كفاءة التحليل الصرفي في استرجاع النصوص العربية *

مساعد بن صالح الطيار **

ملخص البحث:

لقد ظهرت في السنوات الأخيرة محاولات عدة من أجل تحسين أداء استرجاع النصوص العربية الكاملة. ولعل من أهم هذه المحاولات هو تمكين الباحث أو مستخدم نظام الاسترجاع من استرجاع معظم أو جميع صيغ الاستفسار المتوافرة في قاعدة النصوص الكاملة. ولقد استخدمت تقنيات متعددة في هذا المجال مثل المحللات الصرفية وتقنية البتر وغيرها. وتعد موسوعة الحديث النبوي (من إنتاج شركة العالمية) من أهم أنظمة استرجاع النصوص التي استخدمت تقنية التحليل الصرفي في استرجاع النصوص العربية.

ولقد تم استخدام هذه الموسوعة في إجراء عشرة استفسارات من أجل مقارنة أداء ثلاثة مناهج من مناهج البحث في أنظمة الاسترجاع العربي وهي (البحث بمستوى الكلمة، البحث بمستوى الكلمة مع اللواصق، والبحث بمستوى جذر الكلمة). هدف هذه الدراسة هو تسليط الضوء على أداء هذه المناهج الثلاثة مستخدمين قياس التحقيق والاستدعاء.

ولقد كشفت الدراسة فيما يتعلق بأداء التحقيق أن البحث بمستوى الكلمة حقق أعلى نسبة تحقيق مقارنة بالمنهجين الآخرين عند مستوى ٩٦٪، وحقق البحث بمستوى الكلمة مع اللواصق نسبة ٩٤٪، وكانت أدنى نسبة تحقيق هي من نصيب البحث بمستوى الجذر، وذلك عند مستوى ٧٥٪. أما فيما يتعلق بأداء الاستدعاء فحقق البحث

* قدم هذا البحث (باللغة الإنجليزية) كورقة بحث إلى المؤتمر والمعرض الدولي السادس للحاسبات متعددة اللغات والمنعقد في الفترة من ١٧ - ١٨ أبريل ١٩٩٨م في جامعة كمبردج (بريطانيا).

** مبتعث من قسم المكتبات والمعلومات - كلية العلوم الاجتماعية - جامعة الإمام محمد بن سعود الإسلامية لتحضير درجة الدكتوراه من جامعة ديمنتفورت (لستر بريطانيا).

بمستوى الجذر أعلى نسبة استدعاء عند مستوى ٧٩٪، يليه البحث بمستوى الكلمة مع اللواصق بنسبة ٣٦٪، بينما حقق منهج البحث بمستوى الكلمة أقل نسبة استدعاء مقارنة بالمناهج الأخرى، وذلك عند مستوى ١٨٪.

الكلمات المفتاحية: استرجاع النصوص العربية، التحليل الصرفي، التكشيف الآلي، استرجاع المعلومات.

١ - مقدمة:

إن أغلب نظم الاسترجاع التقليدية قائم على نظام المضاهاة بين الاستفسار (كلمات البحث التي يستخدمها المستفيد عند البحث في قواعد المعلومات أو غيرها) والوثائق أو النصوص المخزنة في قاعدة البيانات. ونظرًا لأن معظم اللغات البشرية قد تستخدم صيغًا عدة في التعبير عن مفهوم واحد فإن مثل هذه النظم قد تخفق في استرجاع جميع أو أغلب صيغ الاستفسار المتوافرة في قاعدة المعلومات. ومن أجل التغلب على هذه المشكلة، فقد تم تدعيم بعض نظم الاسترجاع بتقنيات لديها القدرة على استرجاع صيغ الاستفسار المختلفة مثل تقنية البتر، وتقنية المحلل الصرفي؛ وذلك من أجل رفع كفاءة الاسترجاع لهذه النظم. ويرى جمع من المهتمين بقضية استرجاع المعلومات أن تصنيف الكلمات ذات الجذر الواحد قد يعزز من كفاءة النظام في عملية الاسترجاع (Montgomery, 1972; Porte, 1980; Harman, 1990; Savoy, 1993). وفي وقت مبكر من تاريخ استرجاع المعلومات حظيت اللغة الإنجليزيّة بمزيد من الاهتمام فيما يتعلق برفع كفاءة الاسترجاع؛ وذلك بالتعامل مع اللغة الإنجليزية وتحليلها آليًا (انظر: Salton, 1971; Sparck Jones, 1973; Walker, 1987).

أما الوضع بالنسبة للغة العربية فإن هذا المجال يعد من المجالات الجديدة في بيئة الاسترجاع العربي. وبدأ هناك نوع من الاهتمام في هذه القضية المهمة يبرز في بعض الدراسات كما هو عند (الأطرم، ١٤١٠هـ؛ الصوينع، ١٤١٥هـ؛ علي، ١٩٨٨م). ولا زال هناك جدل دائر في الساحة العربية حول كفاءة التحليل الصرفي في عملية الاسترجاع بين بعض المختصين باللغويات الحاسوبية من جهة، وبين بعض من المختصين بقضية الاسترجاع باللغة العربية من جهة أخرى. حيث يرى (علي، ١٩٨٨م) و (Plessis; 1990) وهما من المهتمين باللغويات الحاسوبية أهمية بناء نظام استرجاع المعلومات باللغة العربية على التحليل الصرفي من أجل أن يسترجع النظام جميع مشتقات الاستفسار أو كلمات البحث. ويرى الفريق الآخر (الأطرم، ١٤١٠هـ) و(الصوينع، ١٤١٥هـ) - وهما من المختصين في مجال المكتبات والمعلومات - أنه ليس هناك كبير صلة بين إرجاء

الاستفسار (كلمات البحث) إلى الجذر الأصلي وبين عملية الاسترجاع. وكلا الفريقين يدعم رأيه بما يعززه من الأدلة والتعليل. وفي واقع الأمر فإن الجذر كمنهج لاسترجاع المعلومات باللغة العربية له مميزاته وعيوبه كما ستبين هذه الدراسة التي بين أيدينا.

