

# Thésaurus et systèmes de traitement automatique de la langue Thesauri and Automatic Language Processing Tesauros y sistemas de tratamiento automático de la lengua

Lyne Da Sylva

Volume 52, numéro 2, avril-juin 2006

Les langages documentaires

URI : <https://id.erudit.org/iderudit/1030018ar>

DOI : <https://doi.org/10.7202/1030018ar>

[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

Da Sylva, L. (2006). Thésaurus et systèmes de traitement automatique de la langue. *Documentation et bibliothèques*, 52(2), 149-156.  
<https://doi.org/10.7202/1030018ar>

Résumé de l'article

Cet article expose le rôle que le thésaurus documentaire classique est appelé à jouer dans certains systèmes de traitement automatique de la langue. Ce type de ressource lexicale est très prisé par le domaine du traitement automatique de la langue, puisqu'il permet d'appréhender, au moins en partie, la sémantique des documents. Cette capacité peut être mise à contribution dans un grand nombre d'applications différentes. Sont présentés les thésaurus les plus utilisés pour le traitement automatique, suivis des exigences particulières qui s'appliquent aux thésaurus pour le traitement automatique. Enfin, sont brièvement abordés les efforts de construction automatique de thésaurus.

# Thésaurus et systèmes de traitement automatique de la langue

LYNE DA SYLVA

École de bibliothéconomie et des sciences de l'information

Université de Montréal

lyne.da.sylva@umontreal.ca

## RÉSUMÉ | ABSTRACTS | RESUMEN

*Cet article expose le rôle que le thésaurus documentaire classique est appelé à jouer dans certains systèmes de traitement automatique de la langue. Ce type de ressource lexicale est très prisé par le domaine du traitement automatique de la langue, puisqu'il permet d'appréhender, au moins en partie, la sémantique des documents. Cette capacité peut être mise à contribution dans un grand nombre d'applications différentes. Sont présentés les thésaurus les plus utilisés pour le traitement automatique, suivis des exigences particulières qui s'appliquent aux thésaurus pour le traitement automatique. Enfin, sont brièvement abordés les efforts de construction automatique de thésaurus.*

### *Thesauri and Automatic Language Processing*

*The role of the classic thesaurus in certain systems of automatic language processing is the central theme of this article. This type of lexical resource is important in the field of automatic language processing because considers the semantics of documents. This capacity can become an asset with a wide range of applications. The thesauri used for automatic processing and their requirements are described. Finally, the projects to automatically construct thesauri are briefly discussed.*

### *Tesauros y sistemas de tratamiento automático de la lengua*

*Este artículo expone el papel que el tesoro documental clásico está llamado a desempeñar en algunos sistemas de tratamiento automático de la lengua. Este tipo de recurso léxico es muy apreciado por el campo del tratamiento automático de la lengua, porque permite comprender, al menos en parte, la semántica de los documentos. Esta capacidad puede ser aprovechada por numerosas aplicaciones diferentes. Se presentan los tesauros más utilizados para el tratamiento automático, seguidos de las exigencias particulares que se aplican a estos tesauros. Finalmente se aborda de manera breve los esfuerzos de construcción automática de tesauros.*

LE DOMAINE DE LA GESTION DOCUMENTAIRE voit apparaître de plus en plus de logiciels qui prétendent alléger ou même remplacer le traitement humain des documents. On a raison d'être sceptique devant leurs promesses de performance. Comment penser que ces systèmes de traitement automatique, qui ne traitent que des chaînes de caractères, puissent arriver à saisir le contenu conceptuel des documents qu'ils manipulent? Cet article vise à illustrer le rôle que joue le thésaurus documentaire, dans sa version la plus classique, dans certains systèmes de traitement automatique. L'exposé témoignera du grand intérêt porté à ce type de ressource lexicale par le milieu du traitement automatique de la langue (TAL). L'avantage principal du thésaurus dans ces systèmes est de permettre d'aller au-delà de la simple forme graphique des mots pour s'approcher d'une représentation du sens. Cette capacité accrue peut être mise à contribution dans un grand nombre d'applications différentes.

Notons déjà que la notion de thésaurus dans les contextes documentaires est souvent bien différente de celle qui est adoptée dans les systèmes informatiques. Dans le premier contexte, l'objet est bien défini et son développement est normalisé sur les plans national et international. Dans le deuxième contexte, le « thésaurus » est un concept flou, faisant référence à des dictionnaires de synonymes, des dictionnaires analogiques, des thésaurus documentaires, des réseaux sémantiques ou à des structures plus complexes que sont les ontologies.

