Professional journals, Farming
LDOCE : Agriculture, Food, Animal husbandry, Engineering
R.&C. : Tech, Culin, Cine, Aut

Category J — Learned & scientific writings

LOB : Natural sciences, Medicine, Mathematics, Psychology, Sociology, Demography, Linguistics, Education, Politics and economics, Law, Philosophy, History, Literary criticism, Art, Music, Technology and engineering
LDOCE : Science, Linguistics and grammar, Mathematics and arithmetic, Art (not sculpture)
R.& C. : Chem, Phys, Gram, Elec, Tech, Math

Contrairement au chapitre précédent, Robert & Collins semble ici plus «pointu» dans ses étiquettes. Il faut évidemment rappeler que nous nous sommes volontairement limités à l'exploitation des deux premières positions des étiquettes du LDOCE.

D. TRAITEMENT D'UN TEXTE ISOLÉ À CONNOTATION MAJEURE UNITAIRE

Il s'agit d'un article de revue dont nous avons perdu les références, vieux de quelques années, mais traitant déjà de microinformatique. Cet article comportant plusieurs parties, indépendamment sous-titrées, nous avons tenté les opérations suivantes :

1. Indexation de l'article complet, sur base du profil thématique du corpus LOB :

index obtenu :
(nous n'avons retenu que les étiquettes dominantes, et les chiffres cités entre parenthèses sont ceux donnés par la fonction numérique de modélisation)

LDOCE : Data processing and computer technology (37.9), Manufacturing (3.9), Recording (3.3), Printing and publishing (2.5)
R.& C.: Computers (19.5), Tech (2.4)

contre-index obtenu :
(restreint, pour simplifier, aux termes rejetés à la fois par LDOCE et Robert & Collins)
Allen, America(n), Booz, Britain, CPT, English, Europe(an), Exxon, Flexowriter, Friden, Hamilton, Hewlett, IBM, Intel, Japanese, Lanier, Mercedes, Microwriter, Motorola, NBI, NEC, Nippon, Olivetti, Packard, Pascal, Redactron, Selectric, Sony, TRS, Toshiba, Vydec, Wang, Zilog

et ...

alphanumeric, businessmen, byte(s), micro, microchip, microprocessor(s), minicomputers, processor(s), processors, teletext, typesetting.

2. Indexation indépendante de chaque partie de l'article, sur base du profil thématique de l'article complet :

index obtenus :
(restreints, pour simplifier, aux deux premières étiquettes)

partie A. «Last word for the word processor».

LDOCE : Business, Science
R.& C. : Ind, Comm

partie B. «Thanks for the memory».

LDOCE : Printing and publishing, Household
R.& C. : Typ, Tennis

partie C. «Sharing the work».

LDOCE : Data processing and computer technology, Sounds
R.& C. : Press, Computers

partie D. «Star-studded future ?»

LDOCE : Building, Clothing
R.& C. : Naut, Tech

partie E. «The 64k question».

LDOCE : Track and field (athletics), Occupations
R.& C. : Sport, St-Ex

E. NOTRE APPRÉCIATION DES INDEXATIONS PRÉCÉDENTES

L'indexation semble suivre les notations thématiques majeures associées à des gros volumes de textes où la variabilité du sujet est cependant certaine.

Sur des textes plus unitaires, les étiquettes majeures semblent très bien mises en évidence, et le contre-index s'avère particulièrement pertinent !

L'expérience de détection d'orientations spécialisées dans un texte à caractère unitaire ne semble pas convaincante. On heurte évidemment ici la question de la hiérarchisation thésaurique qui est absente dans les dictionnaires utilisés. Peut-être aussi la taille de ces parties d'articles est-elle insuffisante (la loi des grands nombres ne pouvant exercer son pouvoir régulateur).

IX. CONCLUSIONS

Pour conclure, réexaminons, à la lumière de ce qui a été dit, certains objectifs que nous nous étions assignés.

Pertinence de résultats au travers d'idées et de techniques simples

Nous laisserons au lecteur le soin d'apprécier personnellement les quelques résultats présentés ici. Qu'il sache cependant que nous n'avons pas cherché à ne présenter que les moins mauvais, pour ne pas dire les meilleurs. Nous avons au contraire observé une constance rassurante dans le niveau de qualité obtenu, qui semble lui-même peu dépendant du dictionnaire utilisé.