٢ - التحليل الصرفي واسترجاع النصوص العربية:

تعد اللغة العربية من فصيلة اللغات السامية، التي تتميز - خصوصًا العربية منها - بالقدرة على توليد مئات الكلمات من جذر واحد؛ وذلك عن طريق استخدام الصيغ (القوالب) الصرفية العربية. ونظرًا لهذه الخاصية فإن التحليل الصرفي لأي نظام مبني على اللغة العربية يجب أن لا يغفل هذه الخاصية. ويؤكد على هذا النوع من التحليل الصرفي (علي، ١٩٨٨م) و(Plessis; 1990). حيث يرى الأخير أن كلمة (ولد) قد ترد وتتكرر في نص من النصوص بأشكال مختلفة مثل الولد، أولاد، ولدان.. إلخ. وهذا يعني أن النظام الذي لا يعتمد على التحليل الصرفي لن يسترجع إلا ما تم إدخاله من قبل المستفيد، وهذا - بدوره - يؤدي إلى إهمال الصيغ الأخرى ذات العلاقة. ويشير (Al Naim, 1991) إلى أن أهمية التحليل الصرفي للغة العربية تنبع من النقاط التالية:

١ - تكاد تصل نسبة الكلمات المشتقة باللغة العربية إلى ٨٠٪ من جميع الكلمات العربية.

٢ - أن الجذور العربية، والتي يشتق منها الكلمات ليست كثيرة العدد (٧٠٠٠ أو أكثر بقليل)، مما يساعد ويسهل عملية معالجة اللغة العربية آليًا.

٣ - أن الاشتقاق يتم - غالبًا - عن طريق الصيغ الصرفية، وهذه الصيغ تكاد لا تتجاوز ٤٠٠ صيغة صرفية.

ونظرًا لهذه الخاصية فقد ظهر في السنوات الأخيرة عدد من المحاولات الجيدة فيما يتعلق بمعالجة اللغة العربية آليًا. حيث تم نشر دراسات متنوعة عن عدد من المحللات الصرفية بعضها استخدم في مجال استرجاع المعلومات والبعض الآخر تم استخدامه في تطبيقات أخرى مثل الترجمة الآلية، والتصحيح الإملائي وغيرها من التطبيقات الحاسوبية. انظر على سبيل المثال (علي، ١٩٨٨م) و (Al Fedaghi and AlAnzi, 1989) و (Talouth and Al Dannan, 1989) و (Beesley and Alpnet, 1990).

إن المهتم بقضية استرجاع المعلومات باللغة العربية يعرف ما تسببه السوابق واللواحق للكلمات من مشكلات عدة. ومن أجل التغلب على هذه المشكلات يرى (Al Khrash 1991) أن نظام الاسترجاع ينبغي أن يصمم بحيث تكون لديه القدرة على نزع السوابق واللواحق من الكلمات الكشفية (الكلمات الدالة) قبل أن تخزن في الملف المنقلب (الملفات الكشفية)، بينما يذهب (البخيت، ١٤١٤هـ) إلى أن استخدام البتر بكل صوره سيحل من مشكلة السوابق واللواحق في نظم استرجاع المعلومات

العربية. وعلى أية حال، فإن البتر يعد مناسبًا بشكل أكبر للغات اللصقية (اللغة الإنجليزية) أكثر منه مناسبة للغات غنية الاشتقاق والتصريف (اللغة العربية) نظرًا لوجود الإعلال والإبدال وجموع التكسير وغيرها من المسائل الصرفية التي لا يمكن معالجتها بالبتر وحده؛ لذا يرى (علي، ١٩٨٨م) أن حل مثل هذه المشكلات ممكن، وذلك باستخدام المحلل الصرفي القادر على استرجاع جميع أشكال الكلمة، والتخلص من السوابق واللواحق وحتى من الحشو الذي يكون وسط الكلمة. وهناك عدة فوائد يمكن تحقيقها باستخدام المحلل الصرفي في استرجاع المعلومات لعل من أهمها:

١ - تمكين المستفيد من استرجاع جميع صيغ الكلمة المدخلة، دون الحاجة إلى التفكير في إدخال عدة صيغ للكلمة نفسها من قبل المستفيد.

٢ - توسيع نطاق البحث.

٣ - باستخدام المحلل الصرفي في عملية استرجاع المعلومات فإن الاستدعاء سيرتفع.

وعلى الرغم من أهمية التحليل الصرفي في عملية استرجاع المعلومات إلا أنه في أحيان كثيرة يتم استرجاع مواد غير صالحة أو غير مطابقة للاستفسار المدخل من قبل المستفيد.

٣ - الدراسات السابقة:

هناك الكثير من الدراسات التجريبية ناقشت نظم استرجاع المعلومات، والعوامل المؤثرة في كفاءة هذه النظم. ومعظم هذه الدراسات كانت حول نظم الاسترجاع المصممة خصيصًا للغات اللاتينية (أغلبية هذه النظم تتعامل مع اللغة الإنكليزية). وقد سبقت الإشارة إلى بعض هذه الدراسات في الفقرات السابقة. ونحن في هذه الفقرة من الدراسة سنركز - بشكل موجز - على بعض الدراسات المهمة التي عالجت استرجاع المعلومات باللغة العربية. ففي دراسة (الأطرم، ١٤١٠هـ) المطولة تمت مناقشة بعض القضايا اللغوية ذات العلاقة باسترجاع المعلومات العربية خصوصًا السوابق واللواحق. بينما أشار (الصوينع، ١٤١٥هـ) بشيء من الإيجاز إلى علاقة التحليل الصرفي وأثره في استرجاع المعلومات. ولقد ناقش (علي، ١٩٨٨م) أغلب القضايا المتعلقة باللغويات الحاسوبية ذات العلاقة باللغة العربية. كما تطرق لبعض المسائل الصرفية والاشتقاقية ذات الأثر في استرجاع النصوص العربية الكاملة.

وهناك ثلاث دراسات أكاديمية ناقشت مناهج الاسترجاع الثلاثة باللغة العربية (كلمة، كلمة مع اللواصق)، (ساق الكلمة)، (وجذر الكلمة). فالدراسة الأولى قام بها الخراشي (Al Kharashi, 1991) ، حيث صمم نموذجًا لاسترجاع المعلومات باللغة العربية، وقام بإجراء عدد من التجارب (استخدم خلالها ٣٥٥ عنوانًا فقط) من أجل مقارنة أداء ثلاثة مناهج بحثية (كلمة، ساق، جذر).

