مشكلات استخدام اللغة العربية فى نظم استرجاع المعلومات الببليوجرافية

أسامة لطفى محمد

مدرس مساعد: قسم المكتبات

كلية الآداب جامعة المنوفية ـ مصر

email: OS - LOTFY @ FRCU. EUN. EG

أولاً: المعايير

أ: شفرات تمثيل الحروف العربية.

منذ اختراع الحاسب فى بداية الأربعينيات، وحتى منتصف السبعينات اقتصرت معالجة الحاسبات على البيانات المعتمدة على الحروف الرومانية، وبدأ فى منتصف السبعينات يظهر اهتماماً كبيراً فى استخدام الحاسب مع النصوص غير الرومانية، وتنقسم الحروف غير الرومانية إلى قطاعين:

1 - **القطاع الأول:** لا تزيد حروفه عن 256 حرف، ومنها الحروف العربية والحروف الروسية

2 - **القطاع الثانى:** والذى تزيد حروفه عن 256 حرف، مثل: الحروف الصينية والكورية.

ومما ييسر عملية التعريب هو عدم حاجة الشفرة العربية إلى أكثر من 256 حرف[1].

ولكن ظهرت مشكلة حيث اختلفت الأكواد Code Pages باختلاف شركات الإنتاج لبرامج التعريب، ولكن منذ عام 1976 بدأت مجموعة من الجهود العربية فى محاولة تطوير كود عربى موحد، نتج عنها ما يعرف بالأسمو 449 (ASMO 449) ، وهى الشفرة العربية للحروف ذات السبع محارف (7 Bits)، والمستخدمة فى الحاسبات الصغيرة (Mini - Computers)، و (ASMO - 708) ككود عربى للثمانية محارف[2].

ويلاحظ أن استخدام ASMO - 708 فى بناء قواعد البيانات الببليوجرافية كاد أن يقتصر على قواعد البيانات المبنية باستخدام الإصدارة الثالثة من برنامج CDS / ISIS وذلك لعدم التزام مصممو النظم فى العالم العربى باتباع هذه الشفرة، بل تم استخدام الشفرات الخاصة بنظم التعريب المستخدمة مثل (نافذة ومساعل العربى وفجر. . . إلخ).

هذا إلى أن ظهرت بيئة Windows المعربة ونتيجة لطغيان Windows كبيئة للتشغيل أصبحت الشفرة الحروف العربية الخاصة بها هى معيار الأمر الواقع للثمان محارف.

ومن أهم التطورات الأخيرة هو ظهور الشفرة الموحدة للحروف UNICODE والتى تغطى جميع حروف

اللغات الحية ومنها العربية ـ حيث تسمح بتمثيل 65000 حرف وقد ظهر منها الطبعة الثانية حتى الآن إلا أنها لم تنتشر الانتشار المتوقع حتى الآن سواء على المستوى العالمى أو على المستوى العربى ويخشى أن تلقى نفس مصير ASMO - 708.

ب ـ بنية التسجيلة الببليوجرافية:

أما بالنسبة لاستخدام الحاسب الآلى، نجد أن أهم الجوانب التى تعتمد على المعايير: هو جانب إنشاء بنية التسجيلات الببليوجرافية، وذلك لأهمية هذا النوع من المعايير عند التعاون بين المكتبات، أو عند شراء التسجيلات الببليوجرافية سواء على الخط المباشر، أو على أقراص مدموجة، أو على أقراص ممغنطة.

وتنقسم هذه المعايير إلى ثلاث مستويات، هى:

أولاً: مستوى الهيكل العام للتسجيلة الببليوجرافية:

ويتناوله معيار واحد حالياً وهو ISO 2709.

ثانياً: المستوى التفصيلى لتيجان وأسماء الحقول والحقول الفرعية

وتغطيه مواصفة MARC و UNIMARC و CCF.

ثالثاً: مستوى محتويات حقول التسجيلة الببليوجرافية

والذى تغطيه قواعد الفهرسة الأنجلو أمريكية وخطط التصنيف المختلفة وقوائم رؤوس الموضوعات أو المكانز.

