CHAPITRE QUATRIÈME

Traitement automatique des langues pour l'accès au contenu des documents ⋆

Christian Jacquemin⋆, Pierre Zweigenbaum†

⋆ LIMSI-CNRS
BP 133
91403 ORSAY Cedex
FRANCE
jacquemin@limsi.fr et
http://www.limsi.fr/Individu/jacquemi/

† DIAM : Service d'informatique médicale, DSI/AP-HP
et Département de Biomathématiques, Université Paris 6
91, bd de l'Hôpital, 75634 Paris Cedex 13
pz@biomath.jussieu.fr et
http://www.biomath.jussieu.fr/~pz/

1 INTRODUCTION

Le traitement automatique des langues et la recherche d'information sont deux disciplines dont l'interaction, déjà identifiée depuis longtemps, s'est renforcée ces dernières années.

Le *traitement automatique des langues* (TAL) s'intéresse aux traitements informatisés mettant en jeu du matériau linguistique : analyse de textes, génération de textes, traduction automatique sont parmi les grands types de traitements ; correction orthographique et grammaticale en sont d'autres. La conférence francophone TALN ou les conférences internationales COLING et (E)ACL, la revue francophone Traitement automatique des langues et la revue internationale Computational Linguistics permettent de préciser cette caractérisation.

La *recherche d'information* (RI), ou recherche documentaire, vise à retrouver des documents textuels répondant à un besoin informationnel (un *thème*), spécifié par une requête. Les conférences internationales RIAO et SIGIR, la revue internationale Information Processing & Management sont quelques

⋆ Publié dans « Traitement automatique des langues pour l'accès au contenu des documents. In Jacques Le Maître, Jean Charlet et Catherine Garbay, éditeurs, *Le document en sciences du traitement de l'information*, chapitre 4, pages 71–109. Cepadues, Toulouse, 2000 ».

uns des forums de la discipline. Les moteurs de recherche disponibles sur la « Toile » ont donné une nouvelle jeunesse à la RI, tout en montrant les limites de ses techniques classiques.

La recherche d'information, dans la mesure où elle travaille aussi sur des textes, s'apparente au TAL. Leurs liens sont anciens (voir par exemple les travaux de Karen Sparck-Jones [A.0.8]) et leurs frontières perméables. Les méthodes classiques en RI effectuent néanmoins des traitements linguistiques assez limités. On peut donc espérer qu'une meilleure prise en compte de la nature linguistique du matériau traité, comme le TAL cherche à l'obtenir, apporte une amélioration des résultats de la RI, voire amène à d'autres applications que la simple recherche d'un document pertinent au sein d'une base volumineuse. Nous allons examiner ici les principales techniques concernées.

Nous rappelons d'abord les notions de base du traitement automatique des langues (section 2), puis celles de la recherche d'information (section 3). Le lecteur connaissant les bases du TAL sautera sans regret la section 2 ; celui familier avec la RI fera de même avec la section 3. La section 4 recense diverses techniques de TAL directement pertinentes en recherche d'information. La section 5 présente deux exemples d'applications en RI qui illustrent le rôle de ces techniques. Une bibliographie structurée complète ce texte. Elle est organisée en sections et sous-sections et va au-delà des références citées dans le corps du texte ; les appels de références bibliographiques comme [A.0.2] suivent cette organisation.

2 GRANDS DOMAINES DU TRAITEMENT AUTOMATIQUE DES LANGUES

Nous brossons ici les grands domaines du TAL, en nous appuyant sur un découpage méthodologique classique dans le domaine et en linguistique :

La morphologie (section 2.1) concerne l'étude de la formation des mots et de leurs variations de forme ;

La syntaxe (section 2.2) s'intéresse à l'agencement des mots et à leurs relations structurelles dans un énoncé ;

La sémantique (section 2.3) se consacre au sens des énoncés ;

La pragmatique (section 2.4) prend en compte le contexte d'énonciation.

Dans ce qui suit, nous présentons ces domaines sous l'angle de l'analyse automatique des textes.

2.1 Morphologie

D'un point de vue informatique, un texte est une chaîne de caractères. La première étape de l'analyse d'un texte est la reconnaissance, dans cette chaîne de caractères, d'unités linguistiques de base, les mots, et la mobilisation des informations associées, puisées dans un lexique.

Pour commencer, la chaîne de caractères d'entrée doit utiliser un encodage déterminé (typiquement, pour le français, l'encodage ISO-latin-1), les caractères de contrôle (fin de ligne, etc.) étant eux aussi normalisés. On élimine généralement les caractères non répertoriés.

Il s'agit ensuite de segmenter la chaîne d'entrée en unités élémentaires (en anglais, *tokens*). Différents choix peuvent être effectués à ce stade, selon les séparateurs choisis : tous les caractères non alphabétiques (espaces, apostrophes, tirets...) ou les espaces seulement ; et selon que l'on prend en considération les « mots composés » (*« pomme de terre »* = une unité) ou pas. En tout état de cause, on est généralement amené à distinguer la notion d'unité minimale (« token ») et celle de mot (associé à une information lexicale).

Le *lexique*, en première approximation, est la liste des mots de la langue, et associe à chaque mot les informations linguistiques correspondantes : catégorie syntaxique, traits morphosyntaxiques (genre, nombre, etc.), etc. Plusieurs phénomènes amènent à préciser cette définition du lexique.

- Un mot peut exister sous plusieurs formes : en français, formes fléchies des noms, adjectifs, etc., conjugaison des verbes. On peut alors considérer une *forme canonique*, ou lemme, pour chaque mot, qui sert d'entrée dans le lexique pour l'ensemble de ses formes fléchies (singulier pour le nom, masculin singulier pour l'adjectif, infinitif pour le verbe).

- Un mot peut avoir plusieurs sens (*polysème*) : *« avocat »*, *« coup »*, *« livre »* en sont des exemples ; selon le cas, plusieurs entrées ou sous-entrées sont alors distinguées.

- Plusieurs mots peuvent se trouver partager une forme commune (*homographes*) : *« montre »* est une forme du nom *« montre »* aussi bien

que du verbe *« montrer »* ; *« pu »* est le participe passé du verbe *« pouvoir »* mais aussi de *« paître »*.

– Un mot peut être construit à partir d'un autre : par dérivation (*« penser »* ↦ *« pensable »* ↦ *« impensable »*) ou par composition (*« compter »* + *« gouttes »* ↦ *« compte-gouttes »*, *« un »* + *« jambe »* ↦ *« unijambiste »*, *« sclérose »* + *« artère »* ↦ *« artériosclérose »*).

Enfin, pour de multiples raisons, tous les mots possibles d'une langue ne sont ou ne peuvent être répertoriés a priori dans un lexique. D'une part, les noms propres constituent un inventaire ouvert. D'autre part, de nouveaux mots sont créés régulièrement (néologie) par dérivation et composition, mais aussi par siglaison, abréviation, emprunt, etc. [D.0.1].

2.2 Syntaxe

Pour repérer quels mots fonctionnent ensemble dans une phrase, un premier niveau de modélisation consiste à constituer des classes de mots (catégories syntaxiques, parties du discours) possédant un fonctionnement similaire : Nom (N), Verbe (V), Adjectif (A), etc.

Certaines unités, par accident (homographes : *« la »*, *« est »*) ou de façon plus systématique (*« normale »* : A ↦ N, *« coronarographie »* N ↦ *« coronarographier »* V ↦*« coronarographie »* V), peuvent être *ambiguës* entre plusieurs catégories (ambiguïté catégorielle ou lexicale). Par exemple, chacune des unités de la phrase *« La coronarographie est normale. »* est ambiguë, ce que l'on peut noter :

$$\text{« } La_{/DET,N,PRO}\ coronarographie_{/N,V}\ est_{/A,N,V}\ normale_{/A,N}.\ \text{»}$$

On remarquera que dans le contexte de la phrase entière, aucune de ces unités n'est ambiguë.

Les relations syntaxiques entre les mots d'une phrase peuvent se représenter de plusieurs façons. Le modèle en constituants considère des groupes de mots, ou syntagmes, généralement centrés sur un mot de tête (N, V, etc.), et les modélise par des catégories spécifiques (syntagme nominal ou SN, syntagme verbal ou SV, syntagme adjectival ou SA, etc.). Ces syntagmes peuvent eux-mêmes être éléments d'autres syntagmes, et la structure d'une phrase est alors un *arbre de constituants* (figure 1a).

Le modèle en dépendance considère directement les mots de tête (recteurs, ou régissants), et leur attache les mots qui en dépendent (régis). La structure

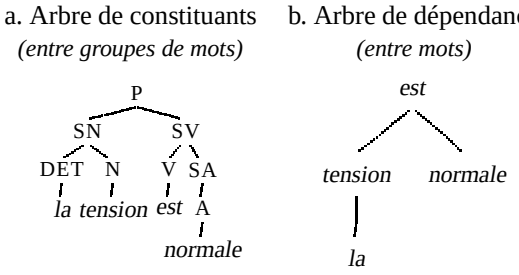

FIG. 1 – *Représentations syntaxiques d'une phrase.*

d'une phrase est alors un arbre de dépendance (figure 1b). Des équivalences existent entre les deux modèles.

