Corpus issus du Web : analyse des pertinences thématique et informationnelle

Louisette Emirkanian[1], Christophe Fouqueré[2], Fabrice Issac[3]

[1]Département de linguistique et de didactique des langues – UQAM – CP 8888 succursale 'Centre-ville'
Montréal (QC) H3C 3P8 – Canada – emirkanian.louisette@uqam.ca
[2]LIPN, CNRS UMR 7030 – Université Paris 13 – Villetaneuse – France
[3]LLI, CNRS UMR 7546 – Université Paris 13 – Villetaneuse – France

Abstract

The purpose of our study is to identify useful reformulation strategies for queries used in information retrieval. As part of an on-going research project, only one aspect of which is discussed here, we conducted a detailed study of five basic queries and of their variants obtained by morphological and synonymic enrichment. Although in general the use of synonymic and morphological variants improves recall, we have found that specifying the syntactic link between the terms in the query improves precision. We first describe the methodology used to assemble the corpus and comment on the data collected. We then examine the syntactic links between query terms and their correlation with thematic and informational relevance.

Résumé

Notre étude a pour objectif principal de mettre à jour des pistes pour la spécification de mécanismes de reformulation de requêtes facilitant la recherche d'information. Dans un projet de recherche dont nous présenterons ici un aspect, nous avons étudié de façon détaillée les résultats de cinq requêtes de base et de variantes obtenues par enrichissement morphologique et synonymique. Alors qu'en général la prise en compte de variantes synonymiques ou morphologiques permet d'améliorer le rappel, nous tenterons de montrer que la spécification du lien syntaxique entre les termes apporte une plus grande précision. Après avoir décrit la méthodologie de constitution de corpus issus du Web et commenté les données recueillies, nous examinerons les liens syntaxiques entre les termes des requêtes et leur rapport avec les pertinences thématique et informationnelle.

Mots-clés : recherche d'information, Web, pertinence thématique et informationnelle, syntaxe

1. Introduction

Nous sommes confrontés tous les jours à une masse importante d'informations. Le problème majeur qui se pose alors est d'être capable de repérer, dans cette masse d'informations, celle qui nous sera utile, celle qui répondra le plus précisément possible à notre besoin d'information. Le repérage d'information (RI) sur le Web présente des défis particuliers, en raison de la grande variété de domaines, genres et styles des documents. Les résultats d'une recherche sont souvent très nombreux et peu pertinents, dans le cas de requêtes constituées d'une courte liste de mots isolés, ou à l'inverse, trop peu nombreux dans le cas où l'on spécifie une requête par une coordination d'un trop grand nombre de termes, ou par l'utilisation d'expressions trop précises.

Jusqu'à récemment, les recherches dans le domaine du RI et celles dans le domaine du traitement automatique des langues n'entretenaient que peu de liens, chaque domaine développant

des outils spécifiques. Depuis peu, la convergence des deux domaines s'est faite et elle a porté ses fruits dans chacun d'eux (Jacquemin, 2000 ; Jacquemin et Zweigenbaum, 2000 ; Gaussier *et al.*, 2000 ; Strzalkowski, 1995 et 1999 ; Spärck Jones, 1999 ; Woods *et al.*, 2000). Des travaux récents tendent à prouver que l'utilisation de techniques de TAL en morphologie, syntaxe et sémantique permet d'améliorer la performance des systèmes de RI tant au niveau de l'indexation qu'à celui du repérage.

Ces connaissances linguistiques sont utilisées pour le découpage en unités linguistiques, pour l'étiquetage, l'analyse en constituants (et l'indexation de syntagmes), la reconnaissance de termes complexes, la désambiguïsation en contexte et l'extension et la reformulation de requêtes (Bouillon et *al.*, 2000).

Notre étude a pour objectif principal de mettre à jour des pistes pour la spécification de mécanismes de reformulation de requêtes facilitant la recherche d'information.

