مقاربة جديدة للتشكيل الآلي باستخدام برنامج الخليل للتحليل الصرفي[*]

محمد ولد عبد الله ولد بباه، عز الدين مزروعي[**]،
عبد الحق لخواجة، عبد الوافي مزيان، شنوفي أمين

ملخص :

نقدّم في هذا البحث مقاربة جديدة للتشكيل الآلي لنصوص اللغة العربية الفصحى. وقد أطلقنا على برنامج التشكيل الآلي المنجز الذي يعتمد هذه المقاربة اسم **برنامج الخليل للتشكيل الآلي**. وتتكوّن عملية المعالجة التي تتم في هذا البرنامج من مستويين. يعتمد المستوى الأول على مخرجات برنامج عربيّ مفتوح المصدر هو برنامج الخليل للتحليل الصرفيّ، وتحتوي هذه المخرجات على التشكيلات الممكنة للكلمة خارج سياقها من النص. بينما يعتمد المستوى الثاني على منهج إحصائيّ باستخدام نماذج ماركوف الخفية (Hidden Markov Models)

(*) تم تقديم هذا البحث في صيغته الأولى مرفقا ببرنامج التشكيل الآلي إلى جائزة المنظمة العربية للتربية والثقافة والعلوم للإبداع والابتكار التقني للباحثين الشبان بالوطن العربي سنة 2012 ونال عليه الباحث محمد ولد عبد الله ولد بباه المركز الثاني في هذه المسابقة.

(**) فريق المعالجة الآلية للغة العربية، كلية العلوم - جامعة محمد الأول – وجدة/المغرب.

حيث قمنا باقتراح تصنيف جديد للكلمات العربية لتمثّل هذه الأصناف (classes) مشاهدات النموذج الماركوفي (observations) في حين تمثّل تسلسلات علامات التشكيل الممكنة الحالات المخفية للنموذج (hidden states).

ولضبط نتائج المستوى الأول من المعالجة عمدنا إلى إنشاء معجم للكلمات العربية شائعة التكرار، قُمنا بدمجه في إطار قواعد معطيات المحلل الصرفي. وقد تم استقاء هذا المعجم من ثماني مدونات عربية متاحة على الشابكة (الإنترنت) وتغطي مختلف المجالات. كما قُمنا بعملية تدريب البرنامج وتقييمه على نصوص متنوعة تشمل نصوصًا معاصرة من مدونة نيملار (NEMLAR Written Arabic Corpus) ونصوصًا تراثية من مدونة "تشكيلة".

1. مقدمة

يتميز نظام الكتابة العربية، في أغلب النصوص، بغياب علامات التشكيل، وهي: الصوائت القصيرة (short vowels)؛ أي الفتحة والضمة والكسرة، بالإضافة إلى علامات التنوين والشدّة والسكون. ويؤدي غياب هذه العلامات إلى زيادة معتبرة في غموض النص العربيّ تصل بحسب (Debili and Achour 1998) إلى التباس أكثر من 90% من النص. وعلى الرغم من كون القارئ المتعلّم يتمكن بسهولة من استرجاع علامات التشكيل الغائبة عن النص اعتمادًا على سياق الكلمات وعلى مدى معرفته بالصرف والنحو العربيين، إلا أن النص غير المشكول يمثّل عقبة بالنسبة لمتعلمي العربية من غير الناطقين بها والأشخاص ذوي الصعوبات في التعلم، كما يمثّل مصدرًا أساسيًا للغموض بالنسبة للحاسوب، وخصوصًا في حالة تطبيقات من قبيل تحويل النص المكتوب إلى كلام (Text to Speech) (Zitouni et al. 2006). فخلافًا للّغات الأوروبية التي يَسهل فيها نوعًا ما إيجاد التقابل بين المقاطع المكتوبة والفونيمات المنطوقة، فإن النص العربيّ غير المشكول لا محيد عن استرجاع علامات تشكيله لهذا الغرض (Vergyri and Kirchoff 2004). وتشير بعض الأبحاث من جهة

أخرى إلى أهمية استخدام النصوص المشكولة للرفع من كفاءة نظم التعرّف على الكلام (Messaoudi et al. 2004). كما تبرز أهمية التشكيل الآليّ في تطبيقات أخرى من قبيل المدونات الشجرية (Treebanks) والمحللات النحوية (Mammouri et al. 2006).

ونظرًا لأهمية استرجاع علامات التشكيل فقد ظهرت محاولات من قبل جهات عدة، طيلة العقود الماضية، من أجل تطوير مشكّل آليّ ذي كفاءة مقبولة. ويمكن أن نقسّم هذه المحاولات، على غرار أغلب تطبيقات المعالجة الآلية للّغة، إلى نوعين : أولهما محاولات أنجزتها شركات تجارية بهدف إنشاء مشكّل آلي مستقل أو مشكّل آلي يمثل جزءًا من تطبيق آخر، كالناطق الآلي أو المدقق الإملائي. ومن أبرز هذه المحاولات يمكن أن نذكر شركة "RDI" ومشكّلها الآلي "ArabDiac"، المشكّل الآلي لشركة صخر، والمشكّل الآلي لشركة "CIMOS". كما أطلقت شركة غوغل منذ سنتين خدمة "Tashkeel"، وهي خدمة مجانية للتشكيل الآلي للنصوص العربية بالنسبة لمتصفحي الشابكة (الإنترنت) غير أن هذه الخدمة لم تكتب لها الاستمرارية لأسباب نجهلها. وبالرغم من الأهمية البالغة لهذه المحاولات، إلا أن طبيعتها التجارية الاحتكارية تحول دون الحصول على تفاصيل كافية عن آلية عملها والمصادر اللغوية التي تستند عليها، فضلا عن شيفرتها المصدرية مما يجعل من تحسينها ودمجها في تطبيقات أخرى أمرًا بعيد المنال.

أما النوع الثاني من هذه المحاولات فيتمثل في جهود الباحثين ضمن مشاريع مراكز البحث الأكاديمية، وقد أسفرت هذه الجهود طيلة العقود الماضية عن ظهور محاولات متعددة في هذا المجال. ويساعد توفر أوراق بحثية منشورة حول هذه الأعمال في التعرف على المقاربات المنتهجة فيها، وهي مقاربات إحصائية في أغلب الأحيان، كما يساعد على أخذ فكرة عن كفاءة البرمجيات المنجزة فيها تحت ظروف الاختبار التي أخضعها لها مطوروها.

وسنتناول بالتفصيل في الفقرة الموالية من هذا البحث أهم الأعمال السابقة المنجزة في مجال التشكيل الآليّ للّغة العربية مع التركيز على المقاربات المعتمدة في هذه الأعمال. وفي الفقرة الثالثة نقدّم برنامج التشكيل الآلي الذي أنجزناه، عارضين مختلف مراحل المعالجة التي يقوم بها والنموذج الماركوفي الذي يعتمد عليه. ثم بعد ذلك نعرض تباعًا لمرحلة تدريب البرنامج ونتائج تقييمه والخلاصات التي خرجنا بها من هذا العمل.

2. الأعمال السابقة

يمكن تقسيم المقاربات المنتهجة من خلال الأعمال السابقة في التشكيل الآلي للنصوص العربية إلى ثلاثة أقسام:

2.أ. المقاربات اللغوية (rule-based approaches):

وفي هذا الإطار ظهرت بعض الأعمال التي تهدف إلى برمجة القواعد اللغوية الصوتية والصرفية والنحوية والإملائية بغية تشكيل الكلمات العربية. ومن بين بواكير هذه الأعمال يمكن أن نذكر الطريقة الواردة في (El-Sadany and Hashish 1988) التي تعتمد على استخدام القواعد الصرفية من أجل تشكيل نصف آلي (semi-automatic) للأفعال العربية. وفي (Debili and Achour 1998) يقدّم المؤلفان دراسة عن التشكيل الآليّ للنصوص العربية تناولاها من جانب غموض الكلمات المكتوبة ومدى إسهام كل من التحليل المعجميّ والتحليل الصرفيّ والعنونة النَحْوية في كشف الغموض، وبالتالي في تشكيل كلمات النصوص العربية.