Même sans ambiguïté lexicale, une phrase peut donner lieu à plusieurs structures syntaxiques (ambiguïté structurelle). Un exemple classique est la phrase *« je vois un homme avec un télescope »* (figure 2), dans laquelle *« avec un télescope »* peut désigner la manière dont je vois l'homme (2a, attachement au verbe *« vois »*, complément circonstanciel de manière) ou au contraire une caractéristique de l'homme (2b, avec un attachement au nom *« homme »*, complément de nom). Des informations sémantiques, voire pragmatiques (comme ce serait le cas ici), sont nécessaires pour déterminer l'interprétation la plus appropriée de ce genre de phrase.

Des relations plus précises entre mots ou syntagmes sont utiles à l'interprétation des phrases. Les relations grammaticales classiques (sujet-verbe, verbe-objet, verbe-objet-indirect, nom-modifieur, etc.) permettent de représenter la fonction des groupes de mots les uns par rapport aux autres. Les relations entre pronom et antécédent, et plus généralement entre anaphore (pronom, mais aussi nom) et antécédent, mobilisant encore davantage sémantique et pragmatique, assurent des mises en relation qui peuvent se situer à distance plus grande et qui sont très utiles en recherche d'information (section 4.3.2).

Les propriétés intrinsèques des mots restreignent le type de relations syntaxiques qu'ils peuvent avoir. C'est en particulier le cas des verbes, qui ré-

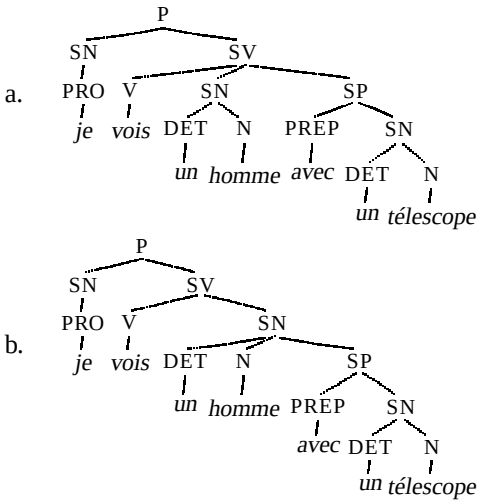

FIG. 2 – *Ambiguïté structurelle*

gissent ou *sous-catégorisent* de zéro à trois ou quatre « arguments » :

« Il pleut. »	*pleuvoir()*
« Jean dort. »	*dormir(X)*
« Jean prend un livre. »	*prendre(X, Y)*
« Jean donne un livre à Marie. »	*donner(X, Y, Z)*
« Jean interdit à Médor de sortir. »	*interdire(X, Y, Z)*

2.3 Sémantique

De même que pour la syntaxe, un premier niveau de modélisation consiste à constituer des classes de mots (*catégories sémantiques*). Ces classes regroupent des mots dont le sens est proche, ou au minimum (pour des classes générales) des mots qui possèdent certaines propriétés sémantiques communes.

Cependant, si en syntaxe on arrive à s'accorder sur des jeux de catégories relativement consensuels (il s'agit d'une vue d'ensemble ; de près, le tableau est beaucoup plus polychrome) [E.1.5], en sémantique aucune classification universelle n'existe (la constitution d'une classification universelle risque

même d'être théoriquement impossible). Les classifications que l'on pourra utiliser (par exemple, les catégories générales de WordNet [F.1.5, G.1.5]) reflètent nécessairement un point de vue, une prise de position culturelle ou ontologique spécifique.

Un mot, même syntaxiquement non ambigu, pourra posséder plusieurs sens. Par exemple, on pourra distinguer l'*« artère »*–vaisseau sanguin de l'*« artère »*–avenue, même si le second est étymologiquement un sens figuré du premier. Le contexte permet en général de déterminer quel sens est à l'œuvre dans un énoncé.

Les mots d'une langue entretiennent un réseau riche de *relations sémantiques paradigmatiques* : hyperonymie / hyponymie (*« vaisseau »*/*« artère »*), méronymie (partie d'un tout : *« vaisseau »* / *« système cardiovasculaire »*), antonymie (*« malin »* / *« bénin »*) et autres contraires, etc. [F.1.3, F.1.4, F.1.1, F.1.2].

Dans un énoncé, les relations grammaticales sont le support de *relations sémantiques syntagmatiques*. Par exemple, les différents actants d'un événement jouent différents *rôles thématiques* : agent, thème, source, destination, etc. Ainsi, dans *« Jean donne un livre à Marie. »*, les rôles par rapport à l'événement *« donne »* pourront être :

$$\text{« } Jean_{/agent,\ source}\ donne\ un\ livre_{/thème}\ à\ Marie_{/destination}\cdot \text{ »}$$

Un mot qui désigne un événement possède des propriétés combinatoires restreintes : il sélectionne comme actants certains types de mots (*restrictions de sélection*). Ces types restreints peuvent être exprimés en termes de classes sémantiques. On pourra par exemple poser pour le verbe *« donner quelque chose à quelqu'un » donner(animé, objet, animé)*, ou encore pour *« interdire » interdire(animé, animé, événement)*.

La représentation sémantique finale que l'on vise à associer à un énoncé dans un système de TAL dépend de l'objectif de ce système. Cet objectif peut être l'extraction d'informations spécifiques (section 3.3), comme c'est le cas dans les tâches définies dans les campagnes MUC d'évaluation de systèmes d'analyse de textes [F.3.1, F.3.2]. Par exemple, l'évolution des postes d'une personne dans une ou plusieurs entreprises (figure 3) était l'une des tâches de la campagne MUC6 [F.3.2].

Un éventail d'informations plus large peut aussi être recherché. La représentation doit alors être plus complète, comme dans le système MENELAS [F.3.3]. La figure 4 montre une représentation (simplifiée pour des raisons de place) de la phrase *« Patient âgé de 62 ans, hospitalisé pour angor spontané*

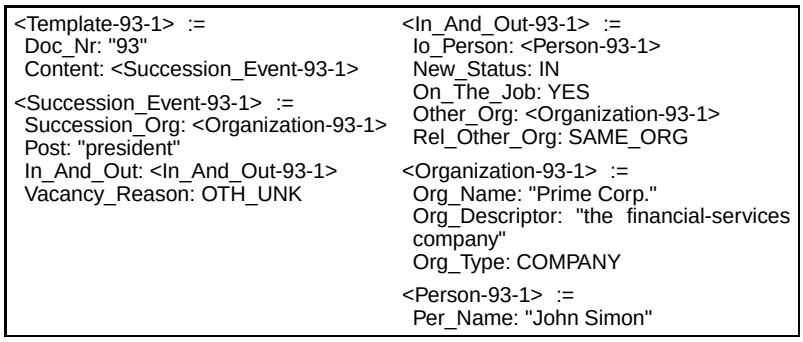

FIG. 3 – *Une représentation à la MUC.*

à répétition. ».

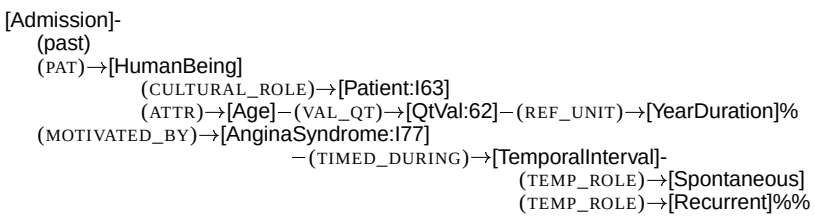

FIG. 4 – *Une représentation en Graphes Conceptuels.*

Les formalismes de représentation employés [F.4.3, F.4.5] sont en général issus de l'Intelligence artificielle, comme les logiques de description [F.4.2, F.4.1] et les Graphes Conceptuels [F.4.4, F.4.6].

2.4 Pragmatique

L'interprétation d'un énoncé dépend de son contexte. Dès que l'on veut traiter plus d'une phrase (et même pour une seule phrase), cette dimension intervient.

Le *co-texte* désigne le texte qui précède (et suit) la phrase courante. Deux

facteurs concourent à faire qu'une phrase s'insère bien dans un texte.

- La *cohésion* régit la continuité du texte. Elle est assurée par l'emploi d'anaphores (section 4.3.2), l'homogénéité du thème, un emploi judicieux d'ellipses, etc.
- La *cohérence* détermine l'intelligibilité du texte. Elle s'appuie sur des structures de discours ainsi que sur les relations causales, temporelles, etc., entre les événements décrits.

Au-delà du texte lui-même, les conditions d'énonciation et les connaissances partagées complètent le contexte d'un énoncé. L'interprétation devra donc faire appel à des connaissances sur le monde (scénarios, plans, etc. [F.6.4, F.6.3]). L'identification de structures de discours (structure de dialogue, structure argumentative, etc.) est également nécessaire selon le type de texte [F.6.5]. De façon générale, une représentation de la situation décrite par un énoncé demande d'effectuer des inférences à partir de représentations initiales (par exemple, « littérales ») de cet énoncé et de représentations du contexte [F.6.1, F.6.2].

3 GRANDS DOMAINES DE LA RECHERCHE D'INFORMATION

Nous brossons d'abord un tableau rapide des principes généraux de la recherche d'information (RI) (section 3.1). Nous passons ensuite en revue différents raffinements qui peuvent lui être apportés (section 3.2). Nous présentons finalement la notion d'extraction d'information, une tâche du traitement automatique des langues proche de la RI (section 3.3).

3.1 Principes généraux de la recherche d'information

La recherche d'information cherche des documents répondant à un besoin informationnel, ou *sujet* (figure 5), exprimé à l'aide d'une *requête*. Les documents sont au préalable *indexés* : chaque mot de chaque document est répertorié dans une table inverse, avec ou sans conservation des positions des mots dans le texte d'origine. L'appariement entre la requête et l'index va déterminer les documents qui sont considérés comme répondant le mieux au besoin informationnel initial.

