مهام الصرف الحاسوبي
Computational Morphology Tasks

في هذا الفصل، سنناقش مجموعة من مهام الصرف الحاسوبي الشائعة، ونبين الاتجاهات المختلفة للتعامل معها. معظم هذه المهام ليست غاية في حد ذاتها، على سبيل المثال، تعيين أقسام الكلام (POS) أو استخراج الجذر. فهي تقنيات دعم وتمكين وتعتبر مهمة بالنسبة للتطبيقات العليا مثل الترجمة الآلية (MT)، واسترجاع المعلومات (IR) أو التعرف التلقائي على الكلام (ASR). وهناك عدد قليل من المهام التي تخدم كلا الدورين (كنهاية ومعنى)، على وجه الخصوص التشكيل التلقائي، وهو ما يمكن اعتباره كتطبيق مستقل يسمح للمستخدمين بتحويل نص غير مشكل إلى نص مشكل وكأداة لتحويل النص إلى كلام.

في القسم التالي، سنعّرف عدداً من هذه المهام ونربط بعضها مع بعض. وفي الأقسام الثلاثة المقبلة، سنناقش بتفصيل أكبر ثلاث مجموعات من المهام: التحليل/التوليد الصرفي، والتقطيع وتعيين أقسام الكلام. أما في القسم الأخير في هذا الفصل، سنقارن ونقابل بالتفصيل اثنين من الأدوات المستخدمة لمعالجة اللغة العربية التي تتعامل مع مجموعات فرعية مختلفة من هذه المهام.

٥,١ مفاهيم أساسية

في هذا القسم، سنقدم أولاً بعض مهام الصرف الحاسوبي الأكثر بحثاً والتقنيات المساندة. ثم نناقش بعض المواضيع المشتركة فيما بينها.

يشير **التحليل الصرفي** (Morphological analysis) إلى العملية التي من خلالها يتم تحديد جميع الأشكال الصرفية الممكنة للكلمة (التي تعرف عادة هجائياً). ويشمل كل تحليل للكلمة أيضاً اختيار أحد أساسيات أقسام الكلام (مثل الاسم أو الفعل، فتحديد المجموعة مسألة اختيارية). يمكن أن يكون التحليل الصرفي شكلياً، وفي هذه الحالة تقسم الكلمة الواحدة إلى جميع المورفيمات المكونة لها، أو يكون وظيفياً، وفي هذه الحالة يتم تفسير المورفيمات. على سبيل المثال، في جمع التكسير، التحليل الشكلي قد لا يتعرف على حقيقة أن الكلمة هي في صيغة جمع لأنها تفتقر إلى وجود مورفيم الجمع المعتاد في حين أن التحليل الوظيفي يستطيع التعرف على ذلك.

يعتبر **التوليد الصرفي** (Morphological generation) في الأساس عكس عملية التحليل الصرفي. وهي العملية التي يتم فيها مقابلة التمثيل الأساسي لكلمة ما بتمثيل سطحي (سواء هجائياً أو صوتياً). والسؤال المهم في التوليد هو: ما هو التمثيل الذي ننطلق منه؟ فكلما كان التمثيل أقل عمقاً، سهلت مهمة التوليد. قد يكون بعض التمثيل أقل تقييداً من غيره مما يؤدي إلى نتائج صحيحة متعددة. ويعتقد أحياناً أن التمثيل الوظيفي هو نقطة انطلاق نموذجية للتوليد.

يشير **فك اللبس الصرفي** (Morphological disambiguation) إلى اختيار التحليل الصرفي من السياق. ويشار إلى هذه المهمة للغة الإنجليزية على أنها تعيين أقسام الكلام (POS tagging) حيث إن مجموعة سمات أقسام الكلام (POS tag set) المعيارية تضم ٤٦ سمة فقط، وتعمل على فك اللبس الصرفي كلية في الإنجليزية. أما في اللغة العربية، فنظرياً قد تضم مجموعة السمات (tag set) الممكنة ما يزيد على ٣٣٠,٠٠٠ سمة [51]، ومن ثَم فإن مهمة فك اللبس الصرفي ستكون أصعب بكثير. وقد اقترح البعض خفض مجموعة السمات للغة العربية، بحيث تُدمج بعض الاختلافات الصرفية، مما يجعل مهمة

فك اللبس الصرفي أسهل. وعادة ما يستخدم مصطلح تعيين أجزاء الكلام (POS tagging) في العربية لتشير لمجموعة أصغر من مجموعة السمات.