غير أن ارتفاع نسبة الغموض وكثرة القواعد الصرفية والنحوية وتشابكها وعدم توفر مُحلّل نَحْوي فعّال، فضلا على أن بعض حالات الالتباس تتطلب تحليلا دلاليا، كل ذلك أسبابٌ تعوق إيجاد مُشَكِّل آليّ للنصوص العربية لا يعتمد سوى على القواعد اللغوية.

2.ب. المقاربات الإحصائية (statistical approaches)

بالنظر إلى النجاح الكبير الذي حقّقته الطرق الإحصائية في مجالات عدة من المعالجة الآلية للّغة، كالتعرف الآلي على الكلام والترجمة الآلية واسترجاع المعلومات وغيرها (Manning and Schütze 1999)، فقد انصبت أغلب الجهود في التشكيل الآلي للغة العربية على استعمال هذه الطرق. وقد تنوعت الأساليب المتبعة في هذا الباب، حيث اهتم بعض الباحثين بنماذج إحصائية تعالج التشكيل على مستوى الحروف، وركز آخرون على التشكيل على مستوى الكلمات، بينما اتبع فريق ثالث أسلوب المزاوجة بين الطريقتين السابقتين.

فقد سجل إمام وفيشر (Emam and Fischer, 2005) براءة اختراع مشكِّل آلي للغة العربية يستند في فكرته على طريقة الترجمة الآلية المُعتمدة على الأمثلة (exemple-based machine translation)، حيث يتم البحث ضمن قواعد معطيات الأمثلة بطريقة تراتبية (hierarchical) عن الجملة المراد تشكيلها ثم عن أجزاء الجملة وصولا إلى الكلمات لاختيار أكثر التشكيلات رجحانًا. كما يتم تطبيق نموذج إحصائيّ (n-gram model) على مستوى الحروف لتشكيل الكلمات التي لم يتم العثور عليها خلال عملية البحث التراتبية السابقة.

غير بعيد عن هذه الفكرة نجد في (Schlippe et al, 2008) مقاربة تعتمد على تقنية الترجمة الآلية الإحصائية (statistical machine translation) انطلاقًا من مدونة متوازية (parallel corpus) مكونة من النصوص المشكولة ومثيلتها غير المشكولة.

كما قدّم (Gal 2002) مقاربة ماركوفية لتشكيل نصوص اللغتين العربية والعبرية. وقد استخدم المؤلف في هذا العمل نصوص القرآن الكريم بالنسبة للعربية ونصوص الكتاب المقدس بالنسبة للغة العبرية. وتُمثّل الكلمات غير المشكولة مشاهدات النموذج المقترح في عمل "Gal" بينما تُمثّل الكلمات المشكولة

الحالات الخفية للنموذج. غير أن عملية التشكيل الآليّ للّغة العربية في هذا العمل تقتصر على الصوائت القصيرة فقط دون غيرها من علامات التشكيل.

وفي إطار المقاربات الماركوفية للتشكيل الآليّ نُسجّل أيضا أعمال كلٍّ من (Deltour 2003) و(Elshafei et al. 2006)، حيث استعرضت الباحثة في (Deltour 2003) جملة من الطرق الإحصائية للتشكيل الآليّ على مستوى الحروف والكلمات خَلُصت فيها إلى أن أفضل نتائج التشكيل في بحثها كانت باستخدام نماذج ماركوف الخفية. كما قدّم الباحثون في (Elshafei et al. 2006) مقاربة ماركوفية للتشكيل الآليّ تتميز عن مقاربة (Gal 2002) في كونها تتعامل مع جميع علامات التشكيل، كما تمت عملية التدريب والتقييم فيها على نصوص متنوعة لا على النص القرآني فقط.

وقدّم (الغامدي وآخرون 2006) برنامج التشكيل الآليّ المطوّر في مدينة الملك عبد العزيز للعلوم والتقنية. ويعتمد هذا البرنامج على التسلسلات الرباعية (quad-grams) للحروف حيث يقوم بتشكيل كل حرف من النص استنادًا إلى أعلى احتمالية لتشكيله ضمن التسلسلات الرباعية التي يظهر فيها.

2.ج. المقاربات الهجينة (hybrid approaches)

ونعني بها المقاربات التي تُزاوج بين القواعد اللغوية والنماذج الإحصائية بغية استغلال نقاط القوة في كلتا الطريقتين. ومن أهم الأعمال التي أنجزت في هذا الإطار نذكر المشكّل الآلي "ArabiDaic" المطوّر من قبل شركة "RDI" (عطية 2007) حيث يعتمد هذا البرنامج على المحلل الصرفي "ArabMorpho" وعلى واسم لتحديد أجزاء الكلام "ArabTagger"، بالإضافة إلى نموذج إحصائي لاختيار التشكيل الأنسب للكلمات باستخدام خوارزمية A*.

كما قدّم المؤلفان في (Nelken and Shieber 2005) طريقة تعتمد على تقنية الآلات محدودة الحالات (finite state automata) لتشكيل النصوص العربية. وتُزاوج هذه الطريقة بين نماذج إحصائية ثلاثية على مستوى الكلمات

(word tri-gram) ورباعية على مستوى الحروف (quad-gram letter) وبين نموذج صرفيّ مُبسّط يتعرف على سوابق ولواحق الكلمات.

وفي (Zitouni et al, 2006) عَرَض المؤلفون مشكّلا آليا للنصوص العربية، وذلك باستخدام مصنّف إحصائي يعتمد طريقة الأنتروبيا القصوى (Maximum Entropy Classifier) (Berger et al. 1996). حيث تُستخدم الخصائص الصرفية ووسوم أجزاء الكلام لاستنتاج التصنيف الأرجح للكلمات.

وفي (Vergyri and Kirchoff 2004) يتم استرجاع علامات التشكيل بمزاوجة التحليل الصرفيّ والمعلومات السياقية مع نموذج عنونة إحصائية. إذ يذهب المؤلفان إلى اعتبار التشكيل مسألة عَنونة غير مسيرة (Unsupervised Tagging) يجري خلالها اختيار أرجح الحلول التي يعطيها محلل باكوالتر الصرفي (Buckwalter 2004). وغير بعيد عن هذه الفكرة يقدّم المؤلفان (Habash and Rambow, 2007) طريقة للتشكيل الآلي تعتمد من ناحية على محلّل باكوالتر الصرفيّ وعلى نموذج إحصائيّ لاختيار أرجح الحلول ضمن مخرجاته.

كما سبق لنا أن قدمنا (بباه وآخرون، 2011) مقاربة صرفية إحصائية للتشكيل الآليّ تعتمد على نموذج ماركوفي لتمثيل تسلسل أوزان الكلمات العربية. ولا يفوتنا أن نشير هنا إلى أنها تشترك مع المقاربة المقدمة في هذا البحث في كونهما تعتمدان معا على برنامج الخليل للتحليل الصرفي، بينما تختلفان في نقاط عدّة، أهمها: مشاهدات النموذج الماركوفي وحالاته الخفية (Observations and Hidden States). وتختلف أيضًا مدونتا التدريب المستخدمتان فيهما، كما أننا أجرينا في عملنا الحالي تعديلات على برنامج الخليل للتحليل الصرفيّ بإضافة قاعدة معطيات معجمية تتضمن الكلمات الأكثر تكرارًا بُغية ضبطها وتسريع مرحلة المعالجة الصرفية.

3. برنامج التشكيل الآلي المُنجز

نعرض في هذه الفقرة لبرنامج التشكيل الآلي الذي قمنا بإنجازه، وقبل أن نتناول الجوانب المتعلقة بآلية عمل البرنامج ونتائج تقييمه، لا بأس أن نشير بإيجاز إلى دواعي تسميته **"برنامج الخليل للتشكيل الآلي "**.

