Morphologie et recherche d'information

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel
Jacques.Savoy@unine.ch

Résumé

Dans cet article, nous décrivons comment concevoir des algorithmes de dépistage de l'information adaptés pour d'autres langues que l'anglais, et plus précisément pour le français, l'allemand, l'italien et l'espagnol. En effet, si de nombreuses études ont proposé diverses approches pour la langue de Shakespeare, la morphologie des autres langues européennes présente une plus grande richesse dont les systèmes de dépistage devraient tenir compte. Les campagnes d'évaluation CLEF s'intéressent particulièrement à ces questions et elles ont permis la création de collections de documents en plusieurs langues. Sur la base de ces corpus, nous proposons différentes approches visant à éliminer les marques liées à la morphologie flexionnelle et dérivationnelle pour le français, l'allemand, l'italien et l'espagnol. En se basant sur dix stratégies de recherche, nous avons évalué ces stratégies afin de dégager quelques lignes directrices concernant l'importance de la morphologie en recherche d'information.

INTRODUCTION

Actuellement, nous stockons un volume sans cesse croissant d'information sur support électronique et la Toile représente un des exemples les plus connus. Les documents mémorisés étant rédigés dans différentes langues, nous cherchons actuellement le moyen le plus efficace de les indexer et de permettre une recherche en-ligne efficace. Ainsi, dans les pays multilingues à l'image de la Belgique, du Canada ou de la Suisse, dans les organisations internationales comme l'OMC ou le Parlement européen ou dans les firmes multinationales, le dépistage automatique en plusieurs langues correspond à une demande de plus en plus pressante.

Dans le but de proposer de tels systèmes pour différentes langues européennes, nous devons savoir si les algorithmes disponibles pour la langue anglaise peuvent être adaptés facilement pour d'autres langues comme le français, l'italien, l'espagnol ou l'allemand. En effet, ces dernières présentent une morphologie plus riche [27] que celle de l'anglais et la langue de Goethe utilise, de manière très fréquente, la concaténation de mots pour générer des mots composés. De plus, nous entendons privilégier des approches automatiques n'exigeant pas la présence de dictionnaires, outils souvent pas disponibles et dont la consultation ralentit le traitement informatique. Comme critère de succès, nous avons choisi non pas la qualité linguistique des traitements suggérés mais la qualité lors du dépistage de l'information.

Afin d'apporter une réponse à ces différentes questions, la première section proposera, pour chaque langue, une liste de mots-outils que la machine peut ignorer lors de l'indexation d'une part, et, d'autre part, un algorithme de suppression des marques liées à la morphologie afin de regrouper sous le même terme d'indexation des formes flexionnelles et dérivationnelles différentes. Dans une deuxième section, nous présenterons différentes stratégies d'indexation et de recherche pouvant gérer diverses langues européennes. Afin de valider nos propositions, la troisième section présentera la collection de documents multilingues du forum CLEF-2001 [16] nous permettant de dégager quelques lignes directrices dans la conception de système de dépistage de l'information œuvrant dans des langues diverses.

MORPHOLOGIE ET RECHERCHE D'INFORMATION

La plupart des langues européennes (incluant le français, l'italien, l'allemand ou l'espagnol) partagent plusieurs caractéristiques propres à la langue de Shakespeare comme, par exemple, la séparation des mots marquée de manière conventionnelle, l'adjonction de suffixes pour indiquer le genre, le nombre ou le temps. Pour ces langues, nous nous appuierons donc sur les stratégies de dépistage de l'information mises au point pour l'anglais mais en les adaptant. Cette dernière étape comprend l'élaboration d'une liste de mots-outils, sujet de notre première sous-section. Dans la deuxième sous-section, nous aborderons le problème de l'élimination des flexions en français, allemand, italien et espagnol tandis que la troisième sous-section présentera les questions liées à la morphologie dérivationnelle des langues française et allemande.

Mots-outils

Afin d'adapter les algorithmes d'indexation disponibles pour la langue anglaise [21], [29] aux autres langues étudiées dans cet article, nous devons d'abord proposer un ensemble de mots-outils peu ou pas porteurs de sens que la machine peut ignorer lors de l'indexation. Dans ce but et suivant les indications de Fox [6], nous avons d'abord retenu les 200 formes les plus fréquentes des langues des corpus de CLEF [17]. De ces listes, nous avons éliminé tous les nombres (comme « 1 » ou « 1994 ») et certains noms comme le mot allemand « Prozent » (classé en 69e position), le nom italien « Italia » (classé 87) ou le terme espagnol « política » (classé 131). Ensuite, nous avons rajouté à ces listes les déterminants, pronoms, conjonctions et prépositions. Ces listes de mots-outils[1] comprennent 217 mots pour le français, 431 pour l'italien, 294 pour l'allemand et 351 pour l'espagnol. Pour l'anglais, nous avons retenu la liste fournie par le système SMART (incluant 571 mots). Notons que pour la langue allemande, nous nous sommes également appuyés sur les études de linguistique quantitative d'Ortmann [13], [14].

L'emploi de telles listes possède l'avantage de réduire les appariements peu opportuns entre requêtes et documents basés sur des formes très fréquentes. Ainsi, dépister un document parce qu'il possède des formes comme « voici », « un », « je » ou « la » en commun avec la requête ne constitue pas une décision très pertinente. Mais la présence de tels mots permet parfois de définir très précisément le sens rattaché à un terme comme dans les expressions « la Poste », « une poste », « un poste » ou « je poste » dans lesquels la sémantique précise dépend fortement du mot précédant le mot « poste ». L'emploi de telles listes permet de plus de réduire la taille des fichiers inverses d'environ 30 % d'une part et, d'autre part, de diminuer sensiblement le temps de réponse. Nos expériences ont démontré que l'emploi d'une telle liste permet d'accroître la précision moyenne et que la différence d'efficacité moyenne entre deux listes demeure marginale.

Morphologie flexionnelle

Afin de fournir une bonne qualité de réponse, la machine doit procéder à une analyse morphologique des différentes formes rencontrées. En effet, les formes « fleur » ou « fleurs » possèdent un sens très proche et si l'usager indique l'un de ces deux mots, la machine devrait être capable de dépister des documents possédant l'une ou l'autre de ces formes. Dans ce

[1] Ces listes peuvent être consultées à l'adresse http://www.unine.ch/info/clef/

but, la morphologie d'une langue nous indique comment délimiter les morphèmes d'un mot (par exemple, la forme « hiboux » obtenue par l'adjonction de la flexion « -x » au morphème « hibou »).

La recherche d'information s'intéresse en premier lieu aux morphèmes lexicaux ou lexèmes possédant une identité sémantique. Sur la base de ce lexème ou radical, les variations de formes les plus répandues proviennent de la morphologie flexionnelle servant à indiquer les variations de genre, de nombre, de temps ou de personne. De plus, lors de l'indexation, nous accordons une importance majeure aux noms et adjectifs cernant mieux le sens d'un document que les verbes. Nos approches n'aborderont donc pas la flexion des verbes qui s'avère plus riche et plus difficile à implanter de manière automatique et ceci sans recourir à des dictionnaires électroniques.

Afin de tenir compte des formes fléchies, plusieurs algorithmes ont été proposés pour la langue anglaise [7]. En nous limitant pour l'instant à l'élimination des flexions, Harman [8] a proposé une approche nommée « S-stemmer » visant à supprimer uniquement la marque du pluriel (voir figure 1).

```
Pour les mots de quatre lettres ou plus
 si la séquence finale est «-ies» mais pas «-aies» ou «-eies» alors
   remplacer «-ies» par «-y»; fin              (queries -> query)
 si la séquence finale «-es» mais pas «-aes», «-oes» ou «-ees» alors
   supprimer le «-s» final; fin                (phrases -> phrase, degrees)
 si la séquence finale «-s» mais pas «-us» ou «-ss»   (corpus, stress)
   supprimer le «-s» final                     (kings -> king)
```

FIGURE 1 – *Algorithme « S-stemmer » de suppression de la marque du pluriel pour la langue anglaise*

Pour la langue française, nous pouvons également limiter l'élimination des séquences terminales correspondant, pour l'essentiel, au pluriel comme l'indique l'approche décrite dans la figure 2.

Pour les langues italienne et espagnole possédant des caractéristiques assez similaires à celles du français, l'annexe 1 présente une proposition pour l'italien (figure A.1) ou pour l'espagnol (figure A.2). Dans les grandes lignes, ces propositions se basent sur les analyses suivantes. En italien, la principale règle de flexion du nombre indique que l'on doit remplacer la dernière lettre (par exemple, « -o », « -a » ou « -e ») par une autre (soit « -i » ou « -e »). Comme deuxième règle, la morphologie de l'italien modifie parfois les deux derniers caractères (par exemple, « -io » en « -o », « -co » en « -chi », « -ga » en « -ghe »). Pour l'espagnol,

la principale règle de flexion du nombre exige d'ajouter une ou deux lettres pour indiquer la forme plurielle des noms et adjectifs (par exemple, « -s » ou « -es » comme dans « amigo » et « amigos » (ami) ou « rey » et « reyes » (roi)), voire de modifier le dernier caractère (« -z » en « -ces » dans le mot « voz », « voces » (voix)). Le code écrit en C de ces différentes procédures est disponible[2] sur la Toile.

```
Pour les mots de six lettres ou plus
 si la lettre finale est «-x» alors
   si la séquence finale est «-aux» alors
     remplacer «-aux» par «-al»; fin                (chevaux -> cheval)
     autrement éliminer la lettre finale «-x»; fin  (hiboux -> hibou)
   autrement     (mots ne se terminant pas par «-x»)
     si la lettre finale est «-s» alors éliminer «-s»   (chantés -> chanté)
     si la lettre finale est «-r» alors éliminer «-r»   (chanter- > chante)
     si la lettre finale est «-e» alors éliminer «-e»   (chante -> chant)
     si la lettre finale est «-é» alors éliminer «-é»   (chanté -> chant)
       (règle simple de correction, baronn-> baron)
     si les deux lettres finales sont identiques alors éliminer la lettre finale
 autrement ne pas modifier les mots de quatre lettres ou moins
```

