IMPACT DE LA PRISE EN COMPTE DE LA MORPHOLOGIE SUR LE REPÉRAGE D'INFORMATION SUR LE WEB

Emmanuel CHIEZE[1], Louisette EMIRKANIAN et Lorne BOUCHARD
Université du Québec à Montréal

La prise en compte de la morphologie des langues naturelles en Repérage d'Information n'est pas une idée nouvelle, mais elle n'est pas intégrée aux moteurs de recherche sur le Web. La présente étude se propose donc d'évaluer la prise en compte de la morphologie du français à l'aide de requêtes typiques de celles soumises sur le Web. Étant donné que la lemmatisation ne peut être employée dans ce contexte, nous proposons une méthode en deux étapes consistant à transformer la requête originale en une série ordonnée de requêtes booléennes, puis à effectuer un enrichissement morphologique de ces dernières. Nous avons développé un outil permettant d'automatiser le processus, du traitement des requêtes à leur exécution et jusqu'au rapatriement des documents identifiés pour évaluation ultérieure. Nous présentons dans cette étude les résultats de cette évaluation qui rendent compte de l'apport des deux étapes de la transformation des requêtes à l'amélioration de la performance des recherches. La transformation de la requête originale en requêtes booléennes, l'affranchissement des spécificités du moteur de recherche et l'amélioration des résultats obtenus constituent certainement l'apport principal de la recherche. L'enrichissement morphologique des requêtes booléennes, qui se limite ici à la flexion des adjectifs et substantifs, améliore encore les résultats, mais de façon moins marquée. Cependant, le jeu de requêtes utilisé faisait en sorte que les résultats absolus étaient peu satisfaisants. Nous en expliquerons les raisons.

INTRODUCTION

Les moteurs de recherche plein-texte disponibles gratuitement sur le Web permettent de repérer les documents selon certaines expressions qu'ils contiennent. Ils ne tiennent cependant pas compte des variations

1 http://www.chieze.emmanuel@uqam.ca/

morphologiques, flexionnelles notamment, des substantifs, adjectifs et verbes. On peut donc s'attendre à ce que de tels outils omettent des textes pertinents dans les résultats des requêtes qui leur sont soumises.

L'on pourrait alors penser à développer un nouvel index du Web, basé sur la lemmatisation, démarche classique d'intégration de la morphologie d'une langue au Repérage de l'Information (RI). Cette technique réduit les termes liés morphologiquement à une forme canonique. Ainsi *manger*, *mange*, *mangeâmes*, *mangeoire* et toutes les autres variantes peuvent être représentées par *mang*, *manger* ou toute autre forme spécifique à cette classe de termes. Différents algorithmes, toujours approximatifs pour des raisons d'efficacité, ont été proposés dans le cadre du RI, dont un pour le français par Savoy (1993). Cette technique est appliquée aux textes du corpus préalablement à leur indexation, et aux requêtes des utilisateurs préalablement à leur exécution contre l'index. Le problème avec cette technique est triple : il faut tout d'abord s'assurer de la validité de l'algorithme utilisé, car le moindre changement qui y serait apporté nécessiterait une réindexation complète du corpus. Il faut ensuite concevoir un moteur de recherche intégrant l'algorithme et fonctionnant efficacement, ce qui est faisable mais coûteux. Et surtout, il faut indexer la portion francophone du Web, ce qui demande des ressources extrêmement importantes. Il est clairement préférable d'utiliser les moteurs existants, si possible, plutôt que d'en concevoir un nouveau.

Nous nous proposons donc ici d'évaluer une démarche automatisée d'intégration de la morphologie du français à l'utilisation de moteurs de recherche du Web au moment de l'exécution de la requête. Pour cela, nous préciserons la démarche d'intégration adoptée et les conditions expérimentales. Nous donnerons ensuite les résultats de l'expérience et conclurons en fournissant des pistes d'approfondissement de cette recherche.

INTÉGRATION DE LA MORPHOLOGIE AU RI SUR LE WEB

Description du corpus et des requêtes utilisés

Le corpus utilisé était la portion francophone du Web indexée par le moteur de recherche sélectionné. Les requêtes utilisées se basaient sur celles conçues directement en français pour les expériences de RI multilingue des conférences TREC-6, TREC-7 et TREC-8 (TREC, 2000). Elles avaient en commun de requérir comme résultats des documents de type informationnel,

donc rédigés, et non des pages promotionnelles. Les requêtes exécutées ont été obtenues à partir du titre des sujets TREC après transformation selon les consignes de Sanderson et Wilkinson (1997) : les articles et les prépositions ont été systématiquement supprimés du titre original, deux fautes d'orthographe ont été corrigées, et l'initiale du premier terme de la requête a été convertie en minuscule[2]. Ainsi l'une des requêtes fournies par TREC était :

```
<num> Number: 51
<F-title> Tremblement de terre au Yunnan

<F-desc> Description: Quels sont les rapports qui décrivent les
conséquences du tremblement de terre au Yunnan (Sud-Ouest de la
Chine)?

<F-narr> Narrative: Les documents disponibles traitent des
conséquences du tremblement de terre au Yunnan. Les rapports sur
les opérations d'aide aux victimes du tremblement de terre ont
aussi de l'intérêt.
```

La requête de base utilisée était donc tremblement terre Yunnan. Les requêtes utilisées figurent au Tableau 1 plus loin dans le texte.

Processus d'expansion morphologique

Au lieu de réduire les termes de la requête et ceux de l'index du corpus à leurs lemmes, l'expansion morphologique utilise l'index plein-texte du corpus et enrichit la requête au moment de son exécution. Elle ne constitue pas une stricte réciproque de la lemmatisation. Peu de recherches ont été publiées sur cette technique prometteuse. En effet, comme le mentionnent Lewis et Sparck-Jones (1996), l'inclusion de traitements de raffinement des requêtes au moment de leur exécution (*late-binding*) donne plus de souplesse quant à leur utilisation. Plusieurs techniques peuvent être utilisées de façon concurrente selon le contexte et les besoins de l'utilisateur, ou l'enrichissement peut être partiel.

