Dissymétrie entre l'indexation des documents et le traitement des requêtes pour la recherche d'information en langue arabe

Ramzi ABBÈS(1), Malek BOUALEM (2)

(1) ICAR-CNRS/ Lyon 2,
ramzi.abbes@univ-lyon2.fr
(2) Orange Labs – France Telecom R&D
malek.boualem@orange-ftgroup.com

Résumé
Les moteurs de recherches sur le web produisent des résultats comparables et assez satisfaisants pour la recherche de documents écrits en caractères latins. Cependant, ils présentent de sérieuses lacunes dès que l'ont s'intéresse à des langues peu dotées ou des langues sémitiques comme l'arabe. Dans cet article nous présentons une étude analytique et qualitative de la recherche d'information en langue arabe en mettant l'accent sur l'insuffisance des outils de recherche actuels, souvent mal adaptés aux spécificités de la langue arabe. Pour argumenter notre analyse, nous présentons des résultats issus d'observations et de tests autour de certains phénomènes linguistiques de l'arabe écrit. Pour la validation des ces observations, nous avons testé essentiellement le moteur de recherche Google.

Abstract
Web search engines provide quite good results for Latin characters-based languages. However, they still show many weaknesses when searching in other languages such as Arabic. This paper discusses a qualitative analysis of information retrieval in Arabic language, highlighting some of the numerous limitations of available search engines, mainly when they are not properly adapted to the Arabic language specificities. To argue our analysis, we present some results based on quite sufficient observations and tests on various Arabic linguistic phenomena. To validate these observations, we essentially have tested the Google search engine.

Mots clés : recherche d'information, langue arabe, indexation, lemmatisation, Google

Keyword : information retrieval, Arabic, indexation, lemmatization, Google

1 Introduction

Avec 90,83% de part de marché en décembre 2007[1], Google est probablement le moteur de recherche le plus puissant sur le marché, ou du moins, le plus utilisé, car il existe bien une corrélation entre les deux constats.

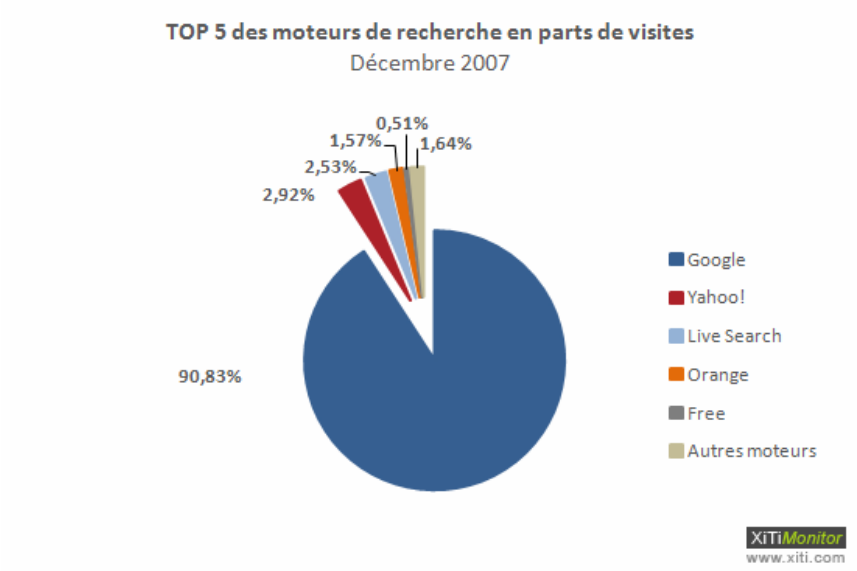

Figure 1. Top 5 des moteurs de recherche en parts de visites

Le système de recherche de Google ainsi que le PageRank (S. Brin and L. Page, 1988) restent très sensibles au comportement des internautes. En effet, les algorithmes de Google pondèrent positivement ou négativement les pages en fonction du nombre de clics sur les liens retournés. Le PageRank de Google note les pages en fonction des liens hypertextes qu'elles contiennent (P. Chen et al ; 2007). Selon que les pages aient été visitées ou non après une recherche, elles gagnent ou perdent des points dans le classement par rapport aux mots clés recherchés. Dans d'autres cas, selon que la page ait servi de routeur aux internautes ou pas (c'est-à-dire que la page ait été pertinente pour la recherche et aurait contenu des liens qui auraient été suivis par l'internaute), elle gagne des points dans les priorités d'indexation.