وتوصل إلى أن أداء الاستدعاء بالنسبة للجذر حقق النسبة الأعلى مقارنة بالمناهج الأخرى.

وفي دراسة مشابهة قام أبو سالم (Abu Salem) بإعادة تجريب عينة الخراشي، ولكن استخدم في دراسته ١٢٠ عنوانًا (في مجال الحاسب الآلي) مع مستخلصاتها. واستخدم أبو سالم في إجراء تجاربه نظام الاسترجاع نفسه الذي قام بتصميمه الخراشي ويعرف باسم microcomputer-based AIRS . وتوصل إلى نتائج قريبة من تلك النتائج التي توصل إليها الخراشي في دراسته. والإضافة التي أتى بها أبو سالم أنه قام بمقارنة أداء مناهج البحث الثلاثة بأداء البحث عبر استخدام المكنز الموضوعي. ولمزيد من التفصيل انظر دراسة أبي سالم في (Abu Salem, 1992) .

وفي دراسة حديثة لحمدي (Hmeidi, 1995)، وصف للتكشيف الآلي وأثره في استرجاع المعلومات باللغة العربية. ومرة أخرى قام حمدي بمقارنة مناهج البحث الثلاثة آنفة الذكر، وحسب زعمه توصل إلى نتائج قريبة لتلك النتائج التي توصل إليها الخراشي وأبو سالم، وهنا نقطة مهمة يجب التنبيه عليها ألا وهي أن عملية التكشيف (المستخدم في الدراسات الثلاث) لجذور الكلمات الدالة تمت بطريقة يدوية وليست آلية. بمعنى آخر، أن إرجاع الكلمة إلى جذرها تم عبر الجهد البشري، وليس هو النظام الآلي (المحلل الصرفي) الذي قام بهذا الجهد. ولا يخفى على المتخصص في هذا المجال الفرق الكبير بين التكشيف المعتمد على الجهد البشري، وبين التكشيف المعتمد على الآلة (التكشيف الآلي). والمجال هنا لا يتسع لبسط الكلام حول هذه الدراسات. وهناك دراسات مماثلة اهتمت بالتكشيف الآلي لنظم استرجاع المعلومات العربية مثل دراسة (Morafeq, 1991) ودراسة (Al Naim, 1989).

٤ - منهجية الدراسة (الإطار التجريبي):

سبق وأن ذكر أن هدف هذه الدراسة هو تسليط الضوء على أثر التحليل الصرفي في أداء استرجاع النصوص العربية. ومن أجل التحقق من ذلك تم تصميم وتطبيق الإطار التجريبي التالي:

٤ - ١ - البحث في موسوعة الحديث النبوي:

لقد تم إجراء تجارب هذه الدراسة باستخدام موسوعة الحديث النبوي الشريف، وهذه الموسوعة من إنتاج شركة العالمية للبرامج (صخر). وتضم هذه الموسوعة تسعة من أهم كتب الحديث النبوي الشريف. حيث يقدر العدد الإجمالي لأحاديث هذه الموسوعة بـ ٦٥,٠٠٠ حديث. وهذه الدراسة استخدمت واحدًا من هذه الكتب التسعة ألا وهو صحيح البخاري، والذي يبلغ عدد أحاديثه ٧٠٠٠ حديث تقريبًا. ويتراوح طول كل حديث من هذه الأحاديث بين ١٠ - ٨٠ كلمة تقريبًا. وتتميز هذه الموسوعة بميزات عدة ليس هنا مجال تفصيلها، ولكن لعل من أهمها:

١ – الخيارات المتعددة فيما يتعلق بمستوى البحث (كلمة، كلمة مع اللواصق، جذر..).

٢ – استخدام تقنية التحليل الصرفي في استرجاع النصوص العربية.

٣ – استخدام نصوص كاملة، وليست مستخلصات أو عناوين فقط.

٤ – استخدام الواصفات الكشفية لنصوص الأحاديث. وهذا بدوره يخبرنا عن عدد الأحاديث في هذه الموسوعة المتعلقة بأي استفسار يتم البحث عنه. وهذا سهل من عملية الحكم على بعض الأحاديث ومدى صلاحيتها للاستفسارات العشرة التي تم بحثها. ولقد تم تحديد عدد الأحاديث المطابقة لكل استفسار من خلال الرجوع إلى هذه الواصفات في أحيان كثيرة.

٤ – ٢ – استفسارات البحث:

إن عينة استفسارات البحث المستخدمة في هذه الدراسة عبارة عن مصطلحات تكشيفية تم اختيارها عشوائيًا من عدة قوائم (تم اختيار أربعة كتب من كتب الأحاديث الشاملة مثل جامع الأصول، وتم تصوير قائمة المحتويات وترقيمها). بلغ عدد المصطلحات التكشينية ٢٠٠ مصطلح تكشيفي. وتم اختيار عشرة مصطلحات تكشيفية بشكل عشوائي، وهي التي تم استخدامها في إجراء تجارب هذه الدراسة (وهي التي نطلق عليها اسم الاستفسار). انظر الملحق رقم (١). وتم البحث في موسوعة الحديث النبوي عن إجابات لهذه الاستفسارات باستخدام مناهج البحث التالية:

١ – البحث بمستوى الكلمة، من أجل استرجاع الكلمة المدخلة نفسها من قبل المستفيد (في هذه الدراسة، كاتب هذه الورقة هو الذي قام بإجراء البحث).

٢ – البحث بمستوى الكلمة مع اللواصق، لاسترجاع الكلمة مع السوابق أو اللواحق فقط (دون الحشو الذي يكون في وسط الكلمة).

٣ – البحث بمستوى الجذر، من أجل استرجاع جميع صيغ الاستفسار المدخل.

ولقد تم استخدام استراتيجية بحث مبسطة تتمثل في إدخال الاستفسار بصيغتين هما: الاستفسار مع (أل التعريف). انظر الملحق رقم(١).

