Proposition d'un modèle relationnel d'indexation syntagmatique : mise en oeuvre dans le système iota

Jean-Pierre Chevallet, Hatem Haddad

Proposition d’un modèle relationnel d’indexation syntagmatique : mise en oeuvre dans le système iota

Jean-Pierre Chevallet* Hatem Haddad**

Équipe Modélisation et Recherche d’Information Multimédia
Laboratoire CLIPS-IMAG, B.P. 53, 38041 Grenoble Cedex 9, France
*E-mail : Jean-Pierre.Chevallet@imag.fr
**E-mail : Hatem.Haddad@imag.fr

Résumé :

Nous présentons un modèle supportant une indexation à base de syntagmes. Cette modélisation inclut une description formelle des termes d’indexation, un processus de dérivation, une fonction de correspondance, une sémantique du langage d’indexation et une fonction de pondération de la correspondance entre termes d’indexation. Elle met en évidence les éléments qui doivent permettre de guider la conception de Systèmes de Recherche d’Informations à base de mots composés. Nous proposons également un choix de techniques pour mettre en oeuvre ce modèle, particulièrement dans l’extraction automatique des syntagmes et dans leur pondération pour le calcul de la mesure pertinence d’un document par rapport à une requête.

Mots-Clés :

Recherche d’Information, Traitement de la Langue Naturelle, Modèle Relationnel de Recherche d’Information, Indexation des syntagmes.

1 Introduction

On peut noter en examinant les recherches récentes un regain d'intérêt des approches linguistiques pour la Recherche d'Information (RI). Ce réveil de l'intérêt pour la linguistique est aussi mis en évidence par l'importance soudaine qu'a pris la technologie de la RI dans les applications industrielles, et particulièrement sur Internet. Les résultats des rapports de TREC soulignent à l'unanimité le rôle important et l'impact considérable que les traitements linguistiques peuvent avoir sur un Système de Recherche d'Information (SRI) [STRZ94, STRZ95, STRZ96].

Les fondements du domaine de la RI depuis plusieurs dizaines d'années, ont été définies principalement par Gerald Salton [SALT71]. La plupart des SRIs actuels se basent toujours sur l'hypothèse initiale qu'un document doit partager les termes d'une requête pour être identifié comme pertinent.

Bien entendu, la force de cette relation de pertinence calculée[1] est proportionnelle à l'intersection des termes entre le document et la requête. Un poids affecté à un mot clé précise l'importance de ce dernier dans le document. Que le modèle soit vectoriel, probabiliste, ou logique, ce poids est une fonction du nombre d'occurrences du terme dans le document. Le problème de la RI semble alors se résumer à un simple calcul de correspondance entre un ensemble de mots clés de la requête de l'utilisateur avec l'ensemble des mots clés représentant le document. Dès l'origine, il a été proposé un minimum de traitements linguistiques se limitant à la troncature des mots extraits du corpus et à l'élimination des mots outils de la langue. Ces traitements rudimentaires sont tous suffisamment adéquats à la tâche de RI pour être encore aujourd'hui utilisés dans bon nombre de travaux de recherche sur l'indexation automatique. La *variation morphologique*[2] est donc traitée sous la forme d'une troncature des suffixes et des flexions, et la *variation lexicale*[3] apparaît sous la forme d'un thésaurus. Par contre, la *variation syntaxique*[4] est rarement intégrée aux systèmes, tout simplement à cause du choix des mots isolés pour supporter les termes d'indexation. Nous avons montré dans [CHEV97] que l'intérêt de termes composés comme termes d'indexation peut se mesurer en terme de précision et de rappel. On peut également le comprendre en examinant le texte : *"des quartiers avec des architectures modernes ont vu le jour à côté de la vieille ville"* qui ne traite pas de *"l'architecture moderne de la vieille ville"* même si les même mots apparaissent ensembles dans le document.

En fait, un texte n'est pas seulement un "sac de mots", mais il est bien véritablement un ensemble fortement structuré de termes qui permettent de communiquer des informations d'une grande précision. Cette richesse et cette complexité doivent être pris en compte dans les SRI. Nous pensons plus précisément, que seule la voie menant vers l'utilisation d'index structurés permettra d'augmenter de manière significative la qualité des SRI actuels. La difficulté de cette approche tient dans la puissance d'expression du langage d'indexation qu'il convient de mettre en place, et

[1] appelée pertinence système par opposition à la pertinence utilisateur que le système tente d'approcher.

[2] Cette variation est de nature flexionnelle qui ne change pas la catégorie grammaticale du terme, comme la variation pluriel/singulier et dérivationnelle qui elle fait dériver un terme d'une catégorie à une autre comme par exemple "polluer" et "pollution".

[3] Elle concerne le glissement d'un terme vers un autre sémantiquement proche comme, par exemple, "voiture" vs "automobile".

[4] Elle concerne les différentes constructions syntaxiques ayant un sens voisin comme "pollution de l'air du fait des moteurs diesel" avec "pollution de l'air par les gaz d'échappement des moteurs diesel".

particulièrement, dans son adéquation à un type de recherche d'informations orienté vers la précision des réponses. Une seconde difficulté réside ensuite dans la mise en oeuvre d'un processus d'indexation fiable, capable de produire des index structurés de qualité suffisante, sur des quantités importantes de documents.

Nous proposons donc dans cet article une approche permettant sur les deux points cités plus haut de réaliser un SRI à indexation automatique orienté vers la précision des réponses à des requêtes complexes exprimées en langue naturelle. Pour ce faire, nous proposons une modélisation des termes d'indexation à partir de syntagmes extraits des textes. Cette proposition se base sur le fait établi que l'information dans les textes (plus particulièrement les textes scientifiques et techniques) se trouve localisée de façon privilégiée dans les groupes nominaux.

Nous présentons d'abord dans la section 2 les principaux travaux traitant de l'extraction et de l'utilisation des groupes de mots dans l'indexation ou l'interrogation des SRI. Dans la section 3, nous proposons notre modèle d'indexation basé sur des syntagmes, avec les directions que nous suivons pour sa mise en oeuvre dans notre système **iota** dans la section 4.

2 Indexation par des multi-termes

L'hypothèse sous entendue dans la mise en place d'un système à indexation automatique à l'aide de multi-termes, est que ces derniers sont plus aptes à désigner des entités sémantiques ou des concepts que les mots uniques et constituent alors une meilleure représentation du contenu sémantique des documents.

Ces groupes de mots peuvent être sélectionnés *statistiquement* ou *syntaxiquement*. Les techniques statistiques permettent de découvrir des séries de mots ou des combinaisons de mots qui occurrent fréquemment dans un corpus alors que les groupes de mots découverts syntaxiquement, sont en fait les *syntagmes nonimaux* utilisés dans les documents à indexer.

La comparaison des résultats des deux approches n'a pas aboutit à des conclusions claires en ce qui concerne leur utilité en RI [FAGA87, MITR97, KRAA98, CHEV97]. La plupart des travaux montrent que les syntagmes offrent un avantage quantitativement insignifiant. Certains travaux ont même trouvé de moins bons résultats, et justifient leur échec par le fait que la fonction de correspondance utilisée n'était pas adaptée [SMEA94], ou que la découverte de syntagmes n'a été appliquée qu'aux requêtes [CROF91, SMEA88]. Fagan qui utilise une approche statistique [FAGA89], montre qu'il y a trop de bruit dans les combinaisons des mots trouvés. Cette approche statistique se base sur l'hypothèse que l'emploi de deux mots en cooccurrence est l'expression d'une relation sémantique entre ces mots.