3.أ. دواعي التسمية

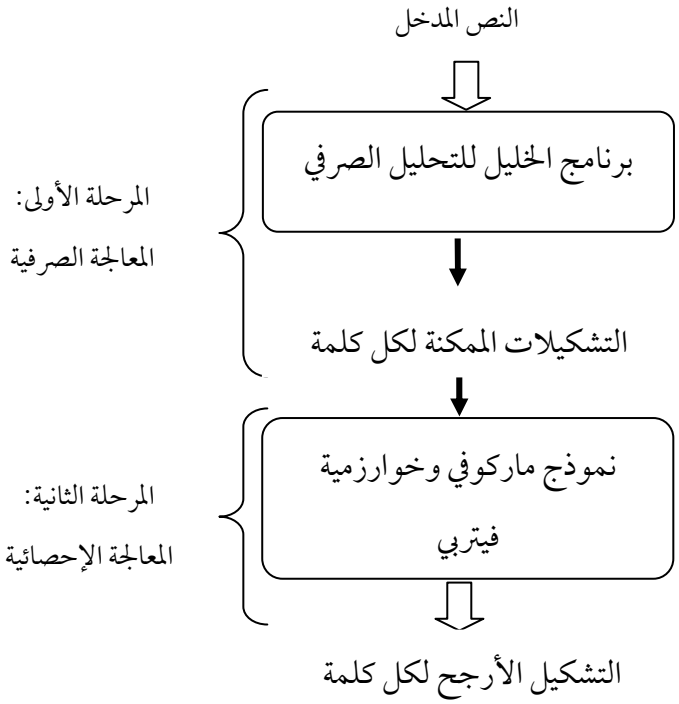
وقع اختيارنا لإطلاق هذه التسمية على برنامج التشكيل الآلي لسببين: أولهما، للبرنامج علاقة وثيقة ببرنامج الخليل للتحليل الصرفي[1]. حيث يعتمد المشكّل الآلي على مُخرجات هذا البرنامج من ناحية وعلى نموذج ماركوفي لاختيار أكثر التشكيلات رجحانًا من ناحية أخرى. أما السبب الثاني فلا يقلّ وجاهة عن سابقه، إذ إن الفضل في ابتكار علامات التشكيل المعروفة حاليًا يرجع إلى الخليل بن أحمد الفراهيدي (ت 170 ه)، وهي الطريقة التي أثبتت فعاليتها وعمليتها عبر الزمن حتى حصل الاتفاق على كتابة المصاحف بها بدلا من طريقة النقط المنسوبة إلى أبي الأسود الدؤلي (ت 69 ه)[2].

3.ب. مراحل التشكيل

تتمّ عملية التشكيل الآليّ في البرنامج المُنجز عبر مرحلتين رئيسيتين كما هو مبين في الشكل 1.

(1) www.sourceforge.net/projects/alkhalil

(2) يذكر أبو عمرو الداني (ت 444 ه) في كتاب المحكم في نقط المصاحف ما نصه: "وقال أبو الحسن بن كيسان قال محمد بن يزيد الشكل الذي في الكتب من عمل الخليل وهو مأخوذ من صور الحروف فالضمة واو صغيرة الصورة في أعلى الحرف لئلا تلتبس بالواو المكتوبة والكسرة ياء تحت الحرف والفتحة ألف مبطوحة فوق الحرف.. إلخ".



شكل 1: مراحل المعالجة في برنامج الخليل للتشكيل الآلي

وفي ما يلي نعرض لهذه المراحل بالتفصيل:

1. مرحلة المعالجة الصرفية

وتتمّ المعالجة في هذه المرحلة باستخدام برنامج الخليل للتحليل الصرفي، وتَهدف إلى استخراج جميع التشكيلات الممكنة لكل كلمة من كلمات النص المُدخل. ولمزيد من إيضاح هذه المرحلة نُقدّم نبذة موجزة عن برنامج الخليل للتحليل الصرفيّ وآلية عمله قبل أن نتناول التفاصيل المتعلقة بقاعدة المعطيات المعجمية التي أضفناها إليه.

نبذة عن برنامج الخليل للتحليل الصرفي

يعتبر برنامج الخليل للتحليل الصرفي (مزروعي وآخرون، 2011) أحد أهم المحللات الصرفية العربية مفتوحة المصدر. وقد تم إنجازه في إطار مشروع شراكة بين كل من جامعة محمد الأول والمنظمة العربية للتربية والثقافة والعلوم ومدينة الملك عبد العزيز للعلوم والتقنية. كما روعيت عند إنجازه التوصيات والمعايير اللغوية والتقنية التي خرج بها المشاركون في اجتماع خبراء المحللات الصرفية العربية (إبريل 2009) بدمشق (حماده والبواب 2009).

ويقوم البرنامج بإعطاء تحليل صرفيّ للكلمات العربية وذلك بتحديد:

- حالات التشكيل الممكنة للكلمة.
- الزوائد التي تلحق بها (السوابق واللواحق).
- نوع الكلمة (اسم أو فعل أو أداة).

o يشمل نوع الكلمة في حالة الاسم: الأسماء الجامدة، أسماء الأعلام، المصادر (الأصلي، الميمي، المرة والهيئة)، المشتقات (اسم الفاعل، اسم المفعول، اسم الآلة ....)

o وفي حالة الفعل: الماضي مبنيًّا للمعلوم وللمجهول، المضارع مبنيًّا للمعلوم وللمجهول (مرفوعًا ومنصوبًا ومجزومًا ومؤكّدًا)، والأمر والأمر المؤكّد.

- الوزن (في حالة الأسماء والأفعال).
- الجذع.
- الجذر (في حالة الأسماء والأفعال).
- الحالة الإعرابية (في حالة الأسماء والأفعال)، وتشمل: الرفع والنصب والجر، الإفراد والتثنية والجمع، التذكير والتأنيث، التعريف والتنكير، التعدي واللزوم.

وتتلخص مراحل المعالجة الصرفية لكل كلمة خارج سياقها من النص في برنامج الخليل في ما يلي:

أولا : مرحلة التقطيع: وتهدف إلى تحديد السوابق واللواحق المحتملة للكلمة، وذلك من خلال عملية تقطيع (segmentation) يجري خلالها استخدام قواعد معطيات خاصة بالسوابق واللواحق. ومن المعروف أن هذه العملية تؤدي غالبًا إلى أكثر من تقطيع نظرًا للطبيعة الإلصاقية للغة العربية.

ثانيًا : تحليل أسماء الأعلام: ويتمّ خلالها معالجة كل تقطيع من التقطيعات المتحصّل عليها في المرحلة السابقة للتحقق من إمكانية أن تكون الكلمة المدخلة اسمًا علمًا. وفي هذه المرحلة يستعين البرنامج بلائحة من أسماء الأعلام تضم 6000 اسم علم.

ثالثًا : تحليل الأدوات: وفيها يُقارن البرنامج الجذوع المحتملة للكلمة المدخلة بلائحة الأدوات. وتضمّ هذه اللائحة حروف الجر والنصب والجزم والعطف والنداء والاستثناء وأدوات الشرط والاستفهام والنواسخ والضمائر المنفصلة والظروف وأسماء الإشارة والأسماء الموصولة.

رابعًا: معالجة الأسماء: وفيها يتحقق البرنامج من إمكانية أن تكون الكلمة المدخلة اسمًا. وتتم هذه المرحلة عبر الخطوات التالية:

- تحديد الأوزان المُحتمَلة للكلمة، وذلك بمقارنتها بالأوزان الموجودة في قاعدة البيانات الخاصة بالأوزان الاسمية.
- استخلاص الجذور المُحتمَلة للكلمة والتحقق من وجودها في لائحة الجذور العربية.
- التحقق من توافق الجذور والأوزان المُحتمَلة، وذلك بالرجوع إلى قاعدة معطيات تضم الجذور العربية مرفقة بأوزان الأسماء المشتقة منها.

خامسا : معالجة الأفعال: وهي مرحلة مماثلة للمرحلة السابقة حيث تجري عبر الخطوات التالية:

- تحديد الأوزان المُحتمَلة للكلمة، وذلك باستخدام قاعدة البيانات الخاصة بالأوزان الفعلية.
- استخلاص الجذور المُحتمَلة للكلمة والتحقق من صحّتها.
- التحقق من توافق الأوزان والجذور المُستخلَصَة.