٤ – ٣ – تقرير مدى الصلاحية:

إن مصطلح الصلاحية Relevanvce يعد من أهم المصطلحات التي عالجها علماء المعلومات في تاريخ نظم استرجاع المعلومات، وبالرغم من ذلك، إلا أنه لم يخلص إلى تعريف موحد لهذا المصطلح (Schamber, 1994) . ومصطلح الصلاحية يستخدم عادة للتعبير عن عدد من العلاقات (مثل: احتياجات

المعلومات، طلب المعلومات، الموضوع...). وعلى أية حال فإن مصطلح الصلاحية في دراستنا هذه ينصرف إلى العلاقة الموضوعية (المفهومية) بين النص الحديثي (الوثيقة) وبين الاستفسار. بمعنى أن الصلاحية تكون صحيحة إذا كان هناك توافق وتطابق بين الاستفسار وبين النص الحديثي (الوثيقة). ومن أجل تحديد مدى الصلاحية بين الاستفسارات العشرة وبين نتائج البحث لهذه الاستفسارات فقد تمت الاستعانة بثلاثة من المحكمين المختصين في مجال الحديث (يحملون الشهادة الجامعية في تخصص علوم الحديث، وأعمارهم تتراوح ما بين ٣٠ - ٣٥ سنة). وتمت مراعاة النقاط التالية عند تقرير مدى الصلاحية بين الاستفسار ونتائج البحث:

١ - تم تصوير وترقيم نتائج البحث (الاستفسارات العشرة)، وتم إرفاق كل استفسار مع نتائجه وسلمت لكل واحد من المحكمين.

٢ - لم يبلغ المحكمون عن مناهج البحث المستخدمة في البحث، وهذا يعني أن نتائج البحث للمناهج الثلاثة دمجت وسلمت للمحكمين.

٣ - لقد تم سؤال المحكمين الثلاثة أن يعطوا أحكامهم لتحديد مدى الصلاحية بين الاستفسار والنص الحديثي وفق درجات المقياس التالي (١ - علاقة قوية ٢-علاقة جزئية ٣-لا يوجد علاقة). إن استخدام هذا المقياس شائع في مجال تقييم نظم استرجاع المعلومات. ومن أجل استخراج نسب الاستدعاء والتحقيق فإن درجات المقياس الثلاث تم دمجها بشكل ثنائي، بمعنى أن النص إما أن يكون ذا علاقة أو لا يكون. وبالرغم من أن دمج درجات هذا المقياس بشكل ثنائي يتسبب في فقدان بعض المعلومات، إلا أن هذا الدمج مقبول ومدعم من قبل المهتمين بتقييم نظم استرجاع المعلومات (Schamber, 1994).

٤ - ٤ - قياس كفاءة الأداء:

هناك مقياسان مشهوران يستخدمان في تقييم أداء نظم استرجاع المعلومات. هذان المقياسان يعرفان باسم مقياس الاستدعاء (Recall) ومقياس التحقيق (Precision). حيث يستخدم مقياس الاستدعاء في اختبار قدرة النظام على استرجاع جميع الوثائق أو النصوص الصالحة في نظام ما. بينما يستخدم مقياس التحقيق من أجل التأكد من قدرة النظام على استرجاع الوثائق أو النصوص الصالحة فقط، وحجب غيرها من الوثائق أو النصوص غير الصالحة. وكلا المقياسين تم استخدامه في هذه الدراسة كما هو موضح أدناه.

مقياس الاستدعاء = عدد الوثائق (الأحاديث) الصالحة والمسترجعة / عدد الوثائق (الأحاديث) الصالحة في النظام

مقياس التحقيق = عدد الوثائق (الأحاديث) الصالحة والمسترجعة / عدد الوثائق (الأحاديث) المسترجعة

٥ - نتائج الدراسة:

يوضح الجدول رقم (١) جميع النصوص المسترجعة الصالحة وغير الصالحة لكل منهج (كلمة، كلمة مع اللواصق، جذر) للاستفسارات العشرة. ويلاحظ بشكل عام - من هذا الجدول - أن منهج البحث بمستوى الجذر يسترجع نصوصًا أكثر من المنهجين الآخرين (كلمة، كلمة مع اللواصق)؛ ولعل هذا يعود إلى قدرة البحث بمستوى الجذر على استرجاع جميع الصيغ للاستفسار، والتي تشترك في الجذر نفسه. ويلاحظ أيضًا أنه إذا استخدم منهج البحث بمستوى الجذر فإن ثلاثة أرباع (٨١٪) النصوص الصالحة تم استرجاعها عن طريق البحث بمستوى الجذر. أما النصوص المسترجعة من طريق البحث بمستوى الكلمة فإنها تعد أقل النصوص المسترجعة مقارنة بالمنهجين الآخرين. وبإيجاز فإن الجدول رقم (١) يوضح لنا أن البحث بمستوى الكلمة استرجع نسبة ١٨٪ فقط من النصوص الصالحة، واسترجع البحث بمستوى الكلمة مع اللواصق نسبة ٢٩٪ من النصوص الصالحة؛ بينما نجد أن البحث بمستوى الجذر استرجع ٨١٪ من النصوص الصالحة.

أما فيما يتعلق بالنصوص المسترجعة وغير الصالحة فإن الجدول رقم (١) يوضح أن البحث بمستوى الجذر استرجع نسبة ١:١٧ من النصوص غير الصالحة مقارنة بالبحث على مستوى الكلمة، ونسبة ١:٦ مقارنة بالبحث على مستوى الكلمة مع اللواصق. ومن ناحية أخرى فإن البحث بمستوى الجذر استرجع ١:٤ من النصوص الصالحة مقارنة بالبحث على مستوى الكلمة، ونسبة ١:٣ من النصوص الصالحة مقارنة بالبحث بمستوى الكلمة مع اللواصق. ويشير الجدول رقم (١) إلى أن نسبة ٢٣٪ من النصوص الصالحة لم تسترجع بواسطة أي منهج من المناهج الثلاثة. كما يلاحظ أن كل نص صالح تم استرجاعه عن طريق البحث بمستوى الكلمة أو الكلمة مع اللواصق فإنه حتمًا سيتم استرجاعه عن طريق البحث بمستوى الجذر.

