Document généré le 5 juil. 2025 23:11

Documentation et bibliothèques

L'apport des dictionnaires électroniques pour l'élaboration de thésaurus
The Contribution of Online Dictionaries in Thesaurus Construction
El aporte de los diccionarios electrónicos para la elaboración del diccionario de sinónimos

Serge Houde

Volume 38, numéro 2, avril–juin 1992

Analyse et gestion de l'information textuelle

URI : https://id.erudit.org/iderudit/1028613ar
DOI : https://doi.org/10.7202/1028613ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Houde, S. (1992). L'apport des dictionnaires électroniques pour l'élaboration de thésaurus. *Documentation et bibliothèques*, *38*(2), 91–95. https://doi.org/10.7202/1028613ar

Résumé de l'article

Les dictionnaires lisibles par machine ont fait l'objet de plusieurs recherches orientées vers leur utilisation pour la construction automatique de thésaurus et de bases de données lexicales. Le compte rendu de ces recherches fait état des méthodes utilisées pour l'extraction automatique des informations contenues dans ces dictionnaires et précise la nature des données ainsi recueillies. L'auteur présente un projet de recherche utilisant le *Robert électronique* sur CD-ROM et mené à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal dans le profil « Analyse de l'information et bases de données ». Ce projet consiste à construire, à l'aide du logiciel SATO et à partir d'une liste de termes du thésaurus du Centre des Données sur les émissions du Service de l'Information de Radio-Canada, un thésaurus contenant seulement les informations fournies par le dictionnaire et à le comparer au thésaurus-source.

Tous droits réservés © Association pour l'avancement des sciences et des techniques de la documentation (ASTED), 1992

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 en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

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.

https://www.erudit.org/fr/

Analyse

L'apport des dictionnaires électroniques pour l'élaboration de thésaurus

Serge Houde
Service documentaire de l'information
Société Radio-Canada

Les dictionnaires lisibles par machine ont fait l'objet de plusieurs recherches orientées vers leur utilisation pour la construction automatique de thésaurus et de bases de données lexicales. Le compte rendu de ces recherches fait état des méthodes utilisées pour l'extraction automatique des informations contenues dans ces dictionnaires et précise la nature des données ainsi recueillies. L'auteur présente un projet de recherche utilisant le Robert électronique *sur CD-ROM et mené à l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal dans le profil « Analyse de l'information et bases de données ». Ce projet consiste à construire, à l'aide du logiciel SATO et à partir d'une liste de termes du thésaurus du Centre des Données sur les émissions du Service de l'Information de Radio-Canada, un thésaurus contenant seulement les informations fournies par le dictionnaire et à le comparer au thésaurus-source.*

The Contribution of Online Dictionaries in Thesaurus Construction

Much research has been done on the use of online dictionaries in the automatic construction of thesauri and glossaries. An account of this body of research outlines the methodologies used to extract the information contained in these dictionaries and clarifies the nature of the data retreived. The author presents a research project using the CD-ROM version of the dictionary Robert, *which he carried out for the course « Analyse de l'information et bases de données » at the* École de bibliothéconomie et des sciences de l'information *at the* Université de Montréal. *Using the software SATO and a list of terms from the thesaurus of the* Centre des Données sur les émissions du Service de l'Information de Radio-Canada, *the project's aim was to build a thesaurus using the information found in the dictionary and to compare it to the source-thesaurus.*

El aporte de los diccionarios electrónicos para la elaboración del diccionario de sinónimos

Los diccionarios legibles con máquinas han sido muy estudiados en investigaciones orientadas hacia su utilización para la elaboración automática del diccionario de sinónimos y base de datos léxicos. El resumen de esas investigaciones da muestra de los métodos utilizados para extraer automáticamente las informaciones de los diccionarios mencionados y precisa la naturaleza de los datos recopilados de esa manera. El autor presenta un proyecto de investigación que se vale del Robert electrónico *en DC-ROM y se llevó a cabo en la Escuela de biblioteconomía y de las ciencias de la información de la Universidad de Montreal: « Análisis de información y bases de datos ». Este proyecto consiste en elaborar, mediante el programa SATO y a partir de una lista de términos del diccionario de sinónimos del Centro de Datos sobre las emisiones del Servicio de Información de Radio Canadá, un diccionario de sinónimos que contenga únicamente las informaciones suministradas por el diccionario al tiempo que se compara con el diccionario de sinónimos que sirvió de fuente.*