وسوف نتناول معايير المستوى الأول والثانى فيما يلى:

1 - معيار ISO 2709:

والخاص بتبادل التسجيلات الببليوجرافية على الأشرطة الممغنطة، حيث يعطى صيغة عامة لتبادل التسجيلات الببليوجرافية وتسجيلات ذات العلاقة بها، مثل تسجيلات الاستناد، وهو لا يحدد طول أو محتويات التسجيلات بينما هو يختص بتقديم معيار دولى يصف هيكل عام، أو إطار مصمم خصيصاً ليستخدم فى الاتصال بين نظم تجهيز البيانات، وليس للاستخدام فى التجهيز داخل تلك النظم، وعلى الرغم من أنه صمم مبدئيًا ليعمل على الشرائط الممغنطة، إلا أنه يمكن استخدامه مع أى من وسائط البيانات الأخرى Data Carriers، وهو بصفة عامة يقسم التسجيلة الببليوجرافية إلى:

معرف التسجيلة (الفاتح) Record Labels

دليل المحتويات Directories

حقول البيانات Data Fields

فواصل التسجيلات Record Separators

ولا توجد مشكلة للتعامل مع البيانات العربية عند هذا المستوى حيث أن المعيار لا يقتصر على شفرة محارف معينة.

2 - مواصفة UNIMARC و USMARC :

ثانى هذه المعايير: هى مواصفة Universal MARC UNIMARC، وهذه الصيغة تحدد المعرفاتTags ومؤشرات Indicators، ورموز الحقول الفرعية Subfield Codes المستخدمة فى التسجيلات الببليوجرافية، وهى تعمل تحت مظلة المعيار ISO 2709، ولكنها تقوم بوظيفة أكثر تفصيلاً وهى تحديد حقول ثابتة التيجان والمواصفات للتسجيلات الببليوجرافية، وقصد بها أن تستخدم كصيغة عامة لتبادل التسجيلات الببليوجرافية على المستوى الدولى أو الوطنى.

3 - تركيبة التراسل المشتركة CCF :

ثالث هذه المعايير وأحدثها، هو مواصفة CCF، والتى بنيت اعتماداً على المعايير السابقة لها، ومنها UNIMARC، وتتضمن أيضًا تعريف المعرفات، والمؤشرات، ورموز الحقول الفرعية، وهى تتميز عنه بتغطيتها الأشمل لكل من المنفردات والأوعية غير المستقلة التى لا تغطيها MARC، وقد صدر منها ثلاث طبعات، آخرها تمتد تغطيها إلى قواعد البيانات غير الببليوجرافية للأشخاص والمؤسسات ومشروعات البحوث، وقد صدرت الطبعة الثالثة عام 1992 فى جزءين الأول التسجيلات الببليوجرافية، والثانى البيانات الحقائقية Factual Data.

وعند هذا المستوى تختفى المعايير العربية، ومما يستغرب له أن أكثر المعايير استخدامًا فى العالم العربى هو معيار USMARC والذى فرضته النظم التجارية المعربة.

ثانياً: خصائص اللغة العربية

ينبغى أن نؤكد أن عملية تعريب برنامج ما لا تعنى فقط مجرد ترجمة للرسائل والقوائم التى يعمل من خلالها النظام، حيث أن للغة العربية خصائص تجعلها متميزة ومختلفة عن اللغات الأخرى الأجنبية، وتؤثر هذه الاختلافات على قدرة برامج استرجاع النصوص بصفة خاصة، أى يجب عند التعريب مراعاة هذه الخصائص للغة العربية حتى نقلل من تأثيرها فى جودة الاسترجاع[3].