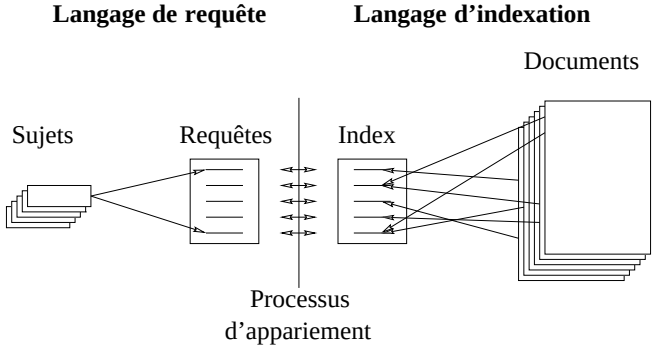

FIG. 5 – *Schéma général de la recherche d'information*

Une extension de ce schéma permet d'effectuer de la *recherche d'information interlangue* : le sujet de recherche est formulé dans une langue (par exemple, français) différente de celle des documents (par exemple, anglais). Dans ce cas, le système de RI inclut une étape de traduction du sujet en une requête dans la langue cible. Les documents trouvés peuvent en retour être également traduits dans la langue source.

3.1.1 Simplification de documents

Avant d'être traités, la requête comme les documents sont « simplifiés ». Cette simplification vise à rendre plus pertinent et plus efficace le processus d'appariement entre requête et index. Elle s'effectue selon les étapes suivantes :

1. Suppression des « mots stop » (mots grammaticaux, mots fréquents, mots sans pouvoir discriminateur...) ;
2. racinisation (*stemming*) (section 4.1.2, réduction des mots de la même famille morphologique à une racine commune),
3. transformation du texte en un sac (ou un ensemble) de mots,
4. amalgame (*conflation*) des mots synonymes.

Dans un modèle d'appariement sur la base de mots communs, tel que le modèle *vectoriel* (voir ci-dessous, section 3.1.3), la suppression des mots

fréquemment partagés tend à éloigner et mieux séparer les documents considérés comme différents. À l'inverse, le regroupement des mots synonymes ou cooccurrents tend à rapprocher les documents semblables.

3.1.2 Indexations

L'indexation peut se faire sur mots simples ou sur syntagmes. Dans ce dernier cas, des groupes de mots constituent des index du document. Ces syntagmes peuvent être obtenus par des techniques symboliques : par étiquetage et filtrage sur la base de patrons syntaxiques (spécification d'une structure syntaxique plus ou moins précise) ou par analyse syntaxique de surface (section 4.2.2). Ils peuvent aussi être obtenus par des techniques statistiques, en étudiant les mots cooccurrents dans des documents ou dans des fenêtres ; ou grâce à des patrons appris sur corpus et combinant des informations syntaxiques et lexicales (par exemple, pour les *entités nommées* des campagnes MUC, voir la section 4.2.4).

3.1.3 Traitement et appariement des requêtes

On retrouve les mêmes traitements que sur les documents. Cependant, en raison de leur petite taille, les requêtes peuvent être analysées par des procédures plus lentes et complexes que celles traitant les documents ; et en raison de leur syntaxe pauvre, les requêtes sont analysées par des procédures symboliques aux contraintes syntaxiques lâches.

Une fois traitées, les requêtes sont appariées avec l'index des documents. Le *modèle booléen* suit une approche du type base de données : les documents sont recherchés sur la base d'une formule logique sur les descripteurs, et les réponses sont de la forme Oui/Non. C'est le modèle classique en recherche bibliographique, où l'on interroge sur le contenu des champs Auteur, Titre, etc. Dans le *modèle vectoriel*, plus un document partage des descripteurs avec la requête, meilleur il est. Les documents sont présentés par ordre décroissant de proximité avec la requête : les réponses sont qualifiées par un pourcentage exprimant leur pertinence. Le *modèle probabiliste* effectue un apprentissage sur les documents : il complète le modèle vectoriel en calculant la pertinence de chaque index pour un document en fonction de documents répondant à des requêtes sur une base documentaire comparable. Ici aussi, un pourcentage quantifie la pertinence des réponses.

3.2 Autour de la recherche d'information

Nous mentionnons ici diverses extensions du modèle de recherche d'information présenté ci-dessus.

L'*expansion de requête* consiste à ajouter des mots à une requête pour lui donner de meilleures chances de ramener les documents pertinents pour le sujet de recherche formulé. Ces mots peuvent être des mots sémantiquement liés à ceux de la requête, des mots trouvés dans des documents pertinents, des mots définis explicitement par l'utilisateur, etc.

Des modèles de recherche d'information plus sophistiqués commencent à être proposés. On peut citer la recherche de « passages » (segments de textes plutôt que documents entiers), les systèmes de question-réponse (section 5.2) qui doivent fournir une information encore plus spécifique (par exemple, *What was the monetary value of the Nobel Peace Prize in 1989?*), le résumé automatique (section 5.1).

Le développement de bases documentaires non exclusivement textuelles demande de mettre en place de nouvelles approches prenant en compte la structure des documents composites et des modes de recherche dans des modalités non textuelles : documents semi-structurés (HTML, XML), bases de données, documents multimédia, etc. (voir les autres chapitres de ce livre).

L'accès à l'information s'effectue dans plusieurs modes. À la recherche documentaire par requête, on peut opposer l'envoi automatique d'informations sur la base d'un profil, sorte de requête personnelle permanente (« push »). La veille technologique, scientifique et autre « intelligence économique » consiste également à surveiller l'apparition de nouveaux documents correspondant à un sujet d'intérêt donné.

Plusieurs tâches se sont également greffées autour de la recherche d'information. La *segmentation thématique* consiste à découper un document en passages traitant d'un seul thème. L'*identification de la langue* s'appuie sur des méthodes statistiques de surface pour assigner une langue à un document (des techniques différentes doivent être utilisées pour les documents multilingues). La *classification et la catégorisation* consistent à regrouper des documents similaires pour faciliter la navigation dans la base documentaire, voire pour permettre le reclassement des documents. Le cas de la catégorisation inclut l'ajout d'étiquettes aux documents.

3.3 Extraction d'information

Alors que la recherche d'information recherche des documents (ou des « passages » de documents), l'*extraction d'information* vise à extraire des informations spécifiques et structurées d'un texte sur un domaine particulier. Elle a été popularisée par une série de compétitions organisées entre systèmes d'analyse de textes par l'agence américaine DARPA, les conférences MUC (*Message Understanding Conferences*), qui ont également promu une évaluation rigoureuse des systèmes de traitement automatique des langues. Par exemple, dans le domaine des fusions entre entreprises [F.3.1], on cherche à identifier le nom des entreprises qui fusionnent, le nom de l'entreprise créée, le type de produit qu'elle commercialise, etc. La figure 3 (section 2.3) montre une représentation de l'évolution des postes d'une personne dans une ou plusieurs entreprises.

Cette tâche nécessite de savoir effectuer différentes sous-tâches, dont certaines ont également fait l'objet d'évaluations pour elles-mêmes : la reconnaissance des *entités nommées* (noms de personnes ou d'organisations, mesures, noms de lieux... – voir la section 4.2.4), le repérage des liens d'anaphore (section 4.3.2), la reconnaissance des abréviations et acronymes.

4 Techniques de TAL pour la recherche d'information

Nous examinons plus particulièrement les techniques suivantes, qui ont un impact actuel ou attendu sur la recherche d'information. Au palier morphologique, la segmentation en unités linguistiques (section 4.1.1) et la « racinisation » (section 4.1.2). Au palier syntaxique, l'étiquetage (ou désambiguïsation) syntaxique (section 4.2.1), l'analyse syntaxique « surfacique » (section 4.2.2), l'indexation sur des syntagmes (section 4.2.3) et la reconnaissance d'entités nommées (section 4.2.4). Aux paliers sémantique et pragmatique, l'étiquetage (ou désambiguïsation) sémantique (section 4.3.1) et la résolution d'anaphores (section 4.3.2). Enfin, deux techniques transversales : la statistique textuelle (section 4.4.1) ainsi que la traduction automatique et la recherche d'information interlangue (section 4.4.2).

4.1 Palier morphologique

4.1.1 Segmentation en unités linguistiques

Nous laisserons de côté la segmentation en paragraphes, en la considérant comme essentiellement déterminée par des critères typodispositionnels (ligne sautée), et nous nous concentrerons sur la segmentation en phrases et en mots.

Le découpage d'un texte en phrases se fait selon les « ponctuations fortes » *« . ! ? »* (augmentées éventuellement du point-virgule et des deux-points). Le problème essentiel est celui de l'ambiguïté du point, qui s'utilise également pour marquer une abréviation (et aussi, en anglais, dans les nombres décimaux) : *« Le voyage aux U.S.A. de J. M. G. Le Clézio. »*. Pour réduire cette ambiguïté, on observera qu'une phrase commence par une majuscule ; le fait qu'une phrase puisse commencer par un nombre (*« 1998 a été une bonne année »*) doit aussi être pris en compte. Enfin, il faut également gérer correctement les incises qui peuvent elles-mêmes constituer des phrases ((...) « ... »). Ces contraintes peuvent s'exprimer à l'aide d'automates à états finis.