Les dictionnaires constituent une source quasi inépuisable d'informations. Dans les domaines spécialisés de la linguistique et de l'analyse documentaire, ces informations sont nécessaires, entre autres, aux diverses applications du traitement automatique des langues naturelles. Les dictionnaires sont également très utiles pour l'élaboration de thésaurus, outils coûteux et longs à concevoir. Un thésaurus ainsi construit peut, par exemple, servir à la recherche en plein texte ou dans des bases de données bibliographiques.

Des recherches importantes ont été entreprises un peu partout dans le monde, particulièrement au cours des dix dernières années, pour extraire automatiquement l'information des dictionnaires lisibles par machine. Il y a malheureusement peu de travaux dans ce domaine du côté des pays francophones. Il faut cependant signaler la présence, sur le marché, d'un dictionnaire français sur CD-ROM, le ***Robert électronique*** (ROBERT-E). Le fait que ce dictionnaire soit disponible au laboratoire d'informatique de l'École de bibliothéconomie et des sciences de l'information de l'Université de Montréal nous a incité à tenter de reproduire à petite échelle ces recherches à l'aide du logiciel SATO (Système d'analyse de textes par ordinateur) et à partir d'un extrait du thésaurus du Centre des Données sur les émissions du Service de l'Information de Radio-Canada.

Avant de présenter notre propre expérimentation, nous faisons un compte rendu des principaux travaux de recherche menés au cours de la dernière décennie dans le domaine de la linguistique computationnelle concernant l'apport des dictionnaires lisibles par machine pour la construction de thésaurus et de bases de données lexicales. Nous avons porté une attention particulière aux études

mentionnées par Martha Evens[1] ainsi qu'aux conférences annuelles de l'*Association for Computational Linguistics* (ACL).

LES TRAVAUX DE RECHERCHE SUR LES DICTIONNAIRES ÉLECTRONIQUES

Tous les dictionnaires présentent l'information de façon structurée selon des règles précises. La méthodologie de base de toutes les recherches dont nous avons pris connaissance repose sur cette caractéristique.

Structure de l'information dans les dictionnaires

Dans les dictionnaires, le repérage automatique des informations est rendu possible grâce à la régularité du lexique et des formules utilisées dans les définitions[2]. À titre d'exemple, une convention fréquemment utilisée dans les dictionnaires consiste à définir un NOM par une phrase incluant un autre NOM, un VERBE par une phrase incluant un autre VERBE[3]. Ex.: EXPROPRIATION = OPÉRATION ADMINISTRATIVE...

Les dictionnaires permettent donc d'établir des relations entre le mot défini et les mots de la définition. En contrepartie, les dictionnaires ne fournissent pas toutes les informations sémantiques relatives aux mots qui sont définis parce que leur conception est établie en fonction de leur utilisation par l'être humain et non par une machine[4].

Organisation de l'information en vue de son traitement

La première étape commune à toutes les recherches a comme objectif de segmenter l'information du dictionnaire et de la transposer dans des fichiers distincts. Le mot qui est défini constituera la clé d'accès à ces fichiers et les données seront les mots contenus dans la définition. Il y aura une clé d'accès pour chaque forme orthographique du mot à l'exception des formes fléchies, en autant qu'elles soient prévisibles. Cette façon de faire facilite la recherche, le mot recherché ne se trouvant pas à l'intérieur d'une autre définition[5].

D'après les chercheurs de l'équipe de l'*Illinois Institute of Technology* sur le *Webster's Seventh Collegiate Dictionary* (W7), 21% des termes utilisés dans le W7 n'ont pas d'entrée directe dans le dictionnaire[6]. C'est souvent le cas des mots avec préfixe ou suffixe. Leur première tâche a donc consisté à s'assurer que chaque mot soit relié à sa propre définition. Par la suite, les chercheurs ont réparti les données dans plusieurs fichiers selon les différentes parties du discours: adjectifs, adverbes, noms et verbes[7].

L'étape suivante consiste à finaliser le processus de transformation des données du dictionnaire en identifiant et étiquetant chaque type d'information, de manière à pouvoir y accéder facilement. Les formules de définitions constituent un outil important pour repérer et étiqueter les relations[8].

