استخدام الرومنة لتحسين الربط بين الكلمات في المدونات المتوازية

المدونات الفرنسية والعربية نموذجا

هدى سعدان[1]، سمار نصر الدين[2]، وفاء بجاوي[3]

الملخص:

يهدف هذا المقال إلى دراسة رومنة الأعلام العربية وكيفية استغلالها، وذلك بهدف تحسين نتائج مقاربة لغوية تعتمد على ربط الكلمات البسيطة والمركبة في النصوص المتوازية التي يتم دراستها انطلاقا من مدونات باللغتين الفرنسية والعربية. تعتمد هذه المقاربة على أساسين يشكلان فحوى منهجها، ألا وهما: استخدام قاموس ثنائي اللغة مع مراعاة الخصائص اللغوية للألفاظ المستعملة والكلمات المشتقة من أصل واحد (أي التي تشترك في أصل المعنى) والتي تسهم في ربط الكلمات. كما يتم تحليل الترابط النحوي والتركيبي بهدف ربط ورصف الكلمات المركبة التي يتم ترجمتها حرفيا. هذا، وقد قمنا بتقييم رابط الكلمات الإلكتروني من خلال تجريبه على مدونة متوازية، بحيث تحصلنا على نتائج جد مشجعة. الجدير بالذكر، لقد قمنا بالاعتماد على برنامج ARACADE II في هذه التجربة، والذي يسمح لأي باحث يسعى إلى وضع نظام ربط كلمات بتقييمه.

الكلمات الدالة: رومنة، الأعلام العربية، مقاربة لغوية، النسخ

Abstract:

This study discusses the way to improve the phonological transfer from Arabic to Latin languages and vice versa. To reach this aim, texts from Latin and Arabic languages are used. Besides, the programme ARACADE II is being used as a basis for our experimentation.

Key Words: Phonology, ARCADE II, Transfer, Lexis, Experimentation

[1]- (LIDILEM)، جامعة قرونوبل (Grenoble)، فرنسا.

[2] - LVIC, CEA LIST, 91 191 Gif sur Yvette.

[3]- (URNOP)، جامعة الجزائر 2، الجزائر.

يهدف هذا المقال إلى دراسة رومنة الأعلام العربية وكيفية استغلالها، وذلك بهدف تحسين نتائج مقاربة لغوية تعتمد على ربط الكلمات البسيطة والمركبة في النصوص المتوازية التي يتم دراستها انطلاقا من مدونات باللغتين الفرنسية والعربية. تعتمد هذه المقاربة على أساسين يشكلان فحوى منهجها، ألا وهما: استخدام قاموس ثنائي اللغة مع مراعاة الخصائص اللغوية للألفاظ المستعملة والكلمات المشتقة من أصل واحد (أي التي تشترك في أصل المعنى) والتي تسهم في ربط الكلمات. كما يتم تحليل الترابط النحوي والتركيبي بهدف ربط ورصف الكلمات المركبة التي يتم ترجمتها حرفيا. هذا، وقد قمنا بتقييم رابط الكلمات الالكتروني من خلال تجريبه على مدونة متوازية، بحيث تحصلنا على نتائج جد مشجعة. الجدير بالذكر، لقد قمنا بالاعتماد على برنامج ARACADE II في هذه التجربة، والذي يسمح لأي باحث يسعى إلى وضع نظام ربط كلمات بتقييمه.

والرومنة مصطلح يطلق على النقل الكتابي لأصوات كلمة ما الى نظام كتابة اللغات اللاتينية (اللغات الأوروبية المتفرعة عن اللغة اللاتينية) وذلك بغض النظر عن نطقها. تشهد الرومنة أوعملية النسخ باللغات الرومانية طفرة كبيرة نظرا لانتشار استعمال كم هائل من اللغات في الشبكة العنكبوتية ونظرا للاحتياجات الهائلة في مجال البحث عن المعلومات عبر اللغات.

وهذا صحيح بشكل خاص فيما يتعلق عن البحث عن أسماء الأعلام (أسماء الأشخاص، الأماكن، الشركات والمنظمات، وما إلى ذلك)، ولكن هذه الأخيرة تتميز بعدد وافر من أشكال مكتوبة والهجاء والنسخ في اللغات المختلفة. وتوضح حالة أسماء الأعلام العربية هذا الوضع المعقد ومتعدد الجوانب.

في هذ المقال، نبدأ بتقديم الخطوط العريضة للمسائل النظرية والصعوبات العملية التي تنشأ في مجال نسخ ورومنة الأسماء والألقاب مع اقتراح الحلول الممكنة التي قد تسهم في حل هذه الصعوبات. نقدم بعد ذلك نظامنا الآلي لرومنة الأسماء العربية المشكلة وغير المشكلة نحوالكتابات المختلفة بالحروف اللاتينية. ومن أجل إظهار أهمية الرومنة في تطبيقات المعالجة الآلية للغة، استخدمنا نتائج هذه الرومنة لتحسين آلية ربط الكلمات ألا وهي المحاذاة.

