Revue québécoise de linguistique

Corpus issus du Web : constitution et analyse informationnelle

Christophe Fouqueré and Fabrice Issac

Volume 32, Number 1, 2003

TALN, Web et corpus

URI: https://id.erudit.org/iderudit/012246ar
DOI: https://doi.org/10.7202/012246ar

See table of contents

Publisher(s)

Université du Québec à Montréal

ISSN

0710-0167 (print)
1705-4591 (digital)

Explore this journal

Cite this article

Fouqueré, C. & Issac, F. (2003). Corpus issus du Web : constitution et analyse informationnelle. *Revue québécoise de linguistique*, *32*(1), 111–134. https://doi.org/10.7202/012246ar

Article abstract

Compared to other information sources (technical documents, news items), the Web offers almost unlimited access to an formation of all kinds. This advantage may be lost if relevant information is buried in the mass of texts. Our research attemps to evaluate how automated language analysis techniques can aid in the search for information in unorganized textual databases. Specifically our study examines the reformulation of search strings. We outline the method for constructing our corpus and then analyse the relevance of web pages retrieved when the initial search string is varied.

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal

CORPUS ISSUS DU WEB : CONSTITUTION ET ANALYSE INFORMATIONNELLE

Christophe Fouqueré
Fabrice Issac
Université de Paris XIII

1. Accès au web par reformulation de requêtes

Nous avons tous été confrontés au problème du choix des mots constituant une requête. Effectuer, sur le web, une recherche sur un mot isolé aboutit très fréquemment à une masse peu pertinente de documents. A contrario, utiliser une phrase complète en tant que requête a peu de chances d'offrir quelque résultat que ce soit. C'est dans l'objectif de cerner l'aide que peut apporter la reformulation que nous avons effectué une étude sur cinq cas de recherche d'informations de type locatif (p. ex. voyage au Tibet). Ce type de recherche d'informations présente un double avantage de notre point de vue : d'une part, les noms utilisés ont une sémantique assez précise tout en autorisant une certaine variabilité (que ce soit par paraphrasage ou par dérivation verbale ou adjectivale); d'autre part, la préposition joue un rôle sémantiquement important. Nous avons donc cherché à connaître l'importance de ces différents éléments dans la (re)formulation de requêtes lorsque l'objectif est la précision des résultats, sans négliger pour autant le taux de rappel d'informations pertinentes. Cette analyse est complémentaire de celle effectuée par Emirkanian et Chieze dans le cadre du même projet[1] et présentée dans ce volume. Emirkanian et Chieze ont plus particulièrement porté leur attention sur le lien syntaxique existant entre les termes de la requête dans les documents, et à son rapport à la pertinence de ces documents. Leur recherche a pu montrer que la prise en compte de ce lien devrait permettre d'augmenter la précision des requêtes sans trop nuire à leur rappel.

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).

Le domaine de la recherche d'information est traditionnellement orienté vers l'utilisation de techniques statistiques ou probabilistes. Même s'il existe des travaux montrant que l'utilisation de techniques à base de traitement automatique des langues permet d'améliorer les performances, et ce à tous les niveaux (Gaussier et coll. 2000), cela reste sujet à débat. En effet, Pincemin 1999 indique que dans certains cas, l'utilisation de formes lemmatisées peut entraîner une baisse de la qualité de la recherche; toutefois, Zweigenbaum, Grabar et Darmoni 2001 montrent qu'il y a amélioration des résultats sous certaines conditions. La plupart des moteurs de recherche fonctionnent pourtant à base de mots clés. La présence ou l'absence d'un élément de la requête détermine dès lors le résultat. On rencontre avec ce type de recherche deux inconvénients majeurs :

1° le silence, quand des documents correspondant à la requête sont absents du résultat. En effet, la recherche exacte d'un mot ne tient généralement pas compte des possibles variations morphologiques ou sémantiques. Par exemple, le mot clé *automobile* ne permettra pas de récupérer les pages contenant *voiture* ou même *automobiles*.

2° le bruit, quand les documents ne correspondant pas à la requête sont présents dans le résultat. C'est cette fois-ci le manque de précision dû à la polysémie qui est en cause. Par exemple, les mots de la requête *quelle est la vitesse du jaguar* ne permettent pas de distinguer s'il est question d'un animal, d'une voiture, d'un avion, d'un système d'exploitation ou d'une console de jeux.

Un moyen de pallier ces inconvénients est de systématiser la reformulation ou l'extension de la requête. Il est en effet possible, en ajoutant un certain nombre de termes, d'élargir et de préciser une requête. Pour cela, on distingue un certain nombre de procédés :

1° méthodes basées sur des concepts : l'extension de requête se fait de manière interactive, au fur et à mesure. Les résultats d'une requête partielle sont présentés à l'utilisateur accompagnés d'un certain nombre de termes que celui-ci peut ajouter ou enlever à la nouvelle requête. On peut voir un outil de ce type à l'adresse http://www.kartoo.com. Il est aussi possible d'ajouter automatiquement des termes à la requête initiale à partir d'une «base de données de concepts» (Klink 2001).

2° méthodes basées sur des modèles probabilistes en deux étapes : on utilise la requête de l'utilisateur pour créer un premier corpus à partir des documents les plus pertinents. Un certain nombre de techniques, dont des techniques probabilistes, sont ensuite utilisées pour introduire de nouveaux termes à la requête. On trouvera une description de ces techniques dans Bigi 2000 et Tauchi et Ward 2001.

3° méthodes utilisant les «journaux» (logs) : l'idée maîtresse de ce type de méthodes est de réutiliser le résultat de recherches précédentes. Le système mémorise pour chaque requête la liste des réponses jugées les plus pertinentes. Pour chaque nouvelle requête, le système effectue un calcul de similarité avec les requêtes précédentes et propose donc les résultats pertinents des requêtes similaires (Hust et coll. 2002).

4° méthodes linguistiques : cette famille de méthodes utilise les mêmes principes que dans le cas de méthodes probabilistes, excepté le fait que le choix de nouveaux termes se fait à partir de bases linguistiques. Il est possible d'utiliser des bases sémantiques généralistes, dont les résultats ne sont pas prouvés, ou de prendre en compte à la fois le contexte et des informations linguistiques comme le font Bouillon et coll. 2000.

Nous nous attacherons dans un premier temps à décrire la méthodologie de constitution de corpus que nous avons utilisée. Dans un deuxième temps, nous commenterons les données de notre corpus afin de mettre en évidence les techniques de traitement automatique du langage naturel (TALN) utiles dans le cadre de l'extraction d'informations via la reformulation de requêtes.