Nous nous sommes restreints à la morphologie des substantifs et des adjectifs, telle que décrite dans (Grévisse, 1993), car aucun verbe ne faisait partie des termes des requêtes. De plus, nous avons exclu de notre étude pour l'instant la morphologie dérivationnelle, plus difficile à appliquer que la morphologie flexionnelle en raison des nombreux changements sémantiques associés. Nous avons utilisé un mécanisme d'expansion non-procédural, basé sur un ensemble de règles déclaratives contextuelles spécifiant des ensembles de patrons de mots apparentés d'un point de vue morphologique. Notons qu'aucune différence n'est établie entre substantifs et adjectifs, puisque leur

[2] La transformation de la première lettre du titre en minuscule pourrait par mégarde transformer une initiale de nom propre. C'est le risque à encourir pour obtenir des procédures systématiques, donc automatisables.

morphologie flexionnelle est identique. En général, l'application de la morphologie flexionnelle n'entraîne aucun changement sémantique. Trois exceptions sont toutefois apparues dans notre étude. Nous en rendons compte dans les résultats.

Intégration de l'enrichissement morphologique aux requêtes

Rappelons que deux types de requêtes peuvent être formulées sur le Web : le premier et le plus populaire, qui consiste à fournir une liste de termes, est le modèle vectoriel. Le moteur de recherche fournit alors les documents correspondant à cette liste par ordre de pertinence, établi selon une formule mathématique. Un exemple est *repérage information français anglais*, dont l'intention est d'identifier des documents sur le repérage d'information en français ou en anglais. Le second type est celui des requêtes booléennes, où les termes identifiant les documents sont reliés entre eux par des opérateurs booléens. Le moteur de recherche retourne alors l'ensemble des documents identifiés par la formule booléenne, sans ordre associé, car la notion de pertinence est binaire dans ce modèle : un document est pertinent ou il ne l'est pas. Un exemple dérivé du précédent pourrait être *repérage AND information AND* (*français OR anglais*). Nous indiquons ci-dessous comment enrichir les deux types de requêtes.

Enrichissement de requêtes booléennes

L'enrichissement morphologique est extrêmement simple dans le modèle booléen : il suffit de remplacer chaque terme de la requête par la disjonction de ses variantes. Ainsi la requête *cheval AND NOT élevage OR équitation* devient, une fois enrichie, (*cheval OR chevaux*) *AND NOT* (*élevage OR élevages*) *OR* (*équitation OR équitations*), en se limitant à la morphologie flexionnelle.

Enrichissement de requêtes vectorielles

L'enrichissement d'une requête vectorielle pose deux problèmes majeurs. Tout d'abord, cet enrichissement peut nuire aux résultats de la requête. Considérons la requête initiale *cheval course*, qui vise à identifier les documents sur les chevaux de course. Une fois enrichie, la requête devient *cheval chevaux course courses*. Des documents ne contenant que les termes *cheval* et *chevaux*, mais en très grand nombre, peuvent surclasser des documents traitant réellement des chevaux de course. Les résultats obtenus ne correspondent plus au besoin initial en information. Le second problème est l'impossibilité d'enrichir correctement des requêtes vectorielles spécifiant des termes obligatoires (Chieze, 2000). En conséquence, nous avons opté pour

une conversion préalable des requêtes vectorielles au modèle booléen suivie de leur enrichissement morphologique.

Conversion de requêtes vectorielles en requêtes booléennes

La démarche de conversion retenue est une extension de la *méthode du quorum* proposée par Cleverdon (1984), qui introduit un ordonnancement dans le modèle booléen afin de «simuler » en partie le modèle vectoriel. L'exécution par la méthode du quorum d'une requête constituée de n termes consiste à exécuter n requêtes. La première identifie les textes contenant les n termes, la seconde ceux en contenant exactement n-1, jusqu'à la dernière étape qui identifie ceux n'en contenant qu'un. Ainsi, si la requête initiale est constituée des trois termes A, B et C, les requêtes formulées pour mettre en oeuvre cette méthode sont :

Étape 1 : A AND B AND C
Étape 2 : (A AND B OR A AND C OR B AND C) AND NOT (A AND B AND C)
Étape 3 : (A OR B OR C) AND NOT (A AND B OR A AND C OR B AND C)

Toutefois, au sein d'une étape donnée, aucun ordonnancement des documents n'est prévu. Nous proposons de détailler les étapes de la méthode précédente en sous-étapes tenant compte de la proximité des termes entre eux, telle que définie par l'opérateur de proximité NEAR. On génère alors pour chaque requête obtenue selon la méthode du quorum l'ensemble des formulations de proximité possibles, selon la même méthode que celle illustrée ci-dessus, où NEAR remplace AND. Nous faisons figurer ci-dessous l'application de cette méthode, que nous appelons méthode du quorum à deux niveaux, à l'exemple précédent.

Étape 1 : Documents contenant les trois termes A, B et C :
- Sous-étape 1.1 : Documents contenant au moins une fois les trois termes A, B et C à proximité l'un de l'autre
  (A NEAR B NEAR C) AND (A AND B AND C)
- Sous-étape 1.2 : Documents contenant au moins une fois deux des trois termes A, B et C à proximité l'un de l'autre
  (A NEAR B OR A NEAR C OR B NEAR C) AND NOT (A NEAR B NEAR C) AND (A AND B AND C)
- Sous-étape 1.3 : Documents où les trois termes apparaissent isolément.
  (A AND B AND C) AND NOT (A NEAR B OR A NEAR C OR B NEAR C)

Étape 2 : Documents contenant exactement deux des trois termes A, B et C :
- Sous-étape 2.1 : Documents contenant au moins une fois ces deux termes à proximité l'un de l'autre
  (A NEAR B OR A NEAR C OR B NEAR C) AND (A AND B OR A AND C OR B AND C) AND NOT (A AND B AND C)
- Sous-étape 2.2 : Documents où les deux termes apparaissent isolément.
  (A AND B OR A AND C OR B AND C) AND NOT (A NEAR B OR A NEAR C OR B NEAR C) AND NOT (A AND B AND C)

Étape 3 : Documents contenant un seul des trois termes A, B ou C :
- Sous-étape 3.1 : (A OR B OR C) AND NOT (A AND B OR A AND C OR B AND C)

L'exécution de ce processus conduit à une explosion combinatoire. On ne peut donc l'appliquer qu'à des requêtes composées de peu de termes, telles que celles typiquement formulées sur le Web.