Les formules de définition

Nature et fonctions

Les formules de définition sont des mots ou des phrases que l'on retrouve fréquemment dans les définitions. Elles constituent en quelque sorte un genre de langage de spécialiste. De fait, certains mots (et, ou, ...) s'avèrent très fréquents par rapport à leur utilisation dans le langage courant.

Prépositions, conjonctions et autres mots du même genre sont relativement peu nombreux mais tout de même essentiels pour l'interprétation des autres mots de la définition. Ils sont considérés comme des éléments primitifs non définis[9,10]. Certaines abréviations, caractères et symboles ont une fonction spéciale: les parenthèses peuvent, par exemple, servir à préciser le nom scientifique du mot qui est défini[11]. Ex.: GRASS = ANY OF A LARGE FAMILY (GRAMINEAE) OF...

Chaque dictionnaire a des signes et des conventions de rédaction qui lui sont propres. En prenant en considération les informations que ces signes introduisent, il faut déterminer ceux qui seront retenus ou mis de côté pour l'étude des relations[12]. Ceci nécessite de bien maîtriser la forme de présentation des définitions dans le dictionnaire utilisé ainsi que les « marques d'usage ».

Les formules de définition marquent une relation sémantique entre les mots de la définition et le mot qui est défini. Elles indiquent les relations les plus importantes. À quelques exceptions près, ces formules n'expriment jamais plus d'une relation. À l'inverse, une relation peut être traduite par plusieurs formules[13]. Ex.: BROWED = ADJ. HAVING A BROW OR BROWS.
BROWED = ADJ. MARKED BY A BROW OR BROWS.

Extraction des formules de définition

La méthode la plus simple pour extraire automatiquement les formules de définition consiste à produire un index KWIC des définitions et à associer chaque mot à sa fréquence d'utilisation. Les formules ainsi identifiées doivent être utilisées un certain nombre de fois pour être jugées significatives.

---

1. Martha Evens, « Computer-Readable Dictionaries », in Martha E. Williams, *Annual Review of Information Science and Technology*, Elsevier Science Publishers B.V., 1989, vol. 24, p.85-117.
2. Nicoletta Calzorali, « The Dictionary and the Thesaurus Can Be Combined », in Martha Evens, *Relational Models of the Lexicon*, Cambridge, Engl., Cambridge University Press, 1988, p. 75-96.
3. Thomas Ahlswede, « A Linguistic String Grammar of Adjective Definitions from Webster's Seventh Collegiate Dictionary », in Stephanie Williams, *Human and Machines*, Norwood, NJ, Ablex, 1985, p. 101-127.
4. Hiroaki Tsurumaru et al., « An Attempt to Automate Thesaurus Construction from an Ordinary Japanese Dictionary », in *Proceedings of the 11th International Conference on Computational Linguistics, 1986 August 25-29, Bonn, FRG*, p. 445-447.
5. Roy J. Byrd et al., « Tools and Methods for Computational Lexicology », *Computational Linguistics*, vol. 13, no. 3/4 (July-December 1987), 219-240.
6. Martha Evens et al., « Lexical-Semantic Relations in Information Retrieval », in Stephanie Williams, *Human and Machines*, Norwood, NJ, Ablex, 1985, p.73-100.
7. *Ibid.*
8. Roy J. Byrd et al., « Tools and Methods ... »
9. Thomas Ahlswede, « A Linguistic String Grammar ... »
10. Thomas Ahlswede and Martha Evens, « Generating a Relational Lexicon from a Machine-Readable Dictionary », *International Journal of Lexicography*, vol. 1, no. 3 (Fall 1988), 214-237.
11. Judith Markowitz et al., « Semantically Significant Patterns in Dictionary Definitions », in *Proceedings of the Association for Computational Linguistics (ACL) 24th Annual Meeting, 1986 June 10-13; New York City, NY*, p. 112-119.
12. Thomas Ahlswede and Martha Evens, « Generating a Relational Lexicon... »
13. *Ibid.*

Grâce aux règles qui prévalent dans leur utilisation, il est possible d'associer certaines formules à la forme des mots qui sont définis. Les lexèmes formés avec le même suffixe (ou avec le même préfixe) ont en commun un certain nombre de formules de définition[14]. Ex.: UN-ACCEPTABLE = NOT ACCEPTABLE. EASI-LY = IN AN EASY MANNER.