هذه بإيجاز أهم مراحل المعالجة في النسخة الرسمية من برنامج الخليل للتحليل الصرفي وهي النسخة المتاحة مجانا على موقع "sourceforge"[3]. غير أن عملية دمج المحلل الصرفيّ في إطار برنامج للتشكيل الآلي تطلبت منا إجراء تعديلات تمثّلت أساسًا في إضافة قاعدة معطيات على شكل معجم يتضمن الكلمات العربية الأكثر تكرارًا في المدوّنات العربية المتاحة، كما قُمنا بتعديل آلية إخراج نتائج التحليل الصرفيّ للاحتفاظ بالتشكيلات الممكنة للكلمة دون غيرها من المعلومات الأخرى، كالنوع والوزن والجذر.. إلخ. وفي ما يلي نستعرض تفاصيل هذا المعجم.

معجم الكلمات الأكثر تكرارًا:

تمّ إنشاء هذا المعجم بُغية تسريع عملية المعالجة الصرفية من ناحية، ومن أجل ضبط نوعية تشكيل الكلمات التي يَكثُر استخدامها في مختلف النصوص من ناحية أخرى. ولهذا الغرض فقد اتبعنا الخطوات التالية:

- تجميع مدوّنة عربية واسعة ومتنوعة تضم أزيد من 250 مليون كلمة، وذلك انطلاقا من المدونات العربية المتاحة على الشابكة (الإنترنت). وقد بلغ عدد المدونات التي استخرجنا منها معجم الكلمات الأكثر تكرارًا ثماني مدونات، هي:

(3) www.sourceforge.net/projects/alkhalil

- مدوّنة الباحث أحمد عبد العالي[4]، وهي مدونة غير مشكولة يبلغ حجمها 147 مليون كلمة (Abdelali et al, 2005). وتتكون هذه المدونة من نصوص صحفية متنوعة تم جمعها من 28 موقعًا صحفيًا بشكل يغطي أغلب البلدان العربية.
- مدوّنة "تشكيلة"[5]، وهي مدونة مشكولة تزيد على 60 مليون كلمة، وتتكون من نصوص تراثية قام بجمعها الباحث طه زروقي انطلاقًا من الكتب الموجودة بالمكتبة الشاملة[6].
- المدوّنة العربية مفتوحة المصدر[7] (OSAC: Open Source Arabic Corpora)، وهي مدونة غير مشكولة جمعها الباحثان معتز خالد سعد ووسام عاشور (Saad and Ashour, 2010) بُغية استخدامها في تطبيقات التنقيب في النصوص (Text Mining). وتحتوي هذه المدونة على قرابة 20 مليون كلمة جُمعت من مواقع إخبارية عربية محلية وعالمية متنوعة.
- مدوّنة الباحثة لطيفة السليطي[8]، وهي مدونة غير مشكولة تضمّ نصوصًا معاصرة وتشمل مجالات مختلفة (سياسية، اقتصادية، دينية، رياضية.. إلخ). وتضم هذه المدونة أزيد من نصف مليون كلمة.
- مدوّنة "Khaleej-2004"[9] وهي مدونة غير مشكولة جمعها الباحث مراد عباس انطلاقًا من موقع اليومية البحرينية أخبار الخليج، وذلك لأجل استخدامها في تطبيقات التصنيف الآلي للوثائق (Abbas and Smaili, 2004). وتضم هذه المدونة قرابة 2,8 مليون كلمة.

(4) http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus/Data/
(5) http://sourceforge.net/projects/tashkeela/
(6) www.shamela.ws
(7) https://sites.google.com/site/motazsite/Home/osac
(8) http://www.comp.leeds.ac.uk/eric/latifa/research.htm
(9) http://sourceforge.net/projects/arabiccorpus/files/

- مدوّنة قرارات الجمعية العامة للأمم المتحدة[10] (UN Parallel Corpora)، وهي مدوّنة متوازية تضم 2100 قرار من قرارات الجمعية العامة للأمم المتحدة مكتوبة باللغات الست الرسمية للأمم المتحدة. وقد اقتصرنا بطبيعة الحال على محتواها العربيّ المكون من 2,5 مليون كلمة، علمًا أن هذه المدونة قد جُمعت في الأصل لغرض استخدامها في تطبيقات الترجمة الآلية الإحصائية (Rafalovitch and Dale 2009).
- مدوّنة RDI[11] وهي مدونة مشكولة تتكوّن أساسًا من كتب تراثية، بالإضافة إلى نسبة قليلة من الكتابات المعاصرة. وقد جُمعت هذه المدوّنة لغرض الاستعمال في مجال التشكيل الآليّ. وتحتوي على 20 مليون كلمة.
- مدوّنة نيملار (NEMLAR Arabic Written Corpus)[12]، وهي مدونة مشكولة ومعنونة صرفيًا ونحويًا، شارك في إعدادها باحثون من مراكز بحثية مختلفة (Attiya et al, 2005) وذلك في إطار مشروع «NEMLAR». نشير إلى أنها مدوّنة غير مجانية تقوم بتسويقها الجمعية الأوروبية للموارد اللغوية (ELRA)، وقد اقتصرنا في هذا البحث على محتواها المشكول الذي يضمّ نصوصًا من مجالات مختلفة تحتوي على قرابة نصف مليون كلمة.

• بعد مرحلة التجميع قمنا بتهيئة المدونة، وذلك بتوحيد ترميزها وطريقة تخزينها، حيث قمنا بحفظها في ملفات نصية (txt.) بترميز Cp1256. كما عملنا على تنقيتها من الرموز والكلمات المكتوبة بحروف غير عربية.

• بعد ذلك قمنا بحساب تكرار كلمات كل مدونة على حدة وترتيبها ترتيبا تنازليًّا. وقد عمدنا في هذه الخطوة إلى عزل كل مدوّنة على حدة، نظرًا لتفاوت أحجام هذه المدونات من جهة، ولكون الكلمات الشائعة تختلف حسب طبيعة

(10) http://www.uncorpora.org/
(11) http://www.rdi-eg.com/RDI/TrainingData/
(12) http://catalog.elra.info/product_info.php?products_id=873

المدونة من جهة أخرى. ففي حين تَكثُر كلمات من قبيل "حدثنا" "قال" و"صلى" في المدونات التراثية، فإننا لا نجد لها أثرًا في مدونة قرارات الجمعية العامة للأمم المتحدة على سبيل المثال.

- وتمثّلت الخطوة الموالية في استخلاص اللائحة الأولية لمداخل معجم الكلمات الأكثر تكرارًا، وذلك بعد دمج اللوائح المستخلصة من كل مدونة وحذف علامات التشكيل التي تظهر في بعضها، وخصوصا لوائح المدونات التراثية، ثم حذف المكرر من هذه المداخل. وقد بلغت هذه اللائحة 16200 كلمة.

- بعد ذلك قمنا بتحديد المقابلات المشكولة لمداخل المعجم وقد تطلبت هذه الخطوة تقسيم المداخل إلى مجموعتين:

  - مجموعة الكلمات التي وجدنا لها مقابلات مشكولة في نصوص المدونات المشكولة (تشكيلة: RDI وNEMLAR). وفي هذه الحالة أرفقناها بالمقابلات المشكولة التي عثرنا عليها.

  - المجموعة الثانية وكانت في حدود 1200 كلمة، قمنا بتحليلها صرفيًّا باستخدام محلّل باكوالتر الصرفيّ لكونها تحتوي كثيرًا من الألفاظ الدخيلة شائعة الاستخدام في النصوص الحديثة وكذلك أسماء الأعلام الأجنبية. ثم أتبعنا ذلك بتدقيق يدوي للكلمات التي حلّلها والكلمات التي لم يتمكن من تحليلها.