1. نبذة تاريخية

شغلت مشكلة النسخ والرومنة بل ولازالت تشغل العديد من المتخصصين المهتمين بمعالجة اللغات، ولكن هذا الاهتمام يعتبر جديد نسبيا لأنه شهد طفرة كبيرة منذ تدفق كم هائل من المعلومات عبر شبكة

الإنترنت واتساع مجال البحث عن هذه المعلومات عبر مختلف اللغات.

في الحقيقة، نلاحظ أن الأنظمة المقترحة تهدف لتعيين رومنة واحدة لإسم معين: وهذا هوالحال للنموذج التوليدي المقترح للأسماء الإنجليزية المكتوبة باللغة اليابانية (Katakana) من خلال البرنامج النصي اللاتيني (Knight, 1997). وقد تم تكييف هذا البرنامج (Stalls, 1998) مع الطريقة التي يتم عبرها نسخ الاسم الإنجليزي المكتوب باللغة العربية إلى اللغة الإنجليزية. هذا ويعتمد نظام توليد الرومنة على قاموس التلقين غير آخذا بعين الاعتبار المنطوقات غيرالمدرجة والمعروفة بالنسبة للقاموس.

وقد أدى ذلك بعض الباحثين للتغلب على هذا النقص باللجوء إلى التقنيات الإحصائية. وهذا هوالحال بالنسبة لبرنامج رومنة أسماء الأعلام الإنجليزية إلى اللغة العربية التي اقترحها (AbdulJaleel, 2003). لكن هذه التقنية أظهرت أيضا حدودها لأنها تقوم على حساب الشكل الأكثر احتمالا، والذي يُفترض أن يكون الشكل الصحيح، وهذا غير صحيح بالنسبة لجميع البلدان العربية أولجميع اللهجات.

للتحايل على صعوبة النطق ومشكلة اللهجات (الغامدي، 2005)، اقترح الغامدي نظام رومنة في الكتابة باللغة الإنجليزية للأسماء العربية المشكلة. ويستند هذا النظام على قاموس الأسماء العربية التي يتم فيها ضبط النطق باستخدام الصوائت التي تضاف إلى الأسماء المدرجة مع الإشارة إلى ما يعادلها في الكتابة باللغة الإنجليزية. ولكن هذا النهج يجمع بين عيوب الطرق السابقة له: ليس فقط أنها لا تشمل المنطوقات غير المدرجة في القاموس، ولكنها أيضا لا تقترح سوى رومنة واحدة فقط لاسم معين. ونحن نرى أن هدف المؤلف هو تعزيز تبني معيار قياسي للرومنة، إلا أن هذا العمل لا يمكن أن يكون نتيجة لمبادرة فردية ومعزولة ولا يمكن أن يقوم عليها وحدها.

في الواقع، المستوى الحالي للبحوث في هذا المجال لا يعبر عن الحالة الراهنة الشديدة التعقيد في مجالي النسخ والترجمة، والتي تؤثر سلبا على كل ما هو شفهي وكتابي في عدة أنظمة لغوية في الوقت نفسه.

في الواقع، تعد عملية رومنة أسماء الأعلام من نظام كتابة مصدر إلى نظام كتابة هدف مهمة حساسة تتطلب عددا من العمليّات والتي تتطلب بدورها النظر في مجموعة من الخصائص الصرفية، والصوتية والدلالية.

وتعتبر هذه العمليّات أمر ضروري لضمان عملية الرومنة، بما في ذلك التطبيقات الأمنية، والتحقق من الهوية، أوإيجاد المعلومات في الإنترنت.

ومع ذلك، لا يوجد حاليا أي بحث في مجال الإعلام الآلي يأخذ بعين الاعتبار الرابط:

- بين التصريف الصوتي (الفونولوجيا) والنسخ بين اللغات.
- بين الغرافيم والرومنة متعددة اللغات.
- بين اللهجات العربية وأنظمة الحروف اللاتينية.

يوجد عدد قليل من الدراسات التي تقترح حلا يأخذ بعين الاعتبار واحدة من هذه القضايا، منها ما هو مخصص للتعرف الآلي على أصل المتكلم وذلك من خلال اللهجة. هذا هو الحال بالنسبة لأعمال (Pellegrino 1999)[1], (Guidère 2004), (de Barkat-Defradas 2004)[2]. وما نقوم به يعد جزءا من البحوث التي نعدها . فالهدف الحالي الذي حددناه وأخذناه على عاتقنا من باب المسؤولية في ضبط الرومنة هوتوفير نظام الرومنة والنسخ الآلي الذي يأخذ في عين الاعتبار جميع الجوانب المذكورة أعلاه، وهي الصلة بين علم الأصوات و الغرافيم واللهجات المختلفة في رومنة الأسماء والألقاب العربية والمكتوبة باللغة العربية نحو الحروف اللاتينية. وللقيام بذلك، قمنا بتحديد عددا من القواعد المستمدة من دراسات التجريبية، والتي تعكس مدى تعقيد هذا المجال.