Affranchissement d'algorithmes d'ordonnancement méconnus

Lorsque l'on a recours au modèle vectoriel sur le Web, on se place à la merci des algorithmes d'ordonnancement des moteurs de recherche commerciaux, dont les détails ne sont pas publics. Plusieurs études comparatives de ces moteurs de recherche, citées par Courtois et Berry (1999), corroborent l'étonnement fréquent des utilisateurs devant le résultat de leurs requêtes. Les auteurs proposent alors trois critères d'évaluation de l'ordonnancement des moteurs de recherche, qui représentent selon eux les attentes courantes des utilisateurs de ces outils. Selon ces critères, le score d'un document devrait être d'autant plus élevé :

1. qu'il contient un nombre important de termes de la requête,
2. que ces derniers sont proches les uns des autres dans le document,
3. et qu'ils sont placés proches du début du document.

Le premier critère a préséance sur les deux autres, qui ne sont pas clairement ordonnés entre eux par les auteurs. Le premier critère est implémenté dans la méthode du quorum originale. Le deuxième critère est implémenté dans la méthode du quorum à deux niveaux, subordonné au premier. Ces critères sont surtout justifiés dans les cas où les termes se complètent les uns les autres selon un schéma de type déterminé-déterminant, comme dans la requête cheval course. Ils le sont moins dans les cas où les termes sont synonymes les uns des autres, comme dans la requête élevage boeuf bovin, ou lorsqu'ils constituent des alternatives, comme dans la requête hôtels Prague Budapest.

Le troisième critère ne peut être pris directement en compte au moyen des moteurs de recherche commerciaux du Web. Toutefois, en utilisant l'option facultative d'ordonnancement vectoriel des résultats d'une requête booléenne, laquelle tient compte en général de la position des termes dans le document, il est possible de l'intégrer indirectement à la méthode proposée. Ainsi on laisse une partie du contrôle de l'ordonnancement du résultat à un algorithme inconnu, mais dans une mesure moindre que lors de l'utilisation directe du modèle vectoriel. La méthode du quorum à deux niveaux permet donc de s'affranchir en partie des algorithmes des moteurs de recherche.

CONDITIONS EXPÉRIMENTALES

Sélection d'un moteur de recherche du Web

Les critères de sélection du moteur de recherche du Web étaient les suivants :

1. la possibilité de spécifier des requêtes booléennes complètes, incluant NEAR et le parenthésage,
2. une large couverture du Web,
3. la possibilité de restreindre les recherches aux documents écrits dans une langue donnée.

À notre connaissance, parmi les moteurs de recherche établis, seul AltaVista répondait au premier critère, crucial pour l'étude, tout en répondant aux deux autres critères également. C'est donc l'outil qui a été retenu pour l'étude. Notons que dans AltaVista, l'opérateur commutatif NEAR impose que ses deux opérandes soient situés dans un voisinage de 10 mots. Nous avons par ailleurs constaté une caractéristique non-documentée, à savoir l'existence d'une taille limite aux requêtes que peut traiter AltaVista : la longueur de l'URL traduisant la requête ne doit pas dépasser 1 113 caractères.

Évaluation de la pertinence des résultats et de la performance des requêtes

L'évaluation de la pertinence des résultats se basait sur la section *Narrative* du sujet TREC associée à chaque requête. Lorsque au moins une portion du document satisfaisait aux critères du *Narrative*, le document était jugé pertinent. Sinon, il était jugé non-pertinent, même s'il semblait contenir des hyperliens vers des documents pertinents. Il faut souligner le fait qu'un document pouvait être jugé pertinent même si seule une faible portion de ce dernier l'était. Pour des raisons économiques et pratiques, l'évaluation de la pertinence des documents a été menée par Chieze (2000).

Les mesures traditionnelles en RI, rappel et précision, développées dans le cadre de corpus relativement petits, statiques et connus de façon exhaustive, ne s'appliquent pas au Web (Buckland et Gey, 1994). L'évaluation du rappel est impossible pour des raisons théoriques, et celle de la précision l'est pour des raisons pratiques. En conséquence, nous avons utilisé la précision à 20 documents, caractéristique de la performance des requêtes à leur début, qui consiste à évaluer le nombre de documents pertinents sur les 20 premiers documents retrouvés (Hawking *et al.*, WWW8), (Leighton, Srivastava, 1997). Cette mesure reflète l'adage selon lequel on ne se préoccupe pas du rappel sur le Web et semble traduire le comportement de

plusieurs utilisateurs de moteurs de recherche du Web. L'utilisation de cette mesure suppose toutefois que le résultat retourné par chaque moteur de recherche soit ordonné. Il est donc indispensable d'utiliser l'option d'ordonnancement vectoriel des résultats d'une requête booléenne.

Exécution des requêtes

Nous avons exécuté chaque requête selon trois modes différents :

1. vectoriel : requêtes issues du titre de TREC,
2. booléen : sous-requêtes issues de l'application de la méthode du quorum à deux niveaux à la requête initiale. Le but est d'évaluer l'impact de cette méthode sur les résultats indépendamment de l'enrichissement morphologique,
3. et booléen enrichi : sous-requêtes issues de l'enrichissement morphologique des sous-requêtes précédentes.