يشير مفهوم **تقطيع النص** (Tokenization)، الذي يسمى أحياناً **التجزئة** (segmentation)، إلى تقسيم كلمة واحدة لمجموعات مروفيمية متتالية، واحدة منها يمثل عادة جذع الكلمة، ومتضمنا المورفيمات التصريفية. ينطوي تقطيع النص على نوعين من القرارات التي تحدد خطة التقطيع [83]. أولاً، يجب أن نختار أي نوع من أنواع المورفيمات لتقسيمها، فليس هناك طريقة واحدة صحيحة للتقسيم. ثانياً، نحن بحاجة إلى أن نقرر بعد فصل بعض المورفيمات إن كان يجب القيام بضبط التهجئة للمقاطع الناتجة، حيث إن تسلسل المورفيمات قد يؤدي إلى تغييرات إملائية على حدودها. على سبيل المثال، التاء المربوطة (ة ħ) تظهر كتاء عادية (ت t) عندما يتبعها ضمير متصل، ولكن عندما نجزئ الزوائد في نهاية الكلمة (enclitic)، قد يكون من المرغوب إعادة التاء المربوطة لشكلها في نهاية الكلمة. وعادة ما تستخدم كلمة "التجزئة" عندما لا توجد تسوية هجائية. كما أن التسوية الهجائية مرغوبة في المعالجة الآلية للغة الطبيعية لأنها تقلل من تناثر البيانات، كما يفعل التقطيع.

التشذيب (Lemmatization)[1] هو مقابلة شكل كلمةٍ ما لمادتها المعجمية (Lemma)، وهو يمثل مُعَيْجِمَتها (lexeme) المعتمدة. والتشذيب تجسيد محدد للمهمة الأكثر عمومية للتعرف على المُعَيْجِمَة (lexeme identification) وهي خطوة يتم فيها حل غموض المادة المعجمية. ولا ينبغي الخلط بين التشذيب وبين التجذيع (stemming) الذي يقوم بمقابلة الكلمة لجذعها. مهمة أخرى ذات صلة هي استخراج الجذر، وهو يركز على تحديد جذر الكلمة.

(١) تسمى أيضاً تعيين المادة المعجمية (توضيح من المترجمة).

Obeikandi.com

التشكيل (Diacritization) هي عملية معالجة التشكيل المفقود من الكلمة (مثل التنوين، وعلامة الإدغام، وغياب علامة الصائت القصير). وهناك علاقة وثيقة بين التشكيل وفك اللبس الصرفي والتشذيب، فمختلف قيم الخصائص الصرفية ومختلف المواد المعجمية لكلمة غير مشكلة قد يؤديان إلى تشكيلات مختلفة، انظر قسم ٢,٣,٤.

إن المهام المختلفة وما يتبعها من مهام فرعية التي نوقشت حتى الآن، يمكن إدراجها في المصطلحات التالية:

- **السياق (Context)**: بعض المهام غير سياقية (أي خارج السياق out-of-context) والبعض الآخر سياقية (في السياق in-context). بشكل عام، تركز المهام غير السياقية على وصف مجموعة القيم الممكنة (مثل أقسام الكلام (POS)، والسمات (tags)، والتشكيل، والمادة المعجمية، والجذر، إلخ) المرتبطة بكلمة ما. في المقابل، تركز المهام السياقية على اختيار القيم المناسبة للسياق (مجدداً، سواء كانت سمات أقسام الكلام، أو تشكيل، أو مادة معجمية، أو جذر، إلخ). ويعتبر التحليل الصرفي وفك اللبس الصرفي نماذج لمهام خارج السياق/في السياق. ويمكن تعريف كل مهمة بناء على هذين الوضعين. على سبيل المثال، عملية التقطيع خارج السياق هي مهمة لتحديد جميع القطع الممكنة المكونة لكلمة ما. والشكل الشائع للتقطيع هو أن تحدد خياراً معيناً في نفس السياق. وهناك طرق حاسوبية مختلفة للتقطيع قد تمثل أو لا تمثل بشكل صريح أو ضمني الخيارات غير السياقية داخلياً.
- **الثراء (Richness)**: تختلف بعض المهام في كونها سطحية أو عميقة، أو مشذبة (fine-grained) أو غير مشذّبة (coarse). على سبيل المثال، هناك عدد كبير من مجموعات تعيين أقسام الكلام التي من الممكن استخدامها مع الصرف الشكلي (في هذه الحالة تعتبر سطحية) أو مع الصرف الوظيفي (في هذه الحالة تعتبر عميقة)،

وقد تركز المهمة على السمة الأساسية (core tag) للكلمة الرئيسة (في هذه الحالة تعتبر غير مشذبة) أو تمتد لتشمل جميع قيم الخصائص التصريفية والزوائد. وبالمثل، يشمل التقطيع الطرق المختلفة لتمثيل الوحدة اللفظية للكلمة (word tokens) بما فيها الجذوع (stems) والمواد المعجمية والجذور أو حتى شكل كلمة معينة تم توليدها، والتشكيل الكامل أو الجزئي.