Dans un projet de recherche[1] dont nous présenterons ici un aspect, nous avons étudié de façon détaillée les résultats de cinq requêtes de base et de variantes obtenues par enrichissement morphologique et synonymique. Nous avons également porté une attention particulière à la dimension syntagmatique ; cette partie du projet fera l'objet de la présentation proposée. Alors qu'en général la prise en compte de variantes synonymiques ou morphologiques permet d'améliorer le rappel, nous tenterons de montrer que la spécification du lien syntaxique entre les termes apporte une plus grande précision.

Nous nous attacherons d'abord à décrire la méthodologie de constitution de corpus que nous avons utilisée en mettant l'accent sur les spécificités du Web ; nous commenterons ensuite les données recueillies pour nos cinq jeux de requêtes, puis nous examinerons les liens syntaxiques entre les termes et leur rapport avec la pertinence thématique et informationnelle.

2. Constitution du corpus & méthodologie

À partir de 5 besoins en information de type X Y, VOYAGE TIBET (où VOYAGE constitue X et TIBET, Y), « FUITE DES CERVEAUX », ÉTATS UNIS, VOL LUNE, MISSION ESPACE, PROMENADE PARIS, nous avons constitué 5 corpus de pages html extraites du Web (désormais corpus FUITE, TIBET, LUNE, ESPACE et PARIS). Des requêtes de la forme « X (NEAR préposition) NEAR Y » ont été exécutées pour constituer chacun des corpus. Ces requêtes ont été construites par variation à partir du besoin initial. Nous avons recherché les pages du Web contenant les mots de base dans l'ordre initial, dans un voisinage, avec ou sans préposition, avec des variations morphologiques et sémantiques sur X et Y ainsi que des variations sur la préposition. 31 prépositions ont été pour cela utilisées (la totalité), de même toutes les variations morphologiques ont été essayées, enfin nous avons identifié tous les synonymes des mots pleins ; par exemple dans le cas de la requête VOYAGE TIBET, *voyage* est remplacé par *voyages*, *séjour*, *voyager*, *trek*, etc. et *Tibet* par *tibétain*, *tibétaines*, etc. Dans le cas où le Y est un adjectif, la préposition pourra être présente ou non (*promenade parisienne* ou *promenade dans les arrondissements parisiens*). Nous avons un total de 14840 requêtes.

2.1. *Principe*

À partir d'une ou plusieurs requêtes, exprimées dans un langage propre à un moteur de recherche, un ensemble de pages est récupéré puis stocké et transcodé. L'outil est en fait

[1] Ce projet a été financé par la coopération franco-québécoise (Ministère des Relations Internationales du Québec et Ministère des Affaires Étrangères de la France).

constitué d'un ensemble modulaire de sous-programmes écrits en langage PERL.

Un premier sous-programme permet de générer une liste de requêtes pour un moteur spécifique à partir de fichiers de configurations. Il effectue toutes les combinaisons possibles entre les différents éléments (mots-clés ou variantes de ceux-ci) de la requête et y associe un ou plusieurs connecteurs spécifiant en particulier la proximité des mots-clés dans la page recherchée. Un deuxième composant interroge alors un moteur de recherche et récupère toutes les adresses des pages correspondant aux différentes requêtes. Toutefois, la liste de pages réellement récupérables n'est pas exhaustive : ainsi, à partir d'une requête sur le moteur de recherche Altavista il n'est possible de récupérer qu'au plus 1010 URL. Les adresses apparaissant en double sont ensuite éliminées. Enfin, un dernier composant récupère la page elle-même (quand celle-ci est effectivement disponible !) et transcode le résultat dans un format XML apte à intégrer des informations supplémentaires (adresse de la page, date de récupération, requêtes associées, etc.). Hormis les quelques cas de fichiers de style (type css), il ne s'agit que de pages au format initial html. L'ensemble de ces pages constitue une photo instantanée des pages récupérables via un moteur de recherche spécifique (en l'occurrence AltaVista).