في الواقع، يوجد هناك العديد من الأمثلة التي ينبغي النظر فيها وفقا لمستوى التحليل المطلوب. نلخص هذه الأمثلة في الجدول التالي، والذي قمنا بتصميمه انطلاقا من ملاحظات تجريبية:

وحدة معالجة الهدف	وحدة معالجة المصدر	نوع المعالجة الآلية
غرافيم اللغات اللاتينية (فرنسية، kh)، (إنجليزية، ḫ) (إسبانية، j)	غرافيم اللغة العربية الفصحى: خ	**رومنة**
غرافيم اللغة العربية الفصحى: خ	غرافيم اللغات اللاتينية (فرنسية، kh)، (إنجليزية، ḫ)، (إسبانية، j)	
غرافيم عربي (لهجة أخرى، مصر) ج=g	غرافيم عربي (لهجة محددة، تونس) ڤ = g	
غرافيم لاتيني (لغة أخرى، لغة إنجليزية): U	غرافيم لاتيني (لغة فرنسية): ou	
فونيم لاتيني (إنجليزية، K)، (فرنسية، k)، (إسبانية، q)	فونيم عربي (وفقا للهجات) مثل: ق	

فونيم عربي (اعتمادا على اللهجات) مثل: غ/ ر	فونيم لاتيني مثل: g/r في الفرنسية	**نسخ**
فونيم عربي (للهجة أخرى) مثل: ق	فونيم عربي (للهجة محددة) مثل: ء	
فونيم لاتيني (لغة فرنسية: z)	فونيم لاتيني (لغة انجليزية: th)	

وبالإضافة إلى ذلك، يجب أن يكون نظام النسخ والرمنة قادرا على إدارة كل من الأسماء العربية التي تحتوي على الصوائت المشكلة وغير المشكلة. كما يجب أن يوفر النظام كل الرومنات الحرفية الصحيحة وفقا للكتابة الأصلية (الغرافيم) أوالكتابة الصوتية أوالنطق (الفونيم)، وليس كتابة واحدة فقط التي تكون على الأرجح الأكثر صحة على المستوى الإحصائي أوالأفضل على المستوى المعياري، إلا أنها لا تعبر عن الواقع اللغوي العربي.

2. رومنة الأسماء العربية إلى اللغات اللاتينية

تستخدم الأبجدية العربية لكتابة العديد من اللغات الآسيوية والأفريقية، وهي الأبجدية الثانية من حيث الاستخدام العالمي بعد الأبجدية اللاتينية. يعود استخدام هذه الأبجدية في تدوين النصوص العربية إلى قديم الزمان، وأشهرهذه النصوص القرآن الكريم.

تكتب الحروف العربية من اليمين إلى اليسار، بنمط يعتمد على وصل حروف الكلمة الواحدة ببعضها، وتشمل هذه الأبجدية 28 حرفا أساسيا. تعتبر بعض الحركات الصوتية جزأ من الأبجدية العربية أيضا، لأنه يشار إلى هذه الحركات برموز اختيارية.

ويوجد أيضا بعض الظواهر الصرفية والصوتية التي يجب أن تؤخذ بعين الاعتبار في الرومنة، مثل ازدواجية الحروف، والتي يعبر عنها في الكتابة العربية ب"الشدة"، تكرار الصوائت التي يعبر عنها في الكتابة العربية ب"التنوين".

3. منهجية وضع نظام الرومنة

لقد قمنا باختيار منهجية "من الأسفل إلى الأعلى" لبناء نظام الرومنة. بعبارة أخرى، بدأنا بالقيام بمعاينة حالية من خلال الرومنات المتواجدة لكل حرف من الأبجدية العربية باللجوء إلى المعايير والممارسات القياسية التي تم ملاحظتها على شبكة الإنترنت. ويستند التحقيق التجريبي على مدونة من النصوص التي تم جمعها في اللغات المستهدفة المختلفة والتي يغطيها نظام الرومنة. ولقد ساهمت هذه العملية في إنشاء مكتبة من الغرافيم المستخدم حاليا في الكتابات التي تستخدم الأبجدية اللاتينية.