2. Le web comme corpus

2.1 Fonctionnement d'un moteur de recherche

Un moteur de recherche sur le web a deux composants principaux articulés autour d'un index (cf. Fig. 1). À partir d'un premier parcours du web, le premier composant du moteur de recherche met en place une représentation des documents (un index), le second composant gère l'accès aux documents (ou en tout cas à leur adresse) à travers l'index pour les utilisateurs finals.

2.1.1 Du document à la représentation du document

L'accès aux différentes pages du web se fait à l'aide de programmes appelés robots. Chaque page est lue et analysée automatiquement. Le résultat de cette analyse est (i) un certain nombre d'informations liées aux mots de la page, (ii) une liste éventuellement vide de liens vers d'autres pages, permettant la poursuite du traitement sur d'autres zones du web. Les informations sont stockées dans un index et la liste des liens est ajoutée à la liste des liens que le robot doit visiter.

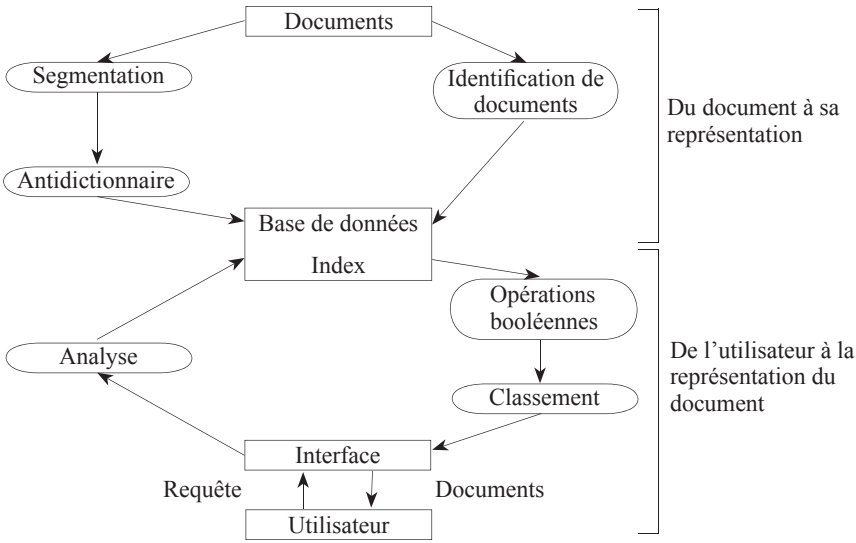

Fig. 1 : Schéma général d'un moteur de recherche

Ce mode de construction de l'index ne permet pas d'avoir un index reflétant exactement le contenu du web à un instant donné. En effet, non seulement le contenu d'une page est sujet à variation, voire à disparition, mais de nombreuses nouvelles pages font leur apparition. De plus, cette méthode ne permet de repérer que les pages liées directement à une page déjà référencée par un robot; ainsi, les pages accessibles uniquement via un formulaire ou à l'intérieur de sites payants («web invisible») ne sont pas dans l'index. L'index résultant de ce parcours ne possède en outre pas d'informations linguistiques.

2.1.2 De l'utilisateur à la représentation du document

L'interrogation d'un moteur de recherche s'effectue à l'aide de requêtes constituées de mots clés et de connecteurs («and», «or», «near») transmis au moteur via une «interface web». Le mode d'interrogation est booléen : le moteur de recherche vérifie la présence ou non d'un des mots clés dans une page. Ce type d'interrogation favorise à la fois le silence, en utilisant le connecteur «and», et le bruit, en utilisant le connecteur «or». Le connecteur «near» autorise une certaine distance entre mots clés (limitée en général à une dizaine de mots); toutefois, les mots clés peuvent ne pas être dans l'ordre initial de la requête. Certains moteurs sont capables de plus de gérer des notions de troncation (notion

de «mot commençant par» : ainsi *march** permet de repérer les pages contenant *marche* ou *marcheur*, ou encore *marchepied*). Le moteur que nous avons utilisé pour notre expérimentation nous permet ces différentes possibilités.

La recherche booléenne n'intègre pas de notion de classement : toutes les pages ont a priori le même statut. Le classement présenté à l'utilisateur résulte d'une comparaison des pages résultats en tenant compte de la proportion de mots clés présents dans la page et de la plus ou moins grande proximité de ces mots clés[2].

2.2 Spécificités du web

Le web, considéré comme corpus, possède un certain nombre de caractéristiques qui impliquent un traitement particulier. On peut mentionner les traits saillants suivants (Baeza-Yates et Ribeiro-Neto 1999) :

1° les données sont **distribuées** : contrairement à la plupart des bases de données ou ressources linguistiques habituelles, les informations sont stockées sur un grand nombre de machines physiques (et par conséquent géographiques) différentes. Le traitement de toutes les informations nécessite un fonctionnement correct pour tous les supports.

2° les données qu'on peut extraire du web sont **volatiles** : potentiellement, n'importe qui peut créer un site web et donc ajouter, supprimer ou modifier le contenu des pages sans en référer à qui que ce soit. De plus, la nature même d'un certain nombre d'informations (articles de journaux, forums, brèves, informations commerciales, ...) fait qu'une part importante de celles-ci est rapidement périmée.

3° la **structure** des pages du web est très difficilement utilisable, voire inexistante : étant donné le nombre important d'éditeurs ou de créateurs de pages, il n'existe pas de norme de structuration. La culture du WYSIWYG[3] étant largement répandue, quand bien même il existerait une telle norme, celle-ci ne serait pas respectée.