2.2. *Choix du moteur*

Le module essentiel au système a pour tâche d'analyser le résultat d'une requête d'un moteur de recherche. Ce résultat est une ou plus généralement plusieurs pages Web. Nous avons choisi pour cela un mécanisme capable tout à la fois d'émettre une requête, de récupérer les pages résultats, et d'extraire de ces pages les URL. L'outil est assez souple pour pouvoir être adapté à de nombreux moteurs. Nous avons choisi d'utiliser Altavista plutôt que Google, qui offre actuellement sans doute le meilleur classement et une meilleure couverture, car ce moteur permet de gérer plus aisément la notion de proximité entre mots clés. Ainsi l'opérateur « NEAR » ajoute à l'opérateur « AND » la contrainte d'une distance maximale de 10 mots entre deux mots-clés d'une requête.

2.3. *Corpus résultant*

Le résultat de l'aspiration des corpus est résumé dans le tableau ci-dessous, où nous avons précisé pour chaque corpus créé le nombre total de mots, la taille (en mégaoctets) du corpus, et enfin le nombre de pages de sites Web qui auront pu être analysées. Les deux dernières colonnes indiquent le nombre moyen de mots par page et son écart type. Il y a lieu de noter la variabilité très importante de ces données selon le thème initial. Cette variabilité n'est pas due à une limitation de nos outils mais bien à la variabilité intrinsèque du Web (pages journalistiques ou économiques, pages à caractère touristique, pages personnelles, etc.).

La variabilité en termes de nombre de pages reflète l'importance d'utiliser le Web comme véhicule de l'information liée à ce thème. Ainsi le thème PARIS renvoie à de très nombreuses pages à caractère journalistique, le thème ESPACE à une très forte volonté de diffusion électronique de l'information scientifique et technique. Si le nombre moyen de mots par page est sensiblement égal, cela résulte du désir des concepteurs de sites Web de découper l'information en pages de taille raisonnable tant relativement à la lecture qu'afin d'éviter un délai de chargement trop important. Toutefois, l'écart type montre que l'affirmation précédente laisse place à de très nombreuses exceptions.

CORPUS	TAILLE (MO)	Nb mots	Nb de pages	Nb de mots / page	Écart Type du nb de mots / page
FUITE	17	908 032	187	3 883	6 705
ESPACE	407	24 905 110	8 291	2 042	6 288
LUNE	187	10 978 288	3 522	2 444	9 186
TIBET	70	3 829 657	1 507	1 779	3 751
PARIS	1 058	60 588 606	20 017	2 364	8 732
TOTAL	***1 739***	***101 209 693***	***33 523***	***3 019***	

Tableau 1. Caractéristique des corpus

2.4. *Post traitement*

Un post traitement est effectué afin de repérer dans le texte :

1. les phrases contenant une séquence correspondant à une des requêtes ayant permis de récupérer la page,
2. les phrases contenant une séquence correspondant à une des requêtes alors que celle-ci n'avait pas permis de récupérer la page,
3. les mots de la requête n'appartenant pas à une séquence.

Pour ce faire la page HTML est transcodée en XML. La structure permet de repérer, outre les informations contenues dans le résumé, les phrases et la liste des requêtes dont le motif syntaxique est inclus dans la phrase. Pour être repérée, une phrase doit avoir en son sein les différents éléments d'une requête dans l'ordre initial. Il est à noter que le découpage en phrases est un exercice non trivial sur ce type de données, par conséquent nous donnons à la notion de phrase un sens relativement lâche. Toutefois, cette étape de découpage en phrases est indispensable dès lors que nous souhaitons prendre en compte la préservation de la structure syntaxique présente initialement dans la requête.

3. Analyse

Afin d'analyser le corpus obtenu du point de vue de la pertinence thématique et informationnelle, nous avons élaboré une interface graphique permettant de choisir son corpus, de le parcourir puis de l'annoter.