- **الاتجاهية (Directionality)**: تعتبر بعض المهام تحليلية في المقام الأول، أي تعمل على مقابلة من شكل سطحي إلى شكل أعمق، والبعض الآخر توليدي، أي تعمل على المقابلة من شكل أعمق إلى شكل سطحي. ويعتبر التحليل والتوليد الصرفي مهمتان نموذجيتان من هاتين الفئتين. تعتبر معظم المهام التي تشتمل على اختيار مجموعة فرعية من خصائص الكلمة، مثل، المادة المعجمية أو الجذر، إلخ، مهام تحليلية. وتركز مهمة تقليص احتمالات اللفظ (Normalized tokenization) على إنتاج شكل سطح طبيعي للكلمة، لذا تعتبر هذه المهمة جزئياً تحليلية وجزئياً توليدية. وتُحلل الكلمة بفعالية لتحديد مكوناتها، ولكن بعد ذلك يُنشأ النموذج الصحيح للكلمة المقطعة. على سبيل المثال، معالجة التاء المربوطة في كلمات تحتوي على الضمائر المتصلة يستلزم إعادة كتابة شكل الكلمة بمجرد قطع الزائدة.

٥,٢ التحليل والتوليد الصرفي

يعتبر التحليل والتوليد الصرفي للعربية محور البحث في مجال معالجة اللغة الطبيعية لفترة طويلة، نظراً للتعقيد الصرفي في العربية.

وهناك بعض المتطلبات المتوقعة من أنظمة التحليل والتوليد الصرفي لأي لغة، وتشتمل هذه على: (١) تغطية للغة المعنية، سواء من حيث تغطية المفردات (بشكل

كبير الحجم) وتغطية الظواهر الصرفية والهجائية (المتانة)، (٢) تُقابل الأشكال السطحية للكلمة من وإلى مستوى عميق من التمثيل الذي يتجرد بقدر الإمكان من الخصائص الصرفية والهجائية للغة معينة[١]. (٣) عكس الاتجاه الكامل للنظام بحيث يمكن استخدامه كمحلل أو مولد، (٤) قابليتها للاستخدام في مجموعة واسعة من تطبيقات معالجة اللغة الطبيعية مثل الترجمة الآلية أو استرجاع المعلومات، وأخيراً (٥) توفرها للمجتمع البحثي. وهذه القضايا أساسية في تصميم أي نظام عربي للتحليل والتوليد الصرفي.

وقد بُنيت العديد من المحللات الصرفية لمجموعة واسعة من مجالات التطبيق بدءاً من الترجمة الآلية إلى استرجاع المعلومات وأيضاً لمجموعة متنوعة من السياقات النظرية اللغوية مثل [84، 85، 86، 87، 88، 23، 89، 90، 91، 92، 80، 93، 67] وغيرها. وتشمل الجهود المنشورة التي تستهدف التوليد الصرفي أو تعامله كجزء من نظام مشترك للتحليل أو التوليد [86، 92، 94، 95، 96، 97] وغيرها. كما أن هناك الكثير من الأعمال التي لن نناقشها هنا، لذا نحث القراء على الاطلاع على بعض المقالات المسحية حول هذا الموضوع [85]. في هذا القسم، سنقدم أولاً أبعاد الاختلاف بين الحلول المختلفة للصرف العربي، ثم سنناقش بمزيد من التفاصيل، أربعة حلول محددة ومتناقضة بطرق مثيرة للاهتمام.

٥,٢,١ أبعاد الاختلاف

فيما يلي قائمة من الجوانب المشتركة من الاختلاف بين المناهج والحلول المختلفة للتحليل والتوليد الصرفي العربي.

---

(١) لمثال على ذلك انظر الشكل رقم (٤,٢) ففي الصف الثاني يظهر المثال مختصراً أما الصف الخامس فيحتوي على توضيح أكثر عمقاً (تعليق من المترجمة).