La difficulté de la segmentation en mots vient du fait que les unités élémentaires (« tokens ») que l'on peut reconnaître avec sûreté ne correspondent pas toujours aux mots. Une méthode progressive consiste à segmenter dans un premier temps de façon excessivement fine (par exemple, jusqu'à 7 unités dans *« c'est-à-dire »*). Les mots contractés peuvent être eux aussi décomposés (*« du »* $\mapsto$ *« de le »*, *« des »* $\mapsto$ *« de les »*, etc.). Dans un second temps, on recompose les unités ainsi obtenues pour identifier des mots. On se fonde pour cela sur le contenu du lexique ou sur des modèles de mots (automates à états finis). Ainsi, la séquence *« c ' est - à - dire »* pourra être identifiée comme un mot, ainsi que *« pomme de terre »*. Cependant, certaines de ces recompositions peuvent être ambiguës : par exemple, dans *« pomme de terre cuite »*, a-t-on affaire à de la *« terre cuite »* ? Ces ambiguïtés sont éventuellement propagées aux étapes suivantes de l'analyse.

En recherche d'information, la segmentation en mots est l'étape de base de l'indexation. La pertinence des unités choisies pour l'indexation influence directement la pertinence des résultats de la recherche. Une segmentation en phrases est utile pour les systèmes de résumé par extraction de phrases (section 5.1) et pour les systèmes de question-réponse (section 5.2), dans lesquels les réponses fournies sont des phrases.

4.1.2 Racinisation

La racinisation est une procédure plus ou moins linguistiquement fondée qui vise à regrouper les mots sémantiquement proches à partir de ressemblances « graphiques » (mots de forme apparentée). Sont généralement regroupés les mots d'un même *paradigme flexionnel* (par exemple les formes conjuguées d'un verbe avec son infinitif), et les mots d'une même *famille dérivationnelle* (par exemple un adjectif avec le substantif associé, comme *« lent »*/*« lenteur »*). En recherche d'information, la racinisation des documents et des requêtes vise à améliorer le rappel. La difficulté de l'opération provient de la complexité et de l'irrégularité plus ou moins grande du système morphologique de la langue étudiée.

La racinisation peut se faire par approximation des phénomènes linguistiques en jeu, ou en recherchant une fidélité linguistique plus grande. Dans la première classe de méthodes figurent les deux algorithmes classiquement utilisés en recherche d'information. Ces algorithmes ont principalement deux fonctions :

désuffixer : supprimer les suffixes qui différencient les flexions d'un mot (par exemple les formes conjuguées d'un verbe) et les mots d'une même famille morphologique (par exemple un verbe comme *« lacer »* et la forme nominale associée comme *« laçage »*),

recoder : regrouper les différentes variantes graphiques d'une même racine (ses allomorphes) comme *« condui-re »* et *« conduct-eur »*.

L'algorithme de Lovins [D.1.1] effectue séparément *désuffixage* puis *recodage*, et l'algorithme de Porter [D.1.2] effectue simultanément ces deux opérations. Ces algorithmes ne sont pas exempts d'erreurs, mais donnent des résultats satisfaisants en RI pour l'anglais.

Dans la seconde classe de méthodes se trouve la racinisation par règles et exceptions [D.2.8]. Chaque suffixe productif est traité par une règle (par exemple, *« -èrent »* marque les verbes du 3ème groupe au passé simple). Les formes pour lesquelles la règle ne s'applique pas (par exemple, *« légifèrent »*) ou celles, très rares, qui sont ambiguës (*« lac-èrent »* / *« lacèr-ent »*), sont listées comme des exceptions à la règle. Comme pour les algorithmes de racinisation sur l'anglais, les *allomorphes* (mots dont les formes fléchies sont bâties sur plusieurs racines) sont réduits à une racine unique (*« cèd- »* $\mapsto$ *« céd- »*). Cette approche permet de refléter très fidèlement les propriétés morphologiques du français.

4.2 Palier syntaxique

4.2.1 Étiquetage ou désambiguïsation syntaxique

L'*étiquetage syntaxique*, ou désambiguïsation syntaxique, vise à associer à chaque mot, en contexte, une « étiquette » syntaxique. Cette étiquette indique la catégorie syntaxique et éventuellement les traits morphosyntaxiques du mot. Par exemple,

$$\text{« } La_{/DETfs}\ coronarographie_{/Nfs}\ est_{/V3spi}\ normale_{/Afs}.\ \text{»}$$

L'étiquetage syntaxique est une étape intermédiaire de nombre de systèmes d'analyse surfacique ou partielle (voir plus bas), c'est pourquoi nous le présentons ici.

La plupart des méthodes cherchent à obtenir cet étiquetage en examinant le contexte immédiat du mot à étiqueter (quelques mots à gauche et à droite). Les méthodes *à base de règles* appliquent aux mots ambigus des règles de désambiguïsation, qui (selon la méthode) interdisent ou autorisent sélectivement certaines séquences d'étiquettes [E.1.3, D.2.10]. Les méthodes probabilistes apprennent des modèles de Markov cachés sur des corpus préalablement étiquetés [E.1.8]. La méthode de Brill [E.1.1, E.1.2] apprend sur un corpus étiqueté des règles de correction d'erreurs d'étiquetage. Enfin, l'application d'un véritable analyseur syntaxique sur une phrase a pour effet de bord de désambiguïser les mots de la phrase [E.2.11]. Ce dernier type de méthode n'est réellement utile dans le contexte de l'étiquetage que si l'analyse syntaxique appliquée n'a pas une complexité trop grande.

Le choix des étiquettes, et en particulier leur finesse, conditionne les performances des étiqueteurs, qui atteignent 90–98 % de mots bien étiquetés selon le jeu de catégories, le corpus, etc. La taille limitée du contexte examiné pour effectuer la désambiguïsation place une limite théorique sur la précision de l'étiquetage effectué [E.1.7]. Par ailleurs, la plupart des mots peuvent changer de catégorie syntaxique (*conversion* d'un adjectif en nom, etc.) ; de ce fait, il est difficile de supposer que toutes les catégories syntaxiques possibles d'un mot se trouvent dans le lexique utilisé.

4.2.2 Analyse « peu profonde », ou « surfacique »

Depuis le milieu des années 1980, le modèle d'analyse syntaxique dominant, fondé sur l'emploi de formalismes grammaticaux élaborés et d'analyseurs mettant en œuvre ces formalismes, a été sérieusement concurrencé

dans l'analyse de grands documents par des méthodes d'analyse simplifiées. Ces méthodes, au moins en première intention, visent des analyses moins « profondes » ou moins complètes que les précédentes.

L'*analyse partielle* ne cherche pas à traiter l'ensemble d'une phrase, mais seulement à analyser certains segments utiles et potentiellement plus faciles à reconnaître (syntagmes nominaux, syntagmes non récursifs et autres « chunks » [E.2.1]). Une méthode souvent employée est l'identification de patrons syntaxiques (typiquement, automates à états finis) dans des textes préalablement étiquetés (voir section 4.2.1).

Une stratégie d'*analyse robuste* fait en sorte de toujours donner un résultat, même incomplet, pour l'analyse d'une phrase. Les analyseurs « classiques » peuvent en général se replier sur une analyse partielle lorsqu'une analyse complète n'est pas obtenue (par exemple, avec les méthodes « tabulaires »). Les analyseurs qui identifient progressivement des segments de phrases « sûrs » (syntagmes non récursifs puis éventuellement syntagmes plus complexes) et les relations entre ces segments sont par nature robustes.

Enfin, l'identification de *cooccurrences* (statistiques), obtenues en recherchant des mots se retrouvant fréquemment conjointement dans une fenêtre, un paragraphe, un document, peut constituer un substitut de l'analyse syntaxique pour détecter des syntagmes élémentaires.

4.2.3 Indexation sur les syntagmes et variation

Une fois que l'on a identifié des syntagmes, on peut s'en servir pour indexer les documents dans lesquels ils apparaissent. L'*indexation sur les syntagmes* (« phrase indexing ») a pour but d'augmenter la précision des index en diminuant leur ambiguïté. L'identification des coocurrences est utilisée en RI pour faire de l'indexation sur des groupes de mots sans avoir recours à des techniques symboliques de TAL plus coûteuses à mettre en œuvre. En concurrence, on trouve des techniques d'analyse robuste et superficielle en TAL appliquées à l'indexation pour la RI [C.1.2, C.2.6, C.3.1, C.3.2]. Ces techniques doivent être capables de de regrouper les variantes d'un syntagme de base qui peut être modifié ou *transformé* pour produire des syntagmes de sens proche. Il est utile de savoir reconnaître ces variations pour pouvoir apparier une requête qui contient l'une des formes avec un document qui en contient une variante [G.3.4]. Par exemple, à partir du syntagme de base *« diffusion de la lumière »*, on repérera *« diffusions de la lumière »*, *« diffusion dépolarisée de la lumière »*, *« diffusion de - lumière »*, *« diffuse une*

lumière » et *« émission de lumière »*. Ces variantes peuvent être obtenues par génération dynamique de patrons de variantes (par exemple, à l'aide de méta-règles) ou par simplification des structures syntaxiques des termes observés. Parmi les enjeux de la reconnaissance de variantes, on peut citer la difficulté à couvrir exactement les variantes pertinentes et le coût computationnel de la production contrôlée de ces variantes.