3.1. *Pertinence informationnelle*

La *pertinence informationnelle* est la pertinence classique du RI, qui identifie les documents répondant complètement ou partiellement au besoin de l'utilisateur, sans nécessairement établir de liens avec la formulation particulière de la requête. Étaient jugés pertinents au niveau informationnel, pour le corpus FUITE, les documents traitant de la fuite des cerveaux vers les États-Unis ; pour le corpus TIBET, tous les documents qui donnaient des informations pratiques pour l'organisation d'un voyage touristique au Tibet ; pour le corpus ESPACE, tous les documents donnant des informations scientifiques sur les missions dans l'espace intersidéral ; pour le corpus LUNE, tous les documents donnant des informations scientifiques ou des détails sur les vols (effectués ou à venir) vers la lune ; enfin pour le corpus PARIS, tous les documents donnant des informations pratiques permettant d'organiser des promenades dans Paris. Nous avons retenu trois niveaux de pertinence, la seule justification de cette extrême limitation est de nous permettre d'avoir des outils qui nous indiqueront les grandes tendances des résultats :

- totalement pertinent (**TP**) : la page concerne en totalité le sujet concerné par la requête. On peut y parler ainsi de vol spatial (corpus ESPACE), de balade en bateaux-mouches dans Paris (corpus PARIS) ;
- partiellement pertinent (**PP**) : la page contient au moins un paragraphe répondant à la requête. C'est le cas par exemple dans les sites d'agences de voyages (corpus TIBET), quand, à coté de safaris kenyans, on mentionne des trekkings au Tibet ;
- non pertinent **(NP)** : la page ne répond pas à la requête. Cela ne signifie pas que les mots constituant la requête (voire le schéma associé) ne soient pas présents dans la page. Il peut s'agir d'un emploi figuratif du thème de la requête.

Pour en avoir fait l'expérience, nous savons que deux facteurs déterminent la "réussite" d'une requête : la pertinence de la page récupérée relativement à l'objectif initial de la requête, la pertinence de l'ordre dans lequel les pages récupérées nous sont exposées.

Le premier de ces facteurs est fondamentalement subjectif : des pages *a priori* très différentes sur le fond peuvent paraître pertinentes pour l'utilisateur qui aura effectué sa recherche. Il en est ainsi par exemple pour le corpus ESPACE, s'agit-il d'aéronautique ? de littérature ? d'ésotérisme ? d'espace architectural ? urbain ? Nous avons pu retrouver toutes ces thématiques lors de l'analyse manuelle que nous avons faite du corpus. Dans cette étude, il n'était pas envisageable de limiter la thématique dès lors que cette limitation n'apparaissait pas dans la spécification de la requête. Toutefois, c'est une optique qu'il est important d'envisager dans cette situation. Pour en revenir à notre projet, nous avons volontairement simplifié l'étude du contenu informationnel des pages.

À partir de cette analyse informationnelle, et des rangs associés à la page par le moteur de recherche, il nous était dès lors possible d'évaluer précisément les divers aspects entrant en jeu lors de la reformulation de requêtes. Il s'agit en effet d'évaluer le rappel, c'est à dire la proportion de pages pertinentes récupérées par chaque requête, et la précision, c'est à dire la proportion de pages pertinentes relativement à l'ensemble des pages récupérables. Plus précisément nous avons cherché à évaluer l'impact de la correction syntaxique de la requête ainsi que celui lié à la détermination de la préposition dans ce cadre. Dans un deuxième temps, nous avons intégré à ce calcul de pertinence le rang associé à la page en cherchant à savoir dans quelle mesure les algorithmes utilisés dans les moteurs de recherche correspondaient à nos premières évaluations. Il va de soi que ces informations sont largement dépendantes du moteur de recherche choisi pour l'expérience. Ainsi la technologie de ramassage, les calculs d'indexation du moteur et les critères utilisés par celui-ci pour le classement des réponses constituent-ils autant de biais relativement à notre expérience. Toutefois, nous considérons que le nombre de pages retenues (1010 au maximum) relativise ce biais.