Si l'utilisateur retrouve quelques réponses satisfaisantes dans la première page des résultats de Google, il est souvent loin de trouver la meilleure réponse. Il s'agit surtout des réponses les plus populaires. En effet, il est clair que, parmi les milliards de pages sur internet, les informations recherchées ne peuvent pas se retrouver uniquement sur une dizaine (ou quelques dizaines) de documents proposés.

Ce constat est encore plus réel lorsque Google est utilisé en langue arabe. En effet, nous avons remarqué que la plupart des réponses figurant en tête de liste, pour une requête en arabe, proviennent des sites de discussions منتديات *mntdyât* ou d'autres sources spécifiques. Le PageRank de Google surclasse ces sources d'informations riches en renvois hypertextes et beaucoup plus utilisées dans le monde arabe au détriment d'autres sources d'informations en provenance de travaux universitaires, journalistiques, etc.

[1] http://barometre.secrets2moteurs.com/index.php/Barometre-1ere-position-xiti

L'essentiel de la population des internautes arabophones maîtrise une seconde langue, le français ou l'anglais. Le web étant très bien développé dans l'une et l'autre de ces deux langues, l'internaute arabe s'oriente vers le multilinguisme pour une grande partie de ses recherches. Cet état de fait n'aide forcément pas au développement des ressources d'information en langue arabe. Par certains aspects de spécialisation de la recherche, comme Google Scholar (C. Neuhaus, E. Neuhaus and A. Asher, 2008), le moteur de recherche donne souvent des réponses satisfaisantes. La mesure du degré de pertinence suppose une connaissance de l'ensemble des documents existants sur le web.

Dans cet article, nous nous intéressons à un certain nombre de phénomènes linguistiques afin de mettre en évidence la nécessité de traitements plus approfondis lors de l'indexation des documents écrits en langue arabe. Dans un premier temps, nos parlerons de la couverture des moteurs de recherche pour l'arabe et de la dissymétrie entre l'indexation et la recherche d'information, particulièrement palpable pour l'arabe non vocalisé. Par la suite, nous insisterons sur l'apport des traitements linguistiques pour l'amélioration de la recherche en arabe. Nous conclurons l'article par une étude statistique sur quelques exceptions orthographiques très répondues auprès des auteurs arabes.

2 La langue arabe sur le web

2.1 Couverture des moteurs de recherche

Les robots des moteurs de recherche (*spiders ou crawlers*) parcourent les sites de la toile, à intervalles réguliers. **L'exploration** est indépendante de l'alphabet, elle dépend surtout des performances en terme de couverture de chacun des moteurs. La problématique de recherche d'information dépend de deux facteurs, le premier concerne l'indexation des pages et le second est lié à la recherche dans les index ou dans les pages elles mêmes.

L'indexation des pages web se fait, pour l'essentiel des moteurs de recherche, par l'une ou la combinaison des méthodes suivantes :

- La récupération des balises « méta » contenant les mots clés décrivant le contenu des pages et proposés par le créateur du site.

- La récupération du contenu de la balise « titre », il est d'ailleurs recommandé de donner des titres différents à chacune des pages web du site pour avoir un maximum de chance de ressortir ces pages parmi les résultats du moteur.

Pour les ressources jugées importantes, les robots peuvent indexer tout le contenu de la page.

La **recherche** est la partie *secrète* des robots, les algorithmes pondèrent les pages en fonction d'un ensemble de critères comme la position du mot dans la page (titre, paragraphe, lien hypertexte) en fonction de l'historique ou encore de la nature de la ressource.