على سبيل المثال لا الحصر :

1 - ترتبط أداة التعريف بالكلمة على عكس اللغة الإنجليزية، فبالتالى يسهل فى الإنجليزية استبعاد أداة التعريف The عن طريق ملف الكلمات الموقوفة، ونتيجة أيضًا لوجود مسافة بينها وبين الكلمة سوف نجد أن جميع تكرارات الكلمة سواء المعرف منها أو غير المعرف تحت مدخل واحد فى الملف المقلوب أو فى الكشافات المطبوعة، أما بالنسبة للغة العربية، فلو اكتفى بمجرد ترجمة البرنامج، سوف يوجد لدينا كم كبير من الكلمات المكررة فى ترتيبها الهجائى تحت حرف الألف واللام مثلاً وكذلك تحت الحرف التالى لها، مثال: العلم، علم، حيث نجد أنها تتكرر مرة تحت حرف الألف، ومرة أخرى تحت حرف العين، وهكذا تشتت تسجيلات المصطلح الواحد تحت مدخلين.

2 - يوجد عدد من البادئات ترتبط بالكلمة، هى: همزة الاستفهام، الواو، الكاف، الفاء، اللام، الباء، السين، الياء، بالتالى يمكن أن تتكرر الكلمة نفسها فى الملف المقلوب تحت حروف تلك البادئات بالإضافة إلى حروفها الأصلية، مثل: بمصر، كمصر، فمصر، ... إلخ، وذلك بالإضافة إلى مصر، فيؤدى ذلك إلى تشتت المصطلح الواحد إلى عدة تكرارات فى الملف المقلوب مما يؤدى إلى زيادة فى الحجم الخاص به، والتأثير على جودة الاسترجاع نتيجة لتشتت المصطلح.

3 - لا تختلف اللغة العربية كثيراً عن اللغة الإنجليزية فى وجود اللواحق، والتى يتم معالجتها عن طريق استخدام أسلوب البتر، ولكنها بالتأكيد تختلف فى تغير حروف الكلمة تبعاً لموقعها من الإعراب، وتختلف أيضًا فى وجود الكثير من صيغ التثنية والجمع للمصطلح على عكس اللغة الإنجليزية التى فى أكثر من 90% من الحالات يكون الجمع بإضافة حرف S أو es لنهاية الكلمة.

4 - تختلف معانى بعض الكلمات فى اللغة العربية باختلاف علامات التشكيل المستعملة معها عِلْم، عَلَم، عَلِمَ أو كَتَبَ، كُتُب... وفى حالة استخدام علامات التشكيل فتواجهنا مشكلة عند الفرز - حيث يتم التعامل معها كحروف مضافة للكلمة ويجب دائمًا البحث بالتشكيل المطلوب.

نتعرف على أكثر هذه المشكلات تأثيراً على نظم الاسترجاع للتسجيلات الببليوجرافية العربية، والتعامل معها، - ونركز هنا على التسجيلات الببليوجرافية، وذلك لاختلاف اللغة المستخدمة فى وصف أوعية المعلومات عن تلك التى تستخدم فى نصوص أوعية المعلومات ذاتها -، وقد تم عمل مجموعة من التجارب للمقارنة بين طبيعة البحث فى كلمات حقل العنوان لتحديد أهم هذه المشكلات:

حيث تم إنشاء ملف مقلوب لجميع الكلمات الواردة فى حقل العنوان لعدد 2500 تسجيلة عربية من قاعدة البيانات لنظام الفهرسة الخاص بمكتبة كلية الآداب جامعة المنوفية، وتم أيضًا عمل ملف مقلوب لعدد 2500 تسجيلة باللغة الإنجليزية من قاعدة بيانات الفهرسة لمكتبة مركز بحوث التنمية والتخطيط التكنولوجى ووجد الآتى:

بالنسبة للغة العربية:

1- الكلمة بدون «ال» فى 37% من الحالات.

2- تظهر الكلمة معرفة «ال» فى 62% من الحالات.

3- تظهر الكلمة مع بادئة من البادئات فى أقل من 1%من الحالات.