- **المعجم والقواعـــد (Lexicon and Rules)**: يعتبر المعجم والقواعد هما جوهر المعرفة الأساسية لأي نظام للتحليل أو التوليد. وعادة ما يحمل المعجم كل المعرفة المعجمية المحددة التي يمكن أن تتضمن السماح بتشكيلات من نمط الجذر– الوزن (معلومات بفئة مفتوحة)، اللواصق (معلومات بفئة مغلقة)، وفئات الكلمة التصريفية (الترتيب الصرفي ومعلومات التوافقية)، علاوة على معلومات إضافية مفيدة مثل دخول المقابلات في لغة أخرى (وهذا ليس بالضرورة للتحليل أو التوليد). من جهة أخرى، نجد أن القواعد عادة ما تكون تعميمية وتعالج ظواهر معجمية مستقلة، مثل إملاء التاء المربوطة من بين ظواهر أخرى. في بعض الطرق، تعتبر القواعد والمعجم في سلسلة متصلة من التعميمات للمعلومات الصرفية: المعجم في الأساس هو لائحة طويلة من قواعد محددة جداً. فماهية المعلومات التي تُمثل في المعجم مقابل القواعد، متروكة تماماً لمصممي النظام. ففي معظم الحالات تكون القرارات واضحة، ولكن البعض منها يمكن أن يذهب لأيٍّ من الاتجاهين. ومن الواضح أنه لوجود نظام يعمل بشكل صحيح، يجب أن تكون القواعد والمعجم متوافقة. لهذا السبب فإنه من الصعب في كثير من الأحيان (أو بشكل غير مباشر) إعادة استخدام المعاجم أو القواعد من نظام لآخر. يمكن إنشاء القواعد والمعجم إما بشكل يدوي أو تعلمه تلقائياً أو بشكل شبه تلقائي. وباعتبارها قاعدة المعرفة للنظام، فإن المعاجم والقواعد تتناقض مع محرك التحليل والتوليد الذي يستخدمهم لإنجاز المهمة. ففي تطبيق محدد للغة معينة، يُضيع هذا التمييز وقد يحتوي المحرك على قواعد ثابتة الترميز (hard-coded). ويمكن لمحلل صرفي بسيط أن يتكون من معجم يسرد كافة الاختلافات الممكنة للكلمة وتحليلاتها المناظرة. مثال على ذلك: معجم اللغة العربية العامية المصرية [52]. قد تسمح بعض المحللات

بناتج قائم على قواعد خلفية (rule-based back-off) لا يظهر في قاموسها. على سبيل المثال، عادة ما ينتج محلل باكوالتر الصرفي العربي [23] قراءات إضافية لاسم العلم مع أنه ليس في قاموسه.

- **التمثيل الخارجي (External Representation)**: ثمة فرق آخر بين الأنظمة في تمثيلها الخارجي (المدخلات/المخرجات). في التحليل، يعد هذا هو هدف التمثيل كمخرج للنظام. ويمكن أن يكون التحليل سطحياً أو عميقاً. على سبيل المثال، استهداف المورفيمات غير المشكلة يعتبر شكل من أشكال التحليل الصرفي السطحي، بينما استهداف المُعَيْجِمَة والخصائص الوظيفية تعتبر نوعاً من أنواع التحليل العميق. أيضا يمكن أن تكون المحللات انتقائية أو شاملة (في مستوياتها العميقة والسطحية)، على سبيل المثال: يركز المحلل الانتقائي على خاصية صرفية محددة أو أكثر من خاصية، مثل التعرف على الجذور أو حرف العطف و+ *wa+*. ويحاول المحلل الشامل التقاط جميع المعلومات الصرفية المختلفة بالعمق المستهدف. أما المحللات الانتقائية فتعتبر أدواتاً مفيدة للغاية لأداء مهام محددة، مثل استرجاع المعلومات [89] أو الترجمة الآلية الإحصائية [98]. وعادة ما تقايض الهدف الغني مع الكفاءة وحتى الدقة. أما في التوليد، فيشمل التمثيل الخارجي كلاً من مدخل/مصدر التمثيل لتوليدها منه والهدف/المخرج لتوليدها إليه، وهذه قد تكون عميقة أو سطحية أو انتقائية أو شاملة، مع كون المخرجات أقل عمقاً من المدخلات. ويعتبر النص السطحي المخرج المولد الأكثر شيوعاً، على الرغم من إمكانية استخدام تمثيلات أخرى داخل نظام أكبر يستخدم التوليد كمكون داخلي (انظر نظام توكن TOKAN في قسم ٥,٥,١). على النقيض من ذلك، فإن المدخلات الأكثر شيوعاً في الغالب عميقة، على سبيل المثال المُعَيْجِمَة والخصائص

[95، 96]. ومثال على التوليد من تمثيل سطحي يظهر في النظم الإحصائية التي تستهدف اللغة العربية، مثل الترجمة الآلية الإحصائية إلى اللغة العربية أو نمذجة للغة العربية. وقد تستخدم مثل هذه الأنظمة بعضاً من التقطيع للغة العربية التي تساعد نماذجها، ولكن المخرجات العربية المقطعة تحتاج إلى إعادة تركيبها (توليدها) إلى شكل سطحي غير مقطع [61، 99، 13].