4.2.4 Reconnaissance des entités nommées

La notion d'*entités nommées*, introduites dans le cadre de l'extraction d'information (section 3.3), se réfère à des concepts uniques et partagés. Les entités nommées comprennent les organisations (entreprises, administrations, musées, etc.), les lieux (villes, régions, fleuves, etc.), les personnes (hommes politiques, vedettes, chefs d'entreprise, inconnus, etc.) et les numériques (poids, longueurs, valeurs monétaires, pourcentages, etc.). Les entités nommées peuvent constituer des index très discriminants, et sont souvent des informations demandées. Par exemple, plusieurs entités nommées sont en jeu pour répondre à la question *« Quel était le nom du PDG de Peugeot en 1987 ? »*.

La reconnaissance des entités nommées s'appuie sur des méthodes symboliques et numériques. Le premier type de méthode repose sur des dictionnaires (de nombreuses listes d'entités nommées sont accessibles en ligne : noms de lieux, annuaires divers, etc.) et des patrons syntaxiques. Ceux-ci sont appliqués sur des textes préalablement étiquetés (section 4.2.1) et peuvent utiliser des repères lexicaux internes (par exemple, unités pour les mesures) ou externes (par exemple, titres honorifiques pour les personnes) [E.3.2, E.3.4]. Le second type de méthode effectue un apprentissage de contextes et de structures, par exemple avec des modèles à apprentissage statistique comme les modèles de Markov cachés [E.3.1, E.3.3].

On notera que l'exhaustivité des listes n'est pas l'enjeu, et que les modèles à apprentissage font mieux que les modèles symboliques. Enfin, la variation intervient également dans l'expression des entités nommées. Abréviations et acronymes (*« MoMA » = « Museum of Modern Art »*), anaphores (section 4.3.2 : pronoms, reprises partielles), variantes graphiques (*« ATT » = « A T T » = « AT&T » = « A T and T »*) et linguistiques (*« Ieltsine » = « Yeltsine » = « Eltsine » = « Ieltsin » = « Yeltsin »*), métaphores (*« IBM » = « Big Blue »*, *« Premier Ministre » = « Lionel Jospin » = « Matignon »*) sont autant de sources de variation qui complexifient la tâche de reconnaissance de ces

entités.

4.3 Paliers sémantique et pragmatique

4.3.1 Étiquetage sémantique

De même que l'étiquetage syntaxique (section 4.2.1) vise à associer à chaque mot une étiquette syntaxique, l'étiquetage sémantique cherche à associer à chaque mot, en contexte, une étiquette sémantique. Cette étiquette sémantique peut être une catégorie sémantique générale (par exemple, animé, événement, mouvement, etc.) ou un sens de mot (par exemple, *« artère »*–vaisseau sanguin *vs « artère »*–avenue). Pour une partie des mots, la désambiguïsation syntaxique peut aider : on en sait davantage sur le sens de *« livre »* si l'on connaît son genre (*« un livre$_{/Nms}$ » vs « une livre$_{/Nfs}$ »*). Par ailleurs, des méthodes similaires à celles employées en syntaxe sont applicables (chaînes de Markov, etc.). Encore plus que pour les travaux en étiquetage syntaxique, le choix des étiquettes a une influence fondamentale sur la nature de la tâche. Les travaux en désambiguïsation sémantique sont relativement récents, mais possèdent une forte dynamique.

4.3.2 Résolution d'anaphores

La résolution d'anaphores consiste à relier entre elles les références à une même entité au sein d'un texte. On distingue plusieurs types d'anaphore. L'*anaphore pronominale* emploie un pronom pour faire référence à une expression antérieure : *« Le dispositif expérimental d'amélioration de l'hybride est rappelé. Il consiste principalement en des tests. »*. L'*anaphore par reprise partielle* reprend une partie de l'expression antérieure, comme dans *« ...sont réalisés grâce à une nouvelle technique d'immobilisation d'enzyme sur électrode de verre. La nouveauté de cette technique réside dans le dépôt d'un agent... »*. L'*anaphore par lien sémantique* ne reprend pas directement un mot de l'antécédent, mais un terme sémantiquement lié (ici, plus générique) : *« La sonde thermique INRA est une résistance de platine... Ce capteur peut ainsi servir à rendre compte du phénomène... »*. De façon générale, la plupart des expressions nominales (syntagme nominal défini, pronom) sont potentiellement des anaphores et potentiellement des antécédents d'anaphores. La résolution d'anaphores requiert des informations aussi bien syntaxiques (genre, nombre) que sémantiques (relation

d'hyperonymie, etc.) et s'appuie sur des considérations pragmatiques (entités les plus saillantes au fil du texte, ou « focus ») [F.5.3, F.5.2, F.5.1, F.5.4].

La résolution d'anaphores est une technique dont l'apport est important dans de nombreuses applications. En extraction d'information, elle permet de garnir une structure d'information avec la référence initiale complète à une entité. Pour répondre à une question (section 5.2), elle permet de donner une référence complète dans la réponse. Dans le résumé automatique (section 5.1), elle permet de rendre cohérentes des phrases issues de segments épars. En traduction ou en compréhension, elle permet de choisir la traduction ou le sens correct d'une anaphore ambiguë.

4.4 Techniques transversales

4.4.1 Statistiques textuelles

Différentes mesures d'indices textuels (mots, chaînes, catégories, patrons sytaxiques, ponctuation, taille des phrases, etc.) sont utiles dans diverses tâches liées à la recherche d'information : citons le typage du corpus, l'identification de la langue, l'ajustement des méthodes d'analyse, la catégorisation, la segmentation thématique, le filtrage et le routage de l'information.

4.4.2 Traduction automatique et recherche d'information interlangue

La traduction automatique, jugée comme un enjeu majeur et accessible dans les années cinquante, est désormais considérée comme une tâche extrêmement complexe. Des sous-produits de la traduction automatique rendent cependant des services appréciables autour de la recherche d'information : les mémoires de traduction, l'alignement de textes bilingues, la recherche d'information interlangue, la constitution de données lexicales multilingues et la traduction par l'exemple.

Un point clé dans la *recherche d'information interlangue*, technique introduite à la section 3.1, est la traduction d'une requête dans la langue cible de la recherche. Cette tâche fournit un exemple prototypique de la façon dont un problème (la traduction de termes complexes) peut être reconsidéré sous un angle différent dans le contexte de la recherche d'information.

Un terme étant donné (par exemple, *« groupe de travail »*), il est bien connu qu'une traduction assemblée mot à mot à l'aide d'un dictionnaire bilingue (*« groupe »* $\mapsto$ *« cluster »*, *« group »*, *« collective »*, ... ; *« travail »* $\mapsto$ *« work »*, *« labour »*, ...) a toutes les chances d'être incorrecte, voire

incongrue (*« work cluster »*, *« labour cluster »*, *« work collective »*, etc.). Cependant, si l'on examine le nombre d'occurrences parmi les documents (du Web par exemple) des termes produits, on peut généralement identifier la ou les traductions correctes. Par exemple, la recherche sur AltaVista d'une chaîne fixe de mots et le relevé des nombres de documents contentant cette chaîne permet de déduire la ou les traductions correctes d'un terme complexe [G.4.2]. Ainsi, les comptages suivants indiquent *« work group »* comme traduction préférentielle de *« groupe de travail »* : (*« work cluster »* : 112, *« labour cluster »* : 0, *« work collective »* 242, *« work group »* : 67 238, etc.).

5 DEUX EXEMPLES D'APPLICATIONS

Nous concluons ce document par l'exposé de deux applications situées à la croisée de la recherche d'information et du traitement automatique des langues : le résumé automatique (section 5.1) et les systèmes de question-réponse (section 5.2).

5.1 Le résumé automatique

Le résumé automatique consiste à partir d'un texte et à produire un texte plus court qui donne les informations principales contenues dans le texte d'origine ou permet tout au moins de s'en faire une idée. Dans le cadre de la recherche d'information, le résumé automatique trouve naturellement sa place lors de la présentation des résultats d'une requête. Au-delà de la lecture des simples titres des documents, un bon résumé, à un facteur de réduction approprié, devrait donner une idée suffisante du contenu d'un document long pour permettre de décider si le document lui-même vaut la peine d'être consulté.

Deux grands types de méthodes ont été employées pour effectuer cette tâche. La première est issue des travaux de l'intelligence artificielle, et pourrait être qualifiée de « méthode du bon élève ». Elle consiste, fort logiquement, (i) à « comprendre » le texte source, c'est-à-dire à constuire une représentation de son sens ; puis (ii) à extraire les informations essentielles de cette représentation ; et enfin, (iii) à générer à partir de cette représentation un texte en langue naturelle.

Cette méthode est confrontée à l'ensemble des difficultés du TAL, en particulier la construction d'une représentation du sens, avec un besoin de connaissances importantes sur le domaine (pour la compréhension et pour l'ex-

traction des informations essentielles) et sur la langue (particulièrement nécessaires en génération de textes). Elle peut être praticable sur des microdomaines, mais son passage à l'échelle est difficile.

La seconde méthode consiste à faire ce qu'un professeur de français pourrait appeler la *méthode du mauvais élève* : extraire du texte initial les phrases les plus importantes et les mettre bout à bout pour construire le résumé. Cette méthode, dite par *extraction de phrases*, a fait l'objet de nombreux développements ces dernières années. Son intérêt est d'être applicable à des documents variés avec des résultats somme toute assez acceptables.