4 - أكثر البادئات تواترًا (باستثناء الواو) فى الظهور فى عناوين الكتب هى «لل» والباء، فعلى سبيل المثال تظهر كلمة المكتبات فى الملف المقلوب بعدة أشكال هى:

المكتبات

مكتبات

بالمكتبات

للمكتبات

المكتبة

مكتبة

بينما يظهر نفس المصطلح فى الملف المقلوب لقاعدة البيانات الإنجليزية بالأشكال التالية:

library

Libraries

Librarian

Librarianship

ويلاحظ أنه فى حالة استخدام اللغة الإنجليزية فإن استخدام أسلوب البتر فى البحث سوف يسترجع جميع المصطلحات، حيث أن جميع التغيرات تظهر فى اللواحق.

أما فى حالة استخدام اللغة العربية لن يصلح أسلوب البتر فى البحث لاسترجاع جميع الأشكال التى ظهر بها المصطلح لأن هناك تغيرات على مستوى البادئات، وفى المثال المذكور لن يتم استرجاع كل من:

المكتبة

المكتبات

بالمكتبة

فيما يلى الحلول التى لجأ إليها الباحث لحل هذه المشكلات من خلال استخدام نظام CDS / ISIS.

أولاً: أداة التعريف «ال»: ـ

وهى أهم هذه المشكلات نظرًا للنسبة العالية لتكرار الكلمات المعرفة بال ويمكن التعامل مع هذه المشكلة بعدة من الطرق:

1 - عن طريق وضع أداة التعريف بين علامتى «< ... >» عند إدخال البيانات ـ حيث أن برنامج- CDS ISIS يتعامل مع الحروف الموضوعة بين هاتين العلامتين «< ... >» بأن يستخدمهما فى العرض والطباعة فقط، ولا يستخدمهما سواء فى بناء الملف المقلوب أو فى عملية الفرز ـ ولكن هذا يمثل جهدًا كبيرًا على مدخلى البيانات ويستهلك الكثير من الوقت.

2 - عمل برنامج يقوم بهذه العملية آليًا، وفى هذه الحالة تظهر مشكلة الكلمات التى يكون فيها حرفا الألف واللام جزءًا أصلياً من الكلمة: مثل: ألمانيا، ألبانيا، ألم... إلخ.

3 - يتم استخدام برنامج لإضافة علامتى «< >» آلياً مع الاعتماد على ملف يتضمن الكلمات المستثناة أو الموقوفة عن هذه العملية هو يختلف بالطبع عن ملف الكلمات الموقوفة لقاعدة البيانات، حيث يمنع الثانى الكلمات من الدخول فى الملف المقلوب كلية بينما يمنع الأول حذف الألف واللام من الكلمات الموجودة به.

وقد استخدم هذا الحل فى إنشاء فهرس المجموعة العربية لمكتبة جامعة الملك فهد للبترول والمعادن، وذلك بوضع ملف إيقاف للكلمات التى تتضمن «ال» أصلية يتضمن حوالى 200 كلمة[4]، ولكن من الأفضل أن يتم بناء هذا الملف تدريجيًا تبعاً لاحتياجات قاعدة البيانات دون الاعتماد على ملف سابق التجهيز، مما يؤكد ذلك هو أن عدد الكلمات التى تضمنت «ال» أصلية فى عناوين جميع الكتب العربية فى مكتبة كلية الآداب جامعة المنوفية هو ثلاث كلمات هى: «الف / الفية / الله».

4 - تعديل برنامج البحث بحيث يقوم المستفيد بإدخال المصطلح ثم يقوم البرنامج بتوليد جميع الأشكال الممكنة للمصطلح وإدخالها كبدائل للبحث باستخدام الرابطة المنطقية «أو».

مثال:

يقوم المستفيد بإدخال مصطلح حاسب فيقوم البرنامج بتوليد الأشكال التالية:

«الحاسب أو بحاسب أو فحاسب أو لحاسب أو للحاسب أو كحاسب أو بالحاسب. . إلخ».

ولكن هذا الأسلوب قد ينتج عنه كلمات بلا معنى أو أخطاء صرفية مثل: المصر، هذا بالإضافة إلى أنه لا يؤثر على عملية الفرز، حيث سوف تخرج الكلمات بنفس الترتيب للبادئات.