Différentes autres approches comme la découverte de combinaisons de mots en fonction de leur régularité d'apparition [CHUR90] ou bien l'utilisation d'une mesure de proximité entre les contextes des termes [BESA99] ou encore l'utilisation des mesures de similarité, telles que l'information mutuelle ou le coefficient de Dice, prennent en considération la distance entre les mots ou l'ordre de leur apparition. Ces approches statistiques imposent l'utilisation de corpus suffisamment important pour obtenir des mesures statistiques valides, c'est-à-dire ne relevant pas de combinaisons accidentelles incompréhensibles par un humain. Elles ont pour avantage d'être techniquement simples à mettre en oeuvre et elles permettent de couvrir de manière exhaustive toutes les combinaisons possibles des termes dans une fenêtre allant de la phrase au document tout entier.

D'autres travaux adoptent une approche hybride qui combinent une approche statistique avec une approche linguistique. Ces systèmes construisent des groupes de mots en sélectionnant les

candidats à partir de schémas syntaxiques puis en les filtrant à l'aide de méthodes statistiques [SIMO00]. Grefenstette dans [GREF92] combine les cooccurrences statistiques des termes et la technique de "tête modifieur" pour extraire des combinaisons de mots relatives à un même contexte. Hearst dans [HEAR92] utilise des patrons syntaxiques pour extraire des syntagmes reflétant une relation sémantique entre les termes alors que Morin dans [MORI99] se base sur les travaux de Hearst en intégrant des mesures statistiques pour le filtrage. Ce dernier met en évidence des schémas lexico-syntaxiques par l'intermédiaire de phrases qui utilisent des couples de termes liés par la même relation sémantique. Il regroupe alors dans une même classe sémantique les expressions partageant le même environnement lexical et syntaxique.

L'expérience menée et acquise dans ces travaux nous conduit aux constats suivants :

Utilité du traitement linguistique : Les différentes conférences de TREC ont conduit vers un consensus reconnaissant à l'usage de techniques linguistiques, un impact positif sur la RI. Il est cependant clair que l'exploitation du potentiel de ces traitements linguistiques est très délicat à mettre en oeuvre, car les phénomènes linguistiques sont variés et complexes. De plus ces traitements doivent s'appliquer à des corpus en croissance permanente[5]. Il est alors difficile de choisir a priori un ensemble de traitements linguistiques applicable aux corpus actuels et capables d'assurer une augmentation significative de la qualité des réponses du système.

Nature du traitement linguistique : La plupart des travaux attribuent les faibles résultats obtenus à l'aide de traitements linguistiques au manque de robustesse et aux faibles performances des analyseurs utilisés. Nous pensons qu'il est important de bien mesurer ce qu'apporte l'usage d'un nouveau traitement plus précis de la langue par rapport à un traitement plus frustre. Il s'agit en fait, de déterminer la couverture des phénomènes linguistiques (et même sémantiques sous jacents) de ce nouveau traitement, par rapport à l'ancien, puis d'évaluer l'*impact effectif* qu'il va avoir sur le SRI dans sa globalité. Par exemple, remplacer la troncature des termes d'indexation par un traitement plus "linguistique" des racines, n'aboutit pas forcément à un meilleur résultat. En effet, cette troncature provoque certes, des incorrections à cause des collisions entre racines de natures différentes, mais ces collisions ne sont pas forcément préjudiciables au système. En effet, elles peuvent intervenir, soit trop peu fréquemment, ou bien dans le contexte des autres termes ; dans le cas de requêtes longues, elles permettent de limiter l'ambiguïté produite par la troncature. On peut donc penser que les modèles hybrides (i.e. linguistique et statistique) sont plus performants que les approches purement linguistiques.

Dépendance entre termes d'indexation : Les travaux actuels se basent sur l'hypothèse d'une indépendance entre les termes d'indexations extraits. Cette indépendance est, par exemple, cruciale pour le bon fonctionnement du modèle vectoriel de RI. Cette indépendance est dans la plupart des cas une hypothèse abusive. Elle peut se justifier à la limite dans le cas de termes simples fortement tronqués, c'est à dire réduits à leur racine. Par contre, l'usage de termes composés comme termes d'indexation renforce leur dépendance. En effet, la composition de terme produit des termes plus spécifiques et donc fortement dépendants des termes composants. De plus, les variations et les constructions elliptiques, lorsqu'elles ne sont pas explicitement traitées, provoquent également une dépendance entre des termes qui en fait sont contextuellement synonymes. La dépendance entre les termes d'indexation complexes doit alors être explicitement étudiée et intégrée dans un modèle d'indexation.

[5]Pour illustration, les corpus de test sont passés de 2 mégabits à plusieurs Gigaoctet pour la conférence TREC, le corpus du WEB est lui estimé à 900 millions de pages.

Pondération des termes d'indexation : Une indexation avec des groupes de mots doit proposer un schéma de pondération adapté doit tenir compte des phénomènes langagiers comme la variation ou l'anaphore. Cette pondération doit être vue, non pas comme une nécessité d'avoir "quelque chose de flou" dans le calcul de la correspondance, mais comme la composition d'un ensemble de mesures distinctes, relatives au sens et à l'importance du terme pour l'usage qui lui est destiné, c'est-à-dire l'indexation pour la résolution d'une requête exprimant un besoin d'information.

Pour répondre aux problèmes posés par les constats précédents, nous proposons un nouveau paradigme d'indexation relatif à la notion de syntagme. Ce paradigme suppose la traduction de syntagmes extraits des corpus à indexer, en terme d'index syntagmatiques possédant une sémantique ensembliste. Cette transformation nous permet de bien séparer les problèmes liés à la définition d'un modèle de recherche d'information, de celui du repérage automatique des termes d'indexation. Nous désirons donc volontairement nous détacher du texte vu comme un signal, pour construire des index plus conceptuels. Le modèle hybride ainsi obtenu fera usage de traitements linguistiques, d'informations de nature sémantique, et pragmatique tout en conservant des mesures statistiques indispensables dans notre contexte de recherche d'information.

3 Un modèle relationnel basé sur les syntagmes

Il est donc clair que notre objectif est de s'éloigner du *signal* pour se rapprocher du *sens*. Notre modèle doit alors manipuler des index structurés se situant au plus près possible de la "signification". Notre objectif est également de proposer un modèle qui puisse être réalisable avec la technologie actuelle. La contrainte du temps d'indexation est une contrainte forte dès lors que l'on désire indexer une masse d'information en rapport avec la production électronique actuelle de documents[6]. La contrainte du temps d'interrogation est une contrainte encore plus forte pour la réalisation d'un SRI interactif.

Nous avons choisi dans un premier temps, de nous intéresser aux informations exprimées dans un syntagme défini sommairement comme *un ensemble de termes respectant des lois de morphologie et de syntaxe, et possédant une signification propre*. Plus précisément, nous centrons notre étude sur les syntagmes nominaux, tout en sachant que pour la recherche d'information, la séparation entre syntagme verbal et nominal n'est par une donnée pertinente. Il faut considérer, par exemple, que l'information contenue dans la formulation "les gaz d'échappement polluent l'atmosphère" doit être rapprochée de la formulation "la pollution de l'atmosphère par les gaz d'échappement". Les systèmes simples à base de troncature ont depuis longtemps fait la preuve que la méconnaissance de la structure syntaxique ou même morphologique des textes n'est pas un obstacle à une indexation opérationnelle. Pourtant, les syntagmes sont des indicateurs plus précis du contenu des documents qu'un ensemble de racines de mots et notre modèle se fonde sur une structure syntagmatique. Pour justifier notre intérêt pour les syntagmes en tant que fondement de la construction des index, il faut noter le rôle important que jouent les syntagmes nominaux des langues naturelles en agissant par exemple comme sujets ou objets. On peut noter également qu'en intelligence artificielle, les syntagmes nominaux sont considérés comme des référents de concepts car ils demeurent une bonne approximation du sens.