Les webmestres de leurs côtés positionnent les mots pertinents pour leurs sites dans les endroits stratégiques pour le robot de recherche. Cet aspect n'a pas de pertinence par rapport aux langues utilisées, mais il est pertinent selon le traitement *linguistique* possible dans chacun des moteurs, en fonction de leur maîtrise de la lemmatisation, de la dérivation ou de leurs ontologies pour relier des mots de la recherche aux mots clés proches sémantiquement ou faisant partie de la même famille morphologique.

2.2 Dissymétrie de l'indexation et la recherche en langue arabe

La recherche d'information en langue arabe montre souvent une certaine dissymétrie entre l'indexation et le traitement des requêtes provoquée particulièrement par l'absence des voyelles dans les textes arabes écrits et aussi par la nature agglutinante de l'écriture arabe.

Par exemple[2], lors de l'indexation d'un document, le verbe "écrire" (كَتَب *katab*), le nom "livres" (كُتُب *kutub*) et le nom "écrit" (كَتْب *katb*) sont tous indexés sous une seule et même entrée كتب *ktb* , car ils ne sont généralement pas vocalisés dans le texte. Il en est de même pour le mot "شعر ŠÝr " dont les différentes vocalisations peuvent avoir des significations différentes (sentir, poème, cheveux, etc.) ou encore le mot "علم" *Ýlm* qui peut désigner plusieurs sens (drapeau, science, connaître, etc.).

Quelles que soient les précisions apportées à la recherche (même si on note le mot entièrement vocalisé), le moteur ne pourra pas séparer ces formes étant donné qu'elles ne sont pas vocalisées à la base. Par conséquent, les mots de l'index ne sont pas vocalisés non plus.

L'agglutination consiste, en figure simplifiée, en l'augmentation de la forme minimale du mot par des proclitiques (pour signifier l'interrogation, la ressemblance, la liaison, etc.) ou des enclitiques (pour rajouter notamment des pronoms). Ces trois exemples illustrent divers cas d'agglutination : (1) Kateb Yassine كاتب يسين *kâtb ysyn* , (2) est-ce que Kateb Yassine ? أكاتب يسين *Þkâtb ysyn* , (3) j'écris à Yassine أكاتب يسين *Þkâtb ysyn* .

2.3 Recherche d'information en langue arabe

Une grande partie des requêtes sur le web, indépendamment des langues, concernent des entités nommées tels que des noms propres. Nos tests sur un échantillon de 2850 requêtes arabes sur un annuaire (M. Boualem et al., 2001) nous ont permis de constater que 94,2% des requêtes concernent les formes nominales, 3,30% concernent les formes verbales et 2,5% concernent les mots grammaticaux. Bien entendu, ces valeurs peuvent être ajustées si nous prenons en compte le contexte non vocalisé des requêtes. En effet, en dehors de quelques formes verbales et de mots grammaticaux non ambigus comme من *mn* , متى *mt* , اخترع *âÌtrÝ (inventé)* , on retrouve beaucoup de formes ambiguës comme نزل *nzl* (descendre ou hôtel) , رقص *rqÒ* , طلب *Ôlb* (demander, une demande) . Notons aussi que les formes verbales rencontrées ne sont pas fléchies.

Mots graphiques		
Formes Verbales	Mots outils	Formes nominales
3,30%	3%	94,20%

La particularité des noms propres arabes est qu'ils sont souvent des dérivées de formes verbales (participe actif, participe passif, etc.). كاتب *kâtb* désigne à la fois l'écrivain et aussi un