Une segmentation correcte en phrases (section 4.1.1) est bien sûr un prérequis pour faire du résumé par extraction de phrases. Les phrases importantes peuvent ensuire être repérées par différentes méthodes, dont plusieurs font appel à des techniques de traitement automatique des langues : par leur position (début de paragraphe, etc.), par la fréquence relative des mots qu'elles contiennent, par comparaison à des patrons de phrases (section 4.2.2), par des chaînes de référence lexicale (anaphores etc.), ou encore par la présence de connecteurs rhétoriques. Le texte reconstruit risque cependant de violer les critères pragmatiques de bonne formation (section 2.4) : cohésion et cohérence. Par exemple, une phrase du texte source qui contient une anaphore (section 4.3.2) ne peut pas être copiée sans un traitement de cette anaphore : remplacement par son antécédent, ou ajout de la phrase contenant cet antécédent.

5.2 Les systèmes de question-réponse

L'objectif d'un système de question-réponse est de fournir une réponse précise à une question posée. Il s'agit d'une tâche plus fine et plus exigeante que la recherche d'information, dans la mesure où il ne s'agit plus de fournir des documents entiers mais des informations spécifiques. Il existe actuellement une demande très forte sur la Toile pour des systèmes réalisant ce type de tâche.

Le processus de traitement d'une question est le suivant [H.0.2] :

1. Les questions sont transformées en requêtes, des sacs de mots (lemmatisés ou non). Un moteur de recherche classique (vectoriel ou booléen) associe à chaque requête un ensemble de documents.

2. Les documents filtrés sont ensuite passés dans un moteur de recherche plus fin, reposant sur des indices linguistiques plus fins donc plus lents

à extraire. On applique une technique de seuillage en fonction de la similitude entre document et question.

3. Au moyen d'un appariement entre les phrases contenues dans les documents et les questions, des *fragments réponses* sont extraits des documents.

Les deux dernières étapes reposent sur des indexations *dépendantes de l'ensemble des questions* alors que l'indexation du premier moteur de recherche (beaucoup plus simple) est indépendante des questions.

Plusieurs pistes sont suivies pour améliorer un tel système. La collecte de données telles que des noms de mesures, des types de questions (pronoms interrogatifs et structures), des entités nommées (section 4.2.4), etc., facilite l'analyse des documents et des questions. Une deuxième piste est l'adaptation ou le développement d'outils tels que des étiqueteurs (section 4.2.1, en les paramétrant pour l'analyse des questions, par exemple), des indexeurs (en les spécialisant pour les entités nommées – section 4.2.4), des analyseurs superficiels (section 4.2.2). Enfin, la combinaison des différents outils doit être affinée : la bonne harmonie entre les différentes composantes est essentielle pour la réussite du projet !

6 CONCLUSION

L'accès au contenu des documents textuels est le domaine de la recherche d'information. Les moteurs de recherche disponibles sur le Web montrent les limites des techniques classiques de la RI. Nous avons vu comment les techniques du traitement automatique des langues peuvent jouer des rôles clé dans des tâches de RI. Gageons que ce rôle sera amené à se développer, et que de même que les correcteurs orthographiques se sont banalisés dans les systèmes de traitement de texte, une panoplie d'outils de TAL va rapidement s'intégrer pour renforcer et agrandir la famille des systèmes de recherche d'information.

BIBLIOGRAPHIE STRUCTURÉE

A Principes et techniques de base en recherche d'information

[A.0.1] Ricardo Baeza-Yates et Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, New-York, 1999.

[A.0.2] Gerard Salton. Experiments in automatic thesaurus construction for information retrieval. In *Proceedings, Information Processing '71*, pages 115–123, Amsterdam, 1971. North Holland.

[A.0.3] Gerard Salton. *Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

[A.0.4] Gerard Salton et Michael E. Lesk. Computer evaluation og indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36, 1968.

[A.0.5] Gerard Salton et Michael E. Lesk. Information analysis and dictionary construction. In Gerard Salton, éditeur, *The Smart Retrieval System — Experiments in Automatic Document Processing*, pages 115—142. Prentice Hall Inc., Engelwood Cliffs, NJ, 1971.

[A.0.6] Gerard Salton et Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, New York, NY, 1983.

[A.0.7] Gerard Salton, C. S. Yang et C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[A.0.8] Karen Sparck Jones. *Automatic Keyword Classification for Information Retrieval*. Butterworth, London, 1971.

[A.0.9] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth, London, 1975.

B Recherche d'information interlangue

[B.0.1] Christian Fluhr, Dominique Schmit, Philippe Ortet, Faza Elkateb, Karine Gurtner et Khaled Radwan. Distributed Cross-lingual Information Retrieval. In Gregory Grefenstette, éditeur, *Cross-Language*

Information Retrieval, pages 41–50. Kluwer Academic Publisher, Boston, MA, 1998.

[B.0.2] Julio Gonzalo, Felisa Verdejo, Irina Chugur et Juan Cigarrán. Using EuroWordNet in a concept-based approach to Cross-Language Text Retrieval. In *Proceedings, COLING/ACL Workshoop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.

[B.0.3] Gregory Grefenstette, éditeur. *Cross-Language Information Retrieval*. Kluwer Academic Publisher, Boston, MA, 1998.

C Traitement automatique des langues et recherche d'information

C.1 Études pionnières en traitement automatique des langues pour la recherche d'information

[C.1.1] A. Andreewsky, Fathi Debili et Christian Fluhr. Computational learning of semantic lexical relations for the generation and automatic analysis of content. In *Proceedings, IFIP Congress*, pages 667–673, Toronto, 1977.

[C.1.2] Fathi Debili. *Analyse Syntaxico-Sémantique Fondée sur une Acquisition Automatique de Relations Lexicales-Sémantiques*. Thèse de Doctorat d'État en Sciences Informatiques, Université of Paris XI, Orsay, 1982.

[C.1.3] Martin Dillon et Ann S. Gray. FASIT : A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108, 1983.

[C.1.4] George S. Dunham. The role of syntax in the sublanguage of medical diagnostic statement. In Ralph Grishman et Richard Kittredge, éditeurs, *Analyzing Language in Restricted Domains. Sublanguage Description and Processing*, pages 175–194. Lawrence Erlbaum Ass., Hillsdale, NJ, 1986.

C.2 Traitement automatique des langues pour l'indexation automatique et la recherche d'information

[C.2.1] Peter Biebricher, Nobert Fuhr, Gerhard Lustig, Michael Schwanter et Gerhard Knorz. The automatic indexing system AIR/PHYS —

from research to application. In *Proceedings, 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'88)*, pages 333–341, 1988.

[C.2.2] Kuang-Hua Chen et Hsin-Hsi Chen. Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation. In *Proceedings, 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, pages 234–241, Las Cruces, NM, 1994.

[C.2.3] George S. Dunham, Milos G. Pacak et Arnold W. Pratt. Automatic indexing of pathology data. *Journal of the American Society for Information Science*, 29(2):81–90, 1978.

[C.2.4] David A. Evans, Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts et Ira A. Monarch. Automatic indexing using selective NLP and first-order thesauri. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'91)*, pages 624–643, Barcelona, 1991.

[C.2.5] David A. Evans et Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings, 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 17–24, Santa Cruz, CA, 1996.

[C.2.6] Joel L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic Methods*. PhD Thesis in Philosophy, Cornell University, 1987.

[C.2.7] Christian Jacquemin. Traitement automatique des langues pour la recherche d'information. *Traitement Automatique des Langues*, 41(2), 2000.

[C.2.8] David D. Lewis, W. Bruce Croft et Nehru Bhandaru. Language-oriented information retrieval. *International Journal of Intelligent Systems*, 4:285–318, 1989.

[C.2.9] Michael L. Mauldin. *Conceptual Information Retrieval : A Case Study in Adaptative Partial Parsing*. Kluwer Academic Publisher, Boston, MA, 1991.

[C.2.10] Renée Pohlmann et Wessel Kraaij. The effect of syntactic phrase indexing on retrieval performances for Dutch texts. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'97)*, pages 176–183, Montreal, 1997.

[C.2.11] Christoph Schwarz. The TINA Project: text content analysis at the Corporate Research Laboratories at Siemens. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'88)*, pages 361–368, Cambridge, MA, 1988.

[C.2.12] Paraic Sheridan et Alan F Smeaton. The application of morpho-syntactic language processing to effective phrase matching. *Information Processing & Management*, 28(3):349–369, 1992.

C.3 Études récentes en traitement automatique des langues pour la recherche d'information

[C.3.1] A. T. Arampatzis, C. H. A. Koster et T. Tsoris. IRENA: Information retrieval engine based on natural language analysis. In *Proceedings, Intelligent Multimedia Information Retrieval Systems and Management (RIAO'97)*, pages 159–175, Montreal, 1997.

[C.3.2] A. T. Arampatzis, T. Tsoris, C. H. A. Koster et Th. P. van der Weide. Phrase-based information retrieval. *Information Processing & Management*, 34(6):693–707, 1998.

[C.3.3] Tomek Strzalkowski. Natural language information retrieval. *Information Processing & Management*, 31(3):397–417, 1995.

[C.3.4] Tomek Strzalkowski et Peter G. N. Scheyen. Evaluation of the Tagged Text Parser. In Harald Bunt et Masaru Tomita, éditeurs, *Recent Advances in Parsing Technology*, pages 201–220. Kluwer Academic Publisher, Boston, MA, 1996.

D Morphologie

[D.0.1] Hervé-D. Béchade. *Phonétique et morphologie du français moderne et contemporain*. Fondamental. PUF, Paris, 1992.

D.1 Deux algorithmes de racinisation classiques pour l'anglais

[D.1.1] Judith Beth Lovins. Development of a stemming algorithm. *Translation and Computational Linguistics*, 11(1):22–31, 1968.