Les traitements linguistiques nécessaires à la mise en oeuvre du modèle doivent permettre non seulement de focaliser l'analyse sur des syntagmes pertinents, cette notion de pertinence restant

[6] Pour se faire une idée de cette masse, il faut avoir en tête les 900 million de pages qui sont estimées être disponibles sur le web.

à définir, mais surtout ces traitements doivent en proposer une structure. C'est justement cette structure qui doit supporter une partie de la signification du terme d'indexation syntagmatique.

Il reste néanmoins très difficile de placer les termes d'indexation réellement à un niveau sémantique. De manière pratique, c'est la syntaxe qui sert de passerelle vers le niveau sémantique [PARA96, KHOO97]. Les travaux actuels qui traitent de l'analyse syntaxico-sémantique ou d'une indexation syntaxico-sémantique se basent plutôt sur un traitement syntaxique que sémantique.

La question pratique que l'on doit finalement se poser sur l'extraction des termes d'indexation à partir des syntagmes, concerne la profondeur de l'analyse à mettre en oeuvre. Une analyse de surface avec des patrons syntaxiques semble suffisante comme l'atteste les travaux de Daille [DAIL94] et de Debili [DEBI82]. Ces patrons ne traitent que les relations homosyntagmatiques c'est-à-dire les relations s'établissant entre éléments appartenant à une même phrase. Cependant il semble que les patrons de courte taille, comme ceux décrits dans ces travaux [7] ne soient pas suffisamment porteurs d'information. Nous verrons plus tard que pour traiter au minimum la variation des termes et les phénomènes d'ellisions, nous allons devoir mettre en place des traitement de surface inter-phrases couplés à des mesures statistiques.

La partie suivante décrit notre modélisation des index en des arbres qui s'inspirent des relations homosyntagmatiques que nous voulons extraire automatiquement avec un degré de réussite suffisamment élevé pour être utilisable dans un SRI.

3.1 Modèle d'indexation syntagmatique

Ce que nous désignons par *terme d'indexation syntagmatique* (TIS), est simplement un terme du langage d'indexation obtenu par une analyse des syntagmes. Autrement dit, un terme d'indexation syntagmatique est la représentation d'au moins un syntagme du corpus. Pour une représentation des TIS en adéquation avec les besoins de la RI, nous proposons un formalisme qui s'affranchit au maximum des détails de l'analyse syntaxique. En effet, l'utilisation directe de l'arbre de dérivation syntaxique ne semble pas donner de résultats satisfaisants comme le montre [SMEA94]. La syntaxe des TIS est directement inspirée des travaux de Arampatziz et All [ARAM98]. Nous ne somme par contre pas d'accord sur leur manière de définir la fonction de correspondance. Cette représentation s'inspire également des travaux de [GREF92, CARB00] qui suggère que chaque syntagme peut être structuré en tête et arguments. La tête exprime le concept central et une liste d'un ou plusieurs arguments modifie la tête et apporte de la précision au syntagme obtenu. Nous avons choisi le formalisme de représentation suivant :

Définition d'un terme d'indexation syntagmatique : Un terme d'indexation syntagmatique I est une structure :

$$I = [TR_1[A_1]R_2[A_2]...R_n[A_n]]$$

où T désigne la tête du terme, $\{A_1, A_2, ..., A_n\}$ désignent les arguments modifieurs, et $\{R_1, R_2..., R_n\}$ qualifie la relation entre la tête et les modifieurs. Cette qualification de la relation est optionnelle. Elle n'apparaît que si cette relation n'est pas générique. La tête est un atome syntagmatique[8]. Les modifieurs peuvent être soit un atome, soit à leur tour des termes d'indexation syntagmatiques. L'ordre entre les modifieurs n'est pas significatif, par contre, tous les modifieurs doivent être différents.

[7] ex. Substantif-Substantif "centre Casino", Substantif-Préposition-Substantif " moulin à vent", Substantif-Adjectif "vache folle", Substantif-Préposition-Substantif-Adjectif "angine de poitrine instable"

[8] Par "atome syntagmatique" nous entendons une succession de mots possédant une signification propre et étant sujet à aucune variation autre que flexionnelle. Ces termes peuvent être des mots composés comme "hot dog" ou "pomme de terre".

Par exemple le syntagme “séparation de la République fédérale tchèque”, peut être traduit par le TIS :

[séparation OBJ [république [fédérale] [tchèque]]]

L’utilisation des relations explicites au cours de l’indexation a été mise en évidence dans plusieurs travaux de recherche d’information dont la plupart restent théoriques [OUNI98, FARR80]. En effet, l’identification automatique correcte de relations sémantiques entre unités textuelles est très difficile. Elle fait intervenir des connaissances générales et beaucoup de connaissances sur le domaine. C’est pourquoi plusieurs travaux se sont orientés vers les relations syntaxiques pour substituer aux relations sémantiques [HEAR92, MORI99].

Ces travaux se sont basés principalement sur les marqueurs sémantiques comme les patrons lexico-syntaxique. On distingue les marqueurs génériques et les marqueurs spécifiques. Un marqueur générique désigne une relation de façon plus ou moins stable à travers différents corpus et domaines techniques alors qu’un marqueur spécifique désigne une relation avec une certaine précision relativement à un domaine.

La plupart des marqueurs et les relations qu’ils désignent ont été bien étudiées dans la littérature ainsi que la résolution des ambiguïtés qu’ils peuvent engendrer. Ainsi on trouve des études concernant la relation de causalité [GARC98], d’ingrédience [JACK96], hyponymie [HEAR92]. L’établissement d’une base de marqueurs ainsi que la définition des relations qu’ils désignent et leurs règles d’application, permet de définir des relations entre les éléments d’un syntagme. L’objectif des relations est de conserver une partie de la précision du syntagme initial lors de sa transformation en index.

Ce modèle très simple de représentation arborescente, va tout de même nous permettre de mettre en place une nouvelle fonction d’indexation. Il nous permet de mettre en évidence des aspects de la recherche d’informations trop rarement étudiés dans les systèmes actuels.

Le modèle d’indexation doit ensuite permettre de valuer un terme d’indexation pour exprimer l’importance qu’il semble avoir par rapport au document tout entier qu’il doit indexer. Cette valeur est le plus souvent calculée à partir de valeur fréquentielle d’occurrences des termes.

Pertinence d’un terme indexation syntagmatique : La pertinence $Pert(I, D)$ d’un TSI I est une valeur exprimant l’importance de ce terme dans le document D.

Comme nous proposons d’utiliser des structures d’indexation complexes extraites automatiquement, cette extraction ne peut pas être considérée comme totalement fiable. L’utilisation dans la correspondance d’un TIS doit alors tenir compte d’un facteur de fiabilité :

Qualité d’un terme d’indexation syntagmatique : La qualité $Qual(I, P, D)$ d’un TSI I, extrait de la phrase P d’un document D, est défini comme la probabilité que ce terme reflète effectivement l’information initialement contenue dans la phrase de ce document. Une valeur moyenne sur toutes les phrases P du document D s’exprime par $Qual(I, D)$.