بعد استكمال هذه الخطوات تحصل لدينا معجم من 16200 كلمة من الكلمات الأكثر تكرارًا في المدونات العربية المتاحة. وقد قمنا بدمج هذا المعجم في إطار عملية التحليل الصرفيّ لبرنامج الخليل بحيث يتعرّف مباشرة على التشكيلات الممكنة للكلمات إن كانت مما يتضمنه معجم الكلمات الأكثر تكرارًا مما يُسرّع من عملية المعالجة.

2. مرحلة المعالجة الإحصائية

بعد أن يقوم البرنامج بمعالجة صرفية لكلمات النص الـمُدخل، وبعد أن يحتفظ بالتشكيلات المحتملة لكلّ كلمة، فإنه يمر إلى المرحلة الثانية من مراحل التشكيل. وتتمثّل هذه المرحلة، كما أسلفنا، في معالجة إحصائية مبنية على نموذج ماركوفي يتمّ عبره اختيار التشكيل الأرجح لكلّ كلمة باستخدام خوارزمية فيتربي (Neuhoff 1975). وفي ما يلي نعرض لهذا النموذج بشيء من التفصيل:

لتكن المجموعة $O = \{o_1, \ldots, o_M\}$ مجموعة منتهية من "المشاهدات" (Observations)

ولتكن $S = \{s1, \ldots, sN\}$ مجموعة منتهية من "الحالات الخفية" (Hidden States)

تعريف:

نُعرّف نموذج ماركوف الخفي من الرتبة الأولى بأنه كلّ زوج من السلاسل $(X_t, Y_t)_{t\geq 1}$ بحيث تكون:

- $(X_t)_t$: عبارة عن سلسلة ماركوف متجانسة (homogeneous Markov chain) تأخذ قيمها في مجموعة الحالات الخفية S بحيث يكون الاحتمال:

$$\Pr(X_{t+1} = s_j / X_t = s_i, \ldots, X_1 = s_h) = \Pr(X_{t+1} = s_j / X_t = s_i) = a_{ij}$$

وتُعبّر a_{ij} عن احتمال المرور من الحالة s_i إلى الحالة s_j .

- بينما تأخذ السلسلة $(Y_t)_t$ قيمها في مجموعة المشاهدَات O بحيث:

$$\Pr(Y_t = o_k / X_t = s_i, Y_{t-1} = o_{k_{t-1}}, X_{t-1} = s_{i_{t-1}}, \ldots, Y_1 = o_{k_1}, X_1 = s_{i_1})$$
$$= \Pr(Y_t = o_k / X_t = s_i) = b_i(k)$$

وتُعبّر $b_i(k)$ عن احتمال مُشاهَدة o_k مع العلم بتحقق S_i .

وبخصوص مشاهدات النموذج وحالاته الخفية في عملنا هذا فإننا نُعرّفها كالتالي:

أ. مشاهدات النموذج (Observations)

نقترح في هذا العمل تصنيفًا جديدًا للكلمات العربية. وتستند عملية التصنيف (classification) هذه على استخلاص **ملامح** (features) من الكلمة المراد معالجتها تتمثّل في:

- طول الكلمة.
- السوابق واللّواحق.
- مواضع أحرف الألف والواو والياء إن وجدت في الكلمة.
- حرف أو حرفان من الكلمة يتم اختيارهما تبعا لطولها.

ولتوضيح هذا التصنيف **نضرب مثالا** بكلمة من قبيل **"التشكيل"** فلدى تصنيف هذه الكلمة يستنتج البرنامج صنفًا على شكل:

F7؛alS؛#؛0؛1-؛5؛ش+ي

وتأويل هذا الصنف راجع إلى كونه اتحاد الملامح المستخلصة التالية:

- **"F7"** : وهو الملمح المشترك بين جميع الكلمات المكونة من سبعة أحرف.
- **"alS"** : ويدلّ على أن الكلمة مبدوءة بأداة التعريف "ال"، وأن أداة التعريف متبوعة بحرف شمسي (S) وهو التاء في هذه الحالة. إن تمييزنا لملمح السوابق التي تتضمن أداة التعريف إلى نوعين شمسي (alS) وقمري (alL) تبعًا للحرف الذي يلي السابق، راجع بطبيعة الحال إلى تأثير ذلك في تشكيل هذا الحرف وحرف اللام الذي يسبقه.

- **"#"** : وتدلّ على أن الكلمة ليست لها لاحقة (suffix).

- **"0"** : وتعني أن حرف الألف موجود في كلمة **"التشكيل"** في الموضع رقم 0 ؛ أي في الحرف الأول من الكلمة. وفي هذا السياق فإنه إذا تعددت مواضع حرف من الحروف (أ و ي)، فإن الملمح الخاص بها يكون على شكل متجهة (vector) تتركب من هذه المواضع. فالملمح الخاص بحرف الواو في كلمة **"الموعودون"** هو (3،5،7) ويعني أن الواو هو الحرف الرابع والسادس والثامن من هذه الكلمة.

- **"1-"** : ويعني أن حرف الواو لا يوجد في مثالنا أي كلمة **"التشكيل"**.

- **"5"** : مما يعني أن حرف الياء هو الحرف السادس من هذه الكلمة.

- **"ش+ي"** : وهما الحرفان اللذان يقعان في الموضعين الرابع والسادس على التوالي، وهما الموضعان اللذان تختار منهما الحروف المميزة للكلمات المكونة من سبعة أحرف.

وتحت الصنف الوارد في المثال السابق تقع الكثير من كلمات اللغة العربية التي تتشابه في كيفية تشكيلها، مثل: **"التشبيه"** و**"التشخيص"** و**"التشريع"** و**"التشغيل"**.. إلخ. إذ يشمل، بعبارة أخرى، كل المصادر التي على وزن **"التفعيل"** إذا كان الحرف الأول من جذرها حرف الشين. وتُعدّ هذه الخاصية أهم إيجابيات هذا التصنيف؛ إذ تكون "بارامترات" النموذج أقل بكثير من مثيلتها في الأعمال السابقة ((Gal, 2002) ، (Elshafei et al, 2006)، (بباه وآخرون 2011) التي تُمثّل الكلمات العربية فيها مشاهدات النماذج الماركوفية.

ب. الحالات الخفية (Hidden states)

تتكون الحالات الخفية في النموذج الماركوفي الذي نقترحه من تسلسلات علامات التشكيل الممكنة في اللغة العربية. وهي عبارة عن متجهات (vectors)

تختلف أبعادها تبعا لأطوال المشاهدات. ولتوضيح هذا التعريف أكثر نضرب أمثلة من هذه الحالات الخفية:

• (◌َ ، ◌ْ ، ◌ُ) : وهو التسلسل المكون من فتحة فسكون فكسرة فضمة، ويُمثّل حالة خفَية تكمن خلف الكثير من المشاهدات عند تصنيف كلمات من أربعة أحرف مثل الكلمات: **"يعرض"** ، **"تعرف"**، **"نصبر"**.. إلخ، بما يسمح بتشكيلها كالتالي: **"يَعْرِضُ"** ، **"تَعْرِفُ"**، **"نَصْبِرُ"** ... إلخ

• (◌ُ ، ◌َ ، ◌ِّ ، ◌َ ، ◌ٌ) : وهو متجه من خمس مُرَكّبات تمثل تسلسل الضمة فالفتحة فالشدة مع الكسر فالفتحة فتنوين الضم. ويقابل هذا التسلسل عددًا من المشاهدات، ما يسمح بتشكيل كلمات مثل **"مُعَلِّمَةٌ"**، **"مُقَدِّمَةٌ"** .. إلخ.

• (◌َ ، # ، ◌َ) : ويدلّ الرمز # على غياب علامة التشكيل على الحرف المقابل. ويسمح هذا التسلسل الثلاثي بتشكيل كلمات مثل: **"جَاءَ"** ، **"نَالَ"**، **"فَازَ"** .. إلخ.

ولا يخفى أن الطابع التجريدي للحالات الخفية في نموذجنا هذا يستدعي عددا أقل من "البارامترات" مقارنة بالنماذج الماركوفية السابقة (Gal, 2002)، (Elshafei et al, 2006) التي تتخذ من الكلمات المشكولة حالات خفية.