Nous décrirons d'abord divers contextes d'utilisation des thésaurus attestés dans les écrits du domaine du TAL avant de recenser les principaux thésaurus utilisés dans ces travaux. Nous énoncerons les propriétés requises pour que les thésaurus puissent être utilisés par les systèmes de traitement automatique, ainsi qu'un certain nombre de problèmes récurrents. Une problématique reliée sera abordée, celle de la construction automatique de thésaurus. Nous conclurons avec quelques remarques prospectives quant à l'avenir du thésaurus pour les systèmes de traitement automatique.



## Autres applications du traitement automatique de la langue

Diverses applications du TAL visent à saisir automatiquement certains liens sémantiques entre les mots, pour servir des applications ultimes comme la traduction ou la condensation automatiques. Pour ce faire, certaines tâches intermédiaires se révèlent nécessaires et pour lesquelles les thésaurus peuvent être utiles.

### Calcul de distance sémantique

On peut vouloir se donner une mesure quantitative de la distance sémantique entre deux mots, pour pouvoir exprimer que « chien » et « molosse » sont très proches de par leur sens, que « chien » et « canidé » le sont aussi mais à un degré moindre, et que « chien » est plus proche de « éléphant » que de « soucoupe », par exemple. Une technique proposée se sert de représentations hiérarchiques des concepts, telles que les thésaurus, pour mesurer cette distance: on attribue la distance entre deux équivalents (synonymes) à 0, entre un spécifique et un générique à 1, entre deux spécifiques d'un même générique à 2, etc. Il s'agit de compter le nombre de liens à traverser dans la hiérarchie pour aller d'un concept à l'autre. Certains chercheurs exploitent cette technique à différentes fins, dont B. Sugato *et al.* (2001) et Z. Zhang *et al.* (2005) dans un contexte d'extraction d'informations, Y. Kim *et al.* (2001) pour la traduction automatique et H. Alani *et al.* (2000) pour choisir certains termes associés utilisés pour étendre une requête. J. Ferlez et M. Gams (2004) en évaluent l'hypothèse de base, en mettant en comparaison des jugements humains sur la similarité des mots. Un des problèmes de cette approche est qu'elle présuppose (à tort) que la distance sémantique est représentée uniformément dans les liens (Resnick, 1995).

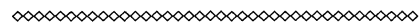
### Résolution d'anaphores

La résolution d'anaphores décrit la façon dont on détermine automatiquement l'antécédent d'un pronom ou d'un autre type d'anaphore (voir Mitkov, 2002). Il s'agit, par exemple, de repérer dans les exemples suivants que « ils », « les minous » et « ces animaux » font tous référence à « les chats ».

- ▷ Josée aime bien les chats. *Ils* sont affectueux et enjoués.
- ▷ Josée aime bien les chats. *Les minous* sont affectueux et enjoués.
- ▷ Josée aime bien les chats. *Ces animaux* sont affectueux et enjoués.

Les cas de reprises lexicales (et non pronominales), comme dans les deux derniers exemples,

## Diverses applications du TAL visent à saisir automatiquement certains liens sémantiques entre les mots.



nécessitent le recours à un thésaurus pour résoudre l'anaphore. En rétablissant les liens sémantiques entre les entités du discours décrites par leurs synonymes ou leurs génériques, la résolution d'anaphores avec thésaurus (Nasukawa, 1994; Denber, 1998) peut aider à repérer dans des textes des réponses à des questions (Litkowski, 2001) ou à effectuer de l'extraction d'information complexe (Putejovskiy *et al.*, 2002).

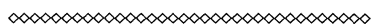
### Préférences sélectionnelles

On nomme « préférences sélectionnelles » les critères sémantiques qu'impose, par exemple, un verbe à ses arguments. Ainsi, le verbe « distribuer » sélectionne un sujet humain ou animé, un objet direct inanimé et, de façon optionnelle, un objet indirect animé. Par exemple: « Marie distribue des sandwichs aux enfants. » Par ailleurs, un nom comme « sandwich » peut avoir un complément de type « aliment » (ou « viande » ou « farce salée »), notamment « au jambon ». La phrase suivante doit être analysée de façon à ce que « au jambon » soit le complément de « sandwichs » et non de « distribuer »: « Marie distribue des sandwichs au jambon. » Une façon d'aiguiller le système vers la bonne analyse dans chaque cas est de prévoir les types de compléments des verbes et des noms à l'aide de termes d'un thésaurus (Sumita *et al.*, 1995): « humain » (et tous ses spécifiques) pour le sujet et pour l'objet indirect, « objet inanimé » pour l'objet direct, etc.