FIGURE 2 – *Algorithme de suppression des marqueurs liés à la morphologie flexionnelle du français*

Pour l'allemand, les variations flexionnelles sont plus complexes et variées car plusieurs morphèmes servent à marquer le pluriel (par exemple, « Sängerin » se transforme en « Sängerin**nen** » (chanteuse), « Boot » en « Boot**e** » (bateau), « Gott » en « G**ö**tt**er** » (dieu) ou « Apfel » en « **Ä**pfel » (pomme)). De plus, cette langue dispose de trois genres (masculin, féminin et neutre) de même que de quatre cas (nominatif, accusatif, génitif et datif) générant des suffixes flexionnels plus diversifiés que les trois autres langues romanes étudiées. Le nombre de flexions possibles s'élève donc à 3 genres x 2 nombres x 4 cas, soit 24 combinaisons. De plus, la flexion liée à un adjectif dépend du fait que cet adjectif est précédé d'un déterminant défini (« der rote Wein »), indéfini (« ein roter Wein ») ou sans déterminant (« roter Wein ») générant ainsi 3 x 24 = 72 possibilités. Mais, les flexions allemandes n'étant pas univoques, nous rencontrons seulement cinq marquages possibles. A titre d'exemple, nous avons repris dans la table 1, les diverses variations de l'adjectif « schön » (beau) et celles liées à son superlatif (« schönst- »). Nous remarquerons

[2] A l'adresse http://www.unine.ch/info/clef/

que tous les marquages flexionnels ne possèdent pas les mêmes fréquences d'occurrence et les suffixes « -en » ou « -e » sont de loin les plus usuels. Notre proposition pour supprimer ces différents marqueurs flexionnels de l'allemand est indiquée dans la figure 3.

schön	1 362		
schöne	1 687	schönste	316
schönem	0	schönstem	0
schönen	1 722	schönsten	478
schöner	501	schönster	0
schönes	463	schönstes	0

TABLE 1 – *Variations possibles autour du radical « schön » (beau) avec leur fréquence d'apparition d'après Ortmann [13]*

```
Pour les mots de cinq lettres ou plus
 éliminer tous les accents
 pour les mots de sept lettres ou plus
   si la séquence finale est «-nen» alors
     supprimer «-nen»; fin                (Sängerinnen-> Sängerin)
 pour les mots de cinq lettres ou plus
   si la séquence finale est «-en» alors
     supprimer «-en»; fin                 (Frauen -> Frau)
   si la séquence finale est «-se» alors
     supprimer «-se»; fin                 (Kenntnisse -> Kenntnis)
   si la séquence finale est «-es» alors
     supprimer «-es»; fin                 (Staates-> Staat)
   si la séquence finale est «-er» alors
     supprimer «-er»; fin                 (Bilder -> Bild)
 pour les mots de cinq lettres ou plus
   si la dernière lettre est «-n», «-s» «-r» ou «-e» alors
     supprimer cette dernière lettre; fin
 autrement ne pas modifier les mots de quatre lettres ou moins
```

FIGURE 3 – *Algorithme de suppression des marqueurs liés à la morphologie flexionnelle de la langue allemande*

Morphologie dérivationnelle

Sur la base d'un lexème, la langue française propose de dériver de nouvelles unités lexicales par l'adjonction d'un suffixe (« blanc »,

« blancheur » ; « volcan », « volcanique ») et / ou d'un préfixe (« visible », « invisible »; « barque », « débarquer »). En recherche d'information, les préfixes sont habituellement ignorés car le sens véhiculé par la forme résultante s'éloigne souvent et sensiblement de la sémantique rattachée au radical. Le caractère assez régulier de ces dérivations en français a permis la mise au point d'un algorithme pour leur élimination [22]. Ce type de construction se rencontre également dans les autres langues européennes comme l'anglais (« white », « whiteness »), l'italien (« ragazza » fille, « ragazzetta » fillette) ou l'espagnol (« árbol » arbre, « arbolito » petit arbre).

```
Pour les mots de dix lettres ou plus
  si la séquence finale est «-emment» alors
    remplacer «-emment» par «-ent»; fin          (prudemment)
  si la séquence finale est «-amment» alors
    remplacer «-amment» par «-ant»; fin          (couramment)
pour les mots de huit lettres ou plus
  si la séquence finale est «-ment» alors
    supprimer «-ment»; fin                       (lentement)
pour les mots de dix lettres ou plus
  si la séquence finale est «-ailler» alors
    supprimer «-ailler»; fin                     (coupailler)
pour les mots de huit lettres ou plus
  si la séquence finale est «-iser» ou «-ier» alors
    supprimer «-iser» ou «-ier»; fin             (cristalliser)
pour les mots de sept lettres ou plus
  si la séquence finale est «-ir» alors
    supprimer «-ir»; fin                         (mentir -> ment)
pour les mots de cinq lettres ou plus
  si la lettre finale est «-s» alors éliminer «-s»     (chantés -> chanté)
  si la lettre finale est «-r» alors éliminer «-r»     (chanter- > chante)
  si la lettre finale est «-e» alors éliminer «-e»     (chante -> chant)
  si la lettre finale est «-é» alors éliminer «-é»     (chanté -> chant)
  (règle simple de correction, baronn-> baron)
  si les deux lettres finales sont identiques alors éliminer la lettre finale
autrement ne pas modifier le mot
```

FIGURE 4 – *Algorithme de suppression des suffixes dérivationnelles pour la langue française*

La langue allemande connaît également ce processus de dérivation mais les règles nous semblent moins strictes que celles du français. Certes, depuis l'adjectif beau « schön », on peut former le nom « die Schönheit » avec l'adjonction du suffixe « -heit » (ou depuis « schnell » (vite), « die Schnelligkeit » avec le infixe « -lig- » suivi du suffixe « -keit »). Mais, en règle générale, la dérivation suffixale allemande n'est pas aussi régulière. Ainsi, depuis l'adjectif « weiß » (blanc), on dispose du nom « das Weiß » sans adjonction ou altération du radical. De manière plus fréquente qu'en français, le radical subit quelques modifications comme, par exemple, avec le verbe jeter « werfen » et le nom relié « der Wurf ».

Afin de tenir compte des formes fléchies et dérivées, plusieurs algorithmes ont été proposés pour la langue anglaise [7] afin de pouvoir retrouver le radical sous-jacent. Ainsi l'approche suggérée par Porter [18] énumère une soixantaine de suffixes flexionnels et dérivationnels que la machine peut supprimer tandis que Lovins [10] en retient environ 260. Pour le français, quelques propositions ont été avancées [22], [24] et nous proposons une version tenant compte uniquement des flexions (voir figure 2) ainsi qu'une approche tenant compte également des dérivations suffixales (voir figure 4).

En langue allemande, en plus du caractère irrégulier de la dérivation suffixale, la morphologie offre la possibilité de former des mots composés par concaténation, processus morphologique que nous rencontrons peu en français (« choux-fleurs », « porte-avion », « basse-cour », « laissez-passer »). Si la langue de Molière laisse clairement apparaître cette concaténation, d'autres langues la cache à l'exemple de l'anglais (« handgun », « worldwide ») ou de l'italien (« capoufficio », chef de bureau) [27]. En langue allemande et contrairement à l'anglais ou l'italien, ce type de construction s'avère très fréquent. Ce processus peut s'opérer sur deux ou plusieurs noms comme, par exemple, pour désigner un employé d'une société d'assurances vie (« Lebensversicherungsgesellschaftsangestellter » formé de « Leben (vie) + S + Versicherung (assurances) + S + Gesellschaft (société) + S + Angestellter (employé) »). De même, le nom « Bankangestelltenlohn » se construit avec les noms « Bank » + « Angestellten » + « Lohn » (salaire). La formation sans adjonction de marqueur morphologique représente la forme la plus courante, soit un peu plus que 60 % des cas [Ortner 91]. Le recours au marqueur « S », parfois précédé d'un « E », constitue environ 15 % des formations de mots tandis que le marqueur « N », parfois précédé d'un « E », représente aussi environ 15 % des cas. Contrairement au français, la langue de Goethe favorise cet amalgame morphologique prioritairement en fusionnant deux noms (représentant environ 77,9 % des cas) tandis que la formation basée

sur un nom et un verbe constitue le deuxième cas le plus fréquent (soit 6,5 %) voire la fusion d'un nom et d'un adjectif (4,6 % des mots composés).

Afin de décomposer automatiquement les mots composés allemands et sans recourir à un dictionnaire, nous avons conçu un algorithme récursif. Ce dernier repose sur des critères quantitatifs et qualitatifs. Dans la description de notre approche, nous noterons par C toute séquence de consonnes et par V une séquence de voyelles (par exemple « llsch »).

Notre approche cherche à décomposer que les mots possédant une longueur initiale supérieure ou égale à 8 caractères. De plus, une coupure ne peut intervenir avant une séquence initiale [O]C, signifiant qu'un mot peut débuter par une série de voyelles qui doit être suivie par des consonnes. Ensuite, l'algorithme cherche l'occurrence d'un des modèles décrits dans la figure 5. Par exemple, le premier modèle « beh be h » signale que lorsque nous rencontrons la chaîne de caractères « beh », l'ordinateur est autorisé à couper le terme en terminant le premier mot par « be » et en débutant le second par « h ». L'ensemble des modèles indiqué dans la figure 5 comprend souvent des séquences de lettres impossibles à trouver dans un mot simple comme « cks », « fff » ou « ksg ». Après avoir détecté un tel modèle, l'ordinateur contrôle que la partie droite comprend au moins quatre caractères débutant potentiellement par une série de voyelles (critère noté [O]) suivie d'une séquence CO. Si une décomposition s'avère possible, l'algorithme recommence son travail sur la partie droite du mot décomposé.