5 - استخدام البتر فى اتجاه اليمين ولكن هذا الحل يعترضه ندرة برامج استرجاع النصوص التى تدعم البتر فى الاتجاهين والبتر الداخلى، ولكن هذا البديل أيضاً لن يؤثر فى عمليات الفرز.

ثانياً: مشكلة البادئات:

لا يمكن استخدام الطريقة الثانية أو الثالثة فى علاج هذه المشكلة، حيث لا يمكن جمع جميع الكلمات التى تحتوى على هذه البادئات ووضعها فى ملف إيقاف، أى معنى ذلك أن نضع جميع الكلمات العربية التى تبدأ بتلك الحروف التسعة، وأيضًا يمثل الحل الأول، والذى يتمثل فى وضع علامتى «< >» حول البادئة «مثل: < ب > مكتبة < ب > المكتبات» عبئاً كبيراً على مدخلى البيانات، وبهذا نجد أنه:

1 - يفضل استخدام الأسلوب الثالث بالنسبة للألف واللام لأهميته فى الفرز، واستخدام نفس الأسلوب بالنسبة ل «لل».

2 - بالنسبة للواو، فيتم كتابتها مفصولة عن الكلمة بمسافة ويتم وضعها فى ملف الإيقاف، كى لا تؤثر على حجم الملف المقلوب المستخدم فى البحث.

3 - ويفضل استخدام الأسلوب الرابع بالنسبة لباقى البادئات حيث لن تؤثر بشكل كبير على جودة الفرز، حيث لم يوجد عنوان واحد من ٢٥٠٠ عنوان يبدأ ببادئة.

أما بالنسبة إلى البديل الخامس وهو استخدام محلل صرفى ليؤدى نفس الوظيفة التى يقوم بها البتر فى اللغة الإنجليزية، وقد تم إنشاء ما يعرف بالمعالج الصرفى متعدد الأطوار يستخدم فى الاسترجاع لكلمات القرآن الكريم، وبعض التطبيقات الأخرى مثل ضغط (compression) الحروف العربية والتدقيق الإملائى العربى[5].

ولكن هنا يجب أن نشير إلى استخدام جميع الجذور للكلمة العربية بدلاً من البتر أو معالجة البادئات سوف يؤدى نتائج غير حميدة فى مجال استرجاع التسجيلات الببليوجرافية، ولتخيل كم الكلمات التى سوف تسترجع بالجذر: علم أو كتب بدلاً من معلومات أو مكتبات، فسوف يسترجع البرنامج عدد كبير من الكلمات التى لاتمت بصلة موضوعية للبحث أكبر من عدد التسجيلات ذات الصلة. حيث أن المحللات الصرفية للغة العربية لم تخرج بعد من إطار توليد الكلمات من الجذر لاستخراج الصيغ الصرفية والتى تستخدم فى البحث.

ثالثاً: استخدام اللغة العربية فى البيئة متعددة اللغات

والمقصود بالبيئة متعددة اللغات فى هذا السياق هو تواجد تسجيلات ببليوجرافية بأكثر من لغة طبيعية مثل العربية والإنجليزية، وقد اعتمدت مراكز المعلومات فى أغلب الأحيان على الفصل بين تسجيلات اللغتين، ولكن هذا الحل غير عملى بالنسبة للمستفيد الذى يريد إجراء بحث موحد فى جميع التسجيلات

الموجودة فى المكتبة فى موضوع معين بغض النظر عن اللغة، وتمثل هذه الطريقة فى المعالجة بأن يتعامل النظام المتكامل مع قاعدة بيانات ببليوجرافية موحدة تتضمن التسجيلات الببليوجرافية باللغتين.