### Désambiguïisation lexicale en contexte

La tâche de désambiguïisation lexicale en contexte (Preiss et Stevenson, 2004) consiste à déterminer, pour un mot polysémique comme « tour », par exemple, quel est son sens dans un énoncé donné. Ainsi, dans la phrase « Ces tours sont un exemple d'architecture gothique », on devrait pouvoir sélectionner automatiquement le sens architectural de « tour » d'après les autres mots du contexte (contrairement au cas de « J'ai fait des tours de voiture »). Cette étape est nécessaire entre autres à la traduction, pour choisir entre les traductions possibles d'un mot polysémique, ainsi qu'à la recherche d'information efficace. Un thésaurus peut faciliter cette désambiguïisation automatique, notamment pour le calcul de distance sémantique entre les divers mots du contexte. Cette technique est utilisée par M. Sussna (1993) pour l'indexation,

## *WordNet n'est pas un thésaurus documentaire, notamment parce qu'il contient un bon nombre d'entrées non nominales.*



par J.M.G. Hidalgo *et al.* (2005) pour la classification et par R. Mandala *et al.* (2000b) et L.A. Urena *et al.* (2000) pour la recherche d'information.

### *Cohésion lexicale, chaînes lexicales et segmentation de textes*

Un certain nombre d'applications du TAL reposent sur la segmentation automatique (Hernandez et Grau, 2002) d'un texte en passages cohérents sur le plan thématique. Le résumé automatique en est un exemple: H. Saggion (2002) identifie les sections d'un article scientifique typique, alors que A. Farzindar *et al.* (2004) découpent des jugements de la Cour fédérale du Canada en sections selon une structure prédéfinie pour le genre. Chacun produit ensuite un résumé qui respecte la segmentation du texte original. Une des techniques proposées pour effectuer cette segmentation automatique exploite la notion de cohésion lexicale (Halliday et Hasan, 1976), assurée entre autres par la récurrence de thèmes dans un discours. Un calcul de similarité peut être effectué pour deux phrases consécutives en tenant compte de la répétition des mots d'une phrase à l'autre (Morris et Hirst, 1991; Hearst, 1997; Harabagiu, 1999; Da Sylva et Doll, 2005). Deux phrases ayant un grand nombre de mots en commun auront un score de similarité plus élevé que deux phrases qui ont peu de mots, voire aucun, en commun. Tant que le score de similarité entre les phrases successives est élevé, on suppose que la suite présente une unité thématique. Un score de similarité très bas indiquerait une rupture dans la thématique et une coupure est alors proposée à cet endroit. Au-delà de la répétition exacte de mots, pour capter la reprise thématique par un synonyme ou un générique, l'utilisation d'un thésaurus est nécessaire. Ainsi, avec un thésaurus approprié, on peut attribuer un score de cohésion élevé pour les deux phrases suivantes, bien qu'aucun mot ne soit répété: «Le père regarda son garçon. L'homme était fier de son fils.» J. Morris et G. Hirst (1991), et M. Hajime *et al.* (1998) font usage d'un thésaurus dans l'algorithme de segmentation. Les chaînes lexicales ainsi créées ([père, homme] ainsi

que [garçon, fils]) peuvent être utilisées entre autres pour identifier les thèmes principaux d'un document (Chali, 2001) ou pour construire un résumé (Barzilay et Elhadad, 1997).

### **Principaux thésaurus utilisés pour le traitement automatique**

Certains thésaurus existants sont privilégiés par les systèmes de traitement automatique.

#### *WordNet*

WordNet (<<http://wordnet.princeton.edu>>), développé au *Cognitive Science Laboratory* de l'Université de Princeton, est de loin le thésaurus le plus utilisé par les systèmes de traitement automatique (sauf dans le domaine médical). Sa conception a été inspirée par les théories actuelles en psycholinguistique. Limité à l'anglais, sa couverture lexicale est toutefois importante: 155 327 mots-formes différents ou 207 016 paires de mots et sens (<<http://wordnet.princeton.edu/man/wnstats.7WN>>). De ceux-là, plus de 117 000 sont des noms, mais WordNet contient aussi plus de 22 000 adjectifs, 11 400 verbes et 4 600 adverbes. Ce sont tous des mots dits «de la langue générale».