Dès lors, il est fondamental d'associer à une analyse statique des pages (ce qui est fait par tout moteur de recherche lors de la constitution de l'index) une analyse dynamique, c'est-à-dire une analyse des pages réellement récupérées lors d'une requête. Cette analyse dynamique est toutefois extrêmement coûteuse en temps, elle n'est donc réaliste que pour un affinement des réponses. Une voie alternative existe néanmoins : l'analyse dynamique de la requête elle-même,

2 En fait, des considérations commerciales peuvent modifier ce classement.

3 What You See Is What You Get : ce qu'on voit à l'écran détermine la structure du document, autrement dit l'information de fond n'est généralement pas séparée de la forme du document.

autrement dit la possibilité de modifier la requête quand cela est souhaitable afin d'améliorer la qualité de l'information retournée à l'utilisateur. C'est l'intérêt de cette approche que nous avons essayé de déterminer dans notre étude. Les travaux effectués jusqu'à présent et qui utilisent le web comme corpus ont des objectifs très différents. Ainsi, celui de Amitay 1999 (sur la langue anglaise) portait sur les 35 mots les plus fréquents présents dans le British National Corpus (BNC) et dans un corpus de pages web personnelles (Home Corpus), et plus particulièrement des mots présents uniquement dans l'un des deux corpus. L'auteur observe plusieurs traits caractéristiques du corpus de pages Internet : 1° l'absence de la troisième personne, et l'emploi très fréquent de la première et de la deuxième personnes; 2° l'absence du verbe *to be* au passé et l'utilisation importante du temps présent; 3° l'absence des connecteurs *but* et *which* entre deux phrases. L'auteure conclut ainsi que la langue utilisée dans les pages personnelles s'apparente à une conversation présentant des faits entre l'auteur et le lecteur. Sans préjuger de l'intérêt de ces résultats, convenons que l'état actuel des travaux sur la structure grammaticale du contenu des pages web n'est pas assez avancé pour pouvoir être utilisé à bon escient dans le cadre de la recherche d'informations. Toutefois, il va de soi qu'il s'agit là d'éléments qu'il faudra associer à nos traitements.

2.3 Constitution du corpus

Notre expérimentation a porté sur cinq cas distincts de recherche d'informations de type locatif : voyage au Tibet, fuite des cerveaux vers les États-Unis, vol sur la lune, mission dans l'espace et promenade dans Paris. Nous avons constitué pour chacun de ces cas un corpus de pages extraites du web (ces corpus seront appelés dans la suite TIBET, FUITE, 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 X et Y 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 au total).

Par exemple, dans le cas de «voyage au Tibet», X a comme valeur initiale *voyage*, Y la valeur *Tibet*. Nous utilisons pour X les variations *voyages*, *séjour*, *voyager*, *trek*, etc. et pour Y *tibétain*, *tibétaines*, etc. Dans le cas où le Y est transformé en un adjectif, la préposition pourra être présente ou non (*promenade parisienne* ou *promenade dans les arrondissements parisiens*). Nous avons

défini un total de 14 840 requêtes (le Tableau 1 récapitule le nombre de requêtes constituées par corpus).

2.3.1 Considérations générales

La constitution d'un corpus, nécessaire pour notre étude, ne va pas de soi. Outre les difficultés techniques inhérentes à ce travail, difficultés sur lesquelles nous reviendrons, se pose le problème crucial de la représentativité du corpus extrait. Le web mondial est en effet constitué actuellement de plus de trois milliards de pages. La récupération des pages elles-mêmes n'est toutefois pas problématique puisqu'il existe de nombreux moyens «d'aspirer» des pages web, que ce soit à partir d'outils dédiés ou via des programmes. Quant à la cohérence du corpus constitué, nous pouvons définir deux cas la garantissant. Soit les données qui vont constituer le corpus sont regroupées sur un ensemble de sites connus et bien référencés; il suffit dès lors de les aspirer. Soit, et c'est notre cas, il faut déterminer l'emplacement des pages qui nous intéressent. Dans un premier temps, il est donc nécessaire de constituer un corpus d'adresses de pages web dont on pense que le contenu répond à un ensemble de critères (syntaxique, sémantique, thématique, etc.) en adéquation avec la recherche initiale d'information. Nous utilisons donc le moteur de recherches comme fournisseur de ces adresses en utilisant toutes les variantes de requêtes exprimables et en tenant compte des critères. Dans un deuxième temps, le corpus est constitué par extraction des pages à partir de la liste d'adresses.

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 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 de la requête (mots clés ou variantes de ceux-ci) 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 techniquement possible de récupérer qu'au plus 1010 adresses de pages (URL) alors même que certaines requêtes en fourniraient plusieurs millions selon AltaVista. 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 ayant permis de récupérer cette page...).

2.3.2 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 le moteur AltaVista 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. Cette contrainte permet d'éliminer un certain nombre de résultats où les mots-clés n'ont vraisemblablement aucun lien syntaxique. Cependant, le moteur travaille sur les mots et non pas sur des structures plus complexes telles que la phrase; il est donc possible que des mots clés proches selon le moteur n'aient toutefois pas de lien syntaxique. Ce problème est en partie géré au moment du transcodage, c'est-à-dire ultérieurement dans la phase de constitution de corpus.

2.3.3 Les requêtes

Notre expérimentation est basée sur cinq familles de requêtes correspondant aux cas de recherche d'information souhaités, nous permettant ainsi d'obtenir cinq corpus. Dans chaque cas, nous avons deux structures de requêtes possibles. Ainsi, dans le cas ESPACE, nous pouvons avoir un nom suivi d'une préposition suivi d'un nom (structure X Prep Y, exemple *séjour dans l'espace*) ou bien un nom suivi d'une forme adjectivale (structure X Y, exemple *vol spatial*)[4]. Dans le Tableau 1, nous avons mentionné entre parenthèses le nombre de valeurs possibles pour chacun de ces éléments (premier mot, préposition, deuxième mot ou forme adjectivale) : ainsi, X(3) Prep(31) Y(1) dans le cas ESPACE signifie qu'il y a trois valeurs de nom (*séjour*, *vol*, *mission*), 31 prépositions possibles, ainsi qu'une seule valeur de deuxième mot, ce qui aboutit à 3*31*1=93 requêtes possibles. Comme nous le verrons ultérieurement, le nombre important de requêtes pour un corpus ne préjuge en rien du nombre de pages accessibles.

4 Dans le cas du corpus FUITE, il y a trois mots en sus de la préposition dans la première structure.

Tableau 1
Structure des requêtes

CORPUS	EXEMPLES	STRUCTURE ET NOMBRE D'ÉLÉMENTS	NOMBRE DE REQUÊTES
FUITE	*cerveaux partant aux États-Unis*	X1(2) X2(20) Prep(31) Y(8)	9920
	fuite des cerveaux vers les États-Unis	X(12) Prep(31) Y(8)	2976
ESPACE	*séjour dans l'espace*	X(3) Prep(31) Y(1)	93
	vol spatial	X(3) Y(1)	3
LUNE	*mission sur la lune*	X(8) Prep(31) Y(2)	496
	vol lunaire	X(8) Y(1)	8
TIBET	*séjour au Tibet*	X(22) Prep (31) Y(1)	682
	voyage tibétain	X(22) Y(1)	22
PARIS	*promenade à Paris*	X(20) Prep(31) Y(1)	620
	séjour parisien	X(20) Y(1)	20

2.3.4 Corpus résultants

Le résultat de notre chaîne de traitement est, pour chaque famille de requêtes, un ensemble de fichiers structurés de la manière suivante : 1° fichier contenant la liste des requêtes utilisées; 2° répertoire des fichiers de données sources en format HTML; 3° répertoire des fichiers de données mises sous forme normalisée (format XML); 4° répertoire des fichiers d'annotations obtenues par analyse manuelle.