Relations établies par les formules de définition

Les principales formules étudiées sont ici présentées en fonction des relations qu'elles introduisent. Nous les avons regroupées autour des trois grands types de relations les plus fréquemment utilisées dans les thésaurus.

a) Relations hiérarchiques

À partir des définitions, il est possible d'établir tout un réseau de termes génériques-spécifiques et de sélectionner des sous-ensembles de termes sémantiquement reliés.

1) Taxonomie

La taxonomie est fondée sur une relation transitive. Par exemple, si la carotide est une artère, et que l'artère est un vaisseau sanguin, alors la carotide est un vaisseau sanguin. Ce genre de relation est essentiel dans un thésaurus puisqu'il permet d'orienter la recherche documentaire vers des termes plus précis (spécifiques) ou à l'inverse vers des termes génériques.

Règle générale, le terme générique est placé au début de la définition. Il peut être précédé d'un quantificateur, d'un modificateur ou d'une locution prépositionnelle. Une définition peut aussi contenir plus d'un terme générique[15].

Les termes partageant un même suffixe sont souvent associés dans les définitions au même terme générique ou à des termes synonymes de ce générique.

Ex.: VOLTAMETRO = STRUMENTO « CHE MISURA » QUANTITA DI ELETTRICITA

DECOMPRESSIMETRO = APPARECCHIO « PER CALCOLARE » DATI DI DECOMPRESSION

L'association des termes spécifiques à un terme générique peut se faire à l'aide d'un programme informatique. Cependant l'intervention humaine demeure toujours nécessaire au niveau de la décision finale.

2) Relations partitives

Les relations partitives sont plus complexes que les relations taxonomiques. Elles peuvent être transitives ou non transitives. Martha Evens et ses collègues[16] établissent au total 7 types de relations partitives. Nous pourrions ne retenir pour la constitution d'un thésaurus que celles qui entretiennent une relation permanente.

Ex.: REPUBLICAN = A MEMBER OF A POLITICAL PARTY

b) Relations d'équivalence

Les recherches ont révélé que les définitions ne fournissent que des quasi-synonymes dans la majorité des cas, la relation ne pouvant s'appliquer dans les deux sens[17]. Cependant, les relations de synonymie et d'antonymie sont souvent mentionnées de façon explicite. Il suffit alors d'identifier les conventions utilisées pour pouvoir les repérer[18].

Ex.: FEMININE = SYN. FEMALE

c) Relations associatives

Elles sont multiples et révélatrices de la richesse des informations contenues dans les dictionnaires. Nous ne mentionnerons que les principales:

1) Relations de dérivation

Le suffixe d'un terme dérivé modifie la signification du terme de base. Puisque nous pouvons associer certaines formules de définition à un suffixe donné, il est possible, par exemple, de repérer tous les mots (nom et verbe) faisant référence à une action.

ASSOCIATION	=	ACTION D'ASSOCIER
(mot)		(formule) (verbe)
définition	=	action de + « base »

Dans le cas où un nom est relié à un verbe d'action qui ne partage pas le même radical, cette stratégie s'avère particulièrement efficace pour la construction d'un réseau sémantique[19].

2) Relations restrictives

Ces relations sont utilisées dans les définitions pour restreindre l'utilisation d'un terme. Elles sont souvent exprimées par des termes et des formules variées. Les expressions exprimant le « but » auquel on associe le terme qui est défini est un moyen de restreindre son sens[20, 21].

Ex.: BIJOU = PETIT OBJET OUVRAGÉ SERVANT À LA PARURE

Problèmes d'analyse

Les recherches ont démontré qu'il est possible d'extraire des données sur les relations à partir des formules de définition seulement. Dans le segment de définition qui doit être analysé, il peut être difficile de repérer le ou les mots clés en relation avec le mot défini. Le mot clé recherché n'est pas nécessairement le premier mot à la suite de la formule de définition. Il faut aussi tenir compte des conjonctions, des énumérations, etc.