Les requêtes booléennes ont été exécutées avec l'option d'ordonnancement des résultats, où les termes originaux de la requête ont été fournis pour l'ordonnancement afin d'assurer la validité de la précision à 20 documents. Par ailleurs, nous n'étions pas complètement assurés du comportement de l'opérateur NEAR dans des expressions telles que *A NEAR B NEAR C*, qui devrait identifier les documents contenant A, B et C dans un voisinage de dix mots indépendamment de leur ordre. Lors de la recherche, l'opérateur NEAR n'était pas commutatif dans cette situation. Plutôt que d'utiliser un opérateur à la sémantique incertaine, nous avons préféré transformer les expressions du type A_1 NEAR A_2 NEAR ... NEAR A_n, où n est supérieur à 2, en (A_1 NEAR A_2) AND (A_2 NEAR A_3) AND ... (A_{n-1} NEAR A_n). Cette dernière expression, version affaiblie de la première, requiert seulement que les termes soient proches les uns des autres deux à deux.

Un outil utilisant la librairie WWW::Search (version 204, xoom.com, 1999) a été écrit en Perl version 5 pour Windows-98 (build 503, ActiveState, 1999) pour automatiser le processus de réécriture de la requête, de son exécution sur AltaVista et de la récupération en local des 20 premiers URLs identifiés comme le résultat de la requête pour permettre leur évaluation ultérieure hors-ligne. L'outil génère pour chaque requête originale la liste des documents obtenus, toutes exécutions confondues, afin d'en permettre l'évaluation à l'aveugle, *i.e.* sans savoir quelle méthode les a identifiés.

RÉSULTATS DE L'ÉVALUATION

Analyse qualitative des résultats absolus

L'analyse qualitative est une compilation des remarques établies lors de l'évaluation de la pertinence des documents. Si l'on s'en tient aux résultats obtenus en valeur absolue, présentés au Tableau 1, ceux-ci ne sont pas fameux, puisque la meilleure méthode en apparence repère moins de trois documents pertinents sur 20 en moyenne. Plusieurs types de facteurs, énumérés ci-dessous permettent d'expliquer en partie ces mauvais résultats :

1. **Facteurs spécifiques au Web** : certaines pages encore identifiées dans l'index d'AltaVista ont disparu du Web ou ont subi une modification majeure. Par ailleurs, il existe des pages identiques sur des serveurs différents ou quasi-identiques sur le même serveur, et dans le cas fréquent où elles n'étaient pas jugées pertinentes, cela diminuait grandement la précision à 20 documents.
2. **Facteurs spécifiques à AltaVista** : la limitation de la taille des requêtes nous a conduit à apposer la mention N/A dans certains cas. Par ailleurs, l'indexation de métabalises invisibles pour le lecteur, mais ayant parfois peu de liens avec le contenu du document, a parfois nui aux résultats.
3. **Facteurs spécifiques à l'outil d'exécution des requêtes** : certaines pages n'ont pas pu être rapatriées, soit qu'elles étaient redirigées, soit qu'elles ne pouvaient être consultées que par un fureteur donné, soit qu'elles contenaient des cadres, non traités par notre outil. Les cadres semblaient cependant être utilisés principalement par des sites promotionnels, ayant surtout recours à une présentation graphique de l'information, alors que nous étions surtout à la recherche de sites informationnels, faisant plutôt appel à des textes rédigés et de taille plus substantielle.
4. **Facteur spécifique au processus d'enrichissement utilisé** : dans le cas des termes *affaire*, *dette* et *stupéfiants*, l'enrichissement morphologique entraînait l'apparition indésirable d'un nouveau sens ou d'un nouvel emploi du terme. Ainsi *affaires* semblait utilisé sur le Web surtout dans son sens économique et non dans son sens politique, au contraire d'*affaire*. *Dettes* faisait surtout référence aux dettes personnelles, alors que *dette* identifiait une collectivité le plus souvent. Enfin *stupéfiant* était surtout l'adjectif synonyme

d'*étonnant* alors que *stupéfiants* était le substantif équivalent à *drogues*. Notons dans ce dernier cas le glissement de catégorie grammaticale.

5. **Inadéquation des requêtes à l'univers documentaire** : les requêtes avaient été conçues pour un corpus d'articles de journaux de la Suisse romande, et non pour le Web, et une partie d'entre elles datent. Il est donc possible que certaines d'entre elles n'aient que peu de documents associés sur le Web, soit que la requête ait été conçue pour tester la capacité d'un outil de repérage d'information à identifier un unique article dans un corpus donné, différent du Web, soit que le Web était moins utilisé à l'époque où le sujet était d'actualité, soit encore que le contenu du Web évolue rapidement en ne gardant essentiellement que les informations les plus récentes.

6. **Imprécision des requêtes** : les requêtes, très brèves, ne traduisaient pas adéquatement le besoin en information sous-jacent, toujours très précis. De plus, l'absence des prépositions entraînait l'apparition de nouveaux sens imprévus. Ainsi *dette Suisse* pouvait référer à la dette de la Suisse envers des pays tiers ou à celle de pays tiers envers la Suisse. La thématique des documents recherchés était donc exprimée de façon très imprécise. Par ailleurs, les critères d'évaluation des requêtes, établis par la section *Narrative*, impliquaient notamment que les documents soient de type informationnel, et non promotionnel. Les évaluations globales de pertinence que nous avons menées ont tenu compte de ces critères, alors qu'ils n'étaient pas exprimés par les requêtes réellement exécutées. Notons que dans le cadre des expériences de TREC, ce problème est marginal du fait que le corpus utilisé est homogène du point de vue du type de documents contenus. Son hétérogénéité se situe uniquement au niveau de la thématique des documents. Nous observons ici une différence majeure de nature entre le RI sur le Web et le RI classique.