[2] Pour la translittération des caractères arabes en caractères latins, nous utilisions la norme ISO 233-2 (Z 46-002, décembre 1993) : [أ, Þ] [ب, b] [ت, t] [ث, ×][ج, Ê] [ح, Î] [خ, Ì] [د, d] [ذ, Æ] [ر, r] [ز, z] [س, s] [ش, Š] [ص, Ò] [ض, Ã] [ط, Ô] [ظ, Û] [ع, Ý] [غ, È] [ف, f] [ق, q] [ك, k] [ل, l] [م, m] [ن, n] [ه, h] [و, w] [ي, y] [, a] [, ∅] [, u] [,] [, ∅] [, i] [, ∅] [ا, ∅]

nom propre comme pour Kateb Yassine. Toutefois, la recherche par كاتب *kâtb* renvoie essentiellement écrivain. Voici quelques exemples de recherche avec des noms :

La recherche du mot كَتَب *katab* sur Google donne, sur les premiers résultats, la notion de «livres». S'agit-il d'une question de «ranking» ou de priorité donné aux noms ? Cependant, nous constatons que **la voyellation du mot clé n'a aucune influence sur la recherche**.

Soit une recherche autour du prénom *Jamel Eddine Elafghani* **جمال الدين الأفغاني** *Êmâl âldyn âlÞfÈâny* et *Jamel Abdennasser le leader* **جمال عبد الناصر الزعيم** *Êmâl Ýbd âlnâÒr âlzÝym* . La recherche par جمال Êmâl renvoie en premier les résultats concernant l'adjectif *beauté (le prénom et l'adjectif ont la même forme graphique non vocalisé)* et pas le prénom. Nous trouvons 5 340 000 pour **جمال** *Êmâl* , 737 000 pour **جمال الدين** *Êmâl âldyn* , 70 700 pour **جمال الدين الأفغاني** *Êmâl âldyn âlÞfÈâny* . Pour la recherche de Nasser, nous obtenons 805 000 pour **جمال عبد**, 293 000 pour **جمال عبد الناصر** *Êmâl Ýbd âlnâÒr* , 253 000 pour **جمال عبد الناصر الزعيم** *Êmâl Ýbd âlnâÒr âlzÝym* . Parallèlement, une recherche avec la mot الزعيم *âlzÝym* , donne 2 100 000 pour **الزعيم** *âlzÝym* . Nous trouvons parmi les premiers résultats des blogs d'amateurs de foot, des informations sur la pièce de théâtre de Adel Imam (742 000 pour **الزعيم عادل** *âlzÝym Ýâdl*). Le premier résultat concernant **جمال عبد الناصر** *Êmâl Ýbd âlnâÒr* arrive en trentième position. **Nous constatons donc une grande faiblesse dans le traitement réservé aux entités nommées.**

Au niveau de notre corpus de travail nous avons réparti les entités nommées identifiées automatiquement entre les trois catégories représentées ci-dessous :

Noms propres		
Pays	Noms/prénoms	villes
74,75%	23,41%	1,87%

3 Apport de l'analyse linguistique pour la recherche d'information en langue arabe

En réalité, la recherche d'information est une tâche dépendante de la langue et son succès est donc lié aux langues des documents et à la manière dont les moteurs prennent en compte les caractéristiques de la langue concernée.

A notre avis, les caractéristiques qui ont le plus d'impact sur la précision des moteurs de recherche concernent principalement la structure morphologique des mots et les variations morphologiques d'un même mot, d'où l'importance accordée par les moteurs de recherche à la lemmatisation et à la troncature.

L'apport de la lemmatisation reste discutable pour l'amélioration des performances des systèmes de recherche d'information dans les documents anglais (D. Harman 1991, 1995), car les règles de formations des mots en anglais sont relativement limitées et systématiques. Les langues à morphologie complexe comme l'arabe (J. Dichy, 1990), présentent un défit aux systèmes de recherche puisque le nombre de règles morphologiques est important et non systématique. Nous allons montrer dans ce qui suit l'insuffisance des traitements de surface et l'apport de la lemmatisation à la recherche d'information en arabe.