Les trois répertoires ont la même architecture : à chaque fichier source correspond au même emplacement un fichier dans chacun des deux autres répertoires. Cela nous permet de plus un accès rapide aux différentes versions et annotations d'une page web. Chaque fichier de format HTML est accompagné d'un fichier «résumé» contenant l'URL d'origine et la liste des requêtes ayant permis de récupérer cette page. On y trouve aussi pour chaque requête le rang de la page tel que celui-ci est donné par le moteur de recherche. Par exemple, dans le Tableau 2, la même page dont l'adresse est donnée en première ligne aura été extraite via les quatre requêtes indiquées ensuite; notons que ces requêtes ont des structures sensiblement différentes :

Tableau 2
Exemple de fichier «résumé»

```
http://perso.wanadoo.fr/france.tibet/campagne_panchen.htm
:::
     march*+NEAR+de+NEAR+tibet*  ;;; rang : 60
     march*+NEAR+du+NEAR+tibet*  ;;; rang : 22
     march*+NEAR+des+NEAR+tibet* ;;; rang : 20
     march*+NEAR+tibetain*       ;;; rang : 26
```

Le résultat de l'aspiration des corpus est résumé dans le Tableau 3, où nous avons précisé pour chaque corpus créé le nombre total de mots, la taille du corpus en mégaoctets, et enfin le nombre de pages de sites web qui auront pu être analysées. Notons 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 : on ne trouvera que des pages journalistiques ou économiques en ce qui concerne la fuite des cerveaux en Californie, alors qu'on ne pourra éviter la multitude de pages à caractère touristique dans le cas du thème concernant la promenade à Paris.

Tableau 3
Éléments quantitatifs des corpus

CORPUS	NOMBRE DE MOTS	TAILLE (MO)	NOMBRE DE PAGES
FUITE	908 032	17	186
ESPACE	24 905 110	407	8 291
LUNE	10 978 288	187	3 522
TIBET	3 829 657	70	1 507
PARIS	60 588 606	1 058	20 017

2.3.5 Posttraitement

Un posttraitement est effectué afin de repérer dans le texte les phrases contenant une séquence syntaxique correspondant à une des requêtes ayant permis de récupérer la page, les phrases contenant une séquence syntaxique filtrant une requête n'ayant pas permis de récupérer la page selon le moteur de recherche, ainsi que les mots de requêtes n'appartenant pas à une séquence syntaxique.

Tableau 4
Page web en format XML

```
<doc>
<url>http://starling.citeglobe.com/cinema/film/himalaya.
htm</url>
<req>aventur*+NEAR+de+NEAR+tibet*</req><rang>52</rang>
<req>aventur*+NEAR+d+NEAR+tibet*</req><rang>16</rang>
<req>aventur*+NEAR+du+NEAR+tibet*</req><rang>18</rang>
<req>aventur*+NEAR+tibetain*</req><rang>1</rang>
<phr num=»0»>Himalaya (L'Enfance d'un Chef), de Eric
Valli, avec Thilen Lhondup (résumé, critique, affiche). [an
error occurred while processing this directive].</phr>
<phr num=»1» alta=»aventur*+NEAR+tibetain*»>Himalaya
(L'Enfance d'un Chef) Aventures tibétaines de Eric Valli,
avec Thilen Lhondup, Karma Wangiel, Lhapka Tsamehoe, Karma
Tensing...</phr>
...
<phr num=»18»>En fait, ça ressemble à du Jean-Jacques
Annaud («Sept ans au <mot>Tibet</mot>»), mais en plus
simple, plus humble, et plus zen.</phr>
...
</doc>
```

Pour ce faire, la page HTML est transcodée en XML (le Tableau 4 contient un extrait de fichier transcodé en format 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 un ordre grammaticalement correct. 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. Le lecteur repèrera aisément les différentes informations sur l'origine de la page, les requêtes ayant permis d'identifier cette page (et le rang affecté par le moteur de recherche), le découpage en phrases tel qu'effectué par notre module, et enfin la zone d'une phrase correspondant à un mot ou bien à la requête elle-même.

2.3.6 Analyse du corpus

La dernière étape du processus consiste à analyser (manuellement) le corpus, soit d'un point de vue syntaxique soit d'un point de vue informationnel. Nous avons pour cela élaboré une interface graphique (à l'aide du langage PHP) permettant de choisir son corpus, de le parcourir (voir la Fig. 2) puis de l'annoter.