وفي ما يلي نُبيّن كيف استخدمنا هذا النموذج لغرض التشكيل الآلي للنصوص العربية.

لنفرض على سبيل المثال جملة عربية على الشكل $W_1,\ldots,W_n$ بعد القيام بتصنيف هذه الكلمات بالطريقة الواردة أعلاه في النموذج فإننا نحصل على سلسلة من المشاهدات $w_1,\ldots,w_n$ ، ولتكن المجموعة $C=\{c_1,\ldots,c_N\}$ مجموعة سلاسل التشكيلات الممكنة في اللغة العربية. انطلاقا من هذه الفرضيات، فإن مسألة تحديد التشكيلات الصحيحة في هذه المرحلة من المعالجة

تتمثّل في إيجاد مجموعة سلاسل التشكيلات $(c_1^*,\ldots,c_n^*)$ التي تحقق المعادلة التالية:

$$(c_1^*,\ldots,c_n^*) = \arg\max_{c_1\ldots c_n\in C} \Pr(c_1\ldots c_n / w_1\ldots w_n)$$

وبما أن:

$$\Pr(c_1\ldots c_n / w_1\ldots w_n) = \frac{\Pr(w_1\ldots w_n / c_1\ldots c_n)\Pr(c_1\ldots c_n)}{\Pr(w_1\ldots w_n)}$$

إن التسلسل $(c_1^*,\ldots,c_n^*)$ يحقق:

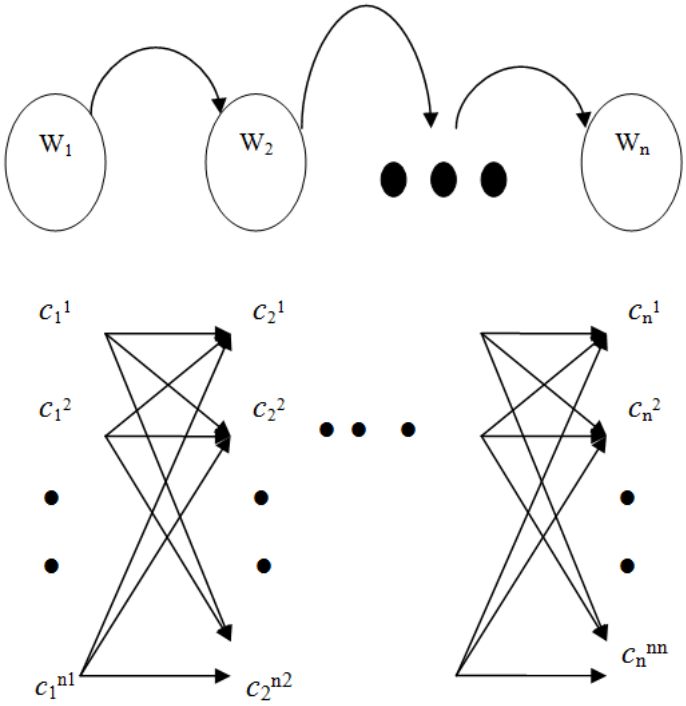
$$(c_1^*,\ldots,c_n^*) = \arg\max_{c_1\ldots c_n\in C} \Pr(w_1\ldots w_n / c_1\ldots c_n)\Pr(c_1\ldots c_n)$$

وهي المعادلة التي يمكن أن تُكتب على الشكل التالي:

$$(c_1^*,\ldots,c_n^*) = \arg\max_{\substack{c_i^{j_i}\in C_i \\ 1\le i\le n}} \Pr(w_1\ldots w_n / c_1^{j_1}\ldots c_n^{j_n})\Pr(c_1^{j_1}\ldots c_n^{j_n})$$

حيث ترمز المجموعة $C_i = \{c_i^1,\ldots,c_i^{n_i}\}$ إلى مجموعة التشكيلات المحتملة الناتجة عن تحليل الكلمة W_i ذات الصنف w_i.

ونقوم بحل هذه المعادلة عن طريق البحث عن المسار الأكثر رجحانًا في شبكة الحلول الناتجة عن التحليل الصرفي للكلمات خارج سياقها كما هو مبين في الشكل التالي:



الشّكل2. شبكة سلاسل التشكيلات المحتملة الناتجة عن تحليل الجملة $W_1,\ldots,W_n$

وتُعتبر خوارزمية فيتربي وسيلة فعالة للبحث عن المسار الأكثر رجحانًا في هذه الشبكة. وتعتمد الخوارزمية على قيم دالتين ϕ و ψ نعرفهما كما يلي:

$$\phi(t,c_t^k) = \max_{\substack{r_i^{ji}\in\Re_i \\ 1\le i\le t-1}} \left[\Pr(\, w_1,\ldots,w_{t-1},w_t \,/\, c_1^{k_1},\ldots,c_{t-1}^{k_{t-1}},c_t^k\,) \times \Pr(\, c_1^{k_1},\ldots,c_{t-1}^{k_{t-1}},c_t^k\,)\right]$$

حيث تُعبّر $\phi(t,c_t^k)$ عن قيمة احتمال المسار الجزئي الأرجح الذي يمر من التسلسل c_t^k (c_t^k ينتمي لمجموعة التشكيلات المحتملة للكلمة W_t).

ويمكن أن نكتب المعادلة السابقة على الشكل التالي:

$$\phi(t,c_t^k)=\max_{\substack{r_i^{j_i}\in\Re_i\\ 1\le i\le t-1}}\prod_{i=1}^{t-1}\left[\Pr(w_i/c_i^{j_i})\times\Pr(c_i^{j_i}/c_{i-1}^{j_{i-1}})\times\Pr(w_t/c_t^k)\times\Pr(c_t^k/c_{t-1}^{j_{t-1}})\right]$$
$$=\Big(\max_{c_{t-1}^{j_{t-1}}\in C_{t-1}}\phi(t-1,c_{t-1}^{j_{t-1}})\times\Pr(c_t^k/c_{t-1}^{j_{t-1}})\Big)\Pr(w_t/c_t^k)$$
$$=\Big(\max_{c_{t-1}^{j}\in C_{t-1}}\phi(t-1,c_{t-1}^{j})\times\Pr(c_t^k/c_{t-1}^{j})\Big)\Pr(w_t/c_t^k)\quad(*)$$

وتسمح لنا المعادلة الأخيرة بحساب قيم الدالة ϕ بصيغة تراجعية.

ومن أجل التعرف على المسار الأرجح فإننا نعرف الدالة ψ التي تسمح في كل لحظة t بتخزين التشكيل الذي يُحقق أكبر قيمة للمعادلة السابقة (*).

نُعرّف ψ كما يلي:

$$\psi(t,c_t^k)=\arg\max_{c_{t-1}^{j}\in C_{t-1}}\phi(t-1,c_{t-1}^{j})\Pr(c_t^k/c_{t-1}^{j})\qquad(**)$$

مع ملاحظة أن: $\psi(t,c_t^k)\in C_{t-1}$

هاتان المعادلتان (*) و (**) تسمحان لنا بإيجاد المسار الأرجح عن طريق خوارزمية فيتربي التراجعية التالية:

- **المرحلة الأولى الابتداء (Initialization) :**

for $1\le k\le n_1$

حساب $\phi(t,c_1^k)$: احتمال أن تبدأ الجملة بالكلمة w_1 ويكون تشكيلها c_1^k .