الجدول رقم (١) مجموع نتائج البحث للاستفسارات العشرة لكل منهج

	جذر					كلمة مع اللواصق					كلمة			
الاستفسار	غ ص	ص ل	ص ر	م س	غ ص	ص ل	ص ر	م س	غ ص	ص ل	ص ر	م س	ص ق
١	٢	٨	٣١	٣٣	٠	٣٦	٣	٣	٠	٣٨	١	١	٣٩
٢	٠	٢	١٦	١٦	٠	٣	١٥	١٥	٠	٧	١١	١١	١٨
٣	٣	٠	١٧	٢٠	٢	١٢	٥	٧	٠	١٦	١	١	١٧
٤	٢	٢	٣٣	٣٥	٠	٣٠	٥	٥	٠	٣٢	٣	٣	٣٥
٥	٢٩	١	١٧	٤٦	١٧	٤	١٤	٣١	٦	٧	١١	١٧	١٨
٦	٢٦	١٢	١٢	٣٨	٠	١٧	٣	٣	٠	١٧	٣	٣	٢٠
٧	٠	٦	٥٤	٥٤	٠	٤٣	١٨	١٨	٠	٥٤	٧	٧	٦١
٨	٠	٤	٦	٦	٠	٤	٦	٦	٠	٦	٤	٤	١٠
٩	٤١	١٢	٢٧	٦٨	٠	٣٥	٤	٤	٠	٣٥	٤	٤	٣٩
١٠	٠	١٢	٢٩	٢٩	٠	٢٨	١٣	١٣	٠	٣٣	٨	٨	٤١
المجموع	١٠٣	٥٩	٢٤٢	٢٤٥	١٩	٢١٢	٨٦	١٠٥	٦	٢٤٥	٥٣	٥٩	٢٩٨

غ ص = عدد النصوص غير الصالحة والمسترجعة.

ص ل = عدد النصوص الصالحة غير المسترجعة.

ص ر = عدد النصوص الصالحة والمسترجعة.

م س = عدد النصوص المسترجعة.

ص ق = عدد النصوص الصالحة في قاعدة النصوص (صحيح البخاري).

ويكشف الجدول رقم (١) أن هناك اختلافًا واضحًا ومهمًا عندما يتم مقارنة جميع النصوص المسترجعة لكل منهج بالنصوص الصالحة. على سبيل المثال فإن عدد النصوص المسترجعة عن طريق البحث بمستوى الكلمة هو ٥٩ نصًا، بينما النصوص الصالحة تصل إلى ٢٩٨ نصًا. وهذا فرق واضح سيتم مناقشته وتوضيح أسبابه في الفقرات التالية من خلال استخدام تحليل سبب الإخفاق.

يوضح الجدول رقم (٢) قيم الاستدعاء والتحقيق لاستفسارات البحث العشرة، حيث قيمة الاستدعاء تتراوح ما بين ٠,٩٤ كحد أعلى إلى قيمة ٠,٠٢ كحد أدنى. بينما تتراوح قيم التحقيق ما بين ١,٠٠ كحد أعلى إلى قيمة ٠,٣١ كحد أدنى.

الجدول رقم (٢) نسب التحقيق والاستدعاء للاستفسارات العشرة لكل منهج

	التحقيق			الاستدعاء		
الاستفسار	جذر	كلمة مع اللواصق	كلمة	جذر	كلمة مع اللواصق	كلمة
١	٠,٩٣	١,٠٠	١,٠٠	٠,٧٩	٠,٠٧	٠,٠٢
٢	١,٠٠	١,٠٠	١,٠٠	٠,٨٨	٠,٨٣	٠,٦١
٣	٠,٨٥	٠,٧١	١,٠٠	٠,٨٥	٠,٣٠	٠,٠٥
٤	٠,٩٤	١,٠٠	١,٠٠	٠,٩٤	٠,١٤	٠,٠٨
٥	٠,٣٧	٠,٤٥	٠,٦٤	٠,٩٤	٠,٧٧	٠,٦١
٦	٠,٣١	١,٠٠	١,٠٠	٠,٦٠	٠,١٥	٠,١٥
٧	١,٠٠	١,٠٠	١,٠٠	٠,٨٨	٠,٢٩	٠,١١
٨	٠,٧٦	١,٠٠	١,٠٠	٠,٦٠	٠,٦٠	٠,٤٠
٩	٠,٣٤	١,٠٠	١,٠٠	٠,٦٩	٠,١٠	٠,١٠
١٠	١,٠٠	١,٠٠	١,٠٠	٠,٧٠	٠,٣١	٠,١٩

ويلاحظ من الجدول رقم (٣) أن متوسط الاستدعاء للجذر هو ٠,٧٩ بينما متوسط الاستدعاء للبحث بمستوى الكلمة هو ٠,١٨ ومتوسط الاستدعاء بالنسبة للبحث بمستوى الكلمة مع اللواصق هو ٠,٣٦ . ويوضح الجدول نفسه أن أعلى نسبة تحقيق هي من نصيب البحث بمستوى الكلمة عند مستوى ٠,٩٦ يليه البحث عن طريق الكلمة مع اللواصق عند مستوى ٠,٩٤ وأخيرًا فإن أقل نسبة تحقيق هي من نصيب البحث بمستوى الجذر، وذلك عند مستوى ٠,٧٥ .

الجدول رقم (٣) متوسط الاستدعاء والتحقيق لمناهج البحث الثلاثة

	جذر	كلمة مع اللواصق	كلمة
متوسط الاستدعاء	٠,٧٩	٠,٣٦	٠,١٨
متوسط التحقيق	٠,٧٥	٠,٩٤	٠,٩٦

٦ - تحليل سبب الإخفاق:

لقد تم في هذه الدراسة تطبيق تحليل سبب الإخفاق لكل من الاستدعاء والتحقيق من أجل الإجابة على الأسئلة التالية: لماذا يوجد إخفاق فيما يتعلق بالاستدعاء أو التحقيق؟ ما أسباب هذا الإخفاق؟ وكما ذكر (Lancaster, 1972) فإن هناك عدة أسباب للإخفاق مثل استراتيجية البحث، سياسة التكشيف، الاستفسار، حاجات المستفيد... وفي هذه الدراسة سيتم التركيز على أسباب الإخفاق المتعلقة بالمسائل اللغوية دون غيرها، ويقصد بالمسائل اللغوية تلك المسائل المتعلقة بـ (الاشتقاق، السوابق واللواحق، التغير الدلالي...).