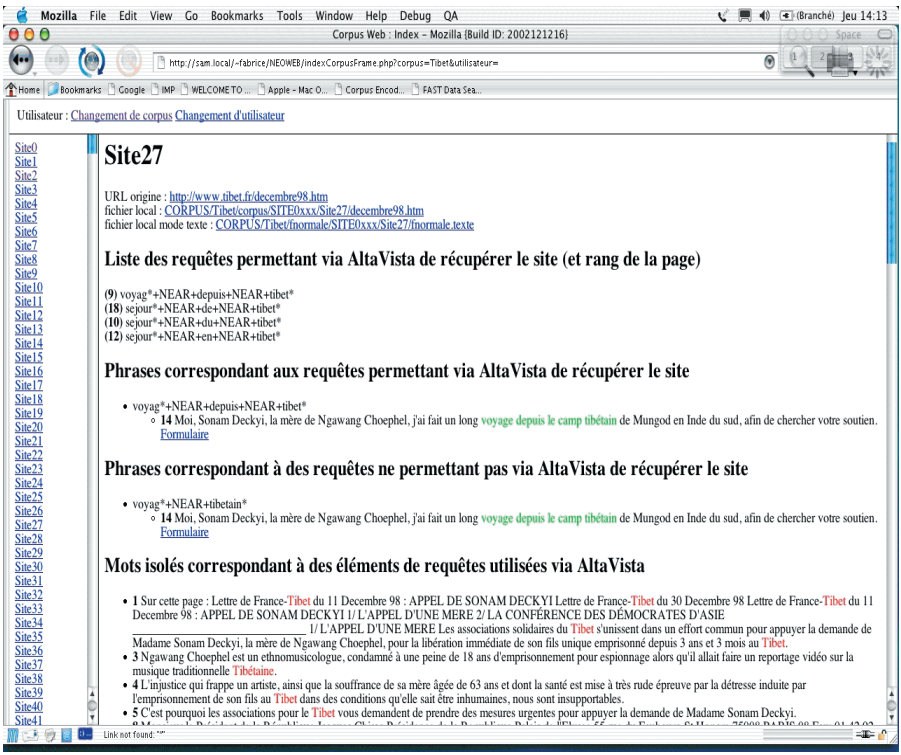

Fig. 2 : Interface de consultation des corpus

3. Analyse factuelle

Nous commenterons dans cette section les données du web recueillies pour les cinq thèmes ESPACE, PARIS, FUITE, LUNE et TIBET. Nous nous intéresserons en particulier à la variabilité quantitative et qualitative de l'information extraite en fonction du thème de la requête. Nous étudierons l'impact quantitatif de l'ajout d'une préposition dans une requête. La section suivante intègrera une analyse informationnelle des pages elles-mêmes. Dans cette section, nous ne

tiendrons donc pas compte de la pertinence thématique ou informationnelle des données extraites. Il va de soi que ces commentaires ne valent que dans le cadre de l'expérience très limitée que nous avons menée. Enfin, nous chercherons dans quelle mesure un ensemble de requêtes peut couvrir un ensemble de pages candidates.

Pour chaque corpus, nous avons à notre disposition la liste des requêtes ainsi que les pages accessibles par moteur de recherche par le biais de ces requêtes. Nous avons effectué par ailleurs une vérification de la présence véritable de la requête dans la page. En effet, l'accessibilité d'une page par un moteur de recherche n'indique pas nécessairement que la requête figure explicitement dans le contenu de la page : celle-ci a ainsi pu être modifiée depuis son analyse par le moteur de recherche. Nous avons de plus réexaminé l'ensemble des pages formant un corpus en y cherchant la présence éventuelle de requêtes (y compris, donc, les requêtes qui ne permettaient pas d'accès aux pages via le moteur de recherche). Cette analyse nous permet de mieux comprendre la couverture potentielle en termes de pages, donc d'apporter quelques éléments relatifs à la variation à effectuer sur une requête initiale.

3.1 Utilité des requêtes en terme d'extraction de pages

La détermination de la liste de requêtes par corpus s'est effectuée sans tenir aucun compte d'informations syntaxiques. Il est clair que nombre de ces requêtes n'ont aucune justification linguistique. Certaines d'entre elles ont toutefois permis d'extraire via le moteur de recherche des pages au contenu pertinent. Cela est dû principalement à la limitation intrinsèque des résultats du moteur de recherche. Parmi toutes ces requêtes, les requêtes **utiles**, c'est-à-dire celles qui auront permis de récupérer au moins une page via le moteur de recherche, ont une proportion relativement stable comme le suggère le Tableau 5, sauf dans le cas du corpus FUITE.

Tableau 5
Requêtes utiles

CORPUS	NOMBRE DE REQUÊTES UTILES	NOMBRE DE REQUÊTES	% DE REQUÊTES UTILES
ESPACE	71	103	69 %
FUITE	95	8 462	1 %
LUNE	79	543	15 %
TIBET	194	705	28 %
PARIS	380	642	59 %

Les résultats du corpus FUITE doivent être interprétés séparément. D'une part, la structure de la requête rend sa variabilité importante. La requête admet trois parties à fortes possibilités de variation, la probabilité d'obtenir une requête syntaxiquement fausse devient un facteur important. D'autre part, le faible nombre de pages récupérables réduit nécessairement la pertinence de ces requêtes. Il n'est toutefois pas inutile de constater que le locatif *États-Unis* est à l'origine de 90 % des pages extraites : la variation de ce locatif en *Californie* ou bien encore *Côte Est* ne permet d'augmenter que très faiblement le nombre de pages récupérables. Quand ces variantes existent, les pages récupérées contiennent de fait le terme *États-Unis* : la variante locative n'est utilisée que pour l'usage paraphrastique littéraire qu'il permet. La prise en compte de cette dimension nous paraît essentielle dès lors qu'on cherche à limiter le nombre de requêtes. Encore est-il nécessaire d'effectuer une analyse linguistique de ce type afin de déceler les termes systématiquement utilisés.

En ce qui concerne les autres corpus, la sélection à effectuer entre les termes «utiles» et les autres s'effectue sur une base de pertinence syntaxique et non sur une variation des mots. Ainsi, dans le cas du corpus PARIS, les variations sur *voyage* qui seraient d'usage peu courant (*pérégrination,* par exemple) permettent toutefois de récupérer certaines pages non accessibles par d'autres mots. A contrario, des requêtes comme *séjour d'un bout à l'autre de Paris* ou encore *arpentage contre Paris* ne permettent aucune récupération de pages. Ce critère de sélection de requêtes peut être mis en place dès lors que les informations de structure argumentale sont connues.

3.2 Information quantitative liée à l'extraction

Nous tenons compte, dans ce paragraphe, du nombre de pages accessibles par requête pour chacun des corpus. Celui-ci rend compte tout à la fois de l'importance du thème dans le web, mais aussi de la potentielle polysémie implicite. En étudiant, à l'intérieur d'un même corpus, la variabilité de cette valeur, nous avons une première indication du rappel pour chaque type de requêtes.

Au vu du Tableau 6 de la page suivante, la variation selon le corpus apparaît clairement. La polysémie des thèmes ESPACE et PARIS induit naturellement un nombre moyen de pages par requête très important. De même, le nombre maximum de pages pour une requête est, du point de vue de l'utilisateur, prohibitif : quel utilisateur aura le courage d'étudier les quelque sept cents pages récupérées par une unique requête? Le nombre total de pages récupérables corrobore ce résultat.