Cette valeur peut, par exemple, être estimée manuellement comme “la proportion d’experts en accord avec la construction de ce terme d’indexation”. La mesure de pertinence d’un TIS doit être comprise comme indépendante de sa qualité. Elle représente sa pertinence en supposant le terme d’excellente qualité.

Finalement, l’information contenue dans la phrase et, plus généralement, exprimée par le document, doit se refléter dans le TIS. Cette grandeur, a priori, indépendante de sa pertinence au document, exprime la quantité d’information “empruntée” au document tout entier. Cependant, on peut remarquer que les TSI porteurs de beaucoup d’information, pourront probablement être jugés fortement pertinents pour le document.

Quantité d'information d'un terme d'indexation syntagmatique : La quantité $Qinf(I)$ d'un TSI I permet d'exprimer le "pouvoir évocateur" de ce terme indépendamment du contexte du document où il se trouve. Cette valeur, subjective[9] exprime une quantité que l'on peut mesurer par approximation et qui permet de comparer les termes entre eux. Les termes comparables sont alors ceux qui partagent de l'information.

Un terme avec une quantité d'information non nulle, reflète alors une partie de l'information exprimée par le document dans lequel il se trouve. Par contre, les mots outils de la langue n'ont quasiment aucune quantité d'information en eux même. On peut également préciser que cette quantité d'information nous semble proportionnelle à la taille du terme, et en rapport avec sa distribution dans un corpus. On peut noter alors que cette quantité augmente quand le terme se spécialise.

Finalement, il nous reste à définir un *index syntagmatique* (IS) comme un ensemble de termes d'indexation syntagmatiques :

Définition d'un index syntagmatique : Un index syntagmatique d'un document D est un ensemble de termes d'indexations syntagmatiques, où chaque terme est muni de sa valeur de pertinence, sa valeur de qualité et sa quantité d'information.

$$IS(D) = \{(I, Pert(I, D), Qual(I, D), Qinf(I))\}$$

Nous allons maintenant détailler comment mettre en place un processus de correspondance, d'abord entre les TIS puis globalement pour l'index syntagmatique.

3.2 Fonction de correspondance entre TIS

Dans notre modèle de RI, une requête est exprimée par un ensemble de termes d'indexations syntagmatiques où la valeur de pertinence n'est pas prise en compte, car nous considérons que, ce terme ayant été volontairement proposé par l'utilisateur dans sa requête, il ne peut être que pertinent.

Définition d'une requête syntagmatique : Une requête syntagmatique RS issue d'une requête en langue naturelle R est un ensemble de termes d'indexations syntagmatiques, où chaque terme est muni de sa valeur de qualité relative à la requête et sa quantité d'information.

$$RS = \{(I, Qual(I, R), Qinf(I))\}$$

La correspondance entre document et requête consiste alors à calculer le degré de correspondance entre deux ensembles pondérés de TIS. Nous proposons de baser le calcul de la correspondance entre ces deux ensembles, sur une mesure de similarité terme à terme.

Dans notre modélisation, la tête du terme est une entité prépondérante. Il semble alors important de privilégier une correspondance basée en priorité sur les têtes de ces deux termes. Les modifieurs, pouvant être eux mêmes des termes, le terme de la requête peut être modifieur d'un terme d'index plus complexe. Par exemple, avec le terme d'indexation :

[séparation [république [fédérale] [tchèque]]]

Il faut pouvoir mesurer une correspondance avec le terme de la requête :

[république [tchèque]]

[9] au sens de "dépendante du sujet". Nous somme d'accord pour considérer que le "sens" d'un terme, comme son "information", appartiennent à la personne qui lit et interprète ce terme.

ou bien avec la requête :

[république [petite] [tchèque]]

Pour solutionner ce problème de correspondance à la fois de manière théorique et élégante, nous proposons d'adopter la vision de Nie dans [NIE90] qui a été reprise par Chevallet dans [CHEV92], et qui consiste à considérer le processus de calcul de la correspondance comme un processus incertain de déduction qui, du document, conduit à impliquer logiquement la requête.

Cette dérivation utilise des schémas de transformation de termes. D'un autre coté, si l'on compare notre formalisme à celui des graphes conceptuels de Sowa [SOWA84], on constate que la dérivation de TIS est un cas particulier de la correspondance des graphes conceptuels. Nous avons déjà montré dans [CHEV91] que la correspondance dans un SRI à base de graphes conceptuels passe par un calcul de projection de graphe. Il nous suffit alors d'adapter ce point de vue au cas particulier des termes que nous avons défini. Les opérateurs algébriques sont alors les suivants :

Jointure : La jointure d'un TIS initial à un autre TIS consiste à ajouter ce terme comme modifieur soit de la tête du terme, soit de l'un des modifieurs du TIS. Le terme ainsi modifié implique d'une part, par spécialisation, le terme initial, mais également le terme qui l'a modifié.

Par exemple :

[séparation [république [fédérale] [tchèque]]] $\sqsubset$ [séparation]
[séparation [république [fédérale] [tchèque]]] $\Rightarrow$ [république [fédérale] [tchèque]]

Il nous faut séparer les deux types d'implications issus de l'opération de jointure. L'implication entre le terme initial générique et le terme spécialisé est une relation

de subsomption de termes. La dénotation ensembliste inspirée des logiques terminologiques [SEBA94] que nous proposons dans la partie suivante, permet d'exprimer cette implication en terme d'inclusion ensembliste des dénotations.

La seconde implication doit être considérée différemment. En effet, il s'agit d'un terme qui implique un de ses modifieurs. Dans une dénotation, il s'agit d'une relation en deux ensembles qui n'ont aucune intersection. Cette implication doit être étudiée plus précisément. Les deux autres opérateurs sont les suivants :

Qualification : La qualification d'une relation consiste à ajouter un qualifieur à un modifieur d'un TIS. Le terme ainsi obtenu est plus spécifique que le terme initial. Il implique donc que le terme initial plus générique.

Nous supposons qu'existent des dépendances entre *TIS simples* (i.e. sans modifieurs). Ces dépendances expriment des relations générique/spécifique (plus ou moins certaines) entre les mots pouvant se placer en tête.

Spécialisation : La spécialisation d'un TIS consiste à remplacer la tête du terme par une tête plus spécifique en utilisant les relations générique/spécifique entre les TIS simples.

Nous allons maintenant donner une sémantique plus précise à notre langage d'indexation.

3.3 Sémantique du langage des TIS

Nous proposons sommairement ici une sémantique dénotationnelle de notre langage des termes d'indexation syntagmatiques. Nous désirons en fait justifier toute mesure de pondération ultérieure par rapport à cette sémantique.

Dénotation : La dénotation d'un TIS est un ensemble fini. Cette dénotation ne peut être vide que dans le cas ou le terme n'a pas de signification dans le contexte d'indexation où l'on se trouve. Par exemple, le terme [république [dauphinoise]] n'a généralement pas de sens. Les termes à dénotation vide sont donc tous équivalents d'un point de vue sémantique. Le singleton est un cas particulier : il représente un terme désignant sans ambiguïté une entité. Par exemple le terme [république [tchèque]] peut être dénoté par un singleton. La dénotation de deux termes peut être identique s'ils désignent la même entité. Le terme [république [fédérale] [tchèque]] peut très bien être associé à la même dénotation que le terme précédent. Alors que le terme [république [fédérale]] peut très bien dénoter au moins l'Allemagne et la Tchéquie.