Tableau 1 Résultats bruts de l'expérience : nombre de documents pertinents et de documents non évalués par requête et par type d'exécution.

Le premier nombre de chaque colonne, en gras, indique le nombre de documents pertinents parmi les 20 premiers résultats de chaque exécution de chaque requête. Le second nombre, en italiques, indique le nombre de documents non-disponibles pour évaluation hors-ligne, quelle qu'en soit la raison.

Ref Trec	Requête	Type d'exécution					
		vectoriel		booléen		booléen enrichi	
CL1	affaire Waldheim	**0**	*4*	**0**	*10*	**0**	*10*
CL2	mariages	**0**	*8*	**0**	*8*	**0**	*9*
CL3	stupéfiants	**1**	*4*	**1**	*4*	**1**	*4*
CL4	recyclage ordures	**5**	*2*	**5**	*2*	**5**	*2*
CL5	acupuncture	**0**	*5*	**0**	*5*	**0**	*5*
CL6	pollution causée automobile	**1**	*7*	**5**	*6*	**8**	*5*
CL7	éducation sexuelle	**1**	*1*	**1**	*1*	**1**	*1*
CL8	vitesse autoroutes Suisse	**0**	*7*	**0**	*10*	**0**	*9*
CL9	effets déforestation	**2**	*5*	**11**	*3*	**10**	*3*
CL10	voitures solaires	**1**	*5*	**1**	*9*	**1**	*9*
CL11	coton écologique	**1**	*8*	**2**	*9*	**1**	*10*
CL12	culture écologique	**0**	*3*	**0**	*3*	**0**	*4*
CL13	processus paix Moyen-Orient	**2**	*3*	**0**	*5*	**0**	*4*
CL14	terrorisme international	**1**	*5*	**8**	*2*	**10**	*4*
CL15	peine mort	**15**	*3*	**18**	*1*	**18**	*1*
CL16	tuberculose	**5**	*6*	**4**	*6*	**5**	*6*
CL17	pommes terre	**5**	*5*	**5**	*5*	**6**	*5*
CL18	parfum	**0**	*5*	**0**	*6*	**0**	*6*
CL19	vins	**0**	*10*	**0**	*9*	**0**	*10*
CL20	maltraitance enfants	**3**	*5*	**7**	*5*	**7**	*5*
CL21	effets chocolat santé	**2**	*5*	**3**	*6*	**3**	*8*
CL22	fast-food Europe	**0**	*13*	**1**	*8*	**1**	*8*
CL23	ours peluche	**0**	*5*	**0**	*3*	**0**	*4*
CL24	ouvriers étrangers Europe	**0**	*8*	**1**	*6*	**0**	*11*
CL25	effets protection éléphants afrique commerce international ivoire	**2**	*3*	**N/A**	*N/A*	**N/A**	*N/A*
26	grogne Lötschberg	**0**	*4*	**0**	*4*	**0**	*4*
27	oléoducs mondiaux	**0**	*4*	**0**	*4*	**0**	*3*
29	destruction forêt tropicale Amérique Sud	**1**	*8*	**5**	*10*	**8**	*7*
29	affaire Ustica	**0**	*4*	**0**	*2*	**3**	*1*
30	famine Soudan	**3**	*4*	**4**	*5*	**4**	*5*
31	conséquences réunification allemande	**2**	*4*	**2**	*7*	**4**	*10*
32	faillite "Banco Ambrosiano"	**0**	*2*	**0**	*2*	**0**	*2*
33	génie génétique	**1**	*8*	**1**	*8*	**1**	*8*
34	nouveau rideau fer	**0**	*4*	**2**	*4*	**2**	*6*
35	limitations importations UE	**1**	*4*	**1**	*11*	**6**	*7*
36	voleurs art	**0**	*4*	**1**	*4*	**1**	*5*
37	unité franco-allemande	**0**	*4*	**0**	*3*	**0**	*4*
38	conversion dette Pologne	**0**	*5*	**1**	*8*	**1**	*6*
39	immigration racisme	**0**	*7*	**1**	*6*	**0**	*7*
40	jet supersonique Concorde	**0**	*4*	**1**	*9*	**1**	*6*
41	statut militaire Allemagne unifiée	**1**	*5*	**1**	*12*	**N/A**	*N/A*
42	compagnies aériennes européennes américaines	**0**	*4*	**1**	*9*	**6**	*7*
43	kidnapping	**0**	*5*	**0**	*5*	**0**	*5*
44	liberté presse Pologne	**0**	*8*	**0**	*3*	**0**	*3*
45	elections Bosnie Herzégovine	**10**	*7*	**13**	*0*	**14**	*0*
46	dette Suisse	**0**	*5*	**2**	*2*	**1**	*4*
47	trou couche ozone	**4**	*4*	**0**	*7*	**0**	*7*
48	trains grande vitesse	**3**	*9*	**1**	*7*	**2**	*10*
49	protection environnement sites	**1**	*8*	**7**	*9*	**N/A**	*N/A*