انظر الجدول رقم (٤) حيث يعرض أنواع سبب الإخفاق لكل من الاستدعاء والتحقيق. وكما هو واضح من الجدول رقم (٤) فإن سبب الإخفاق يمكن تقسيمه إلى ما يلي:

الجدول رقم (٤) سبب إخفاق التحقيق والاستدعاء لكل منهج

	التحقيق			الاستدعاء		
نوع الإخفاق	جذر	كلمة مع اللواصق	كلمة	جذر	كلمة مع اللواصق	كلمة
تعدد أشكال الكلمة	–	–	–	–	١٥٢	١٧٥
المترادفات	–	–	–	٥٦	٦٠	٦١
تعدد المعنى	٤٥	–	–	–	–	–
السوابق واللواحق	–	–	–	–	–	٩
التغير الدلالي	٢٧	٢	–	–	–	–
التشكيل	–	–	٦	–	–	–

٦ - ١ - إخفاق أداء الاستدعاء:

هذا النوع من الإخفاق يخبرنا عن سبب إخفاق منهج معين (كلمة أو جذر) في عدم استرجاع نصوص صالحة على الرغم من وجودها في قاعدة النصوص أو قاعدة المعلومات. وعند فحص كل منهج على حدة تبين أن البحث بمستوى الكلمة أخفق في استدعاء ٢٤٥ نصًا صالحًا من مجموع ٢٩٨. أي أن نسبة إخفاق الاستدعاء بالنسبة للبحث بمستوى الكلمة هو ٨٣٪. بينما نجد نسبة الإخفاق تقل عندما نستخدم البحث بمستوى الكلمة مع اللواصق حيث أخفق هذا المنهج في استدعاء ٢١٢ نصًا صالحًا (٧١٪) من مجموع ٢٩٨ نصًا صالحًا. وبعد فحص هذه النصوص تبين أن سبب إخفاق الاستدعاء يعود إلى النقاط التالية:

٦ - ١ - ١ - تعدد أشكال الكلمة:

ولعل أغلب أسباب الإخفاق بالنسبة للاستدعاء يعود إلى الحقيقة التالية، وهي أن الكلمة المدخلة من قبل المستفيد برزت في نصوص الأحاديث بعدة أشكال (أو قل بعدة صيغ). مثال ذلك الاستفسار رقم (٢) عن (الرهن) تكررت صيغ هذا الاستفسار بعدة أشكال مثل (مرهون، رهنه، مرتهن...) بينما كانت الكلمة المدخلة هي (الرهن، رهن)؛ لهذا السبب أخفق هذا المنهج في استرجاع الصيغ الأخرى التي لم يتم إدخالها من قبل المستفيد. ويوضح لنا الجدول رقم (٤) أن مجموع إخفاق الاسترجاع بواسطة الكلمة هو ٢٤٥ نصًا صالحًا. وبعد فحص هذه النصوص تبين أن ١٧٥ نصًا (٧١٪) من هذه النصوص يعود سبب الإخفاق إلى تعدد أشكال الكلمة المدخلة في قاعدة نصوص الأحاديث. أما بالنسبة لما يتعلق بالبحث بمستوى الكلمة مع اللواصق فإن مجموع الإخفاق هو ٢١٢ نصًا صالحًا، كان من نصيب تعدد أشكال الكلمة المدخلة ١٥٢ نصًا بنسبة ٧٢٪ ومن هذا يتبين لنا أن نسبة إخفاق الاستدعاء - فيما يتعلق بالصيغ الشكلية للكلمة المدخلة - بالنسبة لمنهج البحث بمستوى الكلمة ومستوى الكلمة مع اللواصق متقاربة جدًا. وقبل الانتقال إلى نقطة أخرى، لعله من المفيد التنبيه على أن هذا النوع من الإخفاق (تعدد صيغ الكلمة) يعد نادرًا إن لم يكن مفقودًا بالنسبة للبحث بمستوى الجذر؛ وذلك نظرًا للقدرة الهائلة لهذا المنهج على استرجاع جميع صيغ الكلمة المدخلة دون الحاجة إلى أن يفكر المستفيد من نظام استرجاع المعلومات بإدخال صيغ أخرى للاستفسار، بل تكون هذه المهمة من مهام المحلل الصرفي الذي يقوم بالبحث عن صيغ الكلمة المدخلة نيابة عن المستفيد.

٦ - ١ - ٢ - المترادفات:

وهذا نوع آخر من أسباب إخفاق الاستدعاء. وفي واقع الأمر فإن هذا الإخفاق مرتبط بالتحليل للكلمة وليس بالتحليل الصرفي. لذا فإن هذا النوع

من الإخفاق لن يناقش في هذه الورقة لأنه خارج نطاقها. والجدول رقم (٤) يوضح مرات الإخفاق لهذا النوع.

٦ - ١ - ٣ - السوابق واللواحق:

كان بالإمكان أن يضم هذا النوع من الإخفاق إلى النوع الأول وهو (تعدد صيغ الكلمة). ولكن هناك فرقًا جعلنا نفرد هذا النوع من الإخفاق على حدة. وهو أن عملية التحليل اللغوي بالنسبة للسوابق واللواحق أسهل منه عندما تكون المسألة متعلقة ببنية الكلمة. ولتقريب الصورة نعطي المثال التالي: يكمن نزع السابقة (أل) من الكلمة (الرهن)، وهذا الأمر يحتاج منا فقط إلى كتابة خوارزمية مبسطة تتعرف على السوابق في اللغة العربية من أجل نزعها عند عملية التحليل الصرفي. أما إذا كان التغيير يشمل صلب الكلمة فإن التحليل الصرفي يحتاج مزيدًا من العمق والنفاذ إلى بنية الكلمة من أجل التعرف على صيغتها الصرفية، وعلى الزوائد التي قد تطرأ عليها مثل كلمة (المراهنات). وعلى أية حال فإن الإخفاق بالنسبة لهذا النوع كان قليلاً جدًا في هذه الدراسة حيث بلغ عدد مرات الإخفاق لهذا النوع ٩ مرات من أصل ٢٤٥ إخفاقًا. ونقطة يجب التنبيه عليها، وهي أن في استراتيجية البحث كما هو موضح في الملحق رقم (١) تم استخدام نوعين من الصيغ لكل استفسار وهما البحث باستخدام (أل) التعريف والبحث بدونها. وهذا قد يكون له أثر في تدني الإخفاق بالنسبة لهذا النوع. ونقطة مهمة أخرى، وهي أن هذا الإخفاق عادة ما يوجد عندما يستخدم منهج البحث بمستوى الكلمة فقط دون غيره من مستويات البحث الأخرى.