Tableau 6
Rappel et requête utile

CORPUS	NOMBRE MOYEN DE PAGES PAR REQUÊTE UTILE	NOMBRE MAXIMUM DE PAGES POUR UNE REQUÊTE	NOMBRE TOTAL DE PAGES RÉCUPÉRABLES (DOUBLONS POSSIBLES)
ESPACE	199	760	14 110
FUITE	4	31	408
LUNE	76	490	6 025
TIBET	25	248	4 789
PARIS	116	794	44 205

3.3 Importance de la préposition

La reformulation de requêtes peut être abordée sous différents aspects. Outre le choix de variantes synonymiques, il est encore possible de préciser syntaxiquement la requête dans le but explicite d'augmenter la précision (et ce même pour l'ensemble des requêtes) sans affecter le rappel, au sens où les pages récupérables à partir d'une (unique) structure syntaxiquement lâche peuvent être récupérées par un ensemble de requêtes syntaxiquement plus déterminées. Une des raisons de notre choix de thèmes de requêtes provient justement de cette possibilité d'accroître la précision syntaxique des requêtes en incluant une préposition non sémantiquement vide.

Au vu des résultats obtenus, il apparaît clairement que le nombre de pages récupérées à partir de requêtes syntaxiquement plus précises augmente significativement. Il va de soi que le nombre de requêtes à effectuer augmente dans la même proportion.

Nous avons de plus étudié la présence du schéma associé à une requête dans les pages récupérables par cette même requête. Rappelons qu'une requête, du point de vue du moteur de recherche, n'est pas une structure syntaxiquement correcte. Pour des raisons liées à l'algorithme d'indexation, une requête est constituée d'un ensemble de mots relativement proches dans la page. Le terme «proche» signifie que deux mots de cette requête doivent être séparés par une dizaine d'autres mots au maximum. En aucun cas, l'ordre syntaxique n'est vérifié. Comme le montre le Tableau 7, une simple analyse des pages récupérées en testant l'ordre syntaxique obligatoire permet de constater que les schémas sont présents de manière extrêmement marginale. Concluons de ce fait qu'augmenter la précision syntaxique de la requête (en particulier en figeant l'ordre des mots de la requête de manière à obtenir une structure

syntaxiquement plausible) va radicalement diminuer le rappel (c'est-à-dire le nombre de pages effectivement récupérées parmi celles susceptibles d'intérêt étant donnée la requête initiale).

Tableau 7
Présence du schéma de la requête

CORPUS	% PRÉSENCE DU SCHÉMA	% PAGES SUPPLÉMENTAIRES
ESPACE	7,50 %	60,00 %
FUITE	41,80 %	100,00 %
LUNE	8,50 %	75,70 %
TIBET	6,30 %	28,70 %

4. Analyse informationnelle

Nous nous intéressons, dans cette section, à l'intérêt informationnel des pages récupérées. 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 était ainsi par exemple pour le corpus ESPACE. S'agissait-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. Nous avons analysé une partie (choisie arbitrairement) de chaque corpus en étudiant manuellement le degré de pertinence de chaque page. 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 : 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 : la page contient au moins un paragraphe répondant à la requête. C'est le cas par exemple dans les sites d'agences de presse où, parmi des informations sportives ou économiques, on peut trouver un paragraphe mentionnant une marche (de protestation) pour le Tibet.

– non pertinent : 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.

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 en déterminant une valeur de **pertinence** associée à chaque requête. 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. Enfin, nous avons défini la couverture relative de chaque requête. En effet, il est important de connaître avec précision la quantité minimale de requêtes permettant, sans détériorer la précision, d'obtenir un rappel important (p. ex. 90 % des pages jugées pertinentes). Pour chaque corpus, nous avons analysé manuellement un millier de pages (toutes les pages pour les corpus de faible taille). Cette évaluation reste modeste et mériterait d'être poursuivie. Toutefois, les résultats obtenus deviennent notablement stables, justifiant de considérer les prochaines remarques comme des conclusions temporaires.

Une fois la pertinence manuellement déterminée pour chaque page, une pertinence par requête est calculée comme la moyenne des pertinences sur l'ensemble des pages ayant été récupérées par cette requête. Le Tableau 8 est un extrait des résultats obtenus pour le corpus ESPACE. Nous avons quantifié la pertinence de 3 à 1 : 3 signifie non pertinent, 1 signifie totalement pertinent. Les résultats obtenus suggèrent divers commentaires. En premier lieu, il est notable que la pertinence moyenne est rapidement faible. C'est flagrant au vu du Tableau 8, et on retrouve cette propriété dans le cas des autres corpus. En fait, on trouve la plupart des requêtes non syntaxiquement pertinentes en fin

de tableau (les requêtes ne récupérant aucune page ne sont pas mentionnées). La requête en première ligne du tableau fait figure d'exception : cette requête ne récupère qu'une seule page, qui s'avère, pour des raisons extérieures à la requête elle-même, totalement pertinente. Cela n'invalide toutefois pas la remarque générale faite précédemment. Les requêtes comportant une préposition incorrecte vis-à-vis de X donnent des résultats non pertinents : le rappel est très faible. C'est notoirement le cas pour le corpus FUITE, qui, avec 8462 requêtes pour seulement 186 pages distinctes récupérables, contient un nombre impressionnant de requêtes syntaxiquement incorrectes ne permettant de 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 (p. ex. *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 prépositions reste particulier : le rappel y est (naturellement) très important, mais comme nous l'avons souligné précédemment, la précision est assez faible. Enfin, la grande majorité des pages récupérées l'est 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.

Tableau 8
Pertinence des requêtes (corpus ESPACE)

X	PRÉPOSITION	Y	PERTINENCE
mission*	chez	espace	1
mission*	vers	espace	1,64
vol*	devant	espace	1,73
mission*		spatial*	1,97
mission*	hors	espace	2
séjour*	pour	espace	2,19
vol*		spatial*	2,2

X	PRÉPOSITION	Y	PERTINENCE
vol*	vers	espace	2,31
.....	.....	.....	.....
mission*	aux	espace	2,93
séjour*	chez	espace	3
vol*	près	espace	3
séjour*	par	espace	3
vol*	travers	espace	3
séjour*	du	espace	3
séjour*	en	espace	3
séjour*	d	espace	3
séjour*	a	espace	3

En associant degré de pertinence et rang, nous pouvons calculer un **facteur de qualité** associé à la requête. Ce facteur est égal, pour une requête donnée, à la somme des pertinences obtenues sur chaque page associée à cette requête et pondérées par le logarithme du rang divisé par la somme des logarithmes des rangs. L'utilisation du logarithme permet 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. Le tableau donné en annexe est un extrait des résultats obtenus sur le corpus PARIS. Ce tableau suggère deux commentaires.