Il faut donc analyser les définitions pour en extraire les informations. Un analyseur syntaxique est un outil qui peut permettre une analyse fouillée[22]. Cependant, du point de vue informatique, l'analyseur syntaxique est un « gros système », lent par surcroît, et nécessitant beaucoup de mémoire. À titre d'exemple, le *Linguistic String Parser* utilisé à l'*Illinois Institute of Technology* (IIT) fonctionne sur deux ordinateurs VAX. Malgré tout, le tiers des définitions ne peut être correctement analysé[23].

14. Martha Evens et al., « Lexical-Semantic Relations... »
15. Nicoletta Calzolari, « Detecting Patterns in a Lexical Data Base », in *Proceedings of the 10th International Conference on Computational Linguistics; 1984 July 2-6, Stanford,CA.*, p.170-173.
16. Martha Evens et al., « Lexical-Semantic Relations... »
17. Nicoletta Calzorali, « The Dictionary and the Thesaurus... »
18. Martha Evens, *Relational Models of the Lexicon*, Cambridge, Engl., Cambridge University Press, 1988.
19. Judith Markowitz et al., « Semantically Significant Patterns... »
20. Nicoletta Calzolari, « Detecting Patterns... »
21. Nicoletta Calzorali, « The Dictionary and the Thesaurus... »
22. Thomas Ahlswede and Martha Evens, « Generating a Relational Lexicon... »
23. Thomas Ahlswede and Martha Evens, « Parsing vs. Text Processing in the Analysis of Dictionary Definitions », in *Proceedings of the Association for Computational Linguistics (ACL) 26th Annual Meeting, 1988 June 7-10; Buffalo, NY*, p. 217-224.

Ce constat a eu pour conséquence une orientation des travaux de recherche vers des solutions complémentaires. L'équipe de l'IIT se tourne donc du côté des logiciels d'analyse de texte, sans base de connaissance linguistique, pour repérer les informations explicites (tels les synonymes et les chaînes du genre MOT - RELATION - MOT). En catégorisant la forme lexicale de chaque mot de la définition selon son appartenance à une partie du discours (nom, verbe, adjectif, etc.), les chercheurs ont pu repérer le segment à analyser et le mot clé de la définition avec des taux de succès impressionnants. À ce niveau d'analyse, il nous a semblé qu'un logiciel comme SATO — qui permet justement l'ajout de valeurs de catégories aux formes du lexique — pourrait être utilisé efficacement pour reproduire en partie les travaux dont nous venons de faire état.

LA CONSTRUCTION D'UN THÉSAURUS À L'AIDE DU ROBERT-E ET DE SATO

Pour notre expérimentation, nous avons constitué à partir d'un thésaurus-source, le thésaurus du Centre des Données du Service de l'Information de Radio-Canada, cinq sous-ensembles de termes ayant des caractéristiques communes. En tout, près de deux cents termes désignant des noms d'objets, des disciplines scientifiques, des individus, des établissements ou des mots formés avec le suffixe « -tion ». L'objectif ultime de la recherche consistait à construire un thésaurus contenant seulement les informations fournies par le dictionnaire et à le comparer au thésaurus-source.

Méthodologie générale

L'homogénéité des cinq sous-ensembles visait à repérer les régularités dans les articles de définitions. Nous avons cherché à déterminer si, en attribuant une valeur de catégorie grammaticale à chaque forme du lexique constitué à partir des définitions et en identifiant les formules de définition utilisées, nous étions en mesure de repérer le segment à analyser et les mots clés de la définition.

Nous avons utilisé au maximum les potentialités du ROBERT-E. Il a donc été possible d'extraire facilement bon nombre d'informations en utilisant une fonction de ce dictionnaire électronique permettant de regrouper à l'écran les synonymes selon les divisions de sens et les éléments analogiques, éléments hyponymiques, relations logiques (ex. partitives), relations associatives. Pour chaque terme retenu du thésaurus-source, nous avons par la suite déchargé dans des fichiers de travail une partie de la définition fournie par le ROBERT-E. La portion de la définition qui a été retenue est celle qui détermine le sens, donc épurée des informations secondaires, citations, etc.

L'information extraite du ROBERT-E a été structurée en zones, chacune associée à un type d'information. Cette structuration des données est exigée par les logiciels de gestion de thésaurus comme EDIBASE et s'avère utile pour certains traitements dans SATO.