Relation d'hyperonimie : Un TIS simple I_1 est strictement générique à un autre TIS simple I_2 si et seulement si la dénotation de I_2 est strictement incluse dans la dénotation de I_1. Le terme I_2 est alors strictement spécifique à I_1.

Subsomption stricte : La dénotation d'un TIS I_1 transformé à partir du TIS I_2 par une jointure, une qualification ou une spécialisation par un TIS simple strictement spécifique, est incluse ou égale à la dénotation du terme I_2.

Nous proposons maintenant une utilisation de cette sémantique dans la définition du coût d'une correspondance entre deux TIS.

3.4 Pondération de la correspondance entre TIS

Nous proposons la définition suivante :

Coût d'une dérivation : Le coût d'une dérivation de termes à qualité constante doit être proportionnel à la perte de sa quantité d'information et inversement proportionnelle à la variation de sa dénotation.

Par exemple, la dérivation du TIS [république [tchèque]] en [république] donne à la fois moins d'informations, et dénote plus d'entités. Par contre, si [république [fédérale] [tchèque]] et [république [tchèque]] ont la même dénotation, alors la dérivation de [séparation [république [tchèque]] [séparation [république [fédérale] [tchèque]] a un coût uniquement lié à l'information ajouté par l'introduction de [fédérale].
Nous proposons donc la formule suivante de calcul d'un coût C d'une dérivation élémentaire du terme I_1 vers le terme I_2 :

$$C(I_1 \Rightarrow I_2) = C(I_1 \sqsubset I_2) = \frac{Qinf(I_1) - Qinf(I_2)}{\Delta(\mathcal{D}(I_1, D), \mathcal{D}(I_2, D))}$$

La fonction Δ permet de calculer la différence entre deux ensembles de dénotations $\mathcal{D}$. Le fait de considérer les deux types de dérivations de même manière n'est pas satisfaisant. En effet, dans le cas de la subsomption, la fonction Δ représente le rapport d'inclusion entre les deux dénotations. Dans l'autre cas, on se trouve avec des ensembles de nature différente. Par exemple, [séparation [république [tchèque]]] dénote en fait une séparation particulière, celle de la république Tchécoslovaque en Tchéquie et Slovaquie. Par contre [république [tchèque]] dénote l'actuelle Tchéquie. Pour pouvoir effectuer le calcul du coût, nous simplifions ce cas en disant que seule compte la quantité d'information, la fonction Δ retourne donc la valeur 1.

Pondération d'une correspondance La pondération du calcul de la correspondance entre deux termes correspond au minimum de la somme des coûts de tous les chemins de dérivations possibles. Le coût d'un chemin de dérivation est défini comme la somme des coûts des dérivations élémentaires.

Nous sommes donc maintenant à même de définir une correspondance entre index et requête syntagmatique sur la base de cette correspondance de termes.

3.5 Fonction de correspondance entre index et requête

La correspondance entre un index et une requête utilise donc la correspondance entre deux TIS. La correspondance peut s'établir entre un TIS de la requête et un TIS du document, mais également entre deux TIS du document. Nous devons par cette situation tenir compte de la dépendance entre les termes, au sein même d'un document. Globalement, la confrontation d'une requête avec un document se présente comme l'introduction dans un réseau de dépendance de TIS dans l'index, des nouveaux TIS appartenant à la requête. C'est le réseau obtenu dans sa globalité qui doit être évalué.

Nous calculons de la manière suivante, la distance d en terme de coût entre un TIS I_r de la requête et un index syntagmatique $IS(D)$, c'est à dire un ensemble de TIS I_d qui forment l'index du document D comme décrit en 3.1 :

$$d(I_r, IS(D)) = min_{I_d \in IS(D)}(\frac{C(I_d \sqsubset I_r) * Pert(I_d, D)}{Qual(I_r, R) * Qual(I_d, D)})$$

Cette formule exprime que la distance d'un terme de la requête à l'index du document, se calcule en recherchant le coût de dérivation minimum par rapport à tous les TIS du document, en tenant compte de la qualité des deux termes et de la pertinence du TIS dans son document. Nous supposons bien sur qu'aucun terme n'a une qualité nulle. En pratique, nous ne retenons que les termes dépassant un seuil de qualité. Finalement, la mesure de pertinence P entre un index $IS(D)$ et une requête syntagmatique RS peut se calculer comme une combinaison $\odot$ de la distance d de chaque TIS de la requête :

$$P(IS(D), RS) = \frac{1}{1 + \odot_{I_r \in RS}\, d(I_r, IS(D))}$$

Cette formule assure une mesure de pertinence comprise entre 0 et 1. La combinaison $\odot$ est fonction du sens que l'on veut donner à la requête. Lorsque l'on veut privilégier le meilleur terme de la requête, alors on peut choisir la fonction min. De manière plus neutre on peut se contenter de la somme des distances de tous les termes de la requête.

La partie suivante est consacrée aux problèmes à résoudre quand à la mise en oeuvre de ce modèle.

4 Mise en oeuvre

Le modèle proposé reste une vision théorique de ce qui doit être réalisé en pratique pour faire fonctionner un SRI à base de syntagmes. Dans la partie suivante, nous établissons un bref panorama des phénomènes linguistique que nous prenons en compte dans la mise en oeuvre de l'indexation.

4.1 Phénomènes linguistiques

L'étude des traitements linguistiques à mettre en oeuvre dans un SRI doit concerner principalement le repérage des index syntagmatique, et plus particulièrement ce qui concerne leur décomptage (voir 4.4). Notre but n'est pas de construire un index à partir de tous les syntagmes mais de prendre en compte la variation dans l'expression de ces syntagmes. Cette variation permet d'exprimer un index unique sous des différentes formes syntaxiques. En effet, reconnaître la variation

syntaxique c'est établir des liens entre syntagmes. La difficulté réside dans l'assurance que ce lien soit bien de nature sémantique, car seuls les dérivations pertinentes pour une interrogation, donc seules celles porteuses de sens, peuvent constituer une amélioration à la fois en terme de rappel et précision. Cette affirmation est en fait notre hypothèse de travail.

La normalisation de ces syntagmes est un autre point important dans la mesure ou l'on cherche à expliciter des termes pertinents pour une interrogation : le compromis se trouve entre la richesse des informations extraites, et leur utilisation effective lors de l'interrogation. Par exemple, nous ne retiendrons les variations liées au nombre que si cette perte d'information entraîne une ambiguïté sur un terme discriminant pour le corpus. Dans notre cas, la normalisation consiste à tenir compte des variations flexionnelles, dérivationnelles et syntaxiques. La variation flexionnelle permet d'identifier pour chaque terme, les formes singuliers/pluriels des noms, et les formes infinitives, participes passés et gérondives des noms/verbes. La variation dérivationnelle ou (morpho-syntaxique) permet d'intégrer les phénomènes de nominalisation des adjectifs et des verbes, et adjectivisation des noms.

Dans la suite nous allons examiner les cas de variation qui nous semble prioritaire de traiter dans le cadre de la recherche d'information. On trouvera dans [JACQ97] une étude complète sur la variation. Nous nous bornons ici à présenter les cas que nous traitons dans notre système **iota** .