Prenons un exemple. Face au mot composé « Sterbehilfe » (signifiant euthanasie et composé de « sterben » (mourir) et « hilfe » (aide)), nous constatons que ce mot possède une longueur strictement supérieure à sept caractères. Après cette vérification, la machine débute la recherche de modèles de substitution au troisième caractère. Un appariement sera trouvé avec le premier modèle décrit dans la figure 5 et la machine formera les mots « sterben » et « hilfe ». Cette coupure est validée car le second mot possède une longueur supérieure à quatre caractères. De plus, ce terme répond au critère [O]CO. Finalement, comme le terme « hilfe » possède moins que huit lettres, l'ordinateur ne cherchera pas à poursuivre la décomposition sur la base de ce terme.

séquence à trouver,	fin de mot,	début de mot			
beh	be	h	mt	m	t
cks	ck		rsk	rs	k
dh	d	h	rs	r	s
dp	d	p	rtz	rt	z
ds	d		rv	r	v
dsh	d	h	sb	s	b
enh	en	h	sd	s	d
ens	en		sg	s	g
enw	en	w	sh	s	h
erb	er	b	ssb	ss	b
erm	er	m	sss	ss	s
fff	ff	f	td	t	d
ffs	ff		th	t	h
fp	f	p	tions	tion	
fsf	f	f	tk	t	k
gss	g	s	tm	t	m
ischk	isch	k	tsh	t	h
ksg	k	g	tsv	t	v
lbf	lb	f	ttd	tt	d
lk	l	k	tw	t	w
llm	ll	m	undh	und	h
ll	ll		ungs	ung	
lmt	lm	t	yk	y	k
lt	lt		yn	y	n
lw	l	w	zz	z	z

FIGURE 5 – *Modèles pour la décomposition des mots composés en allemand*

Après de ce traitement, la forme composée et les mots simples sont utilisés pour l'indexation après élimination des divers marquages du cas et du genre selon l'algorithme décrit en figure 3. Différents auteurs suggèrent de procéder à une décomposition de ces termes en s'appuyant sur un dictionnaire allemand [1], sur le système Xelda[3] [9] ou en s'appuyant sur les documents eux-mêmes [3].

Finalement, la majorité des langues étudiées disposent de lettres accentuées qui ne sont pas présentes en anglais (avec quelques exceptions

[3] Voir http://www.xrce.xerox.com/ats/xelda/overview.html

comme "à la carte" ou "résumé"). Pour l'italien, l'espagnol et l'allemand, nous avons simplement supprimé les accents sur les lettres concernées.

Toutes les propositions que nous avons formulées reposent sur une connaissance minimale de la morphologie des diverses langues. Afin de proposer une stratégie indépendante de toute langue naturelle, McNamee & Mayfield [11] suggèrent de découper les phrases en séquences de n lettres consécutives, nommée *n-grams*, pour générer les "termes" d'indexation à retenir, les espaces entre les mots étant retenus. Selon cette démarche, l'expression « das Hausdach » sera représentée par les *6-grams* suivants : « das_Ha », « as_Hau », « s_Haus », « _Hausd », « Hausda », « ausdac » et « usdach ». Cette stratégie apporte une performance intéressante pour le chinois [12] car les mots ne sont pas délimités explicitement dans cette langue. Appliquée aux langues européennes, et à l'allemand en particulier, elle peut nous permettre de dégager le radical présent dans diverses formes voire d'isoler les différents noms entrant dans la composition d'un mot allemand.

STRATÉGIES D'INDEXATION ET DE RECHERCHE

Afin de pouvoir dépister des documents en réponse à une requête donnée, nous devons au préalable les indexer c'est-à-dire extraire une liste de mots-clés caractérisant au mieux leur contenu sémantique. Dans ce but, l'ordinateur identifie les mots tout en ignorant les mots-outils, pour ensuite éliminer des marques générées par la morphologie flexionnelle voire dérivationnelle. Enfin, une pondération est calculée pour chacun des termes d'indexation T_j issus du document D_i. Cette pondération devrait tenir compte des facteurs suivants :

- du nombre d'occurrences du terme T_j (mot simple ou composé, syntagme nominal [4]) dans le document D_i, fréquence notée tf_{ij} ;
- de la fréquence documentaire (notée df_j) c'est-à-dire du nombre de documents dans lesquels le terme T_j apparaît, ou plus précisément de l'idf_j, l'inverse de la fréquence documentaire ($idf_j = \ln [n/df_j]$, avec n indiquant le nombre de documents dans la collection) ;
- de la longueur des documents.

Afin de ne pas alourdir le texte, l'annexe 2 présente une liste complète des formules de pondération utilisées dans cet article dont la dénomination retenue est dérivée du système SMART. Ainsi, pour décrire précisément un

modèle de dépistage, un premier triplet de lettres décrit la pondération utilisée lors de l'indexation des documents et, un second triplet, celle appliquée aux requêtes. Par exemple, une stratégie « nnn-nnn » signifie que seul le nombre d'occurrences est retenu pour pondérer les termes des documents et des requêtes tandis qu'une indexation binaire sera notée « bnn-bnn » (terme présent ou non).

Mais cette fréquence d'occurrence peut être modifiée pour tenir compte du fait que l'apparition de la première occurrence devrait posséder un poids important. De plus, nous devrions accorder une importance décroissante au fil des répétitions d'un même terme dans un document. Ainsi, la différence entre une fréquence d'occurrence de 9 ou de 8 n'apporte pas une information très précieuse tandis que la différence entre une fréquence unitaire ou nulle constitue une information très pertinente. Afin de respecter ces deux principes, nous proposons de pondérer un terme selon l'équation $[0{,}5 + 0{,}5 \cdot tf_{ij}]$, de prendre le logarithme de la fréquence d'occurrence ($\ln(tf_{ij})$) ou de recourir au double logarithme ($\ln(\ln(tf_{ij}))$).

Cette fréquence ne constitue que la première composante d'une formule de pondération. La fréquence documentaire (df_j) peut nuancer nos propos. En effet, si 3 000 documents possèdent le même terme, celui-ci ne permet pas de bien distinguer les articles potentiellement pertinents des autres. Afin de résoudre cette difficulté, nous pouvons tenir compte de la fréquence documentaire ou plus précisément de l'idf d'un terme. Cette valeur favorise les termes apparaissant dans peu de documents au détriment des termes très fréquents dans l'ensemble des articles d'un corpus. A la limite, les mots-outils, apparaissant dans presque tous les documents, possèdent une valeur idf proche de 0 et seront donc ignorés lors de l'indexation.

En combinant ces deux premières sources d'information concernant l'importance relative d'un terme, à savoir sa fréquence d'occurrence d'une part, et, d'autre part, sa fréquence documentaire, nous obtenons les approches notées « atn », « ltn » ou « npn ».

La pondération des termes apparaissant dans une requête suit les mêmes principes et cette formule peut diverger de celle appliquée pour les documents. En notant par w_{ij} le poids du terme T_j dans le document D_i et par w_{qj} le poids attribué au même terme dans la requête, la machine évalue le degré de similarité entre le document D_i et la requête Q (ou le score du document D_i) selon le produit interne exprimé comme suit, avec m in-

diquant le nombre de termes d'indexation communs entre le document D_i et la requête Q :

$$SIM(D_i, Q) = \sum_{j=1}^{m} w_{ij} \; w_{qj}$$

Avant de présenter sa réponse à l'usager, l'ordinateur trie par ordre décroissant du degré de similarité les documents dépistés.

Finalement, dans la conception d'une formule de pondération, nous pouvons aussi recourir à une procédure de normalisation garantissant que tous les poids attribués seront compris entre 0 et 1. Par exemple, la pondération classique $tf_{ij} \cdot idf_j$ normalisée par le cosinus [20] sera notée par le triplet « ntc ». Cette forme tient compte à la fois de la fréquence d'occurrence, de l'inverse de la fréquence documentaire et d'une normalisation. L'approche « ntc-ntc » représente l'état de nos connaissances à la fin des années 80 et cette normalisation par le cosinus peut également être reprise par d'autres modèles décrits ci-dessus pour, par exemple, donner naissance aux approches « lnc », « ltc » ou « dtc ».

Durant ces dix dernières années et sous l'impulsion des différentes campagnes d'évaluation de TREC [28], les techniques de recherche d'information ont dû faire face à des volumes nettement plus importants de documents. De plus, ces collections plus récentes renferment des articles avec un nombre plus important de fautes d'orthographe. Dans de telles conditions, les approches « ltn-ntc », « lnc-ltc » ou « ltc-ltc » voire « ltc-ltc » tendent à apporter de meilleures performances que la stratégie « ntc-ntc ». De plus, de nouvelles formules de pondération plus complexes ont été mises au point, en particulier, le modèle probabiliste Okapi [19], le modèle vectoriel « Lnu-ltc » [2] ou la stratégie « dtu-dtc » [26]. Ces dernières possèdent l'avantage de tenir compte de la longueur des documents en cherchant à pénaliser les longs documents abordant généralement plusieurs sujets et qui répondent, en moyenne, moins bien aux attentes de l'usager. D'un point de vue théorique, des arguments peuvent être avancés pour justifier une stratégie au détriment d'une autre mais une évaluation nous permettra de mieux cerner les mérites relatifs des modèles décrits ci-dessus.