Quant aux idées et techniques mises en œuvre, nul doute que leur simplicité n'aura échappé à personne.

Intérêt de ressources lexicales existantes exploitées par de telles techniques

Si la richesse linguistique de dictionnaires comme le LDOCE ou Le Robert & Collins est bien connue, c'est une autre facette de leur talent que nous avons cherché à exploiter ici par une modélisation statistique, à savoir leurs vertus thésauriques. Le fait d'avoir parfois pu mettre en évidence certaines insuffisances (étiquettes trop générales ou trop peu hiérarchisées) peut être pris comme un hommage indirect à la méthode utilisée. Nous sommes par contre très heureux d'avoir pu montrer que les limitations lexicales inhérentes à leur généralité pouvaient se transformer avantageusement en filtre efficace de termes spécialisés du discours techno-scientifique.

Perspectives pour l'analyse du français

Nous sommes quasi prêts à traiter le français. En effet, du point de vue technique, seule l'analyse morphologique serait à convertir. Des versions antérieures de notre analyseur où l'examen morphologique était inexistant nous ayant d'ailleurs fourni des résultats très honorables, nous estimons même qu'il ne s'agit pas là d'un préalable. De plus, le dictionnaire Robert & Collins que nous avons utilisé pour l'analyse de l'anglais étant un dictionnaire bilingue, nous espérons bien rencontrer dans la partie «français — > anglais» les même qualités lexicales et thésauriques que celles que nous avons exploitées dans la partie «anglais — > français» .

X. OUTILS INFORMATIQUES

L'analyseur a été développé sur microordinateur IBM PC/AT, à l'aide du logiciel «Turbo Pascal» (réf.7) et de la «Turbo Database Toolbox» (réf.8). La préparation des bases de données extraites des dictionnaires a fait appel au logiciel «dBASE III» (réf. 9) ainsi qu'à des prétraitements sur les gros ordinateurs de l'université de Liège.

RÉFÉRENCES

1. *LDOCE, Longman Dictionary of Contemporary English*, (1978), Longman.
2. Robert & Collins (1978) : *Dictionnaire français-anglais, anglais-français*, Paris, Société du Nouveau Littré.
3. LOB (1978), Stig JOHANSSON (in collaboration with G.N. LEECH, Helen GOODLUCK) : Manual of Information — to accompany The Lancaster-Oslo / Bergen Corpus of British English, for use with digital computers, Oslo, University of Oslo, Department of English.
4. MICHIELS, Archibal (1982) : *Exploiting a Large Dictionary Database*, thèse de doctorat, Université de Liège .
5. AMSLER, Robert A. (1984) : «Machine Readable Dictionaries», *Annual Review of Information Science and Technology*, vol. 19, Martha E. WILLIAMS, édit., pp. 161-209.
6. WALKER, Donald E. (1987) : «Knowledge resource tools for accessing large text files» , *Machine translation — Theoretical and methodological issues*, Serge Nirenburg, édit., pp. 247-261.
7. Turbo Pascal 3.0, Borland International, Scotts Valley, California.
8. Turbo Database Toolbox, Borland International, Scotts Valley, California.
9. dBASE III, Ashton-Tate, Culver City, California.

REMERCIEMENTS

Les travaux présentés ont été entrepris à l'Université de Liège à l'initiative et sous la conduite du Professeur Noël et de son collaborateur le plus proche, M. André Moulin, chargé de cours.

Le volet «analyse morphologique» a été réalisé par M. Archie Michiels, chargé de cours à l'Institut Supérieur de Traducteurs et Interprètes de Bruxelles, et à qui nous devons une thèse de doctorat consacrée à l'étude approfondie du fichier informatisé de LDOCE (réf. 4).

Ces travaux auraient été évidemment impensables si nous n'avions pu disposer, dans le cadre de conventions de recherche, des fichiers lisibles par ordinateur de deux dictionnaires importants :

- *le LDOCE, Longman Dictionary of Contemporary English (réf. 1)*
- *Le Robert & Collins, dictionnaire français-anglais / english-french dictionary (réf. 2).*

Nous en remercions chaleureusement les éditeurs.

Enfin nous ne manquerons pas de citer, dans un domaine analogue à celui que nous abordons ici, les remarquables travaux de Robert A. Amsler et Donald E. Walker, plus d'une fois publiés (comme Robert A. AMSLER, 1984, réf. 5, et Donald E. WALKER, 1987, réf. 6).