* entrée = X, numéro séquentiel
* zone = te, terme d'entrée, le descripteur
* zone = na, note d'application, dans le cas présent : la définition
* zone = ep, employé pour, terme synonyme rejeté
* zone = tg, terme générique
* zone = ts, terme spécifique
* zone = ta, terme associé

Ex. : * entrée = 21
* zone = te Psychiatrie
* zone = na Partie de la médecine qui étudie et traite les maladies mentales, les troubles pathologiques de la vie psychique
* zone = ep => Neuropsychiatrie
* zone = ep => Pédopsychiatrie
* zone = ep => Sociopsychiatrie

(=> : identifie dans le ROBERT-E les termes regroupés à l'écran pour leur association avec le terme défini)

Expérimentation

Les données ont été analysées à l'aide du logiciel SATO. Ce logiciel a d'abord permis de déterminer la fréquence des termes du lexique contenus dans les définitions et de repérer les formules de définition.

Par exemple, dans le fichier des termes identifiant des disciplines scientifiques, les mots : BRANCHE, PARTIE, ÉTUDE, SCIENCE, OBJET étaient ceux qui avaient la fréquence d'utilisation la plus élevée parmi les mots catégorisés comme étant des noms communs ou des verbes, après « projection » des bases de données lexicales livrées avec SATO. Le logiciel SATO a la particularité de pouvoir accepter diverses catégorisations au niveau des formes lexicales et d'afficher chaque occurence dans un contexte déterminé par l'utilisateur (la phrase, le paragraphe, la « notice » etc.). En consultant le contexte des mots ayant les fréquences les plus élevées, nous avons pu découvrir les formules de définition utilisées.

Ex. : OCÉANIE = <u>ÉTUDE</u> <u>DES</u> MERS ET OCÉANS

GYNÉCOLOGIE = <u>ÉTUDE</u> <u>DE</u> L'ORGANISME DE LA FEMME

BIOLOGIE = <u>ÉTUDE</u> <u>DES</u> ÊTRES VIVANTS

À partir de ces formules, nous avons recherché la concordance ordonnée des mots formant la formule de définition suivie des mots clés de la définition, c'est-à-dire de ceux qui sont en relation avec le mot qui est défini.

Ex. : DERMATOLOGIE = <u>PARTIE DE</u> LA <u>MÉDECINE</u>
(mot clé) (formule)

Au cours de la recherche, nous avons constaté que notre méthode ne parvenait pas à éliminer le « bruit » associé au repérage de termes non pertinents. La constitution d'un fichier de blocage de termes composés a permis de réduire un peu le bruit au repérage et de simplifier la recherche des mots clés.

De fait, le mot défini est souvent en relation avec un ou plusieurs termes composés de la définition. Ces termes peuvent avoir les formes suivantes :

NOM COMMUN + ADJECTIF
Ex.: CHIRURGIE = PARTIE DE LA THÉRAPEUTIQUE MÉDICALE
NOM COMMUN + (D', DU, DES, DE LA, DE L') + NOM COMMUN
EX.: UNIVERSITÉ = ÉTABLISSEMENT D'ENSEIGNEMENT

La plupart du temps, le mot clé recherché sera un nom commun ou une lexie nominale complexe connue. À la suite d'une formule de définition, nous avons donc cherché à repérer les mots ayant la valeur grammaticale, nom commun ou lexie nominale complexe.

Un fichier de blocage des expressions nominales complexes s'inspirant de la procédure MARQTERM mise au point par Suzanne Bertrand-Gastaldy et Gracia Pagola serait très utile : blocage des mots liés par et, ou, de, du, d', de l', de la.

Ex.: ETHNOLOGIE = SCIENCE DES GROUPES HUMAINS ET DES GROUPES SOCIAUX

Il serait également très utile de lier ensemble les concordances strictes de noms communs ou d'adjectifs rejetés qu'ils soient séparés ou non par une virgule, précédés d'un article ou d'une préposition. De cette façon nous pourrions être assurés de repérer, d'une seule commande, les énumérations fréquentes qu'on retrouve dans un dictionnaire.