ولكن استخدام قاعدة بيانات موحدة لتسجيلات اللغتين يؤدى إلى عدد من المشكلات عند التطبيق تتمثل فى:

أولاً: وجوب توحيد محددات الحقول الفرعية لشكل الاتصال المستخدم:

فعلى سبيل المثال تستخدم الحروف الإنجليزية كمحددات للحقول الفرعية للنسخة الأصلية من CCF، وتستخدم الحروف العربية فى النسخة المعربة من CCF، وهنا تظهر مشكلة عند تصميم جدول اختيار الحقول وصيغ العرض، حيث إذا اتبعنا معيارين فى نفس قاعدة البيانات، سوف نحتاج إلى عمل نسختين من كل من صيغ العرض، ومضاعفة عدد المداخل فى الملف المقلوب.

وبهذا يقترح تعديل مؤشرات الحقول الفرعية فى كل من النسخة العربية والإنجليزية من CCF إلى استخدام الأرقام بدلاً من الحروف الهجائية، سواء الإنجليزية أو العربية، وذلك لأن الأرقام لها شفرات موحدة فى اللغتين، مما يؤدى إلى امكانية توحيد صيغ العرض لتسجيلات اللغتين وعدم تكرار مداخل الملف المقلوب وسهولة البرمجة، بالإضافة إلى توحيد محددات الحقول الفرعية عند إدخال البيانات لتسجيلات اللغتين.

ثانياً: مشكلة توحيد أشكال العرض:

حيث يؤدى استخدام صيغة للعرض تتضمن نصاً مضافاً لبيانات التسجيلة ـ عنوان الحقل مثلاً ـ مع تسجيلة من لغة مخالفة إلى اضطراب وتشويش فى شكل المخرجات على الشاشة أو على الطابعة.

وحلاً لهذه المشكلة قام الباحث بالاستعانة بلغة الصياغة لبرنامج CDS / ISIS فى إعداد مواصفات لأشكال العرض ذات حساسية للغة، حيث تعرض النصوص المضافة العربية مع التسجيلة العربية وتتغير آليا إلى الإنجليزية فى حالة التسجيلة الإنجليزية.

ثالثاً: سياسة التحليل الموضوعى المستخدمة:

ونظراً لطبيعة قواعد البيانات المتعددة اللغات، يجب أن تتبنى المكتبة سياسة محددة تجاه رؤوس الموضوعات حيث توجد خيارات عدة، هى:

1 - استخدام رؤوس موضوعات أو واصفات باللغتين، بمعنى وجود رؤوس موضوعات عربية للتسجيلات العربية، ورؤوس موضوعات إنجليزية للتسجيلات الإنجليزية، مما قد يسبب مشكلة فى اختيار قائمتين مختلفتين لرؤوس الموضوعات تعتمد كل منهما على فلسفة خاصة بها فى تصنيف المعرفة، مما يؤثر على التوافق Consistency فى سياسة التحليل الموضوعى، ويفصل بين التسجيلات ذات الموضوع الواحد المختلفة اللغة.

2 - استخدام رؤوس موضوعات بلغة واحدة العربية أو الإنجليزية لتجميع التسجيلات باللغتين، وسوف يحرم ذلك المستفيدين من البحث باللغة الأخرى والتى قد تكون اللغة الأم لهم.

3 - استخدام مكنز متعدد اللغات: وهنا يجب أن يتقبل البرنامج الأساليب المستخدمة فى إدارة المكانز المتعددة اللغات، والتى تبنى أساساً على قدرة النظام على استدعاء التسجيلات لرأس موضوع بلغة معينة، بغض النظر عن لغة الرأس الذى استخدم فى الوثيقة، أى أن للمكتبة الحرية فى إدخال الرأس بأى لغة من اللغتين، ويقوم البرنامج بالاسترجاع باللغتين.

وقد اختير الحل الثالث، حيث تم تطوير البرنامج الخاص بالمكانز ليسمح بإدارة المكانز متعددة اللغات، حيث يتم إدخال الواصفات باللغتين العربية والإنجليزية إلى المكنز، ويقوم البرنامج باستدعاء المكنز عند الإدخال، حيث يسمح بإدخال المصطلح بأى من اللغتين، وفى حالة البحث يقوم البرنامج بالبحث بالمصطلح وبديله باللغة الأخرى.