Les moteurs de recherche utilisent les parcours de surface pour l'identification des mots, tandis que le mot graphique en arabe présente un caractère complexe et la notion de noyaux est très vague pour la recherche de séquences. L'arabe est une langue flexionnelle où les familles morphologiques (dérivées d'une même racine) peuvent atteindre une taille assez importante. Nous trouvons souvent des formes graphiques proches ou semblables mais n'appartenant pas à la même famille morphologique.

Voyons, à titre d'exemple, ce que donne une recherche de surface des dérivées du mot قال *qâl.* La requête « *قال* » donne prés de 146 formes dans un corpus, ce qui dépasse largement les possibilités dérivationnelles de ce mot. En effet, la requête renvoie des termes comme الانتقال , اعتقالهم , استقالة , العقال , مقاليد اقالتهم , التقاليد , برتقالية , اثقالاً , وقالباً , الأقاليم , قالب … En plus de ce bruit considérable, beaucoup de formes de ce mot restent muettes et il est nécessaire de les rechercher à travers d'autres requêtes : c'est le cas de toutes les formes déclinées de l'inaccompli يقول et des autres déverbaux.

En procédant à une lemmatisation automatique, malgré tous les problèmes inhérents à cette tâche, nous pourrions éviter la plupart des ambiguïtés présentées ci-dessus.

Soit l'exemple de recherche du mot « سماء *smâÞ* » *ciel*, Google renvoie 594 000 pour **سماء *smâÞ*** en appliquant le principe de la complétion. Les résultats contiennent aussi 279 000 pour **أسماء *ÞsmâÞ*** *(des prénoms ou des noms)*, parmi lesquelles nous trouvons aussi **الأسماء *âlÞsmâÞ*** *les prénoms* (les rares cas du pluriel de سماء *smâÞ* sont renvoyés dans des titres comme أسماء السماوات *ÞsmâÞ âlsmâwât* .). ***Cet écart sémantique et morphologique entre les mots de la recherche provient de l'application des règles de lemmatisation des langues à caractères latins sur l'arabe.***

D'un autre côté, la procédure de recherche est souvent confrontée à la nature agglutinante de l'arabe. Par exemple tous les noms communs peuvent être précédé du proclitique de détermination ال *âl* , de politiques de liaison, de ressemblance, d'interrogation, … et peuvent être suivis d'enclitiques pronoms ainsi que de suffixes du pluriel, du duel, du féminin, … Dans un corpus journalistique de deux millions de mots (présenté dans le dernier paragraphe), nous avons reconnus 35 formes augmenté du mot كتاب livre, pour obtenir les formes الكتاب *âlktâb, le livre* / الكتابين *âlktâbyn, les deux livres/* بالكتاب *bâlktâb, avec le livre/* بكتاب *bktâb, avec livre/* بكتابه *bktâbh, avec son(lui) livre/* بكتابها *bktâbhâ, avec son(elle) livre*

Bien entendu, l'ambiguïté « naturelle » de la langue arabe, à tous les niveaux linguistiques, vient s'ajouter pour complexifier tous ces problèmes, déjà nombreux, en recherche d'information en langue arabe.

Lemmatisation

La lemmatisation se définit par l'identification d'une forme canonique correspondant à différentes formes flexionnelles ou dérivationnelles d'un mot donné (dérivation du pluriel, du singulier, du féminin ou du masculin, etc.). L'application de la lemmatisation en recherche d'information en langue arabe ne donne pas toujours les résultats escomptés car le système régissant la dérivation en arabe est plus complexe et ne se résume pas, souvent, à une simple suffixation.

Toutefois la lemmatisation est indispensable en arabe en raison du caractère agglutinant de la langue. Le mot graphique en arabe est augmenté par les proclitiques (de coordination,

d'interrogation, marque de future, de détermination, de préposition, …) et les enclitiques (essentiellement des pronoms).

Dans ce qui suit, nous allons présenter les difficultés de la lemmatisation concernant deux aspects nominaux à savoir (1) le passage du singulier au pluriel ainsi que (2) le passage du masculin au féminin. Nous montrerons l'insuffisance des techniques de suffixation à travers des exemples.