[D.1.2] M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

D.2 Analyse morphologique à grande échelle

[D.2.1] Peter Anick et Suzanne Artemieff. A high-level morphological language exploiting inflectional paradigms. In *Proceedings, 14th International Conference on Computational Linguistics (COLING'92)*, pages 67–73, Nantes, 1992.

[D.2.2] Roy J. Byrd et Évelyne Tzoukermann. Adapting an English morphological analyzer for French. In *Proceedings, 26th Annual Meeting of the Association for Computational Linguistics (ACL'88)*, pages 1–6, Buffalo, NY, 1988.

[D.2.3] Georgette Dal, Nabil Hathout et Fiammetta Namer. Construire un lexique dérivationnel: théorie et réalisations. In *Proceedings, Conférence de Traitement Automatique du Langage Naturel (TALN'99)*, Cargèse, 1999.

[D.2.4] Natalia Grabar et Pierre Zweigenbaum. Automatic acquisition of domain-specific morphological resources from thesauri. In *Actes de RIAO 2000 : Accès à l'Information Multimédia par le Contenu*, pages 765–784, Paris, France, avril 2000. C.I.D.

[D.2.5] Christian Jacquemin. Guessing morphology from terms and corpora. In *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 156–167, Philadelphia, PA, 1997.

[D.2.6] Kimmo Koskenniemi. *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production*. PhD dissertation, University of Helsinki, Helsinki, 1983.

[D.2.7] Robert Krovetz. Viewing morphology as an inference process. In *Proceedings, 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 191–203, Pittsburg, PA, 1993.

[D.2.8] Fiametta Namer. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, 41(2), 2000.

[D.2.9] Jacques Savoy. Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science*, 44(1):1–9, 1993.

[D.2.10] Max Silberztein. *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*. Masson, Paris, 1993.

[D.2.11] Richard Sproat. *Morphology and Computation*. ACL-MIT Press Series in NLP. MIT Press, Cambridge, MA, 1992.

[D.2.12] Évelyne Tzoukermann et Mark Liberman. A finite-state processor for Spanish. In *Proceedings, 13th International Conference on Computational Linguistics (COLING'90)*, Helsinki, 1990.

[D.2.13] Jing Xu et W. Bruce Croft. Corpus-based stemming using cooccurrence of word variants. *ACM Transaction on Information Systems*, 16(1):61–81, 1998.

E Syntaxe

E.1 Étiqueteurs syntaxiques

[E.1.1] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings, 3rd Conference on Applied Natural Language Processing (ANLP'92)*, pages 152–155, Trento, 1992.

[E.1.2] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[E.1.3] Jean-Pierre Chanod et Pasi Tapanainen. Statistical and constraint-based taggers for French. In *Proceedings of the 7 th EACL*, Dublin, Ireland, 1995.

[E.1.4] Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings, 2nd Conference on Applied Natural Language Processing (ANLP'88)*, pages 136–143, Austin, TE, 1988.

[E.1.5] Patrick Paroubek. GRACE : Grammaires et ressources pour les analyseurs de corpus et leur évaluation. page WWW `http://m17.limsi.fr/TLP/grace/`, LIMSI, 1998. consultation 6/11/1998.

[E.1.6] Max Silberztein. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson, Paris, 1993.

[E.1.7] Jacques Vergne et Emmanuel Giguet. Regards théoriques sur le «tagging». In Pierre Zweigenbaum, éditeur, *Actes de TALN 1998*, pages 22–31, Paris, juin 1998.

[E.1.8] Ralph Weischedel, Marie Meeter, Richard Schwartz, Lance Ramshaw et Jeff Palmucci. Coping with ambiguity and unknown

words through probabilistic models. *Computational Linguistics*, 19(2):359–382, 1993. Special Issue on Using Large Corpora: II.

E.2 Analyse superficielle de documents

[E.2.1] Steven P. Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney et Carol Tenny, éditeurs, *Principle-Based Parsing: computation and psycholinguistics*, pages 257–278. Kluwer Academic Publisher, Boston, MA, 1991.

[E.2.2] Salah Aït-Mokhtar et Jean-Pierre Chanod. Incremental finite-state parsing. In *ANLP97*, pages 72–79, Washington, DC, 1997.

[E.2.3] Donald Hindle. Deterministic parsing of syntactic non-fluencies. In *Proceedings, 21st Annual Meeting of the Association for Computational Linguistics (ACL'83)*, pages 123–128, Cambridge, MA, 1983.

[E.2.4] Lynette Hirschman, Ralph Grishman et Naomi Sager. Grammatically-based automatic word class formation. *Information Processing & Management*, 11(1-2):39–57, 1975.

[E.2.5] Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel et Mabry Tyson. FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Emmanuel Roche et Yves Schabes, éditeurs, *Finite-State Language Processing*, pages 383–406. MIT Press, Cambridge, MA, 1997.

[E.2.6] Fidelia Ibekwe-SanJuan. Terminological variation, a means of identifying research topics from texts. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 564–570, Montreal, 1998.

[E.2.7] Satoru Ikehara, Satoshi Shirai et Hajime Uchino. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings, 16th International Conference on Computational Linguistics (COLING'96)*, pages 574–579, Copenhagen, 1996.

[E.2.8] Fred Karlsson. Constraint Grammar as a framework for parsing running text. In *Proceedings, 13th International Conference*

on Computational Linguistics (COLING'90), pages 168–173, Helsinki, 1990.

[E.2.9] Fred Karlsson, Atro Voutilainen, Juha Heikkilä et Arto Anttila, éditeurs. *Constraint Grammar A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, 1995.

[E.2.10] Pasi Tapanainen et Timo Järvinen. Syntactic analysis of natural language using linguistic rules and corpus-based patterns. In *Proceedings, 15th International Conference on Computational Linguistics (COLING'94)*, pages 629–634, Kyoto, 1994.

[E.2.11] Jacques Vergne. Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur – analyse syntaxique automatique non combinatoire. Mémoire d'habilitation à diriger des recherches, Université de Caen, 1999.

[E.2.12] Atro Voutilainen. Designing a (finite-state) parsing grammar. In Emmanuel Roche et Yves Schabes, éditeurs, *Finite-State Language Processing*, pages 283–310. MIT Press, Cambridge, MA, 1997.

E.3 Entités nommées

[E.3.1] Daniel M. Bikel, Scott Miller, Richard Schwartz et Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings, 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 194–201, Washington, 1997. ACL.

[E.3.2] David D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev et James Pustejovsky, éditeurs, *Corpus Processing for Lexical Acquisition*, pages 61–76. MIT Press, Cambridge (Mass.), 1993.

[E.3.3] Andrei Mikheev. Named entity recognition without gazetteers. In *Proceedings, 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, 1999. ACL.

[E.3.4] Nina Wacholder, Yael Ravin et Misook Choi. Disambiguating proper names in text. In *Proceedings, 5th Conference on Applied Natural Language Processing (ANLP'97)*, Washington, 1997. ACL.

F Sémantique et pragmatique

F.1 Sémantique lexicale

[F.1.1] D. A. Cruse. *Lexical Semantics*. Cambridge University Press, Cambridge, 1986.

[F.1.2] Martha Walton Evens, éditeur. *Relational models of the lexicon*. Cambridge University Press, Cambridge, 1988.

[F.1.3] Igor Mel'čuk. *Dictionnaire explicatif et combinatoire du français contemporain*. Presses de l'Université de Montréal, Montréal, 1984.

[F.1.4] Igor A. Mel'čuk et A. Zholkovsky. The explanatory combinatorial dictionary. In M. W. Evens, éditeur, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, pages 41–74. Cambridge University Press, Cambridge, 1988.

[F.1.5] George A. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

F.2 Désambiguïsation sémantique

[F.2.1] Roberto Basili, Michelangelo Della Rocca et Maria Teresa Pazienza. Contextual word sense tuning and disambiguation. *Applied Artificial Intelligence*, 11:235–262, 1997.

[F.2.2] Adam Kilgarriff et Martha Palmer. Special issue on SENSEVAL : Evaluating word sense disambiguation programs. *Computers and the Humanities*, 2000.

[F.2.3] Yorick Wilks, Brian M. Slator et Louise Guthrie. *Electric Words — Dictionaries, Computers, and Meanings*. MIT Press, 1996.

F.3 Analyse sémantique

[F.3.1] Defense Advanced Research Projects Agency. *Fifth Message Understanding Conference (MUC-5)*, San Francisco, Ca, 1993. Morgan Kaufmann.

[F.3.2] Defense Advanced Research Projects Agency. *MUC-6: Proceedings of the Sixth Message Understanding Conference*, Columbia, Maryland, 1996. Morgan Kaufmann. novembre 1995.

[F.3.3] Pierre Zweigenbaum et Consortium MENELAS. MENELAS: coding and information retrieval from natural language patient discharge summaries. In Maria Fernanda Laires, Maria Jůa Ladeira et Jens Pihlkjær Christensen, éditeurs, *Advances in Health Telematics*, pages 82–89. IOS Press, Amsterdam, 1995. MENELAS Final Edited Progress Report.

F.4 Formalismes de représentation des connaissance

[F.4.1] Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, Lori Alperin Resnick et Alexander Borgida. Living with CLASSIC: when and how to use a KL-ONE-Like language. In Sowa [F.4.5], chapitre 14, pages 401–456.

[F.4.2] Ronald J. Brachman et J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[F.4.3] Daniel Kayser. *La représentation des connaissances*. Hermès, 1997.