ويلخص الجدول التالي بعض معادلات الغرافيم بين اللغة العربية واللغات اللاتينية انطلاقا من مدونات الدراسة:

حرف عربي	مايعادله بالحرف اللاتيني	حرف عربي	مايعادله بالحرف اللاتيني
ء	ʾ, a	غ	Gh, gh, Ğ, ğ, ḡ
ا	A, a, ä, â, á, ā, e, ê	ف	F, f, ph
ب	B, b	ق	Q, q, C, c, K, k
ت	T, t	ك	K, k, C, c
ث	Th, th, t, t̲,	ل	L, l
ج	J, j, Dj, dj, g, Ğ, ğ	م	M, m
ح	H, h, Ḥ, ḥ, ḥ, 7	ن	N, n
خ	Kh, kh, ḫ, h̲	ه	H, h
د	D, d	و	W, w, ou, o, u, ô, û, ū, ú, ü
ذ	Dh, dh, D, d, D̲, d̲, Ḍ, ḍ	ي	I, i, y, ï, î, ī,
ر	R, r	آ	A, a, ā, ʾā, ʾâ
ز	Z, z, Z̦, z̦	ة	H, h, T, t, at, a, ẗ
س	S, s	ى	A, a, á, à, ā, ỳ
ش	Ch, ch, Sh, sh, Š, š	أ	A, a, á, à, ā
ص	S, s, Ṣ, ṣ, Ṣ, ṣ	ؤ	U, u, Ou, ou, Û, û
ض	D, d, Ḍ, ḍ, Ḍ, ḍ	إ	I
ط	T, t, Ṭ, ṭ, Ṭ, ṭ	ئ	ʾ, (Blanc)
ظ	Z, z, Ẓ, ẓ, 6’, Dh, dh, D, d	ڤ	G, g
ع	ʿ, ‘, ‘, 3, a, â		

4. معادلات الكتابة بين الأبجدية العربية والأبجدية اللاتينية:

سمح لنا تحليل المدونة باستنتاج أن بعض الحروف العربية لا مثيل لها في الحروف اللاتينية، وقد تم نسخها باستخدام الأرقام العربية في النصوص المكتوبة بالأحرف اللاتينية. هذا النوع من الرومنة يمثل القاعدة الأساسية في كتابة الرسائل القصيرة المستعملة في أوروبا والشرق الأوسط.

ومن خلال الجمع بين هذين النوعين من التمثيل الرمزي، يمكن لنا أن نجد في نصوص الرومنة معادلات للأسماء في اللغات اللاتينية والألقاب الشائعة في العالم العربي :

Nom en arabe **اسم بالحروف العربية**	منى	عدنان	حنان	طارق
Exemple d'équivalents en écriture latine **مثال عن التكافؤ بالحروف اللاتينية**	Mouna ou Mona	Adnane ou 3adnan	Hanane ou 7anan	Tarek ou 6ariq

في الواقع، يعد هذا الاختلاف في الرومنة مصدراً للغموض في المعالجة الآليّة للغات واسترجاع المعلومات، ويمكن تفسيره من خلال ثلاثة أسباب :

أولا: أسباب تاريخية، لأن بعض الدول العربية كان تحت الاستعمار أوتحت الانتداب الفرنسي أو البريطاني لفترات متباينة، وبالتالي، تبقى آثار هذه الفترة في مفرداتهم في النطق التي تأثرت بالنظام الفنولوجي لتلك الدول في نطق الأسماء والألقاب.

وبالتالي، فإن تأثير اللغة الفرنسية ملحوظ في مجال الرومنة المستخدم في بلدان المغرب العربي بدرجات متفاوتة تبعا لخصائص كل دولة. وإن الأمر ذاته بالنسبة لدول الشرق الأوسط فيما يتعلق بالتأثير البريطاني أو الأمريكي في الرومنة.

ثانيا: لأسباب سياسية، حيث لا يوجد معيار موحد أو استراتيجية موحدة في مجال الرومنة بالنسبة للغة العربية. هذا ما أدى كل كاتب أو كاتبة إلى الاعتماد على نطقه للهجات في رومنة وكتابة الأسماء العربية.

المثال الأكثر شهرة هو مثال لورنس العرب الذي قام برومنة اسم مدينة جدة في المملكة العربية السعودية: حيث استخدم 25 مرة (Jeddah) وفي بعض المرات (Jidda)، ومرة واحدة (Jedda)، في نفس الكتاب (1926).

وأخيرا، لأسباب خاصة باللهجات: وجود مجموعة متنوعة من اللهجات الإقليمية والمحلية في العالم العربي، الأمر الذي أدى إلى صعوبة وجود نفس النطق من بلد إلى آخر ومن منطقة إلى أخرى. على سبيل المثال، اسم النبي محمد –صلى الله عليه وسلم– من الأسماء الأكثر شيوعا تم رومنته إلى اللغة الفرنسية وفقا لطرق نطق مختلفة، مما أدى إلى رومنات عربية متعددة وليس رومنة واحدة معيارية:

على سبيل المثال نجد: {*Mohamed, Mouhammad, Muhamed, Mhamed, M'Hamed,* . . . *Muhammad*}. حتى عندما يتم تشكيل الاسم (مُحَمَّدْ)، فإن لديه العديد من الكتابات في النصوص: {Muhamad, Mouhamad, Mohamad, Mehammad, Mehammade}.