Du côté des utilisateurs, ils introduisent toujours des mots minimaux, c'est-à-dire sans clitiques, sauf le proclitique de détermination qui permet souvent de lever les ambigüités entre les noms et les verbes.

Détermination	
Indéterminé	Déterminé
61,03%	37,97%

Etant donné le nombre considérable de réponses renvoyés pour les requêtes avec détermination ou avec clitiques d'une manière générale, nous n'avons pas pu vérifier si le moteur Google procède à la lemmatisation ou pas. En tout cas tous les résultats retournés contiennent la chaîne exacte recherchée.

Singulier pluriel

Soit le mot pluriel كتابات *ktâbât (des écrits)*, la procédure de lemmatisation classique qui considère que le pluriel est obtenu par la suffixation de ات *ât* au nom singulier pour l'obtention du pluriel donnera le lemme كتاب *ktâb* pour كتابات *ktâbât* , alors que le lemme est كتابة *ktâbt* ¨. L'obtention du singulier à partir du duel peut poser quelques difficultés aussi, pour passe du duel فتاتان *ftâtân* au singulier il est nécessaire de retirer le suffixe ان *ân* , mais cela ramène le mot à une forme incorrecte qui est فتات *ftât* , alors que la bonne orthographe est فتاة *ftât* ¨. A ce titre, nos tests sur le moteur Google ont montré son insuffisance dans le traitement de tels phénomènes linguistiques, tels que la différentiation entre les terminaisons **ت** et **ة**.

Pour ce qui est du pluriel en langue arabe, un autre phénomène s'ajoute pour complexifier la lemmatisation : il s'agit de l'existence d'un type de pluriel dit "pluriel brisé" et qui ne suit pas véritablement des règles flexionnelles précises (رجل -رجال *rÊl-rÊâl* , pour homme-hommes et امرأة *âmrÞt* نسوة¨ *nswt* نساء¨ *nsâÞ* pour femme-femmes).

Du côté des utilisateurs nous avons extrait les statistiques suivantes concernant l'utilisation du nombre dans les mots clés

Nombre				
Singulier	Duel	Pluriel		
74,21%	1,77%	24,02%		
		Régulier du masculin	Régulier du féminin	brisé
		71,09%	21,29%	7,52%

Remarque: plusieurs noms duels, d'un point de vue morphologique, désignent en réalité des noms au singulier, c'est le cas, par exemple, du nom de pays البحرين *âlbÎryn Bahrayn* ou encore le prénom محمدين *mÎmdyn* .

Masculin féminin

La règle de suffixation peut-être utilisée aussi pour l'obtention du féminin ou du masculin. Généralement, la règle appliquée consiste à ajouter au masculin le suffixe ة pour l'obtention du féminin, mais cette règle non plus n'est pas systématique, comme dans le cas du mot اثارة *â×ârt¨* ou دراسة *drâst¨* qui sont de noms féminins et n'admettent pas de masculin. Au niveau graphique nous trouvons aussi une multitude de noms masculins se terminant par la ة, marque du féminin comme dans le cas du nom خليفة *Ìlyft ¨*.

Par ailleurs, le genre peut aussi être rendu par des mots ayant des racines différentes, comme pour les mots رجل *rÊl* أب *Þb* ولد *wld* حصان *ÎÒân*جمل *Êml* .