- **المرحلة الثانية الحساب التراجعي (Recursion) :**

for $2\le t\le n$ and $1\le k\le n_t$

$$\phi(t,c_t^k)=\Big(\max_{c_{t-1}^{j}\in C_{t-1}}\phi(t-1,c_{t-1}^{j})\times\Pr(c_t^k/c_{t-1}^{j})\Big)\Pr(w_t/c_t^k)$$
$$\psi(t,c_t^k)=\arg\max_{c_{t-1}^{j}\in C_{t-1}}\phi(t-1,c_{t-1}^{j})\Pr(c_t^k/c_{t-1}^{j})$$

- **المرحلة الثالثة الحالة الأخيرة (Final state):**

$$\psi(n+1) = \arg\max_{c_n^j \in C_n} \phi(n, c_n^j)$$

- **المرحلة الرابعة استنتاج المسار الأرجح (Deducing the best path):**
  - $c_n^* = \psi(n+1)$
  - For t = n-1 :1 $c_t^* = \psi(t, c_{t+1}^*)$

4. تدريب البرنامج

تحتاج "بارامترات" النموذج الإحصائي الوارد أعلاه، والمتمثلة في قيم المصفوفتين (a_{ij}) = **A** و (b_{it}) = **B**، إلى عملية تقييم –خلال مرحلة تدريب البرنامج– بالاعتماد على ذخيرة لغوية كافية.

فلو اعتبرنا ذخيرة لغوية $C = \{Ph_1, \ldots, Ph_K\}$ مكونة من عدد كبير من الجمل، فإن عملية تدريب النظام تتلخص في تقييم "البارامترات" بطريقة (Maximum Likelihood) على الشكل التالي: (Manning and Schütze, 1999)

$$a_{ij} = \frac{\sum_{n=1}^{K} (\text{عدد مرات الانتقال من التشكيل } c_i \text{ إلى التشكيل } c_j \text{ في الجملة } Ph_n)}{\sum_{n=1}^{K} (\text{عدد تكرارات التشكيل } c_i \text{ في الجملة } Ph_n)}$$

$$b_{it} = \frac{(\text{عدد مرات أخذ الصنف } w_t \text{ للتشكيل } c_i)}{(\text{عدد تكرارات التشكيل } c_i)}$$

ولإجراء عملية التدريب هذه قمنا بعزل مدوّنة مشكولة بلغ حجمها 1,24 مليون كلمة مشكولة موزعة بين نصوص مدوّنة "NEMLAR"، التي اخترنا 90% منها (460 ألف كلمة) واستبقينا 10% للاختبار، ومدوّنة "تشكيلة" التي اخترنا منها (780 ألف كلمة). ويرجع السبب في عدم استخدامنا لملايين

النصوص التراثية المشكولة التي لدينا في التدريب إلى محاولة إيجاد توازن في مدوّنة التدريب بين المحتوى المعاصر والمحتوى التراثي. فبالتوازي مع النصوص المعاصرة لمدونة "NEMLAR" قمنا بعزل 30 كتابًا تراثيًا من مدوّنة "تشكيلة" واخترنا بشكل عشوائي من كل كتاب قرابة 10% من محتواه. كما أجرينا على مدونة التدريب سلسلة من العمليات تمثلت في تقسيم محتوياتها إلى جُمل ثم تصنيف جميع كلماتها واستخلاص تسلسلات التشكيل فيها من أجل حساب "البارامترات" باستخدام طريقة الإمكان الأعظم (Maximum likelihood) الواردة أعلاه.

5. اختبار البرنامج

من أجل تقييم كفاءة برنامج التشكيل الآلي فقد اعتبرنا المقياسين المعروفين: نسبة الكلمات الخاطئة (WER : Word Error Rate) ونسبة الخطإ على مستوى الحروف (DER : Diacritic Error Rate). ويتفرع هذان المقياسان في الواقع إلى أربعة مقاييس بالنظر إلى اعتبار تشكيل الحرف الأخير في الكلمة من عدمه. فبذلك كانت المقاييس التي استخدمنا في عملية التقييم كالتالي:

- WER1 : ويقيس نسبة الكلمات الخاطئة مع اعتبار تشكيل الحرف الأخير.
- WER2 : ويقيس نسبة الكلمات الخاطئة مع تجاهل تشكيل الحرف الأخير.
- DER1 : ويقيس نسبة الخطإ في التشكيل على مستوى الحروف بما فيها الحرف الأخير.
- DER2 : ويقيس نسبة الخطإ في التشكيل على مستوى الحروف باستثناء الحرف الأخير.

وتتمثل تجربة التقييم في اختيار خمس عينات من مدونة الاختبار، تتكون كل عينة من 100 جملة، تم سحب كل جملة منها بشكل عشوائي، وذلك بغية

الحصول على القيمة التقريبية لنسب الخطإ الأربع على مستوى كامل مدونة الاختبار.

ويلخّص الجدول التالي النتائج التي حصلنا عليها.

العينة	عدد الكلمات	WER1(%)	WER2(%)	DER1(%)	DER2(%)
T1	2601	27,73	13,42	8,09	4,02
T2	2380	27,57	14,79	8,51	4,75
T3	2525	28,04	13,75	8,23	4,19
T4	2863	31, 34	16,49	9,68	5,24
T5	2356	27,81	13,97	8,34	4,28
المتوسط	**2545**	**%28,5**	**%14,5**	**%8,6**	**%4,5**

جدول1. نتائج تقييم البرنامج على عينات عشوائية من مدونة الاختبار

تُبيّن هذه النتائج أن نسبة خطإ التشكيل على مستوى الكلمات تقارب 29%، وأن تلك النسبة تهبط إلى النصف في حالة تجاهل الحرف الأخير لتصل 14,5%. ولنا على هذه النتيجة بعض الاستدراكات. فالمقياس "WER1" بالشكل الذي اعتمدناه في هذه التجربة هو مقياس **متشدد**، يفترض وجود نسخة من مدونة الاختبار مشكولة بشكل صحيح تام. فعند مقارنة الكلمة الأصلية بنتيجة معالجة البرنامج لنسخة غير مشكولة منها لا يعطي المقياس نتيجة إيجابية إلا في حالة التطابق التام بين الكلمتين. والواقع أن المدوّنات المتاحة لا تَسْلم، من

جهة، من الأخطاء الكتابية كالكلمات المتلاصقة والكلمات المفروقة والتساهل في الهمزات، وغير ذلك من الأخطاء الإملائية التي تُربك المحلّل الصرفي المدمج، كما أنها لا تخلو، من جهة أخرى، من الكلمات المشكولة جزئيًا التي قد يقوم البرنامج بتشكيلها تشكيلا تامّا صحيحًا يبقى رغم ذلك خاطئا بحسب المقياس "WER1".

ويدلّ هبوط نسبة الخطإ إلى النصف (14,5%) في المقياس WER2 على أن نسبة مهمة من أخطاء التشكيل هي أخطاء إعرابية كعدم الاتفاق، أحيانًا، بين حركة المتتابعَيْن (نعتًا وتوكيدًا وعطفًا وبدلًا) أو عدم جرّ المضاف إليه أو الخطأ في حركة بعض معمولات النواسخ وحروف الجرّ. ذلك بأن معيار الترجيح الذي يستند عليه البرنامج هو معيار إحصائيّ يصيب ويخطئ في هذه الحالات الإعرابية تبعًا لقيم الاحتمالات التي استخلصها من مدوّنة التدريب.

وتصل نسبة الخطإ في تشكيل جميع حروف النص "DER1" إلى 8,6%، وتهبط هذه النسبة أيضا إلى حدود النصف 4,5% (DER2) لتتجاوز بذلك نسبة الدقة على مستوى الحروف 95% في حال تجاهل الحرف الأخير.

6. الخاتمة

قدّمنا في هذه البحث برنامجًا جديدًا أطلقنا عليه اسم **"برنامج الخليل للتشكيل الآلي"**، يعتمد هذا البرنامج على مقاربة هجينة لتشكيل اللغة العربية تجمع بين التحليل الصرفي ونماذج ماركوف الخفية. وتختلف هذه المقاربة عن غيرها من المقاربات الهجينة على المستويين اللّغويّ والإحصائيّ.

فعلى المستوى اللغويّ تمّ توظيف أحد أهم المحلّلات الصرفية العربية مفتوحة المصدر، حيث يقوم البرنامج باستخدام نتائج تحليل برنامج الخليل للتحليل الصرفي واستخلاص التشكيلات الممكنة التي يقترحها للكلمات خارج سياقها من النص. وقد تطلَّب دمج المحلل الصرفي في إطار التشكيل الآلي إضافة

قاعدة معطيات معجمية تتضمن الكلمات الأكثر تكرارًا، وذلك بهدف ضبط تشكيلها وتفادي التعميم (over generation) الذي يظهر في كثرة المخرجات المقترحة من قبل المحلّل الصرفي. وقد تم استقاء قاعدة المعطيات المعجمية من مدوّنة تضم أزيد من 250 مليون كلمة جمعت من ثماني مدونات عربية متاحة على الشابكة (الإنترنت).