D'une part, la qualité peut être très variable pour un même degré de pertinence. Toutefois, la qualité augmente en moyenne avec la pertinence, sauf lorsque le degré de pertinence est faible (c'est le sens dans le cas de degré de pertinence entre 2,5 et 3) : il s'agit en fait de cas avec fort rappel, augmentant par là même le rang moyen des pages. D'autre part, le degré de pertinence minimal est assez élevé (2), suggérant une précision assez faible. Toutefois, la grande variabilité des résultats obtenus en tenant compte du rang ne doit pas faire oublier l'importance de la présentation de l'information à l'utilisateur. Cela signifie sans nul doute que la manière dont est calculé le rang par les moteurs de recherche doit être sensiblement modifiée.

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. 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 ayant le facteur 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, c'est-à-dire 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. ex. ajout de la préposition) non seulement permet un rappel significatif, mais encore aura une précision correcte (eu égard à la masse de documents contenus dans le web).

5. Conclusion

Le travail présenté ici porte à la fois sur une problématique de constitution de corpus à partir du web, et sur l'impact de la reformulation de requêtes relativement à la pertinence informationnelle des réponses. La constitution du corpus se faisant via des requêtes, il nous a été possible d'effectuer des évaluations des taux de rappel et de précision des différentes «versions» d'une même requête. Les différentes mesures – pertinence, vitesse de convergence – effectuées sur les corpus résultants montrent l'intérêt de la reformulation de requête associée à une notion de correction syntaxique.

Il va de soi que l'étude dont nous venons d'exposer les premiers résultats doit être poursuivie dans deux directions complémentaires. D'une part, l'investigation que nous avons menée reste quantitativement insuffisante. Il conviendrait d'amplifier l'analyse informationnelle des pages afin de confirmer les conclusions présentées ici. D'autre part, nos thèmes de définition des corpus ont été volontairement ciblés autour de prédicatifs locatifs. L'analyse pourrait être étendue à d'autres thèmes non nécessairement locatifs, mais en conservant un prédicatif possédant une structure argumentale suffisamment riche, condition bien entendu requise pour que l'aide à la recherche d'informations puisse être effectuée.

D'un point de vue plus général, la piste de l'extension de requête, abordée à la fois d'un point de vue sémantique et syntaxique, semble être prometteuse

pour ce qui concerne la recherche d'information sur le web. Cette solution non seulement donne des résultats significatifs, comme le montre notre travail, mais son application concrète est envisageable dès lors que ce traitement intervient en amont de la phase de recherche sur web.

Références

AMITAY, E. 1999 «Anchors in context: A corpus analysis of web pages authoring conventions», dans L. Pemberton et S. Shurville, *Words on the Web - Computer Mediated Communication*, Intellect Books, p. 192.

BAEZA-YATES, R. et B. RIBEIRO-NETO 1999 *Modern Information Retrieval*, New-York, ACM Press.

BIGI, B. 2000 *Contibution à la modélisation du langage pour des applications de recherche documentaire et de traitement de la parole*, thèse de doctorat, Université d'Avignon.

BOUILLON, P. et coll. 2000 «Apprentissage de ressources lexicales pour l'extension de requêtes», *Traitement automatique des langues*, Paris, ATALA et Hermès 41-2 : 367-393.

EMIRKANIAN, L. et E. CHIEZE, 2002 «Variations morphologiques, syntaxiques, sémantiques et repérage d'information sur le Web», communication au colloque TALN, Web et corpus (nov. 2002, Saint-Denis) [texte ici même].

GAUSSIER, E. et coll. 2000 «Recherche d'information en français et traitement automatique des langues», *Traitement automatique des langues*, ATALA/Hermes sciences publications, Paris, 41-2 : 473-493.

HUST, A. et coll. 2002 «Query Expansion for Web information Retrieval», dans S. Schubert, B. Reusch et N. Jesse, *32nd Annual Conference of the German Informatics Society, Web Information Retrieval Workshop,* German Informatics Society, P-19 : 176-180.

KLINK, S. 2001 «Query reformulation with collaborative concept-base expansion», *First International Workshop on Web Document Analysis*, Seattle, p. 19-22.

PINCEMIN, B. 1999 *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, thèse de doctorat, Université de Paris IV-Sorbonne.

TAUCHI, M. et N. WARD 2001 *Searching for Explanatory Web pages using Query Expansion*, PacLing (http://afnlp.org/pacling2001/tauchi.pdf).

ZWEIGENBAUM, P., N. GRABAR et S. DARMONI 2001 «L'apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée», dans D. Maurel, *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours, p. 403-408.

ANNEXE

N	Prép	N	Pertinence	Qualité	N	Prép	N	Pertinence	Qualité
balad*	depuis	Paris*	2	2	deplac*	pour	Paris*	2,87	117,78
deambul*	des	Paris*	2	6	erra*		Paris*	2,87	219,46
deambul*	de	Paris*	2	20	sejour*	en	Paris*	2,88	238,49
march*	autour	Paris*	2	23	aventur*	a	Paris*	2,89	186,14
sillon*	d	Paris*	2	33	deplac*		Paris*	2,89	334,11
arpent*	d	Paris*	2	56	march*	du	Paris*	2,9	209,96
parcour*	pour	Paris*	2	195	parcour*	du	Paris*	2,9	306,15
balad*	des	Paris*	2,05	57,56	erra*	de	Paris*	2,91	136,74
promen*	au	Paris*	2,23	120,93	parcour*	dans	Paris*	2,91	205
deambul*		Paris*	2,26	26,55	erre*	de	Paris*	2,92	221,48
deambul*	dans	Paris*	2,28	14,46	march*	pour	Paris*	2,92	333,13
promen*	par	Paris*	2,31	113,06	travers*	au	Paris*	2,92	345,56
promen*		Paris*	2,32	340,62	aventur*		Paris*	2,92	415,37
promen*	du	Paris*	2,34	211,35	travers*	par	Paris*	2,93	204,5
balad*	dans	Paris*	2,37	148,03	aventur*	d	Paris*	2,93	261,06
sillon*		Paris*	2,4	69,6	march*	en	Paris*	2,93	264,1
promen*	de	Paris*	2,44	500,6	march*	des	Paris*	2,93	412,8
parcour*	par	Paris*	2,46	142,16	aventur*	du	Paris*	2,94	222,59
deplac*	chez	Paris*	2,5	6	travers*	d	Paris*	2,94	373,64
deambul*	d	Paris*	2,5	8	deplac*	devant	Paris*	3	1
march*	%22le+long%22	Paris*	2,5	8	erra*	par	Paris*	3	3
sillon*	sur	Paris*	2,5	11	sillon*	au	Paris*	3	3
deambul*	en	Paris*	2,5	11	erre*	sous	Paris*	3	4
deambul*	a	Paris*	2,5	13	aventur*	chez	Paris*	3	5
promen*	sous	Paris*	2,5	17	parcour*	chez	Paris*	3	5
erra*	sur	Paris*	2,5	24	vagabond*	aux	Paris*	3	5
promen*	travers	Paris*	2,5	36	peregrin*	en	Paris*	3	6
parcour*	vers	Paris*	2,5	43	sillon*	en	Paris*	3	8
march*	chez	Paris*	2,5	75	peregrin*	d	Paris*	3	9
march*	contre	Paris*	2,5	82	errance*	aux	Paris*	3	10
promen*	pour	Paris*	2,5	192,22	peregrin*	de	Paris*	3	12
promen*	en	Paris*	2,5	269,9	sejour*	depuis	Paris*	3	13
promen*	dans	Paris*	2,5	408,25	march*	pres	Paris*	3	14
arpent*	de	Paris*	2,53	134,11	erra*	aux	Paris*	3	15
vagabond*	de	Paris*	2,55	83,37	deplac*	aux	Paris*	3	17
vagabond*		Paris*	2,55	120,23	erra*	pour	Paris*	3	17,33
parcour*	des	Paris*	2,58	337,22	errance*	pour	Paris*	3	18
vagabond*	du	Paris*	2,6	24,3	parcour*	sous	Paris*	3	19