وهناك أكثر من تجربة لإنشاء مكانز ثنائية اللغة فى الوطن العربى، منها:

«المكنز العربى للنشاطات الاجتماعية والاقتصادية والسياسية»

والذى صمم بواسطة مجلس الوزراء الكويتى فى بداية الثمانينات.

وهناك أيضا:

المكنز العربى للبترول

والذى انتهى العمل فيه عام 1987 فى المعهد العربى البترولى «للتدريب» التابع لمنظمة الأوبك[6].

وهذا بالإضافة إلى: مكنز جامعة الدول العربية ثلاثى اللغة: عربى، إنجليزى، فرنسى.

رابعاً: مشكلات إتاحة قواعد البيانات العربية من خلال شبكة إنترنت:

ظهر عدد من المشكلات عند إتاحة قواعد البيانات العربية من خلال شبكة إنترنت من أهمها ما يلى:

1 - الاختيار بين الإتاحة من خلال برنامج متصفح يعمل من خلال mswindows العربية مما يحجب البيانات العربية عن العالم الخارجى لعدم توافر النسخة العربية خارج حدود الوطن العربى أو الاعتماد على برنامج مساعد Plugin يسمح بعرض الحروف العربية على جميع إصدارات Windows الأجنبية ولكن هذا يزيد العبء على المستخدم من ناحية ولا توجد معايير لهذه البرامج من ناحية أخرى.

2 - زيادة عبء عملية التعريب على الأجهزة الخادمة Servers والتى سوف تستخدم لعرض البيانات العربية بشفرة لاتوجد على نظام التشغيل الأصلى لها لكى تعرض على متصفحات موجودة على أجهزة عملاء clients تعمل على أكثر من نظام تشغيل.

3 - معيار Z 39. 50.

يضيف معيار Z 39. 50 بعدًا جديدًا لتعقيدات تعريب قواعد البيانات الببليوجرافية على شبكة إنترنت يتمثل فى تعريب كل البرامج الوسيطة بين الخادم والعميل والتى تتراسل بشكل معيارى من خلال Z 39. 50 وذلك لأن أغلب هذه البرامج الوسيطة تتعامل مع النصوص بشفرات ذات 7 محارف مما يشوهه البيانات العربية ذات الثمانية محارف.

المراجع

(1) - Richard Lee, (1978). "MINISIS: a multilingual information management System". Automated systems for access to multilingual and multiscript library materials problems and solutions. - Munchen: IFLA, 1987. p. 215

(2) - Aman, Mohammed M. (1987). Use of Arabic Script in computerized information systems: Automated systems for access to multilingual and multiscript library materials problems and solutions. - Munchen: IFLA, 1987, p. 129.

(3) تجدر الإشارة فى هذا الصدد إلى مجموعة من الجهود التى تمت لدراسة تأثير اللغة العربية على كفاءة الاسترجاع فى نظم المعلومات البيليوجرافية.

ونستطيع معالجة هذه الدراسات من حيث كيفية تعرفها للمشكلة ـ فهناك الدراسات التى قامت بالرصد النظرى لبعض خصائص اللغة العربية ومدى تأثيرها المتوقع على مدى كفاءة الاسترجاع.

فقد تعرض البعض لاختلاف بنية الكلمة العربية عن الإنجليزية من نواح عدة، ومنها:

بناء الكلمات، الكلمات المتصلة، السوابق، اللواحق... إلخ

ومن هذه الدراسات:

(أ) ـ بيار فيرميل. معالجة المعلومات فى اللغة العربية. «الإعلامية والتعريب». بيروت ـ لبنان: الوكالة الإعلامية الفرنسية، جمعية معالجة اللغة العربية فى الإعلامية، 1984 . ص ص 4 ـ8.