وعلى المستوى الإحصائيّ، قمنا باقتراح نموذج ماركوفي ينطلق من تصنيف جديد للكلمات العربية، بحيث تمثّل هذه الأصناف مشاهدات النموذج، بينما تمثل تسلسلات علامات التشكيل الحالات الخفية له. كما راعينا في تدريب البرنامج استخدام مدوّنة متنوعة تجمع بين النصوص الحديثة والتراثية بلغ حجمها أزيد من مليون ومائتي ألف كلمة.

وتعتبر النتائج التي حصلنا عليها جد مشجعة، إذ وصلت نسبة الدقة على مستوى كامل الحروف إلى 91,4%، وإلى أزيد من 95% بتجاهل الحرف الأخير. وهي نتائج نطمح إلى تحسينها بالعمل على مستويين متكاملين: مستوى المدوّنات، وذلك بتوسيع محتوى المدونة من اللغة المعاصرة وضبط تشكيل أكبر قدر ممكن منها. ومستوى التحليل اللغوي، وذلك بدمج المقاربة في إطار أشمل يتعامل مع باقي المعلومات التي يتيحها محلل الخليل الصرفي سعيًا لإيجاد محلل نحوي ذي كفاءة مقبولة، مما سينعكس إيجابًا على نتائج المشكل الآلي. كما نعتزم فتح مصدر هذا البرنامج لتعميم استخدامه وإتاحة المجال أمام اجتهادات المطورين في تحسينه ودمجه في تطبيقات أخرى.

مراجع باللغة العربية

1 – سلوى حماده، مروان والبواب: **"معايير وضوابط تقييم المحلِّلات الصرفية"**، تقرير من ضمن توصيات المشاركين في اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، 26 - 28 إبريل 2009، دمشق.

2 – عثمان بن سعيد الداني أبو عمرو، **المحكم في نقط المصاحف**، تحقيق: د. عزة حسن، دار الفكر – دمشق، الطبعة الثانية، 1407 هـ.

3 – عز الدين مزروعي، عبد الوافي مزيان، عبد الحق لخواجه، عبد الرحيم بودلال، محمد ولد عبد الله ولد بباه،: **"برنامج الخليل الصرفي "**، المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية، ICCA 2011، 31 مايو ـ 2 يونيو 2011، الرياض.

4 – محمد عطية، **"التشريح البنائي لمشكِّل آلي عربي لتوظيفه في نظام تخليق آلي للصوت المنطوق من النص العربي المكتوب"**، ندوة تقنية المعلومات والعلوم الشرعية والعربية، جامعة الإمام محمد بن سعود الإسلامية، الرياض، 2007.

5 – محمد ولد عبد الله ولد بباه، عبد الوافي مزيان، عز الدين مزروعي، عبد الحق لخواجه، عبد الرحيم بودلال: **"مقاربة صرفية إحصائية للتشكيل الآلي"**، اجتماع خبراء التدقيق الإملائي والتشكيل والتحليل النحوي الآلي، 18 - 20 إبريل 2011، دمشق.

6 – منصور الغامدي؛ محمد خورشيد؛ مصطفى الشافعي؛ فايز الحرقان؛ أبو أوس الشمسان؛ محمد الكنهل؛ سعد القحطاني؛ سيد زيشان مظفر؛ ياسر التويم؛ عدنان يوسف؛ حسني المحتسب؛ **"نظام حاسوبي لتشكيل النص العربي"**، التقرير الفني النهائي، مدينة الملك عبد العزيز للعلوم والتقنية، 2006.

مراجع باللغات الأجنبية

1 – Abbas M., Smaili K.. Comparison of Topic Identification Methods for Arabic Language, International conference RANLP05 : Recent Advances in Natural Language Processing , 21-23 september 2005, Borovets, Bulgary.

2 – Abdelali A., Cowie J., Soliman H.: "Building A Modern Standard Arabic Corpus", Workshop on Computational Modeling of Lexical Acquisition, The Split Meeting, Crotia, 25-28 July 2005.

3 – Attia. M., Yaseen., M., and Choukri., K. :"Specifications of the Arabic Written Corpus produced within the NEMLAR project", www.NEMLAR.org. 2005.

4 – Berger, A., Della Pietra, S., Della Pietra, V., "A maximum entropy approach to natural language processing". Computational Linguistics 22 (1), 39–71. 1996.

5 – Buckwalter T.. Arabic Morphological Analyzer version 2.0. LDC2004L02. 2004.

6 – Debili F., ACHOUR H.: "Voyellation automatique de l'arabe" in Proceedings of the workshop on Computation approaches to Semitic languages, COLING-ACL '98., Montréal, 1998.

7 – Deltour A., « Méthodes statistiques pour la voyellisation des texts arabes », Projet de fin d'études, ENSIMAG-Universität Karlsruhe, 2003.

8 – El-Sadany, T., Hashish, M.,: "Semi-automatic vowelization of Arabic verbs". In: 10th NC Conference, Jeddah, Saudi Arabia. 1988.

9 – Elshafei, M., Almuhtasib H. and Alghamdi M, :"Machine Generation of Arabic Diacritical Marks", The 2006 World Congress in Computer Science Computer Engineering, and Applied Computing . Las Vegas, USA. 2006.

10 – Emam O. and Fischer V.: "Hierarchical Approach for the Statistical Vowelization of Arabic Text. Technical report, IBM Corporation Intellectual Property Law, Austin, TX, US 2005.

11 – Gal Y.: “An HMM Approach to Vowel Restoration in Arabic and Hebrew”, In ACL-02 Workshop on Computational Approaches to Semitic Languages 2002.

12 – Habash N. and Rambow O.: “Arabic Diacritization through Full Morphological Tagging”. In Proceedings of NAACL/HLT 2007. Companion Volume, Short Papers, Rochester, New York, April. 2007.

13 – Maamouri M., Bies A., and Kulick S.: “Diacritization: A Challenge to Arabic Treebank Annotation and Parsing”. In Proceedings of the British Computer Society Arabic NLP/MT Conference, 2006.

14 – Manning C. and Schütze H., Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

15 – Messaoudi A., Lamel L., and Gauvain J. : “The LIMSI RT04 BN Arabic system”. In Proc. DARPA RT04, Palisades NY, November 2004.

16 – Nelken R. and. Shieber S. M.: “Arabic Diacritization Using Weighted Finite-State Transducers. In Proceedings of the ACL 2005 Workshop On Computational Approaches To Semitic Languages”, Ann Arbor, Michigan, USA. 2005.

17 – Neuhoff D.L., The Viterbi Algorithm as an Aid in Text Recognition, IEEE Transaction on Information Theory, pp 222-226, March 1975.

18 – Rafalovitch A, Dale R.: “United Nations General Assembly Resolutions: A Six-Language Parallel Corpus”. In *Proceedings of the MT Summit XII*, pages 292-299, Ottawa, Canada, August.

19 – Saad M. K., Ashour W.: “OSAC: Open Source Arabic Corpora”, 6th ArchEng International Symposiums, EEECS’10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, pp. 118-123, European University of Lefke, Cyprus, 2010.

20 – Schlippe T., Nguyen T. and Vogel S.: “Diacritization as a Machine Translation Problem and as a Sequence Labeling Problem,” 8th AMTA conference, Hawaii, 21-25 October 2008.

21 – Vergyri D. and Kirchhoff K., ‘Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition’. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. COLING 2004, Geneva: 66-73.

22 - Zitouni I., Sorensen J. S. and Sarikaya R : ‘Maximum Entropy Based Restoration of Arabic Diacritics’. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Workshop on Computational Approaches to Semitic Languages, Sydney, Australia. July 2006.