٦ - ٢ - إخفاق أداء التحقيق:

لقد تم استخدام هذا المقياس من أجل التأكد من أداء وكفاءة كل منهج من مناهج البحث الثلاثة (كلمة، كلمة مع اللواصق، جذر) فيما يتعلق باسترجاع النصوص الصالحة فقط، وفي الوقت نفسه حجب غيرها من النصوص غير الصالحة أو غير الموافقة للاستفسار. ولقد بينت هذه الدراسة - وكما هو موضح في الجدول رقم (٤) - أن إخفاق أداء التحقيق يكون شبه نادر عندما نستخدم منهج البحث بمستوى الكلمة أو الكلمة مع اللواصق. ومن ناحية أخرى فإن هذا النوع من الإخفاق يرتفع بشكل ملحوظ عندما يستخدم منهج البحث بمستوى الجذر. وباختصار فإن إخفاق التحقيق يمكن أن يقسم إلى ثلاثة أنواع كما يلي:

٦ - ٢ - ١ - تعدد المعنى:

هذا النوع من الإخفاق يتعلق بمستوى البحث عن طريق الجذر دون غيره. وأرى أن إعطاء مثال أفضل من الدخول في التفاصيل النظرية لهذا الإخفاق، فالاستفسار رقم (٩) يتعلق بموضوع (الأضاحي)، عندما تم بحث هذا الاستفسار عن

طريق الجذر، استرجع النظام ٦٨ نصًا، وبعد فحص هذه النصوص تبين أن النصوص الصالحة والموافقة للاستفسار هي ٢٧ نصًا؛ وهذا يعني أن بقية النصوص ليست صالحة ولا مطابقة للاستفسار. ومن أجل التعرف على سبب الإخفاق تم فحص النصوص غير الصالحة فوجد أن هذه النصوص مرتبطة بصلاة الضحى أو وقت الضحى، بالإضافة إلى نصوص أخرى عن عيد الأضحى. وهذا يوضح لنا أن البحث بمستوى الجذر نجح في استرجاع جميع مشتقات الكلمة المدخلة، ولكن الإخفاق كان مرده إلى تغير المعنى لكلمة (ضحى) التي هي أصل أو جذر الاستفسار (الأضاحي).

٦ - ٢ - ٢ - التغير الدلالي (السياق):

هذا النوع من الإخفاق شبيه بالنوع السابق، ولكن الإخفاق هنا مرتبط بالسياق. فلا يمكن أن يعرف معنى الكلمة إلا بسياقها. على سبيل المثال كان ناتج الاستفسار رقم (٦) - موضوعه عن الوصية - عشرين نصًا مسترجعًا. وبعد فحص هذه النصوص وجد أن ١٢ نصًا غير صالحة أو موافقة للاستفسار، على الرغم من أن معنى الجذر لهذه النصوص معنىً واحد، ولكن عندما تمت قراءة السياق لهذه النصوص وجد بعضها كما يلي:

١ - أوصى بثلث ماله (الوصية المعروفة، وهذا هو المراد من الاستفسار).

٢ - أوصى أمته (نصح أمته).

٣ - أوصاني خليلي (أمرني خليلي).

٦ - ٢ - ٣ - التشكيل:

بالرغم من أهمية التشكيل باللغة العربية لإزالة اللبس الصرفي أو الدلالي، إلا أن أغلب النصوص العربية غير مشكولة. وهذا في رأيي ليس له كبير أثر - إلى حد ما - على استرجاع المعلومات باللغة العربية. ولكن أظهرت هذه الدراسة أن هناك إخفاقًا في أداء التحقيق كان مرده إلى غياب التشكيل. على سبيل المثال الاستفسار رقم (٥) كان عن موضوع (السحر) بتشديد السين وكسرها. لقد استرجع النظام ١٧ نصًا عبر منهج البحث بمستوى الكلمة لهذا الاستفسار. وبعد فحص هذه النصوص تبين أن ستة منها كانت عن (السحر) بفتح السين، وتعني الكلمة بداية الصبح الباكر، وليس كما كان متوقعًا أن يسترجع النظام عن (السحر) بتشديد السين وكسرها.

٧ - الخاتمة:

إن هذه الدراسة التجريبية هدفت إلى مقارنة ثلاثة مناهج من مناهج البحث المستخدمة في نظم الاسترجاع العربية وهي: البحث بمستوى الكلمة، البحث بمستوى الكلمة مع اللواصق، والبحث بمستوى الجذر، ويمكن القول إن النتائج التي تم التوصل إليها في هذه الدراسة تعزز النتائج التي

توصل إليها كل من Al Kharashi (1991) و Abu Salem و Hmeidi (1995) من أن البحث بمستوى الجذر يرفع من أداء النظام خصوصًا فيما يتعلق بالاستدعاء. ونتائج هذه الدراسة لا يمكن بحال أن تؤخذ على أنها نتائج نهائية، على الرغم من المؤشرات المهمة التي تم التوصل إليها بخصوص مناهج الاسترجاع الثلاثة؛ وذلك نظرًا لمحدودية عينة الدراسة، وكذلك طبيعة النصوص التي تم تطبيق الدراسة عليها (الأحاديث النبوية). ويرى صاحب هذه الدراسة أن أداء التحقيق بالنسبة للبحث بمستوى الجذر يمكن تحسينه إذا كان هناك نوع من العمق في التحليل اللغوي للجذور العربية. وعلى ضوء النتائج التي تم التوصل إليها في هذه الدراسة فنحن لا ندعي أن البحث بمستوى الجذر بالنسبة للغة العربية هو المنهج الأكثر مناسبة، وفي الوقت نفسه أيضًا لا نرى أن البحث بمستوى الجذر ليس له علاقة باسترجاع المعلومات باللغة العربية. وموقفنا هذا نابع من عدة نقاط تم ملاحظتها حين القيام بتجارب هذه الدراسة، والتي يمكن تلخيصها في النقاط التالية:

١ – إن كفاءة أداء البحث بمستوى الجذر معتمدة على عمق التحليل للجذور العربية، أو طريقة معالجة هذه الجذور.