	usines chimiques						
50	accidents route	**2**	*6*	**2**	*6*	**2**	*6*
51	tremblement terre Yunnan	**1**	*6*	**1**	*10*	**1**	*8*
52	chômage France	**4**	*5*	**5**	*5*	**4**	*5*
53	libre circulation personnes Europe	**6**	*9*	**8**	*4*	**8**	*4*
54	stocks poisson baisse	**4**	*4*	**14**	*2*	**16**	*1*
55	statistiques avortement	**2**	*1*	**1**	*1*	**1**	*1*
56	exploitation économique fond marin	**0**	*7*	**1**	*5*	**1**	*6*
57	maintien paix OUA	**7**	*7*	**6**	*10*	**7**	*10*
58	baisse tourisme Adriatique	**0**	*8*	**0**	*10*	**0**	*10*
59	exportation médicaments dangereux	**0**	*7*	**0**	*6*	**N/A**	*N/A*
60	vol oiseaux rares	**0**	*6*	**0**	*8*	**0**	*8*
61	armée allemande forces ONU	**0**	*7*	**2**	*6*	**N/A**	*N/A*
62	munitions Seconde Guerre Mondiale	**0**	*8*	**0**	*9*	**0**	*9*
63	rumantsch grischun	**2**	*3*	**3**	*3*	**3**	*3*
64	utilisation engrais chimiques	**1**	*6*	**3**	*1*	**3**	*1*
65	industrie européenne film	**1**	*9*	**9**	*4*	**15**	*4*
66	exportation armes Turquie	**3**	*2*	**4**	*5*	**4**	*4*
67	débris spatiaux	**4**	*5*	**4**	*5*	**4**	*5*
68	homosexualité loi	**4**	*9*	**7**	*4*	**7**	*5*
69	marché cuir	**5**	*6*	**3**	*8*	**3**	*8*
70	séparatistes catalans galiciens	**0**	*7*	**0**	*4*	**0**	*4*
71	sauvetage dauphins	**0**	*4*	**0**	*7*	**0**	*9*
72	traitement déchets	**1**	*6*	**3**	*3*	**3**	*3*
73	normes protection professionnelle	**0**	*5*	**2**	*8*	**1**	*8*
74	traitement déchets nucléaires	**4**	*5*	**4**	*14*	**4**	*14*
75	tremblements terre Sicile	**1**	*8*	**2**	*8*	**1**	*8*
76	premières traces Homme	**0**	*6*	**0**	*5*	**0**	*4*
77	euthanasie illégale	**3**	*2*	**4**	*3*	**5**	*5*
78	politique économique Slovénie	**0**	*5*	**1**	*7*	**1**	*5*
79	dissolution parlement Kosovo	**0**	*9*	**0**	*5*	**0**	*5*
80	communistes Parlement Européen	**0**	*5*	**0**	*3*	**0**	*4*
81	protection animaux	**1**	*8*	**0**	*3*	**1**	*1*
	Médiane	**1**	*5*	**1**	*5*	**1**	*5*
	Moyenne	1.6	5.6	**2.6**	5.7	**3.0**	5.6

Comparaison statistique des types d'exécution

Nous disposons pour l'ensemble des 81 requêtes de la précision à 20 documents mesurée dans chacun des trois types d'exécution. Dans la mesure où le nombre de documents examinés varie selon les requêtes et les types d'exécution, et que ces variations sont dues en partie à l'outil de rapatriement des pages et non seulement au moteur de recherche employé, nous avons préféré ne pas pénaliser une méthode en raison d'un moins grand nombre de documents examinables. Pour pallier ce problème, nous avons plutôt utilisé une variante de la précision à 20 documents, la précision sur le nombre de documents disponibles (PDD), soit P/(20 – ND), où P est le nombre de documents pertinents et ND le nombre de documents non-disponibles.

Nous avons procédé aux comparaisons deux à deux des types d'exécution vectoriel et booléen, et booléen et booléen enrichi, en utilisant le

test unilatéral du rang signé de Wilcoxon (Conover, 1980). Si l'on désigne par E1 et E2 les types d'exécution testés, où E2 a eu une meilleure PDD que E1 pour la majorité des requêtes exécutées, on testera l'hypothèse nulle H_0 selon laquelle la moyenne de la PDD de E1 est supérieure ou égale à celle de E2, associée à l'hypothèse alternative H_A que la moyenne de la PDD de E2 est strictement supérieure à celle de E1. Les résultats statistiques ont été compilés au moyen du logiciel SAS à partir des données de base du Tableau 1, et figurent dans le Tableau 2.

Tableau 2 Comparaison des différences de PDD entre les types d'exécution

%+ : pourcentage des requêtes où le premier type d'exécution a donné de meilleurs résultats que le second
%0 : pourcentage des requêtes où il n'y a pas de différence entre les deux types d'exécution
%- : pourcentage des requêtes où le premier type d'exécution a donné de moins bons résultats que le second
N : nombre de requêtes où une différence de PDD entre les deux types d'exécution a été observée (car seules ces requêtes sont prises en compte dans le test de Wilcoxon)
N' : nombre de requêtes où le premier type d'exécution a donné de meilleurs résultats que le second
T : variable du test
p : degré de confiance associé à T

Comparaison	%+	%0	%-	N	N'	T	p	Conclusion
vectoriel / booléen	11 %	38 %	51 %	50	9	-395	$0{,}5\ 10^{-4}$	H_0 rejetée au niveau 1%
booléen / booléen enrichi	16 %	55 %	29 %	34	12	-149	$4{,}3\ 10^{-3}$	H_0 rejetée au niveau 1%

La conversion d'une requête vectorielle dans le modèle booléen selon la méthode du quorum à deux niveaux permet d'améliorer la performance de la requête. Cela confirme les résultats de Courtois et Berry (1999) selon lesquels AltaVista ne respecte pas tout à fait les critères d'ordonnancement que les auteurs proposent. Toutefois, seule une requête sur deux environ bénéficie d'une amélioration, qui peut alors être suffisamment notable pour justifier l'application de la méthode, compte tenu du faible risque de dégradation des résultats.