EVALUATION

Sur la base des divers modèles de dépistage de l'information décrit dans la section précédente, il s'avère intéressant de savoir si une stratégie s'avère meilleure qu'une autre et ceci indépendamment de la langue utilisée. De plus, comme plusieurs variantes permettant la suppression des flexions et / ou des suffixes dérivationnelles ont été avancées, il serait intéressant de connaître leur efficacité relative. Ces questions seront abordées dans cette section débutant par la présentation des collections créées lors des campagnes d'évaluation CLEF [16], [17]. L'évaluation de dix moteurs de recherche sur des corpus rédigés dans cinq langues formera la deuxième sous-section. Enfin, nous exposerons l'efficacité d'une indexation basée sur des séquences consécutives de cinq lettres (*5-grams*) dans la dernière sous-section.

Les collections de CLEF-2001

Les corpus de documents utilisés dans la campagne d'évaluation CLEF-2001 proviennent de différents journaux tels que le *Los Angeles Times* (Etats-Unis), *Le Monde* (France), *La Stampa* (Italie), *Der Spiegel* et *Frankfurter Rundschau* (Allemagne), d'agences de presse comme *EFE* (Espagne) ou des dépêches de l'agence télégraphique suisse (disponibles en allemand, français et italien). Les documents de ces corpus couvrent approximativement les mêmes thèmes et sont tous extraits de l'année 1994.

Comme la table 2 l'indique, la taille des corpus varie fortement entre les langues avec des volumes plus restreints pour le français et l'italien. Le nombre de termes d'indexation par article reste assez similaire (environ 130) avec une moyenne un peu plus élevée pour la collection anglaise (167,33). Par contre, la variabilité de cette longueur demeure assez forte (écart-type d'environ 120), sauf pour la langue espagnole (écart-type de 60,15) ou pour le corpus italien (écart-type de 97,6).

Le nombre de requêtes disponibles varie très peu entre les collections (environ 48) mais le nombre de documents pertinents par requête se différencie entre les corpus à cause des thèmes abordés par ces dernières. Ainsi, nous rencontrons plus d'articles pertinents par requête pour les langues espagnole (54,97) et allemande (43,47) mais ces corpus sont aussi les plus volumineux. L'écart-type est relativement élevé, indiquant une forte variabilité du nombre d'articles pertinents entre les requêtes. Comme la médiane possède une valeur inférieure à la moyenne, nous pouvons en

déduire que chaque collection dispose d'une majorité de requêtes ayant peu de documents pertinents.

	Anglais	Français	Allemand	Italien	Espagnol
taille	425 MB	243 MB	527 MB	278 MB	509 MB
nb doc.	113 005	87 191	225 371	108 578	215 738
nombre de termes d'indexation différents par document					
moyenne	167,33	140,48	129,26	129,91	120,25
écart-type	126,32	118,61	119,77	97,60	60,15
médiane	138	102	96	92	107
max	1 812	1 723	2 593	1 394	682
min	2	3	1	1	5
nombre de documents pertinents par requête					
total	856	1 212	2 130	1 246	2 694
nb requête	47	49	49	47	49
moyenne	18,21	24,73	43,47	26,51	54,97
écart-type	22,56	24,58	49,20	24,37	63,68
médiane	10	17	27	18	26
max	107	90	212	95	261
min	1	1	1	2	1

TABLE 2 – *Différentes statistiques sur les corpus de CLEF-2001*

En consultant le contenu de ces requêtes, nous constatons que ces dernières ne s'adressent pas à un domaine précis mais couvrent différents besoins d'information comme « Des pesticides dans la nourriture pour bébés », « Embargo sur l'Iraq » ou « Coupe du monde de football ». Suivant le modèle proposé par les campagnes d'évaluation de TREC, chaque requête se subdivise en trois parties logiques comprenant un titre bref, une phrase de description et une partie narrative spécifiant les critères de pertinence (voir table 3). Lors de la campagne CLEF-2001, cinq groupes ont participé à l'élaboration des requêtes et environ dix requêtes proviennent de chaque langue. Ensuite, ces requêtes originales ont été traduites manuellement pour former un ensemble complet dans les cinq langues à savoir le français, l'anglais, l'allemand, l'italien et l'espagnol. Cet ensemble comprend un tiers de requêtes touchant un thème plutôt régional (et ayant le but de dépister des documents pertinents uniquement dans une ou deux collections), un deuxième tiers de requêtes nationales et un dernier tiers de requêtes ayant des articles pertinents dans toutes les cinq collections.

<num> C059 </num>
<title> Les virus d'ordinateur </title>
<desc> Trouver des documents qui parlent des virus d'ordinateur. </desc>
<narr> Sont pertinents les documents énumérant les noms de ces virus ainsi que les dégâts qu'ils peuvent provoquer. </narr>

<title> Computer Viruses </title>
<desc> Find documents about computer viruses. </desc>
<narr> Relevant documents should mention the name of the computer virus, and possibly the damage it does. </narr>

<title> Computerviren </title>
<desc> Suche Dokumente über Computerviren. </desc>
<narr> Relevante Dokumente sollen den Namen des Computervirus nennen und möglicherweise auch die Schäden, die er anrichtet. </narr>

<title> Virus del Computer </title>
<desc> Recupera i documenti che parlano dei virus informatici. </desc>
<narr> I documenti rilevanti devono menzionare il nome del virus informatico, e, possibilmente, devono parlare del danno che provoca. </narr>

<title> Virus informáticos </title>
<desc> Encontrar documentos sobre virus informáticos. </desc>
<narr> Los documentos relevantes deben mencionar el nombre del virus informático, y posiblemente el daño que causa. </narr>

TABLE 3 – *Exemple de requêtes du corpus CLEF-2001 dans les langues française, anglaise, allemande, italienne et espagnole*

Les requêtes utilisées dans nos évaluations ne comprennent que le titre des besoins d'information. Cette représentation correspond à un nombre limité de mots par requête (soit environ 2,6 termes par requête) et reflète mieux la réalité, en particulier celle d'Internet. En tenant compte des parties logiques « description » ou « narrative » (longueur d'environ 16,5 termes par requête), nous pouvons certes augmenter la performance des moteurs de recherche mais nous nous écartons de la réalité de la Toile.

Evaluation de la collection CLEF-2001

Afin de connaître les stratégies d'indexation et de dépistage les plus efficaces, nous avons évalué dix modèles de recherche dont la précision

moyenne, calculée par le logiciel TREC-EVAL, nous servira de mesure de performance. Mais nous nous garderons de tirer des conclusions définitives sur la base d'une différence marginale entre deux évaluations. Afin de savoir si deux modèles présentent une différence de performance pouvant être analysée comme significative, nous avons décidé de recourir à un test statistique basé sur le rééchantillonage aléatoire (*bootstrap*). Cette méthodologie n'impose pas que la distribution sous-jacente des données suive la loi normale [23] et comme souligné par Salton & McGill [21] et démontré dans [23], cette hypothèse n'est pas souvent respectée en recherche d'information, invalidant ainsi différents autres tests statistiques comme ceux basés sur la loi de Student ou de Wilcoxon.

Dans notre processus d'inférence statistique basé sur le rééchantillonage aléatoire, nous posons l'hypothèse Ho que les deux évaluations présentent une performance identique. Une telle hypothèse joue le rôle de l'avocat du diable que nous serons amenés à rejeter (avec un seuil de signification de 5 %) si la précision moyenne, mesurée requête par requête, s'écarte d'une variation considérée comme normale. Cependant, le choix d'accepter Ho ne signifie pas que cette hypothèse est vérifiée mais que, sur la base des résultats obtenus, nous ne disposons pas de suffisamment d'évidence pour contredire cette hypothèse.

Dans la présentation des résultats de nos évaluations, nous cherchons à répondre à deux questions, à savoir quelles sont les stratégies de recherche offrant les meilleures performances d'une part, et, d'autre part, quelles sont les traitements morphologiques les plus efficients. Dans le but de répondre à la première question, nous comparerons les lignes entre elles, tandis que pour répondre à la seconde, nous comparerons les colonnes entre elles. Dans les tables suivantes, nous avons noté en gras la meilleure performance d'une colonne (indiquant ainsi le meilleur moteur pour un traitement morphologique spécifié) et les différences de performance par rapport au modèle probabiliste Okapi seront indiquées par une astérisque (« * »). Pour analyser les différentes approches morphologiques, les pourcentages de différence seront calculés par rapport à la deuxième colonne et les différences statistiquement significatives seront indiquées par un soulignement.

En consultant la table 4a (requêtes courtes, langue anglaise), nous avons évalué trois traitements morphologiques, soit l'absence de toute suppression de suffixes (colonne précédée de l'étiquette « aucun »), le « S-stemmer » et l'algorithme de Lovins [10]. Dans cette table, nous remarquerons que pour six modèles de recherche sur dix, l'approche recourant au « S-stemmer » présente une différence significative par rapport à

une approche ignorant la suppression des suffixes (différence notée par un soulignement). L'approche proposée par Lovins [10] redonne des résultats statistiques similaires. Par contre, aucune différence n'est statistiquement significative lorsque nous comparons le « S-stemmer » avec l'algorithme de Lovins. Dans les deux dernières lignes de la table 4a, nous remarquons que l'approche du « S-stemmer » permet, en moyenne et sur la base de nos dix moteurs, un accroissement de la précision moyenne de 10,77 % tandis que l'algorithme de Lovins apporte un accroissement de 13,4 %. Si nous ignorons les deux stratégies de recherche les moins performantes (soit « bnn-bnn » et « nnn-nnn »), ces deux valeurs varient de manière marginale.