En réalité, WordNet n'est pas un thésaurus documentaire, notamment parce qu'il contient un bon nombre d'entrées non nominales, ainsi que des relations sémantiques additionnelles par rapport au thésaurus traditionnel (par exemple, la relation méronymique «partie-tout»). Les liens de synonymie sont exprimés entre les sens des mots et non entre les mots eux-mêmes. Ainsi, le mot *bank* a plusieurs sens. Le sens «institution financière» est associé à un certain nombre de synonymes, tels que *banking company* par exemple, ce qui définit un *synset*; le sens «berge d'une rivière» est relié à ses propres synonymes et génériques. Les deux sens sont rattachés directement au mot *bank*, mais c'est au *synset* que sont rattachées les relations thésaurales. Il n'y a pas dans WordNet d'équivalent pour l'opposition entre descripteurs et non-descripteurs.

#### *Le thésaurus Roget's*

Le *Roget's International Thesaurus* (Chapman et Roget, 1992) est un dictionnaire analogique ou de synonymes de l'anglais, dont la première version date de 1852. Il a été utilisé notamment pour calculer la cohésion lexicale (Morris et Hirst, 1991) et la désambiguïsation de sens (Yarowsky, 1992). Sa structure conceptuelle est très particulière: il s'agit d'une hiérarchie conceptuelle, un arbre dont les feuilles terminales sont les mots, et les six catégories supérieures regroupent des concepts généraux et non des mots: «*Words*

*Expressing Abstract Relations*»; « *Words Relating to Space*»; « *Words Relating to Matter*»; « *Words Relating to the Intellectual Faculties*»; « *Words Relating to the Voluntary Powers*»; « *Words Relating to the Sentient and Moral Powers* ». Les niveaux intermédiaires regroupent les mots en ensembles basés sur l'analogie. Ce n'est pas un thésaurus documentaire, cependant il semblerait mieux adapté que WordNet pour effectuer certains calculs de distance sémantique (Jarmasz et Szpakowicz, 2003).

## EuroWordNet

Le thésaurus EuroWordNet (Vossen, 1998; <<http://www.illc.uva.nl/EuroWordNet/>>) est un équivalent multilingue de WordNet, parrainé par la communauté européenne. Le projet initial couvrait sept langues, dont le français (chaque langue est liée aux équivalents anglais); d'autres versions linguistiques sont en cours de développement. Il a été peu utilisé jusqu'à maintenant; l'utilisation qu'en font J. Gonzalo *et al.* (1998) pour la recherche d'information translinguistique exploite son atout majeur: la recherche à l'aide de requête dans une langue de documents écrits dans une autre langue.

## UMLS

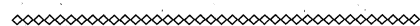
UMLS (*Unified Medical Language System*) est un métathésaurus multilingue du domaine médical (<<http://www.nlm.nih.gov/research/umls/>>) utilisé dans les systèmes de recherche d'information, notamment par D. Eichmann *et al.* (1998) pour la recherche translinguistique. Notons que pour le domaine biomédical, plusieurs recherches reposent également sur l'utilisation des vedettes-matière MeSH (*Medical Subject Headings*).

## **Caractéristiques des thésaurus requises pour l'utilisation automatique et difficultés résiduelles**

Un examen attentif des travaux précités et des observations que l'on y recense suggère que pour être utilisés efficacement par un système de traitement automatique, les thésaurus doivent satisfaire un certain nombre d'exigences. Les plus importantes sont, dans l'ordre: l'accessibilité, la pertinence et la rigueur de conception du thésaurus.

La notion d'accessibilité fait référence à la facilité d'acquisition et d'utilisation du thésaurus. Idéalement, on cherche un thésaurus gratuit et libre de droits. Le format de fichier doit être le plus universel possible: format ASCII (.txt), délimité simplement. À la rigueur, un format de base de données facilement transformable (base de données relationnelle, par exemple) est acceptable, mais pas un format propriétaire (.doc ou

*Pour être utilisés efficacement par un système de traitement automatique, les thésaurus doivent satisfaire un certain nombre d'exigences.*



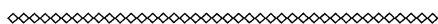
autres semblables) ou basé sur l'image de documents imprimés (PDF). Enfin, il faut avoir accès à la totalité du thésaurus en format numérique, d'une manière qui permette le traitement en lots (*batch processing*) et non limité à une consultation par interface de requête ou par une succession d'hyperliens.