أحيانا، يرافق هذا الاختلاف في الرومنة (نظرا لتعدد اللهجات) استخدام أحرف خاصة في مناطق معينة أودول عربية. ومن الأمثلة على ذلك من المدونات، الأسماء التالية التي تمثل أشكال غير تقليدية في الحروف اللاتينية : Mu`ammar, Mabrūk, aṭ Ṭulayḥah, Bū, Yaḥyá, Ḥammūdah, Musṭafá,Ismā`īl, Hâdî .

5. كيفية عمل برنامج الرومنة:

يستند نظام الرومنة من الكتابة العربية إلى الكتابة اللاتينية على آلة الحالات أوالأمثلة المحدودة. حيث قد يؤدي كل زوج مكون من رموز المدخلات وأحد الحالات إلى عدد من الحالات في الخطوة التالية. يتم تحديد العملية وفقا لطبيعة الكلمة المدخلة: تبدأ آلة الحالات المحدودة في حالة ابتدائية محددة وتبدأ في قراءة متسلسلة من رموز أبجديتها. ويستخدم نموذج التشغيل الذاتي دالة الانتقال لمعرفة الحالة التالية التي تؤدي إليها الحالة الحالية والرمز الذي تمت قراءته.

من خلال القراءة، تعطي الوحدة إجابة "نعم" أو "لا"، بمعنى آخر الوحدة تقبل (نعم) أو ترفض (لا) المدخلات: المشكلة وغير مشكلة. بعد ذلك، يتم معالجة المداخلات على النحو التالي: إذا كانت مشكلة يتم إزالة الصوائت قبل الرومنة، إذا كانت غير مشكلة تتم عملية الرومنة مباشرة.

وفي الأخير، يقدم النظام مجموعة من الحالات، والتي تتميز بأنها حالات قبول تتكون من قائمة مفرزة للأسماء العربية المكتوبة بالأحرف اللاتينية.

يتكون قلب نظام الرومنة من قواعد سياقية. وتهدف هذه القواعد للأخذ بعين الاعتبار الطريقة الأكثر دقة للأشكال الملاحظة لحظة الإدخال: هل هي "كنيّة" ؟ اسم يسبقه مقال؟ أو اسم واحد؟

في هذا الصدد، يتم تعريف اسم الشخص في الثقافة العربية (غودير 2006) على أنه مجموعة من الخصائص التي تعرف الشخص باللغة العربية. ويتكون من حيث المبدأ من أربعة عناصر رئيسية :

- الكنية: تتألف عادة من "أبو" (والد . . .)، يليه اسم الطفل أو "أم" (الأم + اسم الطفل في الأسرة). على سبيل المثال: "أبو عمر" (والد عمر)، "أم محمد" (أم محمد)، الخ.

- الاسم: على سبيل المثال، عمر، علي، محمد، خالد، عبد الله، الخ. وقد يشير الاسم في أحيان كثيرة إلى الأصل العرقي أوالمذهبي للشخص: على سبيل المثال، "عمر" هواسم عادة مستعمل بين أهل السنة. "رستم" هواسم نموذجي إيراني . "أرسلان" هواسم تركي. . . الخ

- النسب: كل اسم يسبق ب "ابن" أو"بن" ("بنت" للنساء). فإنه يدل على النسب الدقيق للشخص. يرجع العرب في بعض الأحيان إلى الأصول في إشارة إلى الأجداد لتفادي الخلط بين الناس. مثلا: محمد بن عبد الله بن صالح بن سعيد . . . الخ

- النسبة: (لاحقة في الاسم تدل على أصل الشخص): كانت تشير في القديم إلى قبيلة أوعشيرة، أما اليوم، أصبحت تشير أساسا إلى مكان ولادة الأفراد: الجزائري (المولود في الجزائر) التونسي (المولود في تونس)، المصري (المولود في مصر)، الخ. و"اسم النسبة" يسبق دائما بأداة التعريف [آل] وينتهي باللاحقة [ي]. كما تدل على الإقامة الإقليمية الأولية للأشخاص وجنسيتهم أثناء الولادة.

وفقا لنموذج المدخلات، نقوم بتطبيق قواعد رومنة الجزء الذي لا يشكل الاسم نفسه، ثم نقوم بتطبيق قواعد رومنة الأسماء. كما يتم تطبيق قواعد رومنة الأسماء بدورها على أساس عدد الصوامت من الاسم ذاته، ووفقا لترتيب أولوي محدد.