N	Prép	N	Pertinence	Qualité	N	Prép	N	Pertinence	Qualité
sejour*	sur	Paris*	2,62	40,57	vagabond*	dans	Paris*	3	20
arpent*		Paris*	2,63	104,89	arpent*	par	Paris*	3	20
promen*	a	Paris*	2,63	187,55	erre*	depuis	Paris*	3	23
sillon*	de	Paris*	2,64	101,59	march*	travers	Paris*	3	25
erra*	d	Paris*	2,65	45,99	peregrin*		Paris*	3	27
balad*	de	Paris*	2,65	242,76	deplac*	au	Paris*	3	27
promen*	des	Paris*	2,65	269,96	erre*	pour	Paris*	3	30
travers*	travers	Paris*	2,65	398,15	travers*	jusqu	Paris*	3	30,5
sillon*	des	Paris*	2,66	41,11	march*	jusqu	Paris*	3	32
sejour*	d	Paris*	2,66	268,07	parcour*	autour	Paris*	3	33
parcour*		Paris*	2,66	363,55	travers*	aux	Paris*	3	37
parcour*	de	Paris*	2,67	328,26	deplac*	des	Paris*	3	44
balad*	au	Paris*	2,68	62,85	balad*	par	Paris*	3	46
sejour*	pour	Paris*	2,68	249,94	erre*	dans	Paris*	3	52
sejour*		Paris*	2,69	159,61	deplac*	sur	Paris*	3	56
parcour*	a	Paris*	2,69	477,93	travers*	sous	Paris*	3	56
balad*	a	Paris*	2,7	112,65	travers*	vers	Paris*	3	65
erra*	dans	Paris*	2,71	41,08	erre*	sur	Paris*	3	74
sejour*	du	Paris*	2,71	157,01	errance*	en	Paris*	3	75
travers*	du	Paris*	2,71	201,09	balad*	pour	Paris*	3	75
travers*	de	Paris*	2,71	461,78	parcour*	depuis	Paris*	3	75,25
travers*		Paris*	2,71	488,91	erre*	des	Paris*	3	80,16
erre*	par	Paris*	2,72	28,34	erre*	du	Paris*	3	80,66
parcour*	en	Paris*	2,72	394,05	aventur*	pour	Paris*	3	81,5
aventur*	aux	Paris*	2,73	58,43	errance*	de	Paris*	3	83,33
balad*	sur	Paris*	2,73	105,68	march*	sous	Paris*	3	85,75
promen*	chez	Paris*	2,74	21,4	errance*	des	Paris*	3	90
arpent*	des	Paris*	2,74	51,03	deplac*	dans	Paris*	3	98,66
promen*	d	Paris*	2,74	329,74	sejour*	par	Paris*	3	105
march*	aux	Paris*	2,76	245,04	aventur*	par	Paris*	3	111,25
erre*	au	Paris*	2,77	59,56	erre*	a	Paris*	3	113,2
travers*	pour	Paris*	2,77	222,97	erre*	d	Paris*	3	113,81
march*	dans	Paris*	2,77	326,02	balad*	du	Paris*	3	125
travers*	des	Paris*	2,78	350,73	erra*	des	Paris*	3	131
travers*	a	Paris*	2,78	378,66	aventur*	en	Paris*	3	131,2
balad*	d	Paris*	2,79	56,33	sejour*	dans	Paris*	3	132,66
deplac*	de	Paris*	2,79	128,25	travers*	sur	Paris*	3	136,57
balad*		Paris*	2,8	317,67	march*	depuis	Paris*	3	144,5
march*	d	Paris*	2,81	311,27	aventur*	au	Paris*	3	145,36

N	Prép	N	Pertinence	Qualité	N	Prép	N	Pertinence	Qualité
erra*	en	Paris*	2,82	58,25	deplac*	en	Paris*	3	152
march*	sur	Paris*	2,82	456,89	erre*	en	Paris*	3	166
march*	de	Paris*	2,83	378,54	aventur*	des	Paris*	3	173,4
march*		Paris*	2,83	414,89	sejour*	des	Paris*	3	191,5
aventur*	de	Paris*	2,84	285,51	march*	par	Paris*	3	205,44
travers*	en	Paris*	2,84	396,78	promen*	sur	Paris*	3	208
march*	a	Paris*	2,84	468,01	errance*		Paris*	3	208,71
march*	au	Paris*	2,85	414,89	aventur*	dans	Paris*	3	226
sejour*	a	Paris*	2,85	422,29	erre*		Paris*	3	253,42
sejour*	de	Paris*	2,85	435,57	sejour*	au	Paris*	3	254,5
balad*	en	Paris*	2,86	152,74	travers*	dans	Paris*	3	273,12
march*	vers	Paris*	2,87	61,53	parcour*	au	Paris*	3	289,66
deplac*	pour	Paris*	2,87	117,78	parcour*	sur	Paris*	3	355
erra*		Paris*	2,87	219,46	parcour*	d	Paris*	3	485