La pertinence dénote la concordance entre les caractéristiques du thésaurus et celles du système de TAL visé. D'abord, le domaine: un thésaurus spécialisé est préférable pour un système spécialisé dans le même domaine, même si les résultats seront alors difficilement transposables à d'autres contextes. Pour un thésaurus de langue générale, la couverture lexicale doit être excellente. Et, bien sûr, le thésaurus doit être disponible dans la bonne langue.

Puisque les systèmes automatiques ne peuvent pas interpréter le sens des chaînes qu'ils manipulent, la rigueur dans la définition des relations est primordiale. La relation hiérarchique doit être utilisée de façon très stricte, en conformité avec les normes d'ailleurs. Pour la relation d'équivalence, les équivalents doivent être de vrais synonymes linguistiques, et non des équivalents documentaires définis contextuellement. Par exemple, dans le thésaurus AGROVOC (<[http://www.fao.org/aims/ag\\_intro.htm](http://www.fao.org/aims/ag_intro.htm)>), « bœuf » et « bouvillon » sont en relation d'équivalence, alors qu'ils ne sont pas strictement synonymes. De manière plus générale, toutes les relations utilisées devraient être univoques, interprétables d'une seule façon. Or c'est rarement le cas pour la relation associative (TA). On voudrait bien l'utiliser dans l'extension de requêtes, mais la variété des relations qu'elle encode peut engendrer rapidement des non-sens. Pour un traitement automatique, le thésaurus sera plus utile si la relation associative est remplacée par d'autres relations plus spécifiques: tout-partie, acteur-action, action-résultat, action-lieu, etc. Ce codage explicite permet de programmer un traitement différencié pour chaque relation.

Même si ces exigences sont respectées, un certain nombre de difficultés demeurent. La forme d'un terme dans le thésaurus n'est pas nécessairement celle que l'on retrouvera dans un texte. Une lemmatisation sera nécessaire, c'est-à-dire qu'il faudra ramener le terme à sa forme « de base », au masculin singulier le cas échéant. Notons que le problème n'est pas le même en français et en anglais: en anglais, il faudra

*L'avenir demeure néanmoins prometteur, même si l'on doit surveiller la montée des ontologies, en particulier dans le domaine biomédical.*



aussi lemmatiser le thésaurus, puisque les termes s'y retrouvent essentiellement au pluriel, conformément aux normes pour le développement des thésaurus multilingues (ISO 5964). Il faudra également faire abstraction des qualificatifs. Par exemple, dans le thésaurus AGROVOC, on trouve à la fois « poisson (aliment) » et « poisson (animal) ». Dans un texte donné, on ne trouvera bien sûr que la chaîne « poisson » et il faudra retirer le qualificatif du thésaurus pour le comparer aux mots du texte. Il faudra aussi effectuer une désambiguïsation lexicale en contexte, pour déterminer si le texte parle de l'animal ou de sa chair comestible, ce qui est un autre problème; celui-ci n'est cependant pas limité aux termes avec qualificatif, mais est présent pour tout terme du thésaurus qui est polysémique.

Il est souvent impossible de dénicher un thésaurus dans un domaine spécialisé (c'est encore plus vrai pour des systèmes bilingues ou multilingues). On s'accommode souvent alors d'un thésaurus de langue générale comme WordNet, mais les résultats sont habituellement décevants.

Certains termes d'un thésaurus documentaire sont accompagnés de notes d'application qui régissent leur utilisation. On pense ici notamment aux notes qui renvoient d'un non-descripteur à plus d'un descripteur, selon le sens; un système automatique sera incapable d'interpréter la note pour choisir le descripteur approprié.

Ainsi, malgré leur intérêt, les thésaurus présentent des difficultés d'utilisation qui ne sont pas toujours simples à résoudre.

### **Une problématique reliée : la construction automatique de thésaurus**

On peut identifier deux motivations à la construction automatique de thésaurus. D'abord, les chercheurs en TAL sentent le besoin d'avoir des ressources lexicales appropriées pour la collection qu'ils sont à traiter: dans la bonne langue et dans le bon domaine. mais l'expertise humaine est coûteuse et les thésaurus existants, même s'ils sont nombreux, ne sont pas nécessairement pertinents. Ensuite, il y a un intérêt théorique à vérifier des hypothèses quant aux connaissances implicites encodées dans des textes

ou des dictionnaires. On trouve bon nombre d'écrits sur le sujet, dont un nombre important sur le chinois (pour n'en nommer que quelques-uns: Foo *et al.*, 2000; Tseng, 2002; Yang et Luk, 2003).