مثال:

إذا كان الاسم مركبا من كلمة عبد + (ال) + اسم (رحيم)، النظام يشرع على النحو التالي :

- رومنة عبد

- رومنة ال
- ربط 'ال' و'عبد ' مع ربطهم بالاسم (الذي تم رومنته أيضا) عن طريق الشرطة أو عن طريق إدراج فارغة بين الاثنين) Abd Al-Rahim ,**عبد الرحيم**
- ثم توليد كل الأشكال الممكنة للرومنة هذه العناصر الثلاثة :

1) Abd Al-Rahim	13) Abd ar-Rahim
2) Abd Al Rahim	14) Abd ar Rahim
3) Abd al-Rahim	15) Abd ar Rahim
4) Abd al Rahim	16) Abdal Rahim
5) Abd El-Rahim	17) Abdarrahim
6) Abd El Rahim	18) Abdel Rahim
7) Abd el-Rahim	19) Abderrahim
8) Abd el Rahim	20) Abdar-Rahim
9) Abd Ar-Rahim	21) Abdul Rahim
10) Abd Ar Rahim	22) ʻAbd Arrahīm
11)`AbdAr-Rahîm	23) 3abd ara7im
12) `Abd Arrahîm	24) …

يتم إضافة خطوة وسيطة من أجل الحصول على علاج آخر، من أجل مواجهة إحدى المشاكلات الصعبة جدا بالنسبة للنسخ، فهناك بعض الأسماء التي تتغير تماما صوتيا لأسباب دينية أو غيرها: هذا هو الحال على سبيل المثال لاسم "موسى" الذي يترجم بـ "مويز"، ويوسف الذي يترجم بجوزيف، ويعقوب باسم جكوب. وتهدف هذه المرحلة إلى تقديم هذه الرومنات أو الترجمات إلى القائمة النهائية.

وبعد التحصّل على قائمة الأسماء المرومنة، يقوم النظام بنوعين آخرين من المعالجة الآلية:

❖ توحيد قائمة الأسماء بالحروف اللاتينية: تهدف هذه المرحلة تنفيذ بعض المعالجة على الأسماء بالكتابة اللاتينية، مثل إزالة أحرف خاصة (علامات التشكيل وأرقام)، وإضافة الحرف الكبير في بداية الاسم، علما أن هذه الظاهرة غير موجودة في الأسماء المكتوبة بالعربية. نظرية الحرف الكبير تستعمل إلا في حالة استخدامها في قواعد البيانات، ولكن لا يتم إضافتها في حالة محركات البحث، التي لا تأخذ بعين الاعتبار هذه الظاهرة.

❖ تحديد قائمة الأسماء بالحروف اللاتينية: تهدف هذه المرحلة تثمين القواعد التي تم استخدامها لوضع القائمة، بهدف عرض النتائج النهائية من المرجح إلى الأقل احتمالا، أو العكس. ومن أجل تحديد عدد ورود الرومنات المختلفة، قمنا باستخدام محركات البحث المتعددة بالإشارة، في كل مرة، لعدد ورودها: على سبيل المثال، الاسم جمال، يولد ثلاثة رومنات مختلفة ومتواجدة في النصوص وأدى حساب الترددات إلى النتائج التالية:

متوسط عدد الورود لاسم على محركات البحث	رومنة بالحروف اللاتينية
4000000	Djamel
5500000	Jamel
500000	Gamel

يُظهر هذا المثال أن الحرف العربي "ج" مكتوب من حيث التردد: (J)، ثم (Dj)، ثم (G).

6. استخدام الرومنة لربط الكلمات

يتم ربط الكلمات البسيطة والمركبة وفقا للخطوات الأربعة التالية:

- الربط فقط باستخدام قاموس ثنائي اللغة لمحرك البحث بين اللغات.
- الربط بالتعرف على الكلمات من أصل واحد والمتشابهة في الجملة المصدر والجملة الهدف.
- الربط باستخدام نتائج التحليل النحوي والتركيبي للجمل المصدر والجمل الهدف.
- ربط الكلمات المركبة التي يتم ترجمتها كلمة بكلمة باستعمال نتائج علاقات التبعية النحوية للكلمات والجمل في اللغة المصدر واللغة الهدف.

1.6 الربط باستعمال قاموس ثنائي اللغة

تهدف هذه المرحلة إلى استخراج ترجمات الكلمات المحورية في جمل اللغة المصدر باللجوء إلى القاموس الثنائي اللغة وقاعدة البيانات المكونة من جمل اللغة الهدف. يمكن شرح هذا الربط كالتالي:

لكل جملة مصدر في المدونة نقوم بـ:

- تحليل لغوي للجملة المصدر

- تحليل لغوي للجملة الهدف التي تم رصفها أوربطها

لكل كلمة في الجملة المصدر، نقوم بـ:

- استخراج الترجمات الممكنة من القاموس الثنائي
- البحث عن الترجمة في الجملة الهدف
- تسجيل الثنائية "كلمة مصدر/كلمة هدف"

2.6 الربط بالتعرف على الكلمات من أصل واحد والمتشابهة

من أجل التعرف على معادلات جديدة، نأخذ بعين الاعتبار الكلمات المتشابهة في اللغة المصدر واللغة الهدف من حيث الخصائص الفونولوجية والمورفوبوجية والسميائية. ونعتبر كل كلمة متشابهة، الكلمة التي تتسم بالخصائص السالفة الذكر. ويمكن شرح طريقة الرصف هذه على أنها العملية التي تسمح باستخراج الكلمات المتشابهة عن طريق البحث في الخصائص الفونولوجية والمورفوبوجية والسميائية، ثم تسجيل وحفظ الثنائية "كلمة مصدر وكلمة هدف":