Les requêtes syntaxiquement incorrectes donnent des résultats non pertinents : le rappel est très faible. C'est notoirement le cas pour le corpus FUITE DES CERVEAUX qui, avec 8462 requêtes pour seulement 182 pages distinctes récupérables, contient un nombre impressionnant de requêtes syntaxiquement incorrectes ne permettant de ne récupérer aucune page. De fait, 95 requêtes seulement sont de ce point de vue "utiles", chaque page étant récupérable en moyenne par 2 requêtes. Dans ce corpus, les prépositions syntaxiquement invalides (*chez les États-Unis*) sont des échecs du point de vue de la récupération. Toutefois, parce que le schéma associé n'est pas toujours respecté dans la phrase réelle, certaines requêtes syntaxiquement incorrectes permettent de récupérer des pages pertinentes non récupérées par d'autres requêtes. Ces cas sont suffisamment rares pour que la sélection de requêtes sur le critère de plausibilité syntaxique soit justifiée. Le cas des requêtes sans préposition reste particulier : le rappel y est (naturellement) très important, mais la précision est assez faible.

Enfin, la grande majorité des pages récupérées le sont aussi par le biais de requêtes avec prépositions. Les requêtes avec prépositions syntaxiquement pertinentes donnent des résultats au moins partiellement pertinents. Dans le cas du corpus ESPACE par exemple, seules 58 des 98 requêtes permettent la récupération de pages.

Enfin, nous avons cherché à analyser la "vitesse de couverture" des pages totalement ou partiellement pertinentes à partir du « facteur de qualité » associé aux requêtes. En associant degré de pertinence et rang, nous pouvons en effet calculer un facteur de qualité associé à la requête. Ce facteur est égal, pour une requête donnée, à la somme des pertinences pondérées par le logarithme du rang divisé par la somme des logarithmes des rangs. L'utilisation du logarithme permettant de minimiser l'effet du rang par rapport à la pertinence. Un résultat se rapprochant de 1 montre une adéquation entre le classement du moteur de recherche et notre propre classement. Pour cela, nous avons réanalysé l'ensemble des résultats en tenant compte du facteur de qualité. Afin de comprendre dans quelle mesure les requêtes pertinentes permettaient d'obtenir les résultats, nous avons cherché à couvrir l'ensemble des pages à partir des pages accessibles des requêtes en commençant par les requêtes de degré de qualité le plus élevé. Il est apparu clairement, pour les cinq thèmes étudiés, que 5 % des requêtes suffisent à récupérer 50 % des pages totalement ou partiellement pertinentes. Cela justifie indirectement la technique de la reformulation de requête. En effet, sous réserve de définir judicieusement ces 5 % de requêtes, la reformulation nous permet à peu de frais, i.e. modulo l'analyse par le moteur de recherche de ces requêtes supplémentaires, d'obtenir une liste significative (pour l'utilisateur) de pages jugées pertinentes. Qui plus est, la reformulation avec précision syntaxique (p.e. ajout de la préposition) non seulement permettra un rappel significatif, mais encore aura une précision correcte (eu égard à la masse de documents contenus dans le Web).

3.2. *Pertinence thématique locale*

La pertinence thématique locale est quant à elle rattachée à un passage contenant les termes de la requête, et non au document dans son ensemble, et identifie si ces termes sont employés dans un sens compatible avec le besoin en information et s'ils constituent l'objet du discours rattaché au passage. Par exemple, pour le corpus TIBET, *voyage au Tibet* sera jugé thématiquement pertinent au même titre que *voyage en terre tibétaine*, ou encore *voyage qui permet de découvrir le Tibet* ; en revanche, *marche pour le Tibet* sera jugé non pertinent.

Nous avons isolé un sous-ensemble représentatif pour chacun des 5 besoins en information. Nous n'avons retenu que les occurrences dans lesquelles les éléments X (préposition) Y apparaissent dans l'ordre de la requête et au sein de la même phrase. Nous avons ainsi obtenu 3174 occurrences réparties dans 1873 pages différentes**.** En effet, nous trouvons fréquemment des documents contenant plusieurs occurrences des termes de la requête. Ces occurrences se répartissent comme suit dans les différents corpus.