Les premiers travaux dans ce domaine ont exploité les dictionnaires de définitions (Shaikovich, 1985; Houde, 1992). Les techniques les plus répandues procèdent par analyse de corpus, selon des approches linguistiques (Bertrand-Gastaldy et Pagola, 1992; Takenobu *et al.*, 1995; Morin et Jacquemin, 2004) ou statistiques (par exemple: Salton, 1972; Guntzer *et al.*, 1989; Crouch, 1990; Grefenstette, 1994; Park et Choi, 1996; Hodge et Austin, 2002; Kar et Yang, 2005; Dejean *et al.*, 2005).

Dans la plupart des cas, les résultats sont en fait proches d'un thésaurus, mais pas identiques, puisqu'ils contiennent des termes reliés de façon erronée. Mais la recherche se poursuit.

### **Conclusion**

Une des frontières reconnues au TAL est le traitement de la sémantique des textes: le traitement formel des aspects lexicaux, morphologiques ou syntaxiques est plutôt bien maîtrisé, comparativement à celui du sens. Les thésaurus sont des outils très appréciés pour rendre possible, au moins partiellement, cette tâche de compréhension du sens, en raison de leur structure formelle et de la clarté des relations sémantiques impliquées. Cette utilité est évidente en regard de l'éventail des contextes d'application des thésaurus présentés ci-haut. Il est presque chose courante, dans des exposés de travaux en TAL, que l'on mentionne en passant que tel thésaurus a été utilisé pour effectuer des généralisations sémantiques.

On remarque aujourd'hui une certaine concurrence entre les notions de thésaurus, de taxinomies (normalement restreintes à la relation hiérarchique) et d'ontologies (permettant une gamme infinie de relations entre les concepts). Chacune a ses applications privilégiées. Les ontologies prendront de plus en plus d'importance dans le contexte de développement du Web sémantique. Pour l'instant, les thésaurus existants sont rapidement happés par les concepteurs de systèmes de TAL, encore preneurs de thésaurus bâtis selon les règles de l'art. L'avenir demeure néanmoins prometteur, même si l'on doit surveiller la montée des ontologies, en particulier dans le domaine biomédical. ☉

### **Sources consultées**

Abuzir, Y. et F. Vandamme. 2001. Automatic e-mail classification based on thesaurus. In *Proceedings of the IASTED International Conference on Applied Informatics. International Symposium on Software Engineering, Databases, and Applications*, 519-524.