وهناك بعض الدراسات التى تناولت تأثير بعض صفات اللغة العربية فى الاسترجاع الفعلى نتيجة لتجارب عملية، مثل:

(ب) ـ ناصر محمد السويدان. «الاسترجاع الموضوعى بواسطة كلمات العنوان». السجل العلمى لندوة استخدام اللغة العربية فى تقنية المعلومات. الرياض: مكتبة الملك عبد العزيز العامة، 1993. ص ص 533 ـ568 .

والذى تعرض فى دراسته لبعض مشكلات بناء كشافات الكلمات المفتاحية الناتجة عن «ال»، وحروف الجر والتصاقها بالكلمات فى اللغة العربية.

كذلك الدراسة المقدمة من:

(ج) ـ سعد عبد العزيز. «نظام الوثائق: نحو نظام ببليوجرافى عربى للوثائق الحكومية فى مكتبات معهد الإدارة العامة». السجل العلمى لندوة استخدام اللغة العربية فى تقنية المعلومات. الرياض: مكتبة الملك عبد العزيز العامة، 1993. ص ص 337ـ344.

حيث تعرض لبعض المشكلات الناتجة عن الخلط بين أشكال كل من: ء، أ، ال، ت، ه، ى، الالف المقصورة.

والدراسة المقدمة من:

(ء) بخيت سليمان البخيت. «البحث فى العنوان فى قواعد البيانات العربية: دراسة تطبيقية على خدمة برمجيات CDS / ISIS». السجل العلمى لندوة استخدام اللغة العربية فى تقنية المعلومات. الرياض: مكتبة الملك عبد العزيز العامة، 1993. ص ص 569 ـ 580 .

والذى قام بحصر تكرارات (الكاف، الباء، اللام فى عناوين 5000 (خمسة آلاف) تسجيلة عربية مخزنة على برنامج CDS / ISIS، وقد اقترح إضافة البتر من اليمين لتحسين كفاءة الاسترجاع فى برنامج CDS / ISIS.

وأخيراً توجد الدراسات التى تعرض لبعض الحلول العلمية التى تم استخدامها فعلياً عند استخدام اللغة العربية ومن هذه الدراسات:

(هـ) - Zahiruddin Khurshid. (1992). "Arabic Online Catalog" Information Technology and Libraries. September 1992, pp. 244 - 251.

حيث قام بعرض بعض الأساليب المستخدمة فى نظام DOBES / LIBIS والتى اعتمدت على وضع علامات خاصة قبل وبعد البادئات واستخدام ملف إيقاف لمعالجة (ال).

والدراسة التى قدمها:

(و) ـ سريع محمد السريع. «نظام ابن النديم فى مكتبات معهد الإدارة العامة». السجل العلمى لندوة استخدام اللغة العربية فى تقنية المعلومات. الرياض: مكتبة الملك عبد العزيز العامة، 1993. ص ص 315 ـ 336.

والذى عرض لبعض مشكلات بناء كشافات الاسترجاع للكلمات العربية واتبع أسلوباً يعتمد على وضع رمز من أربعة رموز قبل بعض الكلمات عند الإدخال لأخذ القرار فى شكل دخول هذه الكلمات إلى الكشاف.

(4) - Zahiruddin Khurshid. (1992). "Arabic Online Catalog". Information Technology and Libraries. September, 1992. p. 249.

(5) نبيل على. (1988). اللغة العربية والحاسوب . ـ القاهرة: دار تعريب للنشر، 1988. ص ص 330 ـ 332.

(6) - Shawky Salem. (1991). "Computerized Bilingual Thesauri". Microcomputer for information management. Mar 1991. p. 29.

(*) تجدر الإشارة إلى وجود البرنامج لدعم إنشاء المكانز العربية تم تطويره فى المملكة العربية السعودية ((راجع «عبد الجبار عبد الرحمن العبد الجبار». «استخدام نظام المستشار فى بناء المكانز العربية». السجل العلمى لندوة استخدام اللغة العربية فى تقنية المعلومات. الرياض: مكتبة الملك عبد العزيز العامة، 1993 . ص ص 631 ـ 647)).