	Nombre de pages	**Nombre d'occurrences**	**Proportion d'occurrences pertinentes au niveau thématique**	**Répartition des occurrences au niveau informationnel**		
				TP	**PP**	**NP**
FUITE	92	116	79 %	23 %	56 %	21 %
TIBET	583	1 051	62 %	46 %	13 %	41 %
LUNE	492	776	70 %	31 %	15 %	54 %
ESPACE	282	540	68 %	34 %	38 %	28 %
PARIS	424	691	43 %	9 %	23 %	68 %
TOTAL	***1 873***	***3 174***	***62 %***	***31 %***	***22 %***	***47 %***

Tableau 2. Évaluation des pertinences informationnelle et thématique

Si l'on compare les proportions d'occurrences pertinentes au niveau thématique à celles des occurrences totalement ou partiellement pertinentes au niveau informationnel, on observe des similitudes sauf dans le cas du corpus LUNE. En effet, dans ce corpus, de nombreuses occurrences du type '*voyage dans la lune, vers la lune, sur la lune*' étaient correctes au niveau thématique mais le document dans lequel elles se trouvaient traitait d'œuvres littéraires ou musicales.

Il est également intéressant d'évaluer la correspondance entre d'une part les occurrences pertinentes au niveau thématique et d'autre part les occurrences totalement pertinentes et partiellement pertinentes au niveau informationnel, en d'autres mots, de vérifier s'il y a correspondance entre pertinence thématique et informationnelle**.** Sur l'ensemble du corpus, sur les 1967 occurrences pertinentes au niveau thématique, 72 % sont totalement ou partiellement pertinentes au niveau informationnel ; la proportion est inversée pour les séquences non pertinentes au niveau thématique ; en effet, seulement 20 % d'entre elles sont totalement ou partiellement pertinentes au niveau informationnel.

3.3. *Liens syntaxiques et pertinences informationnelle et thématique*

Nous avons réparti les occurrences obtenues en six catégories de liens syntaxiques ; elles sont liées au fait d'une part que nous autorisons différentes catégories pour X (noms, verbes) ou Y (noms, adjectifs) et d'autre part que nos requêtes contiennent l'élément NEAR.

- *N1-SP* :

  Y est dans un syntagme prépositionnel directement rattaché à X (*fuite des cerveaux vers les États-Unis, balade dans Paris, mission dans l'espace, vol vers la lune*, etc.)

- *N1-adj* :

  Le terme Y est un adjectif qui modifie directement X (*voyage tibétain, mission lunaire, spatiale, balade parisienne*, etc.)

- *P-N2* :

  Le terme Y est dans un syntagme prépositionnel non relié à X (*Au cours du vol de Gemini 4, Edward White effectua lui aussi une sortie dans l'espace, le 3 juin 1965.)* (http://www.multimania.com/msegret/laconqu.htm)

- *AUCUN LIEN* :

  Les trois termes X, préposition, Y ne sont pas reliés syntaxiquement. (*Un voyage très complet qui permet de découvrir à la fois le Tibet central et l'ancien royaume de Gugé*.) (http://tirawa.com/voyages/himalaya/tibet/tibet-703-lhassa-mont-kailash/index.html)

- *V-SP :*

  Le terme X est un verbe ; le terme Y est dans un syntagme prépositionnel rattaché au verbe (*Voyager au Tibet*).

- *V-SN :*

  Le terme X est un verbe ; Y est dans un syntagme nominal argument du verbe (*Ils ont parcouru le Tibet*).

Le tableau suivant présente une répartition des occurrences selon le schéma syntaxique dans lequel les éléments de la requête apparaissent.

Structures	Nombre total d'occurrences	Proportion d'occurrences thématiquement correctes	Répartition des occurrences au niveau informationnel		
			TP	MP	NP
N1-SP	1 128	77 %	38 %	22 %	40 %
N1-ADJ	280	98 %	35 %	31 %	34 %
V-SN	40	95 %	13 %	37 %	50 %
V-SP	213	79 %	22 %	25 %	53 %
P-N2	890	46 %	30 %	18 %	52 %
Aucun lien	623	33 %	24 %	19 %	57 %
TOTAL	***3 174***	***62 %***	***31 %***	***22 %***	***47 %***

Tableau 3. Évaluation des pertinences thématique et informationnelle selon les structures

Si l'on regroupe les cas où les éléments X et Y sont liés syntaxiquement (i.e. N1-SP, N1-ADJ, V-SN, V-SP) et qu'on les oppose au cas P-N2 (le syntagme prépositionnel n'est pas rattaché au N) et au cas où il n'y a aucun lien entre les termes de la requête, on constate que la proportion d'occurrences pertinentes au niveau thématique est plus élevée lorsque les éléments de la requête ont un lien syntaxique. Pour la pertinence informationnelle, les différences sont moins marquées. De plus, on note que lorsque l'élément X est un verbe, la proportion d'occurrences très pertinentes ou partiellement pertinentes au niveau informationnel est faible.