٢ – من المعروف أن بعض الجذور العربية تتميز بكثرة معانيها المشتركة، بينما هناك جذور تكاد لا تتجاوز المعنى الواحد أو المعنيين.

٣ – من المعروف أيضًا أن بعض الجذور العربية لديها القدرة الكبيرة على توليد مئات الكلمات، بينما هناك بعض الجذور لا تتميز بهذه الخاصية.

٤ – قد يكون استخدام البحث بمستوى الجذر مهمًا في بعض المجالات مثل البحث في النصوص الشرعية أو القانونية، بينما قد لا يكون مناسبًا للنصوص العامة أو الشاملة.

٥ – أن استخدام البحث بمستوى الجذر قد يكون مفيدًا للمستفيد الذي يريد استدعاءً مرتفعًا (جميع الوثائق أو النصوص ذات العلاقة أو القريبة من موضوع البحث)، بينما قد لا يكون كذلك بالنسبة للمستفيد الذي يريد تحقيقًا مرتفعًا (الوثائق أو النصوص ذات العلاقة القوية – فقط – بموضوع البحث).

وأخيرًا، فإن قضية أثر اللغة في استرجاع المعلومات باللغة العربية لا زالت بحاجة إلى مزيد من البحث والدراسة خصوصًا فيما يتعلق بمستوى التحليل الصرفي. وهذا ما يقوم به صاحب هذه الدراسة في الوقت الحاضر. أملاً أن تقدم هذه الدراسة وغيرها من الدراسات المستقبلية ما ينفع ويحسن من أداء نظم استرجاع المعلومات باللغة العربية.

الملحق رقم (١)

الاستفسار	استراتيجية البحث	الاستفسار	استراتيجية البحث
١ – الضيافة	الضيافة أو ضيافة	٦ – الوصية	الوصية أو وصية
٢ – الرهن	الرهن أو رهن	٧ – الميراث	الميراث أو ميراث
٣ – النصيحة	النصيحة أو نصيحة	٨ – زكاة الفطر	زكاة الفطر
٤ – الاعتكاف	الاعتكاف أو اعتكاف	٩ – الأضاحي	الأضاحي أو أضاحي
٥ – السحر	السحر أو سحر	١٠ – الوليمة	الوليمة أو وليمة

المصادر والمراجع

أولاً – المراجع العربية:

الأطرم، محمد بن عبدالله. **كفاءة اللغة العربية في تكشيف واسترجاع الوثائق الربية** .– بحث (مقدم إلى إدارة البحث العلمي بمدينة الملك عبدالعزيز للعلوم والتقنية)، ١٤١٠هـ.

البخيت، بخيت سليمان. **البحث في العنوان في قواعد البيانات العربية** – دراسة تطبيقية على حزمة برمجيات ISIS /CDS. في السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات .– الرياض: مكتبة الملك عبدالعزيز العامة، ١٤١٤هـ.

الصوينع، علي السليمان. **استرجاع المعلومات في اللغة العربية** .– الرياض: مكتبة الملك فهد الوطنية، ١٤١٥هـ.

علي، نبيل. **اللغة العربية والحاسوب** .– الكويت: تعريب، ١٩٨٨م.

ثانيًا – المراجع الأجنبية:

Abu Salem, Hani. **A microcomputer based Arabic bibliographic information retrieval system with relation thesauri** (Arabic- IRS), Ph. D Thesis. Chicago: Illinois Institute of Technology, 1992.

Al Fedaghi, Sabah S. and Fawaz S. Al Anzi. A new algorithm to generate Arabic root-Pattern forms. In: **The 11th National Computer Conference**, 1989. PP. 391-400.

Al Kharashi, Ibrahim. **A microcomputer- based Arabic information retrieval system comparing words, stems, and roots as index terms, ph**. D Thesis. Chicago: Illinois Institute of Technology, 1991.

Al Naim, Faisal. **Text analysis and automatic indexing for Arabic based automated information retrieval system**. M.Sc Thesis, Chico: California State University 1991.

Beesley, Kenneth R. and Alpnet. Finite-state description of Arabic morphology. In: **Second Cambridge Conference Bilingual Computing in Arabic and English**, 1990.

Harman, Donna. How effective is suffixing?. **Journal of the American Society for Information Science**, 1991. 42 (1), PP. 7-15.

Hmeidi, Ismael Ibranhim. **Design and implementation of automatic word and phrase indexing for information retrieval with Arabic documents**, Ph. D. Thesis. Chicago Illinois Institute of Technology, 1995.

Lancaster, F. Wilfrid. Evaluation and testing of information retrival systems. In: Allen Kent and Harold Lancour (ed.), **Encyclopedia of library and information science**, 1972. (8), pp. 234 - 259.

Montgomery, Christine A. Linguistics and information science. **Journal of the American Society for Information Science**, 1972. 23 (3). pp. 195 - 219.

Morfeq, Ali Hussein. **Bayan: a text management system for Arabic engineering documents**. Ph. D. Thesis. Colorado: Colorado University, 1990.

Plessis, B. Du.. Producing Arabic document-dictionaries or concordance: Prospects. In: **Second Cambridge Conference Bilingual Computing in Arabic and English**, 1990.

Porter, M. F. An Algorithm for suffix stripping. **Program**, 1980. 14 (3), pp. 130 - 137.

Salton, Gerard. ed. **The SMART retrieval system: experiments in automatic document processing**, New Jersey: Prentice- Hall, Inc, 1971.

Schamber, Linda. Relevance and information behavior. **Annual Review of Information Science and Technology**, 1994. 29, pp. 3-48.

Spark Jones, Karen and Martin Kay. **Linguistics and information science**. New York: Academic Press, 1983.

Walker, Stephen and Richard M. Jones. **Improving subject retrieval in online catalogues**. London: British Library Board, 1987.