Le figement : Lorsque le sens d'un groupe de mots ne peut pas être déduit du sens des mots qui le composent, nous parlons alors d'expression figée. La notion de figement peut être entendue de deux manières. Au niveau morphosyntaxique, on considère comme figée une séquence de mots qui ne permet pas d'intercalation comme par exemple *pomme de terre*. Ces expressions figées peuvent être identifiées à l'aide d'un dictionnaire. Au niveau sémantique, l'absence de variation d'un syntagme peut être considérée comme un signe de stabilisation d'un concept et le syntagme est considéré alors figé. La notion de figement est importante d'un point de vue sémantique, car elle renvoie le plus souvent à une dénomination univoque. Ces mots des syntagmes figés ne peuvent pas être séparés. Le syntagme dans sa globalité est alors un atome et peut jouer le rôle de tête ou d'argument dans un TIS. Le syntagme *la production de pomme de terre* est alors représenté sous la forme du TIS [production OBJ [pomme de terre]].

La substitution elliptique Notre étude est partiellement motivée par la justesse du décomptage des syntagmes candidats pour l'indexation. Il est donc important de tenir compte des ellisions au sein même des termes. Nous parlons de *substitution elliptique* lorsqu'un syntagme sert d'abréviation à un syntagme plus long, qui représente la forme canonique, et qui, comme telle, doit probablement figurer dans une terminologie du domaine. La substitution elliptique que l'on peut interpréter pour "huile de graine de tournesol" avec "huile de tournesol", peut également s'interpréter comme une métonymie. On peut sous-entendre que l'on fait de l'huile avec la graine du tournesol. La détection de ce type de substitution peut se baser sur une liste pré-établie de formes de référence, ou bien se baser sur une mesure de cohésion des deux termes exprimant le taux de figement du terme dans le corpus. Dans les deux cas, les syntaxes extraits seront de la forme :

[huile [tournesol]]

[huile [tournesol $\overline{PART}$ [graine]]

Le nom de relation surligné exprime un sens inverse de la relation. Par contre si la forme rencontrée est "La graine de tournesol ..." nous proposons bien de produire le syntagme :
[graine $PART$ [tournesol]]

L'anaphore ou la variante elliptique : Dans le discours, une anaphore est la reprise d'un syntagme énoncé (l'antécédent) par un pronom ou syntagme plus court. Par exemple dans les phrases "La république fédérale tchèque est .... Cette république possède ...", la seconde apparition de "république" est une référence anaphorique au syntagme initial. Une anaphore peut ainsi permettre une forme d'abréviation qui permet de désigner une entité sans la nommer explicitement ni la décrire. Les reprises anaphoriques sont exprimées généralement au début d'une phrase par un démonstratif (ce, cette, ces) ou par un pronom personnel (il, elle).

La coordination La coordination permet de factoriser les têtes de deux syntagmes en coordonnant leurs arguments. Elle peut aussi factoriser les arguments et coordonner les têtes. Par exemple, dans la phrase "Il n'est pas vrai que l'unique voie démocratique pour la séparation de la République fédérale tchèque et slovaque ait été le référendum". Le système extrait les syntagmes suivants :

[voie [unique][démocratique] SPEC [séparation OBJ [république [fédérale] [tchèque]] OBJ [république [fédérale] [slovaque]]]

Dans la suite nous détaillons les traitements que nous avons mis en oeuvre dans notre système expérimental **iota** .

4.2 Méthodologie d'extraction des syntagmes

En premier lieu, il faut extraire les syntagmes des textes. Il faut ensuite les structurer, et enfin estimer les valeurs de pertinence, qualité et quantité d'information. Le traitement commence par une analyse morpho-syntaxique qui catégorise chaque mot du corpus. Ce processus réalise une normalisation par lemmatisation. Les syntagmes maximaux (i.e. les plus longs possibles) sont ensuite extraits par repérage de frontières de catégories syntaxiques.

De manière parallèle, un ensemble de multi-termes courts sont extraits de l'ensemble de ces syntagmes maximaux. On calcule un degré de figement de ces termes courts. Cette valeur nous permet de décider si un multi-terme court doit être considéré comme un mot composé figé de la langue, donc dans notre modèle il sera considéré comme un atome syntagmatique. Cette mesure de figement se base sur le nombre d'occurrences de syntagmes maximaux où le multi-terme apparaît sans modification, par rapport au nombre d'occurrences de syntagmes maximaux où les mots de ce multi-terme apparaissent d'une autre manière. Ces atomes sont validés manuellement et viennent compléter une liste de termes et expressions figés.

Chacun des syntagmes maximaux est structuré en tête et arguments. Pour cela nous utilisons un ensemble de patrons syntaxiques. Un patron est composé d'une séquence de catégories, suivie par la structure d'index syntagmatique à produire. Cela permet de déterminer l'élément qui forme la tête et les éléments qui formes les arguments. Par exemple, pour les phrases "La république tchèque est située ...", et "La séparation de la république fédérale tchèque est due à ....", les syntagmes maximaux extraits sont "république tchèque" et "séparation de la république fédérale tchèque". Les patrons suivants permettent de produire les exemples de syntagmes utilisés en 3.2.

SUBC(A) ADJQ(B) $\Rightarrow$ [A [B]]
SUBC(A) PREP(B) ARTD(C) SUBC(D) ADJQ(E) ADJQ(F) $\Rightarrow$ [A [D [E] [F]]

La liste de ces patrons est produite manuellement en fonction des syntagmes nonimaux maximaux produits. La coordination est traitée à part. De manière pratique, lorsqu'aucun patron n'est applicable sur la totalité du syntagme, le système le signale et propose de découper la construction de l'index en utilisant plusieurs patrons. Le système propose toutes les solutions possibles. Une validation manuelle permet d'enrichir la base de patrons.

4.3 Les qualificateurs de relations dans les syntagmes

Notre modèle se concentre sur les relations qui existent entre les éléments d'un syntagme. Ce sont des relations dites locales qui expriment la nature de l'attraction entre la tête du syntagme et ses arguments. Les relations doivent leurs qualifications aux rôles que jouent les arguments dans le syntagme. Dans ce qui suit nous présentons quelques exemples de relations. La qualification de la relation est réalisée après la structuration du syntagme et d'après la nature des termes utilisés.

QUAL : dans une relation de qualification, l'argument joue un rôle de qualifieur de la tête comme c'est le cas lorsque l'argument est un adjectif.

[république QUAL [fédérale]]

SPEC : dans le cas où l'argument est un substantif avec absence de préposition, il s'agit d'une spécification de la tête.

[presse SPEC [papier]]

OBJ, AGT : dans le cas où l'argument est un substantif avec présence de préposition :

- si l'argument dénote une action comme "le vote", "la pollution". Cette dénotation peut être détectée par l'existence d'une forme verbale du substantif de tête. Dans ce cas, les qualificatifs de relations exprimant l'objet et l'agent de l'action peuvent être utilisés (AGT, OBJ). Par exemple, dans la phrase "De plus, lors du vote par l'Assemblée fédérale de la séparation de la Fédération, la coalition a obtenu ...", le syntagme suivant est extrait :

  [vote AGT [assemblée [fédérale]] OBJ [séparation OBJ [Fédération]]]
- si l'argument ne dénote pas une action alors il s'agit d'une spécification de la tête dans "la rubrique du magazine".