في هذه المرحلة، نلجئ إلى رومنة أسماء الأعلام التي تسمح باستنتاج أن اسم العلم باللاتينية "Jackson" هو رومنة الاسم العربي "جاكسون" في الجمل المصدر والهدف واعتبارهما كلمات متشابهة. ومع ذلك، لا تسمح هذه العملية بالتعرف على ثنتيات الكلمات مثل: « blair » et « bleer » التي هي رومنة " بْلِير". ولمواجهة هذه المشكلة، قمنا بتحديد التشابه على مستوى عدد الحروف المتشابهة بدلا من التركيز على المقاطع المتشابهة. الأمر الذي يسمح بتحديد ثنائيات الكلمات التي تم ذكرها من قبل بالإضافة إلى أسماء الأعلام والعبارات الرقمية. المعادلة الخوارزمية التي سمحت بتحديد الكلمات المتشابهة تم تطبيقها بصورة تساعد على اختيار كلمات ذات حجم متقارب وذات حروف متشابهة دون التركيز على ترتيبها في الكلمة الواحدة. تقوم هذه المعادلة على الأسس التالية:

$$\textbf{نسبة الكلمات} = \frac{\text{عدد حروف الكلمات القصيرة}}{\text{عدد حروف الكلمات الطويلة}}$$

$$\textbf{نسبة الكلمات} = \frac{\text{عدد الحروف المتشابهة}}{\text{عدد حروف الكلمة القصيرة}}$$

وإذا كانت نسبة الكلمات بصفة عامة أكبر من 0,8 ونسبة الكلمات المتشابهة تفوق 0,8، نستطيع أن نقول أن الكلمات تشكل ثنائية متطابقة، وبالتالي سيتم تسجيلها على أنها ثنائية اللغة المصدر واللغة الهدف.

3.6 الربط باستخدام نتائج التحليل النحوي والتركيبي للجمل المصدر والجمل الهدف

نلجأ إلى الفئات النحوية التي تحصلنا عليها من خلال التحليل التركيبي والنحوي للنصوص المصدر والهدف من أجل التعرف على ترجمات بعض الكلمات غير الموجودة في القاموس الثنائي اللغة ولكنها محاطة بكلمات مترجمة.

4.6 ربط الكلمات المركبة التي يتم ترجمتها كلمة بكلمة

تهدف هذه الطريقة من الرصف القائمة على ربط الكلمات المركبة التي تترجم كلمة بكلمة إلى إنشاء تطابق تلقائي بين الكلمات المركبة في الجمل الواردة في اللغتين. وتتحقق هذه العملية على مرحلتين:

- استخراج الكلمات المركبة.
- إيجاد التطابق بين بنيات الكلمات المركبة التي تم العثور عليها

تتمثل عملية استخراج الكلمات المركبة في تقسيم الجمل الواردة في اللغتين إلى سلاسل اسمية وفعلية وذلك بتحليل التبعية النحوية من جهة، ومن جهة أخرى باستخراج مصطلحات الجمل المصدر والجمل الهدف باستعمال علاقات التبعية النحوية بين كلمات الجملة الواحدة.

وتهدف استراتجية وضع التطابق بين الكلمات المركبة إلى استعمال مجموعة من القواعد الخاصة بإعادة الصياغة التي يستعملها القاموس الثنائي اللغة مثل القاعدة " ترجمة "أ. ب"= ترجمة" أ". ترجمة "ب" التي تسمح بترجمة الكلمة المركبة «mondialisation néolibérale» (تبديل الصفة بالاسم) كالتالي: ترجمة mondialisation. وترجمة néolibérale= عولمة نيوليبرالية.

وبالتالي تهدف طريقة الرصف هذه الخاصة بالكلمات المركبة التي تترجم كلمة بكلمة إلى البحث لكل مكون للكلمة المركبة في الجملة المصدرعن ترجمة له في قائمة الكلمات المركبة في الجمل الهدف. إلا أن هناك حالات يصعب إيجاد ترجمات لها في القاموس الثنائي.

وإذا تم التعرف على الكلمة المركبة وترجمتها، فيمكن اعتبار الرصف مناسبا. أما إذا لم يتم التعرف على أحد طرفي الكلمة المركبة،، فنحن أمام كلمات مركبة تترجم كلمة بكلمة وكلمات مركبة أخرى لا تترجم كلمة

بكلمة، وفي هذه الحالة، نفترض أن عدد مكونات الكلمة المركبة في اللغتين متطابق.

7. التجربة:

قمنا بالتقييم اليدوي للبرنامج الخاص برصف الكلمات وربطها سواء البسيطة منها أوالمركبة على جزء من مدونة تتكون من نصوص فرنسية وعربية خصعت للربط والرصف بواسطة برنامج ARCADE II .