Nous avons montré, par ailleurs (Emirkanian et Chieze, à paraître), que si l'emploi de dérivés verbaux (pour l'élément X) et adjectivaux (pour l'élément Y) améliorait grandement le rappel, cela se faisait en général au détriment de la précision. Si nous laissons de côté les dérivés verbaux pour X et les dérivés adjectivaux pour Y, et que nous n'examinons que les cas où les termes X et Y sont des noms (voyage(s), trek(king), vol(s), mission(s), promenade(s), balade(s), Tibet, Paris, espace, etc.), nous obtenons les résultats suivants pour les trois schémas 'X préposition Y', 'X … préposition Y' et 'X … préposition … Y'. On remarque que la spécification du lien syntaxique entre les termes de la requête augmente à la fois la pertinence thématique et la pertinence informationnelle.

Structures	Nombre total d'occurrences	Proportion d'occurrences thématiquement correctes	Proportion d'occurrences très pertinentes au niveau informationnel
N1-SP	995	72 %	37 %
P-N2	590	45 %	30 %
Aucun lien	284	30 %	20 %

Tableau 4. Importance du lien syntaxique

4. Conclusion

L'analyse que nous avons menée a été effectuée en parallèle sur 5 corpus de pages extraites du Web. L'ampleur autant que la variabilité des corpus obtenus nous a permis d'effectuer à la fois des analyses automatiques (relatives à la précision et à la couverture de chaque type de requête) et des analyses manuelles (sur la pertinence thématique et informationnelle des pages extraites, sur les structures syntaxiques). Nous avons montré dans cet article les liens qui pouvaient exister entre pertinence thématique et informationnelle. Enfin, les résultats de cette étude indiquent que la spécification du lien syntaxique entre les termes de la requête permet

d'accroître la pertinence au niveau thématique. Il est important de noter que nous retrouvons cette augmentation au niveau de la pertinence informationnelle. La spécification du lien syntaxique par une préposition constitue dès lors l'une des pistes possibles pour la reformulation de requêtes dans l'objectif d'améliorer sensiblement la précision des réponses.

Références

Bouillon P., Fabre C., Sébillot P. et Jacqmin L. (2000). Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL*, vol. (41/2) : 367-393.

Emirkanian L. et Chieze E. (à paraître) Variations morphologiques, syntaxiques, sémantiques et Repérage d'Information sur le Web. *Revue Québécoise de Linguistique*.

Gaussier E., Grefenstette G., Hull D. et Roux Cl. (2000). Recherche d'information en français et traitement automatique des langues. *TAL*, vol. (41/2) : 473-493.

Jacquemin Chr. (éd.) (2000). *Traitement automatique des langues pour la recherche d'information*. *TAL*, vol. (41/2).

Jacquemin Chr. et Zweigenbaum P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In Le Maître J., Charlet J. et Garbay C. (Eds), *Le document en sciences du traitement de l'information.* Cepadues : 71-109.

Spärck Jones K. (1999) The role of NLP in text retrieval. In Strzalkowski T., *Natural Language Information retrieval*. Kluwer : 1-24.

Strzalkowski T. (1995). Natural language information retrieval. *Information Processing & Management*, vol. (31/3) : 397-417.

Strzalkowski T. (Ed.) (1999). *Natural Language Information Retrieval*. Kluwer.

Woods W.A., Bookman L.A., Houston A., Kuhns R.J., Martin P. et Green S. (2000). Linguistic Knowledge can improve information retrieval. In *Proceedings of the 6th Applied Natural Language Processing Conference*.