	Précision moyenne (% changement)		
req=titre modèle	aucun 2,7 termes	S-stemmer [8] 2,7 termes	Lovins [10] 2,7 termes
Okapi	**43,09**	**47,55** (+10,4 %)	**48,80** (+13,3 %)
Lnu-ltc	39,25*	44,82 (+14,2 %)	44,83 (+14,2 %)
dtu-dtc	40,53*	44,35* (+9,4 %)	47,60 (+17,4 %)
atn-ntc	39,72*	43,51* (+9,5 %)	45,21 (+13,8 %)
ltn-ntc	34,38*	36,66* (+6,6 %)	37,58* (+9,3 %)
lnc-ltc	23,72*	27,42* (+15,6 %)	28,66* (+20,8 %)
ltc-ltc	25,12*	27,30* (+8,7 %)	28,32* (+12,7 %)
ntc-ntc	22,58*	24,50* (+8,5 %)	26,27* (+16,3 %)
bnn-bnn	21,39*	23,06* (+7,8 %)	22,98* (+7,4 %)
nnn-nnn	12,01*	14,05* (+17,0 %)	13,05* (+8,7 %)
moyenne		+ 10,77 %	+ 13,40 %
moyenne - (bnn, nnn)		+ 10,37 %	+ 14,74 %

TABLE 4 A – *Précision moyenne de différentes stratégies de dépistage (corpus en langue anglaise, 47 requêtes, 856 documents pertinents)*

Sur la base de ces résultats, nous constatons également que le modèle probabiliste Okapi présente la meilleure performance moyenne. En deuxième position, nous trouvons le modèle vectoriel « dtu-dtc » et au troisième rang les modèles « Lnu-ltc » ou « atn-ntc ». Si statistiquement les différences sont significatives en l'absence de tout traitement morphologique (deuxième colonne), en consultant l'approche de Lovins, les différences de performance entre ces quatre meilleurs moteurs ne peuvent être perçues comme statistiquement significatives.

En analysant les résultats de la table 4b (requêtes longues, langue anglaise), nous retrouvons le modèle probabiliste Okapi comme meilleure

stratégie de recherche. Par contre, la différence entre ce modèle et l'approche vectorielle « Lnu-ltc » n'est jamais statistiquement significative. Si l'on analyse les différences entre le modèle Okapi et les huit autres stratégies, la différence de performance s'avère très souvent statistiquement significative.

	Précision moyenne (% changement)		
req=t-d-n modèle	aucun 14,7 termes	S-stemmer [8] 14,3 termes	Lovins [10] 13,7 termes
Okapi	**51,47**	**54,40** (+5,7 %)	**58,00** (+12,7 %)
Lnu-ltc	50,12	52,53 (+4,8 %)	57,30 (+14,3 %)
dtu-dtc	48,70*	49,71* (+2,1 %)	54,22* (+11,3 %)
atn-ntc	49,05*	51,59 (+5,2 %)	55,06* (+12,3 %)
ltn-ntc	42,86*	44,36* (+3,5 %)	45,86* (+7,0 %)
lnc-ltc	40,41*	45,05* (+11,5 %)	46,85* (+15,9 %)
ltc-ltc	41,33*	41,21* (-0,3 %)	43,23* (+4,6 %)
ntc-ntc	34,65*	34,34* (-0,9 %)	37,31* (+7,7 %)
bnn-bnn	25,32*	24,45* (-3,4 %)	20,36* (-19,6 %)
nnn-nnn	12,33*	11,70* (-5,1 %)	11,22* (-9,0 %)
moyenne		+ 2,30 %	+ 5,72 %
moyenne - (bnn, nnn)		+ 3,94 %	+ 10,73 %

TABLE 4B – *Précision moyenne de différentes stratégies de dépistage (corpus en langue anglaise, 47 requêtes, 856 documents pertinents)*

Dans cette table, nous constatons également que l'approche basée sur le « S-stemmer » présente une différence s'avérant statistiquement significative pour cinq stratégies de recherche sur les dix retenues par rapport à une approche ignorant tout traitement morphologique. Si l'on compare cette dernière avec l'algorithme de Lovins, notre évaluation signale une performance statistiquement différente pour sept moteurs. Enfin, seulement pour deux cas (« Lnu-ltc » et « nnn-nnn »), une différence statistique est décelée entre le modèle de Lovins et celui du « S-stemmer » (différence notée par un double soulignement). Par contre, la dernière ligne de la table 4b indiquent que l'accroissement moyen de performance sur la base des huit meilleures stratégies de recherche présente une valeur moindre que celle observée en présence de requêtes courtes (table 4a). De plus, pour les deux moteurs les moins efficaces (« bnn-bnn » et « nnn-nnn »), tout traitement morphologique apporte une dégradation de la qualité de réponse.

Pour la langue française et en recourant aux requêtes courtes (table 5a), le modèle Okapi propose la meilleure stratégie de dépistage en présence de

deux méthodes de traitement morphologique sur trois. Les autres modèles de recherche performants en français sont les approches vectorielles « dtu-dtc », « atn-ntc» et « Lnu-ltc ». Les différences entre le moteur Okapi et ces trois autres ne sont pas statistiquement significatives. En utilisant des requêtes longues (table 5b), des conclusions très similaires peuvent être tirer.

	Précision moyenne (% changement)		
req=titre	aucun	pluriel (avec accents) CLEF01	dérivationnelle sans accents
modèle	2,8 termes	2,8 termes	2,8 termes
Okapi	34,21	**43,96** (+28,5 %)	**45,26** (+32,3 %)
Lnu-ltc	32,84	39,78* (+21,1 %)	41,52 (+26,4 %)
dtu-dtc	**34,27**	42,70 (+24,6 %)	44,39 (+29,5 %)
atn-ntc	31,94	41,38 (+29,6 %)	42,56 (+33,2 %)
ltn-ntc	32,61	37,61* (+15,3 %)	39,04* (+19,7 %)
lnc-ltc	23,45*	30,54* (+30,2 %)	31,12* (+32,7 %)
ltc-ltc	23,41*	29,59* (+26,4 %)	29,71* (+26,9 %)
ntc-ntc	24,00*	28,03* (+16,8 %)	28,95* (+20,6 %)
bnn-bnn	18,10*	21,90* (+21,0 %)	23,26* (+28,5 %)
nnn-nnn	13,19*	13,75* (+4,2 %)	14,02* (+6,3 %)
moyenne		+ 21,78 %	+ 25,62 %
moyenne - (bnn, nnn)		+ 24,07 %	+ 27,68 %

TABLE 5 A – *Précision moyenne de différentes stratégies de dépistage (corpus en langue française, 49 requêtes, 1 212 documents pertinents)*

En analysant l'efficacité des trois traitements morphologiques avec des requêtes courtes (table 5a), nous constatons que pour neuf modèles de recherche sur dix, une différence statistiquement significative apparaît entre une approche ignorant la morphologie et une approche éliminant la marque du pluriel ou une approche tenant compte également de la morphologie dérivationnelle. Des conclusions similaires peuvent être obtenues lorsque nous analysons les requêtes longues (table 5b). Par contre, en comparant notre approche supprimant simplement le pluriel avec celle éliminant aussi quelques marqueurs liés à la morphologie dérivationnelle, une différence statistique ne peut être détectée que pour deux modèles (« dtu-dtc » et « Lnu-ltc ») et seulement lors de l'emploi de requêtes courtes. Ce phénomène a déjà été mis en lumière pour la langue anglaise (tables 4). Même si le français présente une morphologie plus riche, la suppression des suffixes dérivationnels n'apporte pas toujours l'effet escompté. A titre d'exemple, le moteur de recherche www.Yahoo.fr retourne pour la requête

« cheval » des sites parlant de « cheval**erie** » ou de « cheval**et** » confirmant ainsi les limites sous-jacente à l'élimination automatique de tels suffixes.

	Précision moyenne (% changement)		
req=t-d-n	aucun	pluriel (avec accents) CLEF01	dérivationnelle sans accents
modèle	19,9 termes	19,2 termes	19,3 termes
Okapi	43,17	**50,29** (+16,5 %)	50,00 (+15,8 %)
Lnu-ltc	42,77	48,68 (+13,8 %)	49,03 (+14,6 %)
dtu-dtc	**43,39**	50,06 (+15,4 %)	**51,12** (+17,8 %)
atn-ntc	42,23	50,38 (+19,3 %)	50,40 (+19,3 %)
ltn-ntc	40,43*	44,14* (+9,2 %)	44,63* (+10,4 %)
lnc-ltc	35,91*	41,40* (+15,3 %)	41,29* (+15,0 %)
ltc-ltc	34,17*	39,09* (+14,4 %)	38,10* (+11,5 %)
ntc-ntc	31,90*	34,45* (+8,0 %)	34,46* (+8,0 %)
bnn-bnn	13,94*	11,71* (-16,0 %)	11,77* (-15,6 %)
nnn-nnn	12,89*	12,64* (-1,9 %)	12,36* (-4,1 %)
moyenne		+ 9,39 %	+ 9,28 %
moyenne - (bnn, nnn)		+ 13,98 %	+ 14,06 %

TABLE 5B – *Précision moyenne de différentes stratégies de dépistage (corpus en langue française, 49 requêtes, 1 212 documents pertinents)*

Comme pour la langue anglaise, l'augmentation moyenne de la performance sur nos dix modèles de recherche s'avère plus marquée lors de l'usage de requêtes courtes (dernière ligne des tables 4a et 5a) qu'en présence de requêtes longues (dernière ligne des tables 4b et 5b).

Pour la langue allemande (table 6a et 6b), l'élimination des marques du pluriel apporte une différence statistiquement significative pour six modèles de recherche (requêtes courtes) ou pour huit modèles (requêtes longues). Comme pour les langues anglaise et française, tenir compte de la morphologie permet une augmentation moyenne de performance plus sensible en présence de requêtes courtes.