L'enrichissement morphologique appliqué au modèle booléen améliore la performance de ce dernier, mais de façon marginale. Seules 30 % des requêtes ont bénéficié de l'enrichissement morphologique booléen dans notre jeu de tests, et 15 % ont vu leur résultat empirer, faiblement cependant, suite à cet enrichissement. Cette dégradation, relative à la mesure utilisée, la PDD, est contre-intuitive, car l'enrichissement morphologique ne peut qu'augmenter

le rappel de la requête. Mais l'enrichissement peut avoir des effets négatifs tant sur la précision globale de la requête que sur sa PDD. Cela s'explique par deux facteurs principalement :

- les changements sémantiques accompagnant certains passages du singulier au pluriel ou vice-versa, comme dans le cas de *dettes*, de *stupéfiants* ou d'*affaire*. Cela concerne la requête 46.
- le fait que l'ordonnancement opéré par le moteur de recherche, hors de notre contrôle, affecte la dernière sous-requête exécutée dans la méthode du quorum à deux niveaux. Dans le cas de la requête *effets déforestation* par exemple, seule la sous-requête *effets NEAR déforestation* est exécutée dans le modèle booléen, et seule *(effets OR effet) NEAR (déforestation OR déforestations)* est exécutée dans le modèle booléen enrichi, puisque ces sous-requêtes identifient chacune plus de 20 documents et que nous ne nous intéressons qu'aux 20 premiers d'entre eux. Or ces sous-requêtes, ordonnancées toutes deux selon les termes de la requête initiale, identifient un ensemble de documents différents, et le rang de chacun d'entre eux peut donc être différent dans l'un et l'autre résultat. Un document identifié dans les 20 premiers dans le mode booléen peut donc se retrouver hors de cette fenêtre dans le mode booléen enrichi.

Par ailleurs, certains résultats semblent à première vue surprenants, et sont liés à des usages rares ou spécialisés d'une variante flexionnelle peu usuelle. Ainsi la requête *tuberculose* donne de meilleurs résultats une fois enrichie. Cela tient au fait que les textes médicaux identifient plusieurs types de tuberculose, et qu'il est donc légitime de parler des tuberculoses. Au contraire, *chômage France* voit ses résultats empirer une fois l'enrichissement effectué sur le terme *chômage*. Cela tient au fait qu'il existe quelques documents sur le Web mentionnant le terme *chômages*, soit en raison de fautes de frappe, soit délibérément comme dans le cas du titre d'un livre « La France des chômages ». Les deux exemples sont tous deux contre-intuitifs car seule une variante des termes est couramment utilisée.

Par ailleurs, les termes d'une requête peuvent en général être combinés de multiples façons, mais l'utilisateur a tendance implicitement à ne considérer que celles correspondant à son besoin en information, d'où son étonnement lorsque des textes identifiés n'ont aucun rapport avec ce dernier. Ainsi l'interprétation *dette de la Suisse envers des pays tiers* de la requête *dette Suisse* peut-elle ne pas être entrevue si le thème recherché est celui de *la dette de pays tiers envers la Suisse*. Inversement, certaines combinaisons apparemment aberrantes, telles que *euthanasie NEAR illégales*, et repérées

par l'enrichissement morphologique adopté ici, permettent néanmoins d'identifier des documents que ni la requête de base ni une requête enrichie de façon apparemment raisonnée, telle que *(euthanasie NEAR illégale) OR (euthanasies NEAR illégales)*, n'auraient trouvés : ainsi l'exemple d'un document ne contenant qu'une seule occurrence des termes de la requête, dans le contexte *En Australie, l'euthanasie active et l'aide au suicide sont illégales...* Il ne faut donc jamais perdre de vue le fait que les moteurs de recherche se limitent à identifier des textes contenant les termes à proximité les uns des autres, et non selon un patron syntaxique déterminé.

Enfin, certaines requêtes du jeu de test se prêtaient peu à un enrichissement morphologique flexionnel, ce qui a pu contribuer au faible apport de la prise en compte de la morphologie flexionnelle. Citons entre autres les requêtes *ours peluche*, *unité franco-allemande*, ou encore *grogne Lötschberg*.

En conclusion, notons donc que l'on peut ordonner les modèles selon leur performance, pour obtenir le classement suivant : vectoriel < booléen < booléen enrichi.

LIMITES DE L'ÉTUDE

Il est extrêmement difficile de s'assurer de la représentativité d'un jeu de requêtes, car l'ensemble infini des besoins en information n'est pas descriptible. L'étude respecte cependant les standards du RI en ce qui concerne le nombre de requêtes utilisées, et ces dernières étaient représentatives de celles typiquement émises sur le Web, tant en ce qui concerne leur structure que leur longueur (Silverstein *et al.*, 1998), mais non nécessairement en ce qui concerne leur sujet. Nous aurions peut-être abouti à des conclusions différentes en utilisant des requêtes traduisant un besoin en information plus général, ou appelant à identifier des sites promotionnels plutôt qu'informationnels. De plus, du fait même que l'expérience a été menée sur le Web, elle n'est pas parfaitement reproductible puisque le corpus et l'index d'AltaVista peuvent évoluer d'une exécution de l'expérience à l'autre. En revanche, l'utilisation du Web nous a permis d'identifier certains problèmes qu'un environnement fixe et artificiel n'aurait peut-être pas comportés. Par ailleurs, le mécanisme de rapatriement des URL pourrait être amélioré afin d'obtenir des résultats plus représentatifs de ceux obtenus par des utilisateurs en ligne. Enfin une autre limite importante et intrinsèque à l'étude était le fait qu'elle ait été menée en français et qu'elle ne tenait compte que d'une partie de la morphologie flexionnelle de cette langue.

CONCLUSION

L'étude entreprise avait pour objectif principal de déterminer l'intérêt de prendre en compte la morphologie du français lors de recherches d'information sur le Web. Pour cela, il était nécessaire de spécifier un mécanisme d'intégration de la morphologie aux requêtes, qui consistait dans un premier temps à transformer les requêtes du modèle vectoriel dans le modèle booléen selon la *méthode du quorum à deux niveaux*, puis à procéder à l'enrichissement morphologique des requêtes obtenues en substituant à chaque terme l'ensemble de ses variantes définies par la morphologie flexionnelle. Un objectif secondaire de la recherche consistait à vérifier si la méthode du quorum à deux niveaux permettait de formuler des requêtes plus performantes que la formulation vectorielle originale.