  [rubrique SPEC [magazine]]

PART : cette relation importante exprime la relation d'holonymie. C'est la relation qui lie un composant à son objet composé.

Relation inverse : Dans la structure d'un TIS ordre entre la tête et ses modifieurs n'est pas permutable. Il nous faut alors la possibilité d'exprimer, en la surlignant, l'inverse de la relation entre une tête et son modifieur (cf exemple 4.1).

4.4 Traitement statistique

Les termes d'indexation syntagmatiques étant constitués, il reste à leur associer leurs valeurs de pondération.

Décomptage statistique : Dans les SRI classiques, le décomptage statistique concerne les termes d'indexations issus de mots isolés. Il se base sur le calcul du nombre d'occurrences des représentants d'un terme d'indexation. Par exemple, dans le cas d'un RI à base de troncature, le terme d'indexation est une racine de mot et le décomptage se réalise sur le texte issus de cette troncature. Cette valeur représente donc le nombre de mots où le terme racine apparaît. Dans le cas des index syntagmatiques, le décomptage doit se faire au niveau des syntagmes. Pour ce décomptage, il s'agit d'associer un syntagme à son représentant terme d'indexation. Pour être correct, cette association doit tenir compte des phénomènes langagiers comme la variation ou l'anaphore. En particulier, l'anaphore est un piège qui peut fausser un décomptage. Nous avons traité l'anaphore directement au niveau des TIS. Lors du traitement d'un document, nous calculons d'abord tous les TIS possibles à partir des syntagmes maximaux. Nous reprenons ensuite ces TIS pour les comptabiliser. Une anaphore n'est possible qu'entre

un TIS I_1 que l'on peut déduire par jointure et qualification à partir d'un TIS I_2. Cette opération construit un forêt de dépendance de TIS. Nous reprenons le texte initial et lorsque dans le texte nous trouvons I_1 suivi de I_2 dans une certaine fenêtre, nous estimons que I_2 est une référence anaphorique de I_1. Dans ce cas, seul le terme I_1 est comptabilisé (deux fois dans ce cas).

Pondération des syntagmes : Le poids d'un TIS est en fonction de sa pertinence, sa qualité et quantité d'information.

Pertinence d'un TIS : Elle est une fonction de pondération qui reflète l'importance d'un TIS dans un document D. Elle est exprimée en fonction des mesures classiques de tf*idf. Il faut bien noter que le calcul de la fréquence d'un TIS est obtenu par un décomptage des syntagmes associés en tenant compte de l'anaphore elliptique comme indiqué précédemment.

Qualité d'un TIS : Elle doit être reliée à une validation manuelle des structures obtenues. En retour, le système doit pouvoir en déduire de manière heuristique comme l'absence de patron pour ce terme. Nous n'avons pas assez de recul pour cette mesure qui est laissée à 1 dans notre système.

Dénotation d'un TIS : Nous ne pouvons pas disposer directement d'une dénotation des termes que nous utilisons pour l'indexation. En pratique, nous avons mis en place des règles heuristiques pour estimer ces dénotations et surtout les "deltas" de variation des dénotations, puisque ce sont ces variations qui justifient la pondération finale. Il faut avant tout postuler que tous les TIS extraits du corpus ont une dénotation non vide. Il ne semble pas raisonnable de lier la taille de la dénotation à la fréquence d'apparition du terme dans le corpus. Cette fréquence est plus en rapport avec la pertinence de ce terme pour le document. Par contre, les informations portées par les articles définis singulier, peuvent induire une estimation portant sur le singleton. et les articles au pluriel sur un ensemble plus grand. Cette information peut servir d'amorce pour estimer des tailles de dénotations par somme des dénotations élémentaires.

Quantité d'information d'un TIS : Nous avons constaté expérimentalement que cette mesure dépend des catégories grammaticales dont sont issus les atomes d'un TIS. Par exemple, les substantifs expriment plus d'informations que les adjectifs. Nous avons utilisé la formule heuristique suivante :

$$Qinf(I) = \sum_{a \in I} Qinf(Cat(a))$$

où a représente un atome syntagmatique de I, $Cat(a)$ la catégorie grammaticale de cet atome et la fonction $Qinf$ la quantité d'information d'une catégorie grammaticale.

Les mesures que nous proposons ici sont celles de notre première expérimentation. Notre objectif est de compléter notre système pas à pas en essayant de déterminer de manière pratique les traitements linguistiques et les choix de mise en oeuvre du modèle qui *influencent réellement* la qualité du système. Notre démarche expérimentale se fonde alors sur les collections test de l'action Amaryllis de l'AUPELF. Cette démarche vise à mettre en place uniquement les traitements qui favorisent le rappel et la précision pour les requêtes proposées dans ces corpus. Le choix de ces traitements se fait en examinant les requêtes et les documents jugés pertinents. Cette démarche peut sembler non objective car fondée sur un examen des "bonnes réponses" que le système doit indiquer. En fait, la difficulté de mise en place de traitements linguistiques adéquats est telle, qu'une démarche plus "en aveugle" nous fait prendre trop de directions. Notre objectif final est de déterminer les choix de traitements qui seront fortement dépendants des référentiels que sont nos collections de test, de ceux généralisables.

5 Conclusion

Nous avons proposé un modèle général décrivant les éléments à prendre en compte lors de l'utilisation des syntagmes comme support des termes d'indexation. L'originalité de notre approche tient dans l'introduction dissociée, au niveau de la modélisation, des notions de pertinence, qualité et quantité d'information d'un terme d'indexation. La plupart des modèles de RI fondent leur calcul de correspondance, au mieux sur une vision probabiliste ou logique, et au pire sur un calcul donné à priori. Nous pensons que les notions de dénotation et de quantité d'information, doivent être le fondement de tout SRI désirant se rapprocher du sens et s'éloigner du signal. Pour cela, nous avons associé à notre langage d'indexation une sémantique dénotationnelle qui, bien qu'embryonnaire, nous fournit une justification pour définir la notion de coût d'une correspondance.

Ces idées sont en cours d'expérimentations dans notre système **iota** . Ce système est issu de divers travaux de notre équipe notamment [CHIA86], [KERK84], [BRUA87],[BRUA97]. Il est actuellement capable d'extraire la terminologie d'un corpus à l'aide de patrons appliqués après analyse morphologique. Cette terminologie est structurée par des relations de dépendances extraites automatiquement (c.f. [HADD00]). Cette extraction a été évaluée au sein de l'action ARC3 de l'AUPELF, et notamment par l'INRA sur un de leurs corpus. Cette dernière évaluation montre que la précision de cette extraction est très bonne car sur les termes que nous avons proposé, seul 9% doivent être considérés comme du bruit. Ces résultats nous laissent à penser que l'intégration de ces termes de bonne qualité dans le processus d'indexation décrit dans cet article, doit faire apparaître une variation significative de la qualité globale du SRI. L'implantation de la correspondance s'inspire des travaux sur la mise en oeuvre de la projection dans le cadre d'un SRI basé sur les graphes conceptuels décrite dans [OUNI98].

Nous remercions A. Lacombe, Responsable du Service Linguistique de l'INRA, pour nous avoir proposé une évaluation manuelle des termes que notre système **iota** est actuellement capable d'extraire à partir de leurs corpus.

Références

[ARAM98] A. T. Arampatzis, T. Tsoris, C. H. A.Koster, and Th. P. Van Der Weide. Phrase-based information retrieval. *Information Processing and Management*, 34(6) :693–707, 1998.