Pour la langue de Goethe, le modèle Okapi se retrouve à nouveau très régulièrement en tête suivi des approches vectorielles « dtu-dtc », « Lnu-ltc » et « atn-ntc». Statistiquement, les différences entre le moteur Okapi et les stratégies « dtu-dtc », « Lnu-ltc » ne peuvent pas être interprétées comme étant significatives.

	Précision moyenne (% changement)		
req=titre modèle	aucun 2,3 termes	pluriel (sans accents) 2,3 termes	décomposition 3,1 termes
Okapi	**28,50**	**35,25** (+23,7 %)	**34,45** (+20,9 %)
Lnu-ltc	26,96	33,37 (+23,8 %)	33,69 (+25,0 %)
dtu-dtc	28,04	33,97 (+21,1 %)	34,16 (+21,8 %)
atn-ntc	26,64*	31,97* (+20,0 %)	32,47 (+21,9 %)
ltn-ntc	27,59	32,26 (+16,9 %)	33,71 (+22,2 %)
lnc-ltc	21,21*	26,46* (+24,8 %)	24,94* (+17,6 %)
ltc-ltc	22,14*	26,51* (+19,7 %)	25,13* (+13,5 %)
ntc-ntc	22,04*	26,23* (+19,0 %)	26,85* (+21,8 %)
bnn-bnn	16,18*	19,91* (+23,1 %)	20,61* (+27,4 %)
nnn-nnn	16,70*	18,15* (+8,7 %)	14,27* (-14,6 %)
moyenne		+ 20,08 %	+ 17,75 %
moyenne - (bnn, nnn)		+ 21,13 %	+ 20,58 %

TABLE 6 A – *Précision moyenne de différentes stratégies de dépistage (corpus en langue allemande, 47 requêtes, 2 109 documents pertinents)*

	Précision moyenne (% changement)		
req=t-d-n modèle	aucun 16,7 termes	pluriel (sans accents) 15,9 termes	décomposition 19,7 termes
Okapi	37,19	**42,18** (+13,4 %)	**41,10** (+10,5 %)
Lnu-ltc	36,61	40,47 (+10,5 %)	39,95 (+9,1 %)
dtu-dtc	**37,22**	41,12 (+10,5 %)	37,96 (+2,0 %)
atn-ntc	36,19	39,80* (+10,0 %)	38,94* (+7,6 %)
ltn-ntc	34,56*	36,04* (+4,3 %)	36,11* (+4,5 %)
lnc-ltc	31,80*	35,70* (+12,3 %)	32,82* (+3,2 %)
ltc-ltc	31,22*	34,08* (+9,2 %)	31,57* (+1,1 %)
ntc-ntc	29,72*	32,42* (+9,1 %)	31,99* (+7,6 %)
bnn-bnn	14,69*	11,77* (-19,9 %)	12,13* (-17,4 %)
nnn-nnn	12,34*	10,39* (-15,8 %)	8,81* (-28,6 %)
moyenne		+ 4,35 %	- 0,04 %
moyenne - (bnn, nnn)		+ 9,90 %	+ 5,71 %

TABLE 6 B – *Précision moyenne de différentes stratégies de dépistage (corpus en langue allemande, 49 requêtes, 2 130 documents pertinents)*

Comme traitement morphologique propre à la langue allemande, nous avons proposé un algorithme permettant la séparation des mots composés. En présence de telle construction linguistique, la machine retient à la fois

la forme composée et les diverses composantes qu'elle aura détectées. En présence de requêtes courtes (table 6a), cette décomposition permet dans six cas sur dix d'améliorer statistiquement la qualité de la réponse obtenue. Cependant, aucune différence significative sépare cette approche de notre algorithme d'élimination des marques du pluriel. En face de requêtes longues, notre proposition de décomposition ne s'écarte pas d'une solution ignorant tout traitement morphologique bien que les performances obtenues présentent un léger accroissement.

Indexation par les *n-grams*

La génération de termes d'indexation par coupure des mots rencontrées après cinq lettres consécutives constitue une alternative aux différentes approches décrites. Cette démarche possède l'avantage de fournir une stratégie réellement indépendante de la langue. Ainsi, si l'ordinateur rencontre l'expression « recherche d'information », il génère les termes suivants : « reche », « echer », « cherc », ... « erche », « infor », ... « mation ». Dans notre approche et contrairement à McNamee & Mayfield [11], la séparation entre les mots est prise en compte.

Dans nos évaluations, nous avons porté notre choix sur la valeur de cinq pour générer nos *n-grams* car elle permet de ne pas couper trop de mots dans les différentes langues étudiées. Ainsi, pour l'anglais, le pourcentage de mots rencontrés possédant une longueur inférieure ou égale à cinq s'élève à 68,14 %, pour le français à 63,6 %, pour l'allemand à 57,08 &, pour la langue italienne à 59,78 % et pour l'espagnol à 61,28 %.

En analysant les résultats de nos évaluations (tables 7, table A.2 et A.4), nous constatons qu'avec ce type d'approche, le modèle Okapi demeure toujours en tête au niveau de la performance moyenne. De plus, les différences de précision moyenne au regard des dix stratégies de recherche et des cinq langues sont très rarement significatives lorsque nous les comparons avec une approche ne procédant à aucune modification des mots rencontrés. La seule exception notable se situe au niveau de l'approche « nnn-nnn » qui offre une bonne performance avec les *5-grams* et avec des requêtes courtes. En présence de requêtes longues, l'image est renversée et l'approche « nnn-nnn » ignorant tout traitement morphologique propose une qualité supérieure par rapport à l'indexation *5-grams*.

En comparant cette stratégie des *5-grams* avec une indexation s'appuyant sur un traitement morphologique, les différences de

performance favorisent nettement cette dernière. Par exemple, pour la langue anglaise (requêtes courtes), le modèle Okapi possède une performance de 40,15 comparée à 48,80 (suppression des suffixes via l'algorithme de Lovins) soit une différence de 17,75 %. Pour le français, l'italien ou l'espagnol, des conclusions similaires peuvent être tirées, que les requêtes soumises soient longues ou courtes. En revanche, pour la langue allemande et en présence de requêtes longues, la performance obtenue par le modèle probabiliste Okapi s'avèrent assez proche entre deux stratégies (42,71 vs. 41,10, soit une différence de 1,26 %).

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,7 termes	5-grams 7,9 termes	req=t-d-n aucun 14,7 terms	5-grams 44,3 termes
Okapi	**43,09**	**40,15** (-6,8 %)	**51,47**	**52,29** (+1,6 %)
Lnu-ltc	39,25*	34,38* (-12,4 %)	50,12	48,59 (-3,1 %)
dtu-dtc	40,53*	36,13* (-10,9 %)	48,70*	47,64* (-2,2 %)
atn-ntc	39,72*	38,79* (-2,3 %)	49,05*	50,65 (+3,3 %)
ltn-ntc	34,38*	31,72* (-7,7 %)	42,86*	40,32* (-5,9 %)
lnc-ltc	23,72*	19,36* (-18,4 %)	40,41*	37,68* (-6,8 %)
ltc-ltc	25,12*	21,12* (-15,9 %)	41,33*	36,49* (-11,7 %)
ntc-ntc	22,58*	21,77* (-3,6 %)	34,65*	33,69* (-2,8 %)
bnn-bnn	21,39*	26,14* (+22,2 %)	25,32*	17,08* (-32,5 %)
nnn-nnn	12,01*	13,41* (+11,7 %)	12,33*	8,11* (-34,2 %)
moyenne		+ 4,42 %		- 9,43 %
moyenne -(bnn, nnn)		+ 9,76 %		- 3,44 %

TABLE 7 A – *Précision moyenne de différentes stratégies de dépistage (corpus en langue anglaise, 47 requêtes, 856 documents pertinents)*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,8 termes	5-grams 7,6 termes	req=t-d-n aucun 19,9 terms	5-grams 70,3 termes
Okapi	34,21	**38,84** (+13,5 %)	43,17	**44,82** (+3,8 %)
Lnu-ltc	32,84	34,09* (+3,8 %)	42,77	42,02 (-1,8 %)
dtu-dtc	**34,27**	33,19* (-3,2 %)	**43,39**	43,36 (-0,1 %)
atn-ntc	31,94	36,23 (+13,4 %)	42,23	42,19 (-0,1 %)
ltn-ntc	32,61	34,53 (+5,9 %)	40,43*	37,77* (-6,6 %)
lnc-ltc	23,45*	26,51* (+13,0 %)	35,91*	34,03* (-5,2 %)
ltc-ltc	23,41*	26,80* (+14,5 %)	34,17*	33,32* (-2,5 %)
ntc-ntc	24,00*	26,46* (+10,3 %)	31,90*	33,24* (+4,2 %)
bnn-bnn	18,10*	22,80* (+26,0 %)	13,94*	7,38* (-47,1 %)
nnn-nnn	13,19*	11,55* (-12,4 %)	34,17*	3,11* (-90,9 %)
moyenne		+ 8,48 %		- 14,62 %
moyenne -(bnn, nnn)		+ 8,91 %		- 1,02 %

TABLE 7 B – *Précision moyenne de différentes stratégies de dépistage (corpus en langue française, 49 requêtes, 1 212 documents pertinents)*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,3 termes	5-grams 11,7 termes	req=t-d-n aucun 16,7 terms	5-grams 75,8 termes
Okapi	**28,50**	**33,56** (+17,8 %)	37,19	**42,71** (+14,8 %)
Lnu-ltc	26,96	30,01* (+11,3 %)	36,61	40,07* (+9,5 %)
dtu-dtc	28,04	32,08 (+14,4 %)	**37,22**	36,79* (-1,2 %)
atn-ntc	26,64*	32,44 (+21,8 %)	36,19	39,26 (+8,5 %)
ltn-ntc	27,59	32,80 (+18,9 %)	34,56*	32,41* (-6,2 %)
lnc-ltc	21,21*	23,60* (+11,3 %)	31,80*	32,53* (+2,3 %)
ltc-ltc	22,14*	24,88* (+12,4 %)	31,22*	30,63* (-1,9 %)
ntc-ntc	22,04*	24,83* (+12,7 %)	29,72*	31,10* (+4,6 %)
bnn-bnn	16,18*	21,34* (+31,9 %)	14,69*	5,75* (-60,9 %)
nnn-nnn	16,70*	12,09* (-27,6 %)	12,34*	5,38* (-56,4 %)
moyenne		+ 12,47 %		- 8,68 %
moyenne -(bnn, nnn)		+ 15,05 %		+ 3,81 %