- Alani, H., C. Jones et D. Tudhope. 2000. Associative and spatial relationships in thesaurus-based retrieval. In *Proceedings of Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000 (Lecture Notes in Computer Science vol. 1923, 45-58)*.
- Ardo, A. et T. Koch. 1999. Automatic classification applied to full text Internet documents in a robot-generated subject index. In *Proceedings of the 23rd International Online Information Meeting, 239-246*.
- Bang, S.L., J.D. Yang et H.J. Yang. 2006. Hierarchical document categorization with k-NN and concept-based thesauri. *Information Processing & Management* 42 (2): 387-406.
- Barzilay, R. et M. Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL97/EACL97 Workshop on Intelligent Scalable Text Summarization, Madrid, 11 juillet 1997, 10-17*.
- Bertrand-Gastaldy, S. et G. Pagola. 1992. L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur: applications possibles avec SATO. *Documentation et bibliothèques* 38 (2): 75-89.
- Chali, Y. 2001. Topic detection using lexical chains. In *Engineering of Intelligent Systems: 14th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE 2001: Proceedings (Lecture Notes in Artificial Intelligence vol. 2070), 552-558*.
- Chapman, R. et P. Roget. 1992. *Roget's International Thesaurus*. 5<sup>e</sup> éd. New York, NY: HarperCollins.
- Chartron, G., S. Dalbin, M.G. Monteil et M. Verillon. 1989. Indexation manuelle et indexation automatique: dépasser les oppositions. *Documentaliste — Sciences de l'information* 26 (4-5): 181-187.
- Chu, W.W., Z. Liu, W. Mao et Q. Zou. 2005. A knowledge-based approach for retrieving scenario-specific medical text documents. *Control Engineering Practice* 13 (9): 1105-1121.
- Crouch, C.J. 1990. Approach to the automatic construction of global thesauri. *Information Processing & Management* 26 (5): 629-640.
- Da Sylva, L. et F. Doll. 2005. Information architecture for document description: Semantic thematization of text segments. In K. Tochtermann et H. Maurer (dir.). *Proceedings of I-KNOW '05, 5th International Conference on Knowledge Management, Graz, Austria, 29 June-1 July 2005, 612-620*.
- Dejean, H., E. Gaussier, J.-M. Renders et F. Sadat. 2005. Automatic processing of multilingual medical terminology: Applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine* 33 (2): 111-124 (Special Issue: Information Extraction and Summarization from Medical Documents).
- Denber, M. 1998. *Automatic resolution of anaphora in English*. Technical report, Eastman Kodak Co.
- Dillon, M. 1982. Thesaurus-based automatic book indexing. *Information Processing & Management* 18 (4): 167-178.
- Efthimiadis, E.N. 1996. Query expansion. In M.E. Williams (dir.), *Annual Review of Information Science and Technology* 31: 121-187.
- Eichmann, D., M.E. Ruiz et P. Srinivasan. 1998. Cross-language information retrieval with the UMLS Metathesaurus. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 72-80*.
- Farzindar, A., G. Lapalme et J.-P. Desclés. 2004. Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues (TAL)* 45 (1): 1-21. (Numéro spécial: Le résumé automatique de texte: solutions et perspectives).
- Ferlez, J. et M. Gams. 2004. Shortest-path semantic distance measure in WordNet v2.0. *Informatica* 28 (4): 381-386.
- Foo, S., S.C. Hui, H.K. Lim et L. Hui. 2000. Automatic thesaurus for enhanced Chinese text retrieval. *Library Review* 49 (5-6): 230-239.
- Ginsberg, A. 1993. Unified approach to automatic indexing and information retrieval. *IEEE Expert* 8 (5): 46-56.
- Gonzalo, J., F. Verdejo, C. Peters et N. Calzolari. 1998. Applying EuroWordNet to crosslanguage text retrieval. *Computers and the Humanities* 32 (2-3): 185-207.
- Greenberg, J. 2001a. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science* 52 (6): 487-498.
- . 2001b. Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science* 52 (5): 402-415.
- Grefenstette, G. 1994. *Explorations in automatic thesaurus discovery*. Dordrecht: Kluwer Academic.
- Guntzer, U., G. Juttner, G. Seegmuller et F. Sarre. 1989. Automatic thesaurus construction by machine learning from retrieval sessions. *Information Processing & Management* 25 (3): 265-273.
- Hajime, M., H. Takeo et O. Manabu. 1998. Text segmentation with multiple surface linguistic cues. In *Proceedings of COLING-ACL'98, 881-885*.
- Halliday, M. et R. Hasan. 1976. *Cohesion in English*. London: Longman.
- Harabagiu, S. 1999. From lexical cohesion to textual coherence: A data driven perspective. *International Journal of Pattern Recognition and Artificial Intelligence* 13 (2): 247-265.
- Hearst, M.A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23 (1): 33-64.
- Hernandez, N. et B. Grau. 2002. Analyse thématique du discours: segmentation, structuration, description et représentation. In *Actes de CIDE' 2002, Hammamet, Tunisie, 277-285*.
- Hidalgo, J.M.G., M. de Buenaga Rodriguez et J.C.C. Perez. 2005. The role of word sense disambiguation in automated text categorization. In *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005 (Lecture Notes in Computer Science 3513), 298-309*.
- Hodge, V.J. et J. Austin. 2002. Hierarchical word clustering — automatic thesaurus generation. *Neurocomputing* 48: 819-846.
- Houde, S. 1992. L'apport des dictionnaires électroniques pour l'élaboration de thésaurus. *Documentation et bibliothèques* 38 (2): 91-95.
- Jarmasz, M. et S. Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), 212-219*.
- Kar, W.L. et C.C. Yang. 2005. Automatic crosslingual thesaurus generated from the Hong Kong SAR Police Department Web corpus for crime analysis. *Journal of the American Society for Information Science and Technology* 56 (3): 272-282.
- Kim, Y., B.T. Zhang et Y.T. Kim. 2001. Collocation dictionary optimization using WordNet and k-nearest neighbor learning. *Machine Translation* 16 (2): 89-108.
- Litkowski, K.C. 2001. Syntactic clues and lexical resources in question-answering. In *Information Technology: Ninth Text REtrieval Conference (TREC-9) (NIST SP 500-249), 157-166*.
- Mandala, R., T. Tokunaga et H. Tanaka. 2000a. Query expansion using heterogeneous thesauri. *Information Processing & Management* 36 (3): 361-378.