EX.: SUBSTANCE_ÉLASTIQUE_,-_IMPERMÉABLE

PHÉNOMÈNES_DE_L'_ESPRIT_,_DE_LA_PENSÉE

Résultats

Les résultats de la recherche mettent en évidence le fait que les définitions du dictionnaire ont permis de repérer surtout des termes génériques et associés. Le faible nombre de termes spécifiques est notable. Il est fort possible qu'en analysant l'ensemble de l'article du dictionnaire dans sa version détaillée, nous aurions alors trouvé beaucoup plus de termes clés en relation avec le terme défini. Nous avons tout de même pu exploiter la régularité des formules de présentation des articles du ROBERT-E malgré un nombre relativement restreint de définitions analysées.

Ex.: THÉSAURUS-DICTIONNAIRE

* entrée = 9
* zone = te Génétique
* zone = tg Science
* zone = tg Biologie
* zone = ta Hérédité

* entrée = 10
* zone = te Gérontologie
* zone = tg Médecine
* zone = ta Vieillissement

THÉSAURUS-SOURCE

* Génétique
  EP Hérédité
  EP Manipulation génétique
  XX Biologie
  XX Science
  TA Biotechnologie
  TA Malformation congénitale
  C1 Science - SVS

* Gérontologie
  EP Gériatrie
  XX Science sociale
  TA Personne âgée
  TA Vieillissement
  C1 Science - SVS

Le thésaurus obtenu à l'aide des informations du ROBERT-E (thésaurus-dictionnaire) a une taille 2,5 fois plus petite que le thésaurus-source (804/316 renvois) soit environ 39 % de la taille du thésaurus-source. Cependant, dans certains cas, le nombre de termes de référence est supérieur dans le thésaurus-dictionnaire. Ce fut le cas pour le fichier des termes désignant des disciplines scientifiques pour lequel le nombre total de termes génériques et de termes associés est plus élevé dans le thésaurus-dictionnaire. Il faut mentionner que ce corpus de définitions était relativement riche en information.

CONCLUSION

Il est intéressant de remarquer jusqu'à quel point l'information contenue dans ce genre d'ouvrage de référence « traditionnel » devient accessible lorsqu'associée à la technologie informatique. La combinaison de la technologie du CD-ROM et de l'hypertexte en fait un outil polyvalent et permet une consultation nettement plus en profondeur. La qualité des informations contenues dans le dictionnaire en fait un ouvrage intéressant pour la construction des thésaurus. Nous avons pu constater, au niveau des termes génériques par exemple, la grande rigueur du dictionnaire. La difficulté réside essentiellement dans l'extraction automatique ou semi-automatique des informations du dictionnaire. Les recherches actuelles n'ont pu lever tous les obstacles. Il s'agit néanmoins d'un champ de recherche important dans le contexte des développements informatiques associés à l'information textuelle.

SOURCES D'INFORMATION COMPLÉMENTAIRES

AHLSWEDE, Thomas. « A Tool Kit for Lexicon Building », in *Proceedings of the Association for Computational Linguistics (ACL) 23rd Annual Meeting, 1985 July 8-12; Chicago, Ill., Morristown, N.J.* p. 268-276.

AMSLER, Robert. « Machine-Readable Dictionaries », in WILLIAMS, Martha E. *Annual Review of Information Science and Technology.* White Plains, NY, Knowledge Industry Publications, 1984. vol.19, p.161-209.

MILLER, George. « Dictionary of the Mind », in *Proceedings of the Association for Computational Linguistics (ACL) 23rd Annual Meeting, 1985 July 8-12, Chicago, Ill. Morristown, N.J.* p.305-314.

MOLHOT, Pat and GOLDBOGEN, Geof. « The Use of Inter-Concept Relationships for the Enhancement of Semantic Networks and Hierarchically Structured Vocabularies », in *Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, Proceedings of the Conference, 1990 October 28-30; Waterloo, Ont.* p.39-51.

NEFF, Mary and BOGURAEV, Braninir. « Dictionaries, Dictionary Grammars and Dictionary Entry Parsing », in *Proceedings of the Association for Computational Linguistics (ACL) 27th Annual Meeting, 1989 June 26-29; Vancouver, B.C.* p.91-101.

Le Robert électronique, [manuel d'interrogation]. [Paris, s.n.], 1989.

TREMBLAY, Diane. « Traitement des langues naturelles et développement de banques de données linguistiques ». *Terminogramme*, vol.46 (janvier 1986), 11-13.