TABLE 7 C – *Précision moyenne de différentes stratégies de dépistage (corpus en langue allemande, 47 requêtes, 2 109 documents pertinents)*

CONCLUSION

En nous basant sur l'évaluation de dix stratégies d'indexation et de recherche œuvrant sur des collections rédigées dans cinq langues différentes, nous pouvons tirer les conclusions suivantes :

- le modèle probabiliste Okapi présente souvent la meilleure performance que ce soit avec des collections de documents rédigés en anglais (tables 4), en français (tables 5), en langue allemande (tables 6), italienne (table A.1) ou espagnole (table A.3) ;
- comme deuxième meilleure stratégie, on rencontre les modèles vectoriels « Lnu-ltc », « dtu-dtc» ou « atn-ntc ». La désignation précise de la deuxième meilleure approche dépend de la langue, de la longueur de la requête et du traitement morphologique utilisé. De plus, la différence entre les deux meilleures solutions n'est pas souvent statistiquement significative ;
- l'élimination des marques liées à la morphologie permet généralement d'augmenter la précision moyenne. Cet accroissement s'avère très souvent statistiquement significatif par rapport à une approche renonçant à toute élimination de suffixes et cela en considérant les huit meilleures stratégies de recherche et pour les cinq langues étudiées ;
- pour les cinq langues étudiées, l'élimination des flexions s'avère, en moyenne, plus bénéfique en présence de requêtes courtes que pour les requêtes longues ;
- en comparant l'élimination uniquement des marqueurs flexionnels avec une approche supprimant les morphèmes liés à la flexion ou à la dérivation, on constate généralement que cette seconde méthode présente une performance moyenne supérieure. Par contre la différence entre les deux stratégies s'avère rarement statistiquement significative et ceci que la langue sous-jacente soit l'anglais ou le français ;
- l'élimination des accents semblent apporter une très légère augmentation. Ces variations étant marginales, aucune différence n'est statistiquement significative ;
- la décomposition des mots composés en allemand ne semble pas être la meilleure approche pour cette langue (tables 6) ;
- en recourant à l'indexation par *5-grams*, nos évaluations n'indiquent presque jamais une différence statistiquement significative par rapport à une approche renonçant à toute modification. De plus, le recours au *5-grams* proposent une qualité de réponse inférieure à celle obtenue par des modèles supprimant les flexions.

Dans l'interrogation en-ligne de collections de documents rédigés dans d'autres langues que l'anglais, différentes questions doivent encore obtenir une réponse plus claire et définitive. Par exemple, nos propositions pour supprimer à la fois les suffixes flexionnels et dérivationnels n'apportent pas une amélioration jugée statistiquement significative par rapport à la suppression uniquement des flexions pour la langue française. Certes, d'autres approches pour la suppression des suffixes ont été proposées (en utilisant, par exemple, le système Xelda [9]) mais la question de leur efficacité comparée à nos approches reste ouverte. Nous pouvons également nous interroger sur l'opportunité de conserver la distinction entre majuscule et minuscule, en particulier pour les noms propres. Cette technique semble améliorer la précision moyenne [9] mais la reconnaissance des noms propres n'est pas parfaite ni aisée comme le montre le mot « Pierre » dans la phrase « Pierre qui roule … ». Finalement la méthode utilisée pour décomposer de façon efficace les mots allemands afin d'accroître la performance d'un système de recherche reste ouverte. La question de savoir si cette décomposition peut s'appliquer également de manière bénéfique à d'autres langues germaniques comme le hollandais ou le suédois voire à d'autre langues comme le finnois ou le turc demeure sans réponse claire.

Remerciements

L'auteur remercie C. Buckley de SabIR pour avoir mis à notre disposition le système SMART sans lequel cette étude n'aurait pas pu être envisagée. Cette recherche a été subventionnée, en partie, par le Fonds national suisse (subside 21-58 813.99).

RÉFÉRENCES

[1] M. Braschler, B. Ripplinger et P. Schäuble. Experiments with the Eurospider Retrieval System for CLEF-2001. *Proceedings of CLEF-2001,* pages 45-50, Sophia-Antiplolis : ERCIM EEIG, 2001.

[2] C. Buckley, A. Singhal, M. Mitra et G. Salton. New Retrieval Approaches using SMART. *Proceedings of TREC'4,* pages 25-48, Gaithersburg : NIST Publication #500-236, 1996.

[3] A. Chen. Multilingual Information Retrieval Using English and Chinese Queries. *Proceedings of CLEF-2001,* pages 21-27, Sophia-Antiplolis : ERCIM EEIG, 2001.

[4] J.-P. Chevallet et H. Haddad. Proposition d'un modèle relationnel d'indexation syntagmatique : mise en œuvre dans le système IOTA. *Actes du XIXe congrès INFORSID*, pages 465-483, 2001.

[5] Duden. *Band 4. Grammatik.* Mannheim : Dudenverlag, 1991.

[6] C. Fox. A Stop List for General Text. *ACM-SIGIR Forum*, vol 24, pages 19-35, 1990.

[7] W.B. Frakes. Stemming Algorithms. In W.B. Frakes & R. Baeza-Yates (Eds), *Information Retrieval, Data Structures & Algorithms*, pages 131-160, Englewood Cliffs: Prentice-Hall, 1992.

[8] D. Harman. How Effective is Suffixing?. *Journal of the American Society for Information Science*, vol 42, pages 7-15, 1991.

[9] W. Kraaij. TNO at CLEF-2001: Comparing Translation Resources. *Proceedings of CLEF-2001*, pages 29-40, Sophia-Antiplolis : ERCIM EEIG, 2001.

[10] J.B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, vol 11, pages 22-31, 1968.

[11] P. McNamee et J. Mayfield. JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. *Proceedings of CLEF-2001*, pages 121-131, Sophia-Antiplolis : ERCIM EEIG, 2001.

[12] J.Y. Nie et F. Ren. Chinese Information Retrieval: Using Characters or Words? *Information Processing & Management*, vol 35, pages 443-462, 1999.

[13] W.D. Ortmann. *Hochfrequente deutsche Wortformen I.* München : Goethe-Institut, 1975.

[14] W.D. Ortmann. *Hochfrequente deutsche Wortformen II.* München : Goethe-Institut, 1976.

[15] L. Ortner, E. Müller-Bollhagen, H. Ortner, H. Wellmann, M. Pümpel-Mader, H. Gärtner. *Deutsche Wortbildung, Typen und Tendenzen in der Gegenwartssprache. Vierter Hauptteil.* Berlin : Alter de Gruyter, 1991.

[16] C. Peters, M. Braschler, J. Gonzalo et M. Kluck (Eds). Results of the CLEF 2001 Cross-Language System Evaluation Campaign, Proceedings of CLEF-2001, Berlin : Springer-Verlag, Lecture Notes in Computer Science, 2002.

[17] C. Peters (Ed.). Cross-Language Information Retrieval and Evaluation, Proceedings of CLEF-2000, Berlin : Springer-Verlag, Lecture Notes in Computer Science, 2069, 2001.

[18] M.F. Porter. An Algorithm for Suffix Stripping. *Program*, vol 14, pages 130-137, 1980.

[19] S.E. Robertson, S. Walker et M. Beaulieu. Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management*, vol 36, pages 95-108, 2000.

[20] G. Salton et C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, vol 24, pages 513-523, 1988.

[21] G. Salton et M.J. McGill. *Introduction to Modern Information Retrieval*, New-York : McGraw-Hill, 1983.

[22] J. Savoy. Stemming of French Words Based on Grammatical Category. *Journal of the American Society for Information Science*, vol 44, pages 1-9, 1993.

[23] J. Savoy. Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, vol 33, pages 495-512, 1997.

[24] J. Savoy. A Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science*, vol 50, pages 944-952, 1999.

[25] J. Savoy. Report on CLEF-2001 Experiments. *Proceedings of CLEF-2001*, pages 11-19, Sophia-Antiplolis : ERCIM EEIG, 2001.

[26] A. Singhal, J. Choi, D. Hindle, D.D. Lewis et F. Pereira. AT&T at TREC-7. *Proceedings TREC-7*, pages 239-251, Gaithersburg : NIST Publication #500-242, 1999.

[27] R. Sproat. *Morphology and Computation*. Cambridge: The MIT Press, 1992.

[28] E.M. Voorhees et D. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). *Information Processing & Management*, vol 36, pages 3-35, 2000.

[29] I.H. Witten, A. Moffat et T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. San Francisco : Morgan Kaufmann, 1999.