[F.4.4] John F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, London, 1984.

[F.4.5] John F. Sowa, éditeur. *Principles of Semantic Networks*. Morgan Kaufmann Publishers, San Mateo, Ca., 1991.

[F.4.6] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. PWS Publishing, Boston, 1995.

F.5 Anaphores

[F.5.1] Simon Philip Botley, éditeur. *Approaches to Discourse Anaphora: Proceedings of the DAARC Colloquium*. Lancaster University, Lancaster, 1996.

[F.5.2] Jaime G. Carbonell et Ralf D. Brown. Anaphora resolution: A multi-strategy approach. In *Proceedings, 12th International Conference on Computational Linguistics (COLING'88)*, pages 96–101, Budapest, 1988. ACL.

[F.5.3] Graeme Hirst. *Anaphora in Natural Language Understanding: A Survey*. Springer-Verlag, Berlin, 1981.

[F.5.4] R. Mitkov, L. Belguith et M. Stys. Multilingual robust anaphora resolution. In *Proceedings of the Third International Conference on*

Empirical Methods in Natural Language Processing (EMNLP-3), pages 7–16, Granada, 1998. ACL.

F.6 Pragmatique

[F.6.1] Marc Cavazza et Pierre Zweigenbaum. A semantic analyzer for natural language understanding in an expert domain. *Applied Artificial Intelligence*, 8(3):425–453, 1994.

[F.6.2] Phil N. Johnson-Laird. *Mental Models*. Cambridge University Press, Cambridge, 1983.

[F.6.3] R. C. Schank et C. K. Riesbeck, éditeurs. *Inside Computer Understanding*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1981.

[F.6.4] Roger C. Schank et Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: an Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[F.6.5] T. A. van Dijk et W. Kintsch. *Strategies of Discourse Comprehension*. Academic Press, New York, 1990.

G Terminologie et recherche d'information

G.1 Acquisition automatique de thésaurus

[G.1.1] Houssem Assadi. Knowledge acquisition from texts: Using an automatic clustering method based on noun-modifier relationship. In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97)*, pages 504–506, Madrid., 1997.

[G.1.2] Houssem Assadi et Didier Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. In *Proceedings, 10th Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, pages 505–514, Rennes, 1996. A.F.C.E.T.

[G.1.3] Carolyn J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing & Management*, 26(5):629–640, 1990.

[G.1.4] Carolyn J. Crouch et Bokyung Yang. Experiments in automatic statistical thesaurus construction. In *Proceedings, 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)*, pages 77–88, Copenhagen, 1992.

[G.1.5] Christiane Fellbaum, éditeur. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[G.1.6] Edward A. Fox, J. Terry Nutter, Thomas Ahlswede, Martha Evens et Judith Markowitz. Building a large thesaurus for information retrieval. In *Proceedings, 2nd Conference on Applied Natural Language Processing (ANLP'88)*, pages 101–108, Austin, TE, 1988.

[G.1.7] Gregory Grefenstette. Corpus derived first, second and third-order word affinities. In *Proceedings, EURALEX'94*, 1994.

[G.1.8] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, 1994.

[G.1.9] Benoît Habert, Elie Naulleau et Adeline Nazarenko. Symbolic word clustering for medium-size corpora. In *Proceedings, 16th International Conference on Computational Linguistics (COLING'96)*, pages 490–495, Copenhagen, 1996.

[G.1.10] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings, 14th International Conference on Computational Linguistics (COLING'92)*, pages 539–545, Nantes, 1992.

[G.1.11] Christian Jacquemin. A symbolic and surgical acquisition of terms through variation. In Stefan Wermter, Ellen Riloff et Gabriele Scheler, éditeurs, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438. Springer, Heidelberg, 1996.

[G.1.12] Emmanuel Morin. Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *t.a.l.*, 40(1):143–166, 1999.

[G.1.13] Emmanuel Morin et Christian Jacquemin. Projecting corpus-based semantic links on a thesaurus. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 389–396, University of Maryland, juin 1999.

[G.1.14] Gerda Ruge. Experiments on linguistically based term associations. In *Proceedings, Intelligent Multimedia Information Retrieval*

Systems and Management (RIAO'91), pages 528–545, Barcelona, 1991.

[G.1.15] Padmini Srinivasan. Thesaurus construction. In William B. Frakes et Ricardo Baeza-Yates, éditeurs, *Information Retrieval: Data Structure and Algorithms*, pages 161–218. Prentice Hall, London, 1992.

G.2 Acquisition terminologique

[G.2.1] Didier Bourigault. *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes*. Thèse en Mathématiques, Informatique Appliquée aux Sciences de l'Homme, École des Hautes Études en Sciences Sociales, Paris, 1994.

[G.2.2] Didier Bourigault. LEXTER, a Natural Language tool for terminology extraction. In *Proceedings, 7th EURALEX International Congress*, pages 771–779, Göteborg, 1996.

[G.2.3] Didier Bourigault et Christian Jacquemin. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings, 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 15–22, Bergen, 1999.

[G.2.4] Anne Condamines et Josette Rebeyrolles. CTKB: A corpus-based approach to a Terminological Knowledge Base. In *Proceedings, 1st Workshop on Computational Terminology (COMPUTERM'98)*, pages 29–35, Montreal, 1998.

[G.2.5] Ido Dagan et Kenneth W. Church. *Termight*: Identifying and translating technical terminology. In *Proceedings, 4th Conference on Applied Natural Language Processing (ANLP'94)*, pages 34–40, Stuttgart, 1994.

[G.2.6] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans et Philip Resnik, éditeurs, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. MIT Press, Cambridge, MA, 1996.

[G.2.7] Béatrice Daille. Identification des adjectifs relationnels en corpus. In *Proceedings, Conférence de Traitement Automatique du Langage Naturel (TALN'99)*, Cargèse, 1999.

[G.2.8] Béatrice Daille, Benoît Habert, Christian Jacquemin et Jean Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–258, 1996.

[G.2.9] Chantal Enguehard et Laurent Pantera. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32, 1995.

[G.2.10] Katerina T. Frantzi et Sophia Ananiadou. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings, 16th International Conference on Computational Linguistics (COLING'96)*, pages 41–46, Copenhagen, 1996.

[G.2.11] Éric Gaussier. Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 444–450, Montreal, 1998.

[G.2.12] Benoît Habert et Christian Jacquemin. Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *Traitement automatique des langues*, 34(2):5–42, 1993.

[G.2.13] John S. Justeson et Slava M. Katz. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17(1):1–19, 1991.

[G.2.14] Andy Lauriston. Automatic recognition of complex terms: Problems and the TERMINO solution. *Terminology*, 1(1):147–170, 1994.

G.3 Analyse des variations terminologiques et utilisation éventuelle en indexation

[G.3.1] Henk Barkema. Determining the syntactic flexibility of idioms. In Udo Fries, Gunnel Tottie et Peter Schneider, éditeurs, *Creating and using English language corpora*, pages 39–52. Rodopi, Amsterdam, 1994.

[G.3.2] Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

[G.3.3] Thierry Hamon, Adeline Nazarenko et Cécile Gros. A step towards the detection of semantic variants of terms in technical documents. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 498–504, Montreal, 1998.

[G.3.4] Christian Jacquemin. What is the tree that we see through the window: A linguistic approach to windowing and term variation. *Information Processing & Management*, 32(4):445–458, 1996.

[G.3.5] Christian Jacquemin. Improving automatic indexing through concept combination and term enrichment. In *Proceedings, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 595–599, Montreal, 1998.

[G.3.6] Christian Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 341–348, University of Maryland, 1999.

[G.3.7] Christian Jacquemin, Judith L. Klavans et Evelyne Tzoukermann. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings, 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL'97)*, pages 24–31, Madrid., 1997.

[G.3.8] Karen Sparck Jones et John I. Tait. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66, 1984.

[G.3.9] William A. Woods. Conceptual indexing : A better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA, 1997.

[G.3.10] Fuyuki Yoshikane, Keita Tsuji, Kyo Kageura et Christian Jacquemin. Detecting japanese term variation in textual corpus. In *Proceedings, 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99)*, pages 97–108, Academia Sinica, Taipei, Taiwan, 1998.

G.4 Utilisation de la terminologie en recherche d'information

[G.4.1] Jacek Ambroziak et William A. Woods. Natural language technology in precision content retrieval. In *Proceedings, Natural Language Processing and Industrial Applications (NLP+IA'98)*, Moncton, New Brunswick, CA, 1998.

[G.4.2] Gregory Grefenstette. The WWW as a resource for example-based MT tasks. In *Proceedings, ASLIB Translating and the Computer 21 Conference*, London, 1999. ACL.

H Systèmes de question réponse et résumé automatique

[H.0.1] Regina Barzilay, Kathleen McKeown et Michael Elhadad. Informational fusion in the context of multi-document summarization. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 550–557, University of Maryland, 1999.

[H.0.2] O. Ferret, B. Grau, G. Illouz, C. Jacquemin et N. Masson. QALC - the question-answering program of the language and cognition group at LIMSI-CNRS. In *Proceedings, Workshop Question-Answering track at the Text REtrieval Conference (TREC8)*, Gaithersburgh MD, 1999. NIST.

[H.0.3] L. Hirschman, M. Light, E. Breck et J. D. Burger. Deep read: A reading comprehension system. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 325–332, University of Maryland, juin 1999.

[H.0.4] Inderjeet Mani et Mark T. Maybury, éditeurs. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Mass, 1999.