[BESA99] R. Besancon, M. Rajman, and J.C. Chappelier. Textual similarities based on a distributional approach. In *International Workshop on Similarity Search (IWOSS99), Firenze, Italy*, Septembre 1999.

[BRUA87] Marie-France Bruandet. Outline of a knowledge base model for an intelligent information retrieval system. In *ACM-SIGIR, New Orleans*, pages 33–43, 1987.

[BRUA97] Marie-France Bruandet, Jean-Pierre Chevallet, and Francois Paradis. Construction de thesaurus dans le systeme de recherche d'information iota : application a l'extraction de la terminologie. In *1eres Journees Scientifiques et Techniques du Reseau Francophone de l'Ingerierie de la Langue de l'AUPELF-URF , Avignon - France*, pages 537–544, Avril 1997.

[CARB00] J. P. Carballo and T. Strzalkowski. Naturel language information retrieval : progress report. *Information Processing and Management*, 36(1) :155–178, 2000.

[CHEV91] Jean-Pierre Chevallet. E.l.e.n. : Un système d'interrogation d'une base de logiciels. In *Congrès INFORSID 1991*, pages 1–20, Juin 1991.

[CHEV92] Jean-Pierre Chevallet. *Un modèle logique de recherche d'information appliqué au formalisme de graphes conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels*. PhD thesis, Université Joseph Fourrier, Mai 1992.

[CHEV97] J. P. Chevallet and M. F. Bruandet. Impact de l'utilisation de multi termes sur la qualité des réponses d'un système de recherche d'information à indexation automatique. In *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information Collection UL3 Lilles,ISBN 2-84467-002-4, Lille, France*, pages 223–238, Octobre 1997.

[CHIA86] Y. Chiaramella, B. Defude, M.F. Bruandet, and D. Kerkouba. Iota : a full test information retrieval system. In *ACM conference on research and development in information retrieval., Pisa, Italy*, pages 207–213, Septembre 1986.

[CHUR90] W. K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1) :22–29, 1990.

[CROF91] W. B. Croft, H. R. Turtle, and D. Lewis. The use of phrases and structured queries in information retrieval. In *Proc. of ACM SIGIR*, pages 32–45, Octobre 1991.

[DAIL94] B. Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7, 1994.

[DEBI82] F. Debili. Analyse syntaxico-semantique fondee sur une acquisition automatique de relations lexicales-semantiques, 1982. Habilitation à diriger des recherches.

[FAGA87] J. L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval : A Comparison of Syntactic and Non-Syntactic Methods*. PhD thesis, Department of Computer Science, Cornell University, Ithaca, New York, 1987.

[FAGA89] J. L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Scienc*, 40(2) :115–132, 1989.

[FARR80] J. Farradane. Relational indexing. part i. *Journal of Information science*, 1(5) :267–276, 1980.

[GARC98] D. Garcia. *Analyse automatique de textes pour l'organisation causale des actions, réalisation du système informatique COATIS*. PhD thesis, Université Paris 4, 1998.

[GREF92] G. Grefenstette. Use of suntactic context to produce term association lists for text retrieval. In *Annuel International ACM SIGIR conference on research and development in information retrieval (SIGIR'92), ACM press, Copenhagen, Denmark*, pages 89–97, Juin 1992.

[HADD00] M.H. Haddad, J.P. Chevallet, and M.F. Bruandet. Relations between terms discovered by association rules. In *4th European conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, Lyon France*, septembre 2000.

[HEAR92] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France*, Juillet 1992.

[JACK96] A. Jackiewicz. L'expression lexicale de la relation d'ingrédience (partie-tout). *Faits de Langues*, (7) :53–62, 1996.

[JACQ97] C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus., 1997. Habilitation à diriger des recherches, Université de Nantes.

[KERK84] D. Kerkouba. *Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles. Application à un corpus technique.* PhD thesis, Institut National Polytechnique de Grenoble, 1984.

[KHOO97] C. S. Khoo. The use of relation matching in information retrieval. *Singapore Libraries*, 26(1) :3–22, 1997.

[KRAA98] W. Kraaij and R. Pohlmann. Comparing the effect of syntactic vs. statistical phrase indexing strategies for dutch. In *In Christos Nicolaou and Constantine Stephanidis, editors, Proceedings of Second European Conference on Research and Advanced Technology for Digital Libraries ECDL'98*, pages 605–614, 1998.

[MITR97] M. Mitra, C. Buckley, A. Singhal, and C. Cardi. An analysis of statistical and syntactic phrases. In *In Proceedings of RIAO'97 computer-Assisted Information Searching on Internet, McGill University, Montreal*, pages 200–214, Juin 1997.

[MORI99] E. Morin. *Extraction de lien sémantique entre termes à partir de corpus de textes techniques*. PhD thesis, Institut de recherche en informatique de Nantes, Decembre 1999.

[NIE90] J. Nie. *Un modèle logique général pour les systèmes de recherche d'information. Application au prototype RIME*. PhD thesis, Université Joseph Fourrier, Juillet 1990.

[OUNI98] I. Ounis. *Un modèle d'indexation relationnel pour les graphes conceptuels fondé sur une interprétation logique*. PhD thesis, Université Joseph Fourrier, Février 1998.

[PARA96] F. Paradis. *Un modèle d'indexation pour les documents textuels structurés*. PhD thesis, Université Joseph Fourier, Novembre 1996.

[SALT71] Gerard Salton. *The SMART retrieval system : experiments in automatic document processing*. Prentice Hall, 1971.

[SEBA94] F. Sebastiani. Probabilistic terminological logic for modelling information retrieval. In *In W. Bruce Croft and Cornelis J. van Rijsbergen (eds.), Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin*, pages 122–130, Juillet 1994.

[SIMO00] J. L. Simoni. *Accès à l'information à l'aide d'un graphe de termes construit automatiquement (Intégration de l'interrogation et de la navigation)*. PhD thesis, Université Paris 7, Janvier 2000.

[SMEA88] F. Smeaton and C. V. Rijsbergen. Experiments on incorporating syntactic processing of user queries into a document retrieval strategy. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval, ed. Y. Chiaramella, Grenoble, France*, pages 31–51, 1988.

[SMEA94] A. Smeaton, R. O'Donnell, and F. Kelledy. Indexing structures derived from syntax in trec-3 : System description. In *In Donna K. Harman, editor, The Third Text REtrieval Conference (TREC-3)*, pages 55–67, 1994.

[SOWA84] J.F. Sowa. *Conceptuel Structures-Information Processing in Mind and Machine*. Addison-Wesley Publishing Company, 1984.

[STRZ94] T. Strzalkowski and J. P. Carballo. Natural language information retrieval : Trec-3 report. In *In Donna K. Harman, editor, The Third Text REtrieval Conference (TREC-3)*, pages 39–54, 1994.

[STRZ95] T. Strzalkowski and J. P. Carballo. Natural language information retrieval : Trec-4 report. In *In Donna K. Harman, editor, The Fourth Text REtrieval Conference (TREC-4)*, pages 245–258, 1995.

[STRZ96] T. Strzalkowski, J. Wang F. Lin, L. Guthrie, J. Leistensnider, J. Wilding, J. Karlgren, T. Straszheim, and J. Carballo. Natural language information retrieval : Trec-5 report. In *In E. Voorhees and Donna K. Harman, editor, The Fifth Text REtrieval Conference (TREC-5)*, pages 291–334, 1996.