ANNEXE 1. DÉCOMPOSITION EN ITALIEN ET ESPAGNOL

```
Pour les mots de six lettres ou plus
 éliminer tous les accents
 si la séquence finale est «-ie» ou «-he» alors
   supprimer «-ie» ou «-he»; fin            (amiche -> amic)
 si la séquence finale est «-hi» ou «-ii» alors
   supprimer «-hi» ou «-ii»; fin            (balocchi -> balocc)
 si la séquence finale est «-ia» ou «-io» alors
   supprimer «-ia» ou «-io»; fin            (ufficio-> uffic)
 si la dernière lettre est «-e», «-i», «-a» ou «-o» alors
   supprimer cette dernière lettre; fin     (uffici -> uffic)
 autrement ne pas modifier le mot
```

FIGURE A.1 – *Algorithme de suppression des marqueurs liés à la morphologie flexionnelle de la langue italienne*

```
Pour les mots de cinq lettres ou plus
 éliminer tous les accents
 si la séquence finale est «-eses» alors
   remplacer «-eses» par «-es»; fin         (corteses -> cortes)
 si la séquence finale est «-ces» alors
   remplacer «-ces» par «-z»; fin           (veces -> vez)
 si la séquence finale est «-os», «-as» ou «-es» alors
   supprimer «-os», «-as» ou «-es» ; fin    (hermanos -> herman)
 si la dernière lettre est «-o», «-a» ou «-e» alors
   supprimer cette dernière lettre; fin     (hermano -> herman)
 autrement ne pas modifier le mot
```

FIGURE A.2 – *Algorithme de suppression des marqueurs liés à la morphologie flexionnelle de la langue espagnole*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,8 termes	pluriel 2,8 termes	req=t-d-n aucun 18,1 terms	pluriel 17,5 termes
Okapi	**33,80**	**39,76** (+17,6 %)	**41,66**	**48,18** (+15,7 %)
Lnu-ltc	32,46	38,22 (+17,7 %)	40,16	46,50 (+15,8 %)
dtu-dtc	33,55	39,61 (+18,1 %)	39,38*	46,13 (+17,1 %)
atn-ntc	31,55	36,20* (+14,7 %)	39,28*	45,07* (+14,7 %)
ltn-ntc	29,93*	34,81* (+16,3 %)	36,86*	41,98* (+13,9 %)
lnc-ltc	24,97*	29,10* (+16,5 %)	33,81*	38,91* (+15,1 %)
ltc-ltc	24,73*	28,12* (+13,7 %)	31,71*	35,00* (+10,4 %)
ntc-ntc	25,32*	27,22* (+7,5 %)	31,35*	34,29* (+9,4 %)
bnn-bnn	21,22*	23,31* (+9,8 %)	21,79*	19,90* (-8,7 %)
nnn-nnn	16,75*	16,43* (-1,9 %)	20,54*	18,50* (-9,9 %)
moyenne		+ 13,02 %		+ 9,34 %
moyenne -(bnn, nnn)		+ 15,28 %		+ 14,01 %

TABLE A.1 – *Précision moyenne de différentes stratégies de dépistage (corpus en langue italienne, 47 requêtes, 1 246 documents pertinents)*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,8 termes	5-grams 9 termes	req=t-d-n aucun 18,1 terms	5-grams 61,6 termes
Okapi	**33,80**	**34,70** (+2,7 %)	**41,66**	**42,33** (+1,6 %)
Lnu-ltc	32,46	31,22* (-3,8 %)	40,16	40,97 (+2,0 %)
dtu-dtc	33,55	31,14* (-7,2 %)	39,38*	39,56* (+0,5 %)
atn-ntc	31,55	33,25 (+5,4 %)	39,28*	40,23 (+2,4 %)
ltn-ntc	29,93*	30,12* (+0,6 %)	36,86*	35,07* (-4,9 %)
lnc-ltc	24,97*	22,20* (-11,1 %)	33,81*	32,71* (-3,3 %)
ltc-ltc	24,73*	22,19* (-10,3 %)	31,71*	30,81* (-2,8 %)
ntc-ntc	25,32*	23,03* (-9,0 %)	31,35*	30,17* (-3,8 %)
bnn-bnn	21,22*	22,46* (+5,8 %)	21,79*	11,11* (-49,0 %)
nnn-nnn	16,75*	13,06* (-22,0 %)	20,54*	11,07* (-46,1 %)
moyenne		- 4,89 %		- 20,49 %
moyenne -(bnn, nnn)		- 6,08 %		- 17,32 %

TABLE A.2 – *Précision moyenne de différentes stratégies de dépistage (corpus en langue italienne, 47 requêtes, 1 246 documents pertinents)*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,7 termes	pluriel 2,7 termes	req=t-d-n aucun 16,7 terms	pluriel 16 termes
Okapi	**40,54**	**50,55** (+24,7 %)	**50,37**	**56,78** (+12,7 %)
Lnu-ltc	39,41	46,81* (+18,8 %)	49,17	54,31 (+10,5 %)
dtu-dtc	40,39	48,12* (+19,1 %)	46,85*	52,64* (+12,4 %)
atn-ntc	39,24	47,52* (+21,1 %)	49,25	55,35 (+12,4 %)
ltn-ntc	38,34	45,28* (+18,1 %)	46,93	51,16* (+9,0 %)
lnc-ltc	31,69*	38,89* (+22,7 %)	44,64*	48,84* (+9,4 %)
ltc-ltc	31,77*	36,03* (+13,4 %)	41,75*	44,98* (+7,7 %)
ntc-ntc	31,34*	35,13* (+12,1 %)	39,78*	42,35* (+6,5 %)
bnn-bnn	23,65*	26,59* (+12,4 %)	23,06*	20,78* (-9,9 %)
nnn-nnn	18,79*	20,01* (+6,5 %)	23,00*	23,20* (+0,9 %)
moyenne		+ 16,90 %		+ 7,15 %
moyenne -(bnn, nnn)		+ 18,75 %		+ 10,07 %

TABLE A.3 – *Précision moyenne de différentes stratégies de dépistage (corpus en langue espagnole, 49 requêtes, 2 694 documents pertinents)*

	Précision moyenne (% changement)			
modèle	req=titre aucun 2,7 termes	5-grams 8,7 termes	req=t-d-n aucun 16,7 terms	5-grams 59,8 termes
Okapi	**40,54**	**43,37** (+7,0 %)	**50,37**	**52,01** (+3,3 %)
Lnu-ltc	39,41	38,62* (-2,0 %)	49,17	48,12* (-2,1 %)
dtu-dtc	40,39	39,01* (-3,4 %)	46,85*	46,86* (+0,0 %)
atn-ntc	39,24	41,06 (+4,6 %)	49,25	49,33* (+0,2 %)
ltn-ntc	38,34	38,31* (-0,1 %)	46,93	45,43* (-3,2 %)
lnc-ltc	31,69*	30,23* (-4,6 %)	44,64*	42,31* (-5,2 %)
ltc-ltc	31,77*	30,53* (-3,9 %)	41,75*	40,38* (-3,3 %)
ntc-ntc	31,34*	30,50* (-2,7 %)	39,78*	40,08* (+0,8 %)
bnn-bnn	23,65*	27,56* (+16,5 %)	23,06*	13,84* (-40,0 %)
nnn-nnn	18,79*	16,42* (-12,6 %)	23,00*	15,36* (-33,2 %)
moyenne		- 0,12 %		- 8,28 %
moyenne -(bnn, nnn)		- 0,63 %		- 1,20 %

TABLE A.4 – *Précision moyenne de différentes stratégies de dépistage (corpus en langue espagnole, 49 requêtes, 2 694 documents pertinents)*

ANNEXE 2. FORMULES DE PONDÉRATION

Afin d'attribuer un poids w_{ij} reflétant l'importance de chaque terme d'indexation T_j, j = 1, 2, ..., t dans un document D_i, nous pouvons recourir à l'une des formules décrite dans le tableau ci-dessous. Dans ce dernier, tf_{ij} indique la fréquence (nombre d'occurrences) du terme T_j dans un document (ou dans une requête), n représente le nombre de documents D_i dans la collection, df_j le nombre de documents dans lesquels le terme T_j apparaît (fréquence documentaire), et idf_j l'inverse de la fréquence documentaire ($idf_j = \ln [n/df_j]$).

ntc	$w_{ij} = \dfrac{tf_{ij}\ idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik}\ idf_k)^2}}$	bnn	$w_{ij} = 1$
atn	$w_{ij} = \left[0{,}5 + 0{,}5\ \dfrac{tf_{ij}}{maxtf_{i.}}\right] idf_j$	nnn	$w_{ij} = tf_{ij}$
npn	$w_{ij} = tf_{ij}\ \ln \left[(n - df_j) / df_j\right]$	ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$
lnc	$w_{ij} = \dfrac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} (\ln(tf_{ik}) + 1)^2}}$	Okapi	$w_{ij} = \left((k_1 + 1)\ tf_{ij}\right) / \left(K + tf_{ij}\right)$
dtc	$w_{ij} = \left(\ln\left(\ln(tf_{ij}) + 1\right) + 1\right)\ idf_j \Big/ \sqrt{\sum_{k=1}^{t} \left(\left(\ln\left(\ln(tf_{ik}) + 1\right) + 1\right)\ idf_k\right)^2}$		
ltc	$w_{ij} = \dfrac{\left(\ln(tf_{ij}) + 1\right)\ idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right)\ idf_k\right)^2}}$		
dtu	$w_{ij} = \dfrac{\left(1 + \ln\left(1 + \ln(tf_{ij})\right)\right)\ idf_j}{(1 - slope)\ pivot + slope\ nt_i} \Big/ (1 + pivot)$		

Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij})\right) / \left(1 + pivot\right)}{(1 - slope)\ pivot + slope\ nt_i}$

TABLE A.5 – *Différentes formules de pondération*

De plus, la longueur du document D_i (ou le nombre de termes d'indexation associé à ce document) est notée par nt_i. La constante slope a été fixée arbitrairement à 0,1 et la valeur pivot à 125. Pour le modèle Okapi, K se calcule comme suit :

$$K = k_1 \left[(1 - b) + \frac{b\ nt_i}{advl}\right]$$

et nous avons retenu la valeur $k_1 = 1{,}2$, b = 0,75 et advl = 900. Ces valeurs ont été choisies selon les recommandations de Robertson *et al.* [19].