Du côté des utilisateurs nous avons extrait les statistiques suivantes concernant l'utilisation du genre dans les mots clés

Genre					
Masculin	Féminin				
50,13%	49,84%				
	avec marque		sans marque		Autres
	ayant masculin	Sans masculin	ayant masculin	Sans masculin	Féminin d'un pluriel masculin
	47,11%	11,69%	16,81%	1,01%	23,38%

4 Pratiques d'écriture en langue arabe – problèmes généraux

La recherche d'information en langue arabe doit aussi faire face à des difficultés supplémentaires dues aux habitudes orthographiques des auteurs, dont l'impact sur la recherche d'information n'est pas négligeable. Pour examiner cet aspect, nous avons mené une étude sur un corpus journalistique contemporain et nous donnerons des statistiques relatives aux pratiques d'écriture chez les auteurs arabes contemporains. Pour notre étude, nous avons choisi de travailler sur un corpus de deux millions de mots de la presse écrite. Des analyses détaillées seront publiées dans les Journées internationales d'Analyse statistique des Données Textuelles (R. ABBES, J. DICHY, 2008).

4.1 Hamza et 'alif.

Les scripteurs confondent souvent la *hamza* (أ – إ) et le *'alif* en début de mot. On trouve par exemple, dans le corpus de deux millions de mots, 26.923 fois الى *'ilâ*, « à », et 2.089 fois إلى *Þl ¨*. On trouve également 33.901 ان *ân* indifférencié, contre 50.569 أن *Þn* (conjonctions *'an* ou *'anna*) et 759 إن *Þn* (conjonctions *'in* ou *'inna*). Les estimations auxquelles nous étions parvenus dans cette étude indiquent que le taux des items renfermant cette seule confusion s'élève à 5,79% de l'ensemble des items ou encore à 6,76% des mots. Nos tests sur le moteur Google ont montré qu'il renvoie indifféremment toutes les orthographes de la Hamza au début du mot.

4.2 Yâ' et 'alif maqṣûra.

A la *hamza-'alif* initiale s'ajoute une autre confusion, située pour celle-ci en fin de mot, entre ي (lettre *yâ'* finale) et ى (ou *'alif maqṣûra*). Le mot نادي, *nâdî*, « club », par exemple, peut être noté نادى *nâd ¨*, ce qui correspond à l'usage des typographes égyptiens (mais peut aussi être lu

comme *nâdâ*, « convier, convoquer »). De même, pour cette confusion, le moteur Google renvoie systématiquement les résultats avec ي *y* et ىَ, même si le mot de la requête est écrit entre guillemets. Par exemple, la recherche du mot الأولي *âlÞwly* renvoie en tête des pages contenant الأولى *âlÞwl* ́. Notons qu'il s'agit d'un mot qui cumule les deux difficultés.

4.3 Le caractère ‘ـ’

Les typographes font un usage fréquent du caractère ‘ـ’ (appelé *kashida*), qui permet l’allongement du trait au milieu des mots, pour une meilleure lisibilité et pour limiter les espaces blancs sur une ligne justifiée ou même parfois pour des raisons purement calligraphiques. Ce caractère, ne faisant pas partie de l’alphabet arabe, est souvent une source de confusion pour les systèmes de traitement de la langue arabe. Le moteur Google semble éliminer ce caractère dans le mot de la requête.

4.4 Absence des signes de vocalisation

L’absence des signes de vocalisation dans les textes – à laquelle les lecteurs arabes sont accoutumés –, constitue une source de difficulté considérable pour l’analyse automatique de l'arabe. Certains signes diacritiques relatifs à la base (ou noyau lexical) sont indispensables pour la détection du sens du mot. Ils sont par conséquent indispensables pour le choix du mot pertinent dans la recherche d'information, particulièrement en l'absence de contexte. Les analyses peuvent en effet reconnaître dans un même item plusieurs patrons (وزن), voire plusieurs combinaisons de patrons et de racines.

4.5 Ta marbouta ة et ha ه

Nous avons remarqué, essentiellement sur le web, une confusion fréquente entre les lettres ة et ه en fin du mot. Google semble avoir transformé la fréquence en règle et semble renvoyer exactement les mêmes résultats pour les requêtes مكتبة *mktbt¨* et مكتبه *mktbh* .