هذا، ويلخص الجدول التالي النتائج التي تحصلنا عليها فيما يتعلق بالدقة والتذكير في استعمال الرومنة العربية أولا ودورها في تحديد الكلمات المتشابهة لرصفها وربطها.

	دقة	تذكير	F- القياس
تحديد المتشابهات باستعمال النسخ بالعربية	0. 88	0. 85	0. 86
تحديد المتشابهات بدون استعمال النسخ بالعربية	0. 85	0. 80	0. 82

نتائج تقييم ربط الكلمات ورصفها

الخاتمة:

قدمنا في هذا المقال شرحا لنظام رومنة الأسماء العربية من اللغة العربية إلى اللغات اللاتينية. وقد تم استعمال هذا النظام في عملية ربط ورصف كلمات مستقاة من مدونة تتكون من نصوص فرنسية وعربية، حيث تمر عملية الرصف بمرحلتين هما:

- ربط الكلمات البسيطة باستعمال القاموس الثنائي اللغة وبالاعتماد على الخصائص اللغوية للكلمات والمتشابهات.
- ربط الكلمات المركبة باللجوء إلى علاقات التبعية النحوية.

تعطي هذه العملية نتائج جد مرضية عندما يتم استعمال الرومنة العربية لربط أسماء الأعلام في الجملة المصدر والجملة الهدف. وفي هذا الصدد، تهدف أبحاثنا في المستقبل إلى توسيع نطاق التقييم من أجل تأكيد مقارباتنا الخاصة بربط الكلمات من جهة، ومن أجل إدراج قائمة كلمات ثنائية اللغة يتم وضعها من خلال نظام ترجمة احصائي لتحسين نوعية الترجمة، من جهة أخرى.

الهوامش:

1-http://www.praxiling.fr/corpuspluriels/wp-content/uploads/2009/11/IAL_01.pdf

2-http://www.praxiling.fr/corpuspluriels/wp-content/uploads/2009/11/MIDL_041.pdf

المراجع:

ABDULJALEEL, N. ,& LARKEY, L. (2003). Statistical transliteration for English-Arabic Cross Language Information Retrieval. In *Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management*. (pp. 139-146). New Orleans, LA, New York.

ALGHAMDI, M. (2005). Alghorithms for Romanizing Arabic names. *Journal of King Saud University: Computer Sciences and Information,* 17, 1–27.

ALSALMA, A. , & ALGHAMDI, M. , &ALHUQAYL, K. , & ALSUBAI, S. (2007). Romanization System for Arabic Names. In *The First International Symposium on Computer and Arabic Language* (pp. 214-227), Riyadh.

BARBU A. M. (2004). Simple linguistic methods for improving a word alignment. Actes de la 7th International Conference on the Statistical Analysis of Textual.

BLANK I. (2000). Parallel Text Processing: Terminology extraction from parallel technical texts . Dordrecht: Kluwer.

BROWN P. F. ,& PIETRA S. A. D. ,& PIETRA V. J. D. , &MERCER R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics,* 19 (3).

DAILLE B. ,& GAUSSIER E. , &LANGE J. -M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *Actes de la 15th International Conference on Computational Linguistics (COLING'94).*

DEBILI F. ,& ZRIBI A. (1996). Les dépendances syntaxiques au service de l'appariement des mots. *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle.*

DEMPSTER A. P. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological),* 39, 1.

DENERO J. , & KLEIN D. (2008). The Complexity of Phrase Alignment Problems. *Proceedings of ACL* .

DICHY, J. (2009). La polyglossie de l'arabe illustrée par deux corpus. In M. Bozdemir et L. –J. Calvet (EDS)*, Politiques linguistiques en méditerranée*(82-102). Paris: Honoré Champion.

GAUSSIER E. ,& LANGE J. M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *TAL 36*.

GUIDERE, M. (2004). Le traitement de la parole et la détection des dialectes arabes. In *Langues stratégiques et défense nationale, Publications du CREC* (pp. 53-75). Saint-Cyr.

KNIGHT, K. ,& GRAEHL, J . (1997). Machine transliteration. In *Journal version Computational Linguistics, 24(4),* 599–612.

MACCARTNEY B. ,& GALLEY M. , & MANNING C. D. (2008). A Phrase-Based Alignment Model for Natural Language Inference. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

QUNAIIR, H. (2001). Romanizing Arabic names. In *Journal er-Riyadh, article n°12314 (en arabe)*

OCH F. J. (2003). GIZA++: Training of statistical translation models. http://www-i6. informatik. rwth-aachen. de/Colleagues/och/software/GIZA++. html.

OZDOWSKA S. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. *Actes de la 11ème conférence TALN-RECITAL.*

SMADJA F. ,& MCKEOWN K. ,&HATZIVASSILOGLOU V. (1996). Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics,* 22 (1).

STALLS, B. ,& KNIGHT, K . (1998). Translating names and technical terms in Arabic Text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages,* Montreal, Québec.