- \_\_\_\_\_. 2000b. Improving information retrieval system performance by combining different text-mining techniques. *Intelligent Data Analysis* 4 (6): 489-511.
- Mitkov, R. 2002. *Anaphora Resolution*. London: Longman.
- Morin, E. et C. Jacquemin. 2004. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities* 38 (4): 363-396.
- Morris, J. et G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17 (1): 21-48.
- Nasukawa, T. 1994. Robust method of pronoun resolution using full-text information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Kyoto, Japan*, 1157-1163.
- Park, Y.C. et K.S. Choi. 1996. Automatic thesaurus construction using Bayesian networks. *Information Processing & Management* 32 (5): 543-553.
- Pizzato, L.A.S. 2003. Query expansion based on thesaurus relations: evaluation over Internet. In *Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003: Proceedings (Lecture Notes in Computer Science Vol. 2588)*, 553-556.
- Preiss, J. et M. Stevenson. 2004. Introduction to the special issue on word sense disambiguation. *Computer Speech and Language* 18 (3): 201-207.
- Putejovsky, J., J. Castano, J. Zhang, M. Kotecki et B. Cochran. 2002. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Pacific Symposium on Biocomputing, Honolulu, Hawaii, 4-9 January 2000*, 362-373.
- Resnik, P. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montréal, 20-25 August 1995*, 448-453.
- Saggion, H. et G. Lapalme. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics* 28 (4): 497-526.
- Salton, G. 1972. Experiments in automatic thesaurus construction for information retrieval. In *Information Processing 71: Proceedings of the IFIP Congress 1971, 1, Ljubljana, Yougoslavie, 23-28 August 1971*, 115-123.
- Shaikevich, A.Y. 1985. Automatic construction of a thesaurus from explanatory dictionaries. *Automatic Documentation and Mathematical Linguistics* 19 (4): 76-89.
- Soergel, D. 1999. The Rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science* 50 (12): 1119-1120.
- Srinivasan, P. 1996. Query expansion and MEDLINE. *Information Processing & Management* 32 (4): 431-443.
- Sugato B., R.J. Mooney, K.V. Pasupuleti et J. Ghosh. 2001. Evaluating the novelty of text-mined rules using lexical knowledge. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), San Francisco, California, 26-29 August 2001*, 233-238.
- Sumita, E., O. Furuse et H. Iida. 1995. An example-based disambiguation of English prepositional phrase attachment. *Systems and Computers in Japan* 26 (4): 30-41.
- Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM 93), Washington, DC, 1-5 November 1993*, 67-74.
- Takenobu, T., I. Makoto et T. Hozumi. 1995. Automatic thesaurus construction based on grammatical relations. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), pt. 2, Montréal, 20-25 August 1995*, 1308-1313.
- Tseng, Y.H. 2002. Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology* 53 (13): 1130-1138.
- Urena, L.A., J.M.G. Hidalgo et M. de Buenaga. 2000. Information retrieval by means of word sense disambiguation. In *Proceedings of the Third International Workshop on Text, Speech and Dialogue, TSD 2000. (Lecture Notes in Artificial Intelligence Vol. 902)*, 93-98.
- Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In *17th International Conference on Research and Development in Information Retrieval (SIGIR'94), Dublin, Ireland, 3-6 July 1994*, 61-69.
- Vossen, P. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Dordrecht, Netherlands: Kluwer.
- Yang, C.C. et J. Luk. 2003. Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws. *Journal of the American Society for Information Science and Technology* 54 (7): 671-682.
- Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92, Nantes, France, 23-28 August 1992*, 454-460.
- Zhang, H.P., J. Sun, B. Wang et S. Bai. 2005. Computation on sentence semantic distance for novelty detection. *Journal of Computer Science and Technology (English Language Edition)* 20 (3): 331-337.
- Zhang, Z., L. Da Sylva, C. Davidson, G. Lizarralde, G. et J.Y. Nie. 2004. Domain-specific QA for the construction sector. In *Proceedings of the Workshop on Information Retrieval for Question Answering (IR4QA), SIGIR'04, Sheffield, UK, 29 July 2004*, 65-71.