5 Conclusion

La recherche d’information en langue arabe se heurte inlassablement à ses spécificités linguistiques, notamment par sa morphologie complexe, par l’absence de signes de vocalisation dans les textes publiés et par la richesse des mots graphiques arabes. Ces spécificités sont souvent à l’origine d'une grande dissymétrie entre l’indexation et la recherche. Dans cet article, nous avons présenté une partie de la typologie des contraintes morphologiques de la langue arabe et montré le comportement des outils de recherche d'informations face à ces contraintes. Nous avons montré les limites des parcours de surface et de l’utilisation des caractères de troncature. La forme graphique des mots ne peut être utilisée pour la constitution de familles morphologiques, en raison de l’agglutination et de la richesse dérivationnelle en langue arabe. La solution réside dans l’utilisation de véritables règles linguistiques permettant la lemmatisation pour la constitution des familles morphologiques. Nous avons appuyé notre constat par des résultats d’études expérimentales concernant les pratiques d’écriture des auteurs des documents et les difficultés qu'elles engendrent pour les systèmes intégrant des traitements automatisés de la langue arabe. Une conclusion nous parait évidente à réitérer et à rappeler : elle concerne la nécessité d’intégration de véritables traitements linguistiques, à tous les niveaux (lexical, syntaxique et sémantique) pour améliorer la recherche d'information en langue arabe.

6 Bibliographie

ABBÈS RAMZI (2004) *La conception et la réalisation d'un concordancier pour l'arabe*. Thèse de doctorat en Sciences de l'Information, Lyon : INSA, décembre 2004.

ABBÈS RAMZI, DICHY JOSPH (2008) « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1 », JADT'2008, 12-13 mars 2008, ENS-LSH Lyon, France.

ABBÈS RAMZI, DICHY JOSEPH and HASSOUN MOHAMED (2004). « The Architecture of a Standard Arabic Lexical database: some figures, ratios and categories from the DIINAR.1 source program. » In *COLING'04. Proceedings of the Workshop Computational Approches to Arabic Script-bases Languages*, 28.08.2004, Genève : 15-22.

BOUALEM MALEK, SNEIFER RÉGINA (2001) "Hahooa Arabic Web Directory & Natural Language Processing for Arabic Information Retrieval", ACL 2001, Workshop on Arabic Language Processing, Toulouse, France, July 2001.

BOUALEM MALEK, ZAJAC R. (1999), "Unicode-based Arabic text", ATLAS'99, Arabic Translation and Localisation Symposium, Tunis, May 26-28, 1999.

BUCKWALTER TIM (2004). « Issues in Arabic Orthography and Morphological Analysis. » In COLING'04, Proceedings of the Workshop Computational Approaches to Arabic Script-bases Languages, 28.08.2004, Genève : 31-41.

BRIN S. and PAGE L. (1988), The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems* 30 (1988), pp. 107–117.

CHEN P., XIE H., MASLOV S. and REDNER S. (2007), Finding scientific gems with Google's PageRank algorithm, Journal of Informetrics Volume 1, Issue 1, January 2007, 8-15.

DICHY JOSEPH (1990). L'Écriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse d'État, Université Lumière-Lyon 2.

HARMAN DONNA (1991). *How effective is suffixing*? Journal of the American Society of Information Science. vol. 42, No 1. pp. 7-15, 1991.

HARMAN, DONNA, ed. (1995) TREC-3. Proceedings of the Third Text Retrieval Conference. Washington, DC: GPO.

LALLICH-BOIDIN GENEVIEVE, MARET DOMINIQUE (2005). Recherche d'information et traitement de la langue fondements linguistiques et applications. Villeurbanne : Presses de l'ENSSIB, Les Cahiers de l'ENSSIB.

NEUHAUS CHRIS, NEUHAUS ELLEN and ASHER ALAN (2008), Google Scholar Goes to School: The Presence of Google Scholar on College and University Web Sites, The Journal of Academic Librarianship In Press, Corrected Proof, , Available online 29 January 2008.

VÉRONIS JEAN (2006), Étude comparative de six moteurs de recherche. Rapport d'étude, Université de Provence, février 2006.