Le premier résultat obtenu, probablement généralisable à n'importe quelle langue, est qu'il est préférable d'exécuter une requête vectorielle sur AltaVista selon la méthode du quorum à deux niveaux plutôt que directement dans le modèle vectoriel. Le second résultat est que l'enrichissement morphologique flexionnel de requêtes booléennes est dans l'ensemble un facteur d'amélioration de la performance des requêtes, mais dans une faible mesure. Il est cependant impossible de généraliser ce résultat à d'autres langues, alors qu'en revanche, le choix du moteur de recherche ne devrait pas influencer le résultat. Le troisième résultat, non lié aux techniques d'enrichissement utilisées, consiste à confirmer la différence profonde de nature entre le RI sur le Web et le RI classique.

Il existe au moins trois directions naturelles pour poursuivre la recherche. La première consisterait à procéder à un enrichissement morphologique complet (flexionnel et dérivationnel), mais sélectif, afin d'éviter tout glissement sémantique ou de catégorie. La seconde consisterait à identifier les locutions et à affiner leur expansion morphologique. Ainsi, dans le cas de la requête *destruction forêt tropicale*, *forêt tropicale* forme la base d'une locution, identifiant un concept, et a comme seule variante flexionnelle *forêts tropicales*, alors que *destruction forêt* est plutôt le squelette d'une expression nominale liant deux concepts entre eux, et peut avoir comme variantes flexionnelles *destruction forêts*, *destructions forêt* ou *destructions forêts*. La troisième, qui nous semble la plus importante, consisterait à mieux tenir compte de certaines spécificités du Web comme univers documentaire. Il faudrait au minimum trouver des moyens de caractériser les types de documents ou de textes indexés, et permettre au recherchiste de spécifier ce critère de façon explicite dans ses requêtes. Il s'agit d'un programme de recherche en soi, qui devra préciser en premier lieu la notion de type de

documents et de type de textes, employée de façon très informelle ici. En parallèle, il serait important de réviser la définition de la pertinence, en décomposant cette notion en plusieurs dimensions comme le proposent Cosijn et Ingwersen (2000), plutôt que de la voir comme un bloc monolithique, et en en adaptant les procédures d'évaluation de façon à comparer deux méthodes sur les seules dimensions où elles peuvent différer. Dans le cas présent, cela nous aurait entre autres conduits à ne pas tenir compte du type de document lors de l'évaluation de la pertinence, puisque ni les requêtes de base ni les requêtes transformées ou enrichies n'exprimaient ce critère.

RÉFÉRENCES BIBLIOGRAPHIQUES

ActiveState. 1999. Perl pour Windows 98. Adresse URL :

http://www.activestate.com/ActivePerl, (consulté le 12/08/1999)

AltaVista. 2000. *Accès au moteur de recherche* (page consultée le 21/04/2000), Adresse URL : http://www.altavista.com .

BUCKLAND, Michael et Gey, Fredric. 1994. «The Relationship between Recall and Precision », *Journal of the American Society for Information Science*, vol. 45, no 1, p. 12-19.

CHIEZE, Emmanuel. 2000. *Prise en compte de la morphologie du français dans le repérage d'information sur le Web*, Mémoire de maîtrise, Université du Québec à Montréal, Montréal, p. 26-28.

CLEVERDON, C.W. 1984. «Optimizing Convenient On-line Access to Bibliographic Data Bases », Information Services and Use, vol. 4, no 1, p. 37-47.

CONOVER, W.J. 1980. *Practical Nonparametric Statistics*, John Wiley & Sons, New York, p. 280-283.

COURTOIS, Martin P. et BERRY, Michael W. 1999. « Results Ranking in Web Search Engines », Online, Mai 1999, (consulté le 1999/09/03). Adresse URL : http://www.onlineinc.com/onlinemag/OL1999/courtois5.html.

Grévisse. 1993. *Le bon usage*, Paris, éditions Duculot, 1762 p.

HAWKING David, CRASWELL Nick et THISTLEWAITE Paul. 1999. « Overview of TREC-7 Very Large Collection Track », *TREC-7 Proceedings*, (consulté le 19/11/1999), Adresse URL : http://trec.nist.gov/pubs/trec7/papers/vlc_overview.pdf .

COSIJN, Erica et INGWERSEN, Peter. 2000. « Dimensions of relevance », *Information Processing and Management*, no 36, p. 533-560.

LEIGHTON, H. Vernon et SRIVASTAVA, Jaideep, 1997. « Precision among World Wide Web Search Services (Search Engines) : AltaVista, Excite, Hotbot, Infoseek, Lycos », Technical Report, (consulté le 15/11/1999), Adresse URL : http://www.winona.msus.edu/library/webind2/webind2.htm.

LEWIS, David et SPARCK Jones, Karen. 1996. « Natural language processing for information retrieval », *Communications of the ACM*, vol. 39, no 1, p. 92-101.

SANDERSON, Mark et WILKINSON, Ross, 1997. « Guidelines for generating very short TREC queries », (consulté le 20/11/1999), Adresse URL : http://dis.shef.ac.uk/mark/guidelines.html.

SAVOY, Jacques. 1993. «Stemming of French Words Based on Grammatical Categories », *Journal of the American Society for Information Science*, vol. 44, no 1, p. 1-9.

SILVERSTEIN C., HENZINGER M., MARAIS H., MORICZ M. 1998. « Analysis of a Very Large AltaVista Query Log », *SRC Technical Note, Digital Equipment Corporation*, 1998-014. Adresse URL : ftp://ftp.digital.com/pub/DEC/SRC/technical-notes/SRC-1998-014.pdf (consulté le 13/10/1999).

TREC, 2000. « Text REtrieval Conference (TREC) Data – Non English Test Questions (Topics) File List », (consulté le 09/03/1999).URL : http://trec.nist.gov/data/topics_noneng/index.html.

xoom.com. 1999. WWW::Search, version Windows. Adresse URL : http://members.xoom.com/WWW_Search. (consulté le 29/10/1999).