- **التمثيل الداخلي (Internal Representation)**: يختلف التمثيل الداخلي للمعجم والقواعد اختلافاً كبيراً بين مختلف الأنظمة. على سبيل المثال، تستخدم بعض الأنظمة تمثيلاً بسيطاً عبارة عن بادئة- جذع- لاحقة (prefix-stem-suffix) [23، 96]، في مقابل تمثيل يعتمد على الجذور والأوزان واللواصق [86، 80]. حتى إن استخدام الأوزان قد تتفاوت، على سبيل المثال: الأوزان الصرفية التي تتطلب قواعد حتى تكون متصرفة تماماً أو الأوزان الألومورفية (allomorphic) التي تتطلب مدخلات معجمية أكثر تعقيداً [67]. يمكن أن يكون التمثيل الداخلي، إلى حدٍّ ما، مستقلاً عن التمثيل الخارجي. على سبيل المثال، يمكن للمعجم الذي يستخدم الجذوع في التمثيل الداخلي للبحث والتطابق أن يكون لديه جذراً ثابت الترميز ومعلومات لأوزان مرتبطة بكل جذع. وفي كثير من الأحيان يعتبر التمثيل الداخلي تمثيلاً غير مستقل (ملزم) بحيث يكون غير صالح خارج حدود النظام الذي يستخدمه. على هذا النحو، فإنه لا ينبغي استخدامها في ظل افتراضات تختلف صحتها. على سبيل المثال، استخدام معجم للجذوع كقاموس للتصحيح الإملائي هو استخدام غير صالح لهذا المورد، وذلك لكون كثير من مدخلاته عبارة عن إملاء جزئي للكلمات التي تجمع في داخل نظام التحليل/التوليد المستخدم للمعجم.
- **المحرك (Engine)**: استخدمت مجموعة متنوعة من الأطر ولغات البرمجة للتحليل

والتوليد، مع درجات متفاوتة من المتانة، والتعقيد والكفاءة. بعض من الحلول الأكثر تعقيدا، مثل استخدام [86] (finite state machinery، 80 تقايض التميز وإمكانية الرجوع بالوضع بانخفاض السرعة وكبر حجم النماذج مقارنة مع أبسط الحلول المستندة على البرمجة [23، 96، 67].

- **الاتجاهية (Directionality)**: تركز بعض الأنظمة على التحليل فقط، أو التوليد فقط بدلاً من الاثنين معاً. إلا أن هناك تقنيات معينة بطبيعتها قابلة للعكس مثل (finite state machinery)، ولكن البعض الآخر ليس كذلك، مثل الحلول المستندة على البرمجة. فإذا كان تحليل تمثيل الهدف سطحياً جداً، فإن التوليد لن يكون صعباً أو ذا مغزى.
- **التوسع (Extensibility)**: تتفاوت الطرق المختلفة للصرف في سهولة توسعها. فكلما كانت القواعد والمعجم ثابتة الترميز ومدمجة، كانت عملية التوسع معقدة. واحدة من التحديات الخاصة هو في توسيع أنظمة اللغة العربية المعاصرة للتعامل مع اللهجات العربية، التي تتطلب تحديثات على كل من القواعد والمعاجم.
- **الأداء وقابلية الاستخدام (Performance and Usability)**: هناك أبعاد كثيرة للحكم على الأداء وقابلية الاستخدام. التغطية، سواء من حيث التغطية المعجمية وتغطية الظواهر الصرفية، يعتبر مقياساً مهماً. وينبغي على نظم التحليل والتوليد إخراج تحليلات وتحققات (للنماذج المولدة) صحيحة، ولا شيء غير هذه المخرجات. فالدقة (precision)[١] المنخفضة أو الاسترجاع (recall)[٢] المنخفض غير

(١) في هذا السياق، يتم تعريف الدقة في نظام معين على أنه عدد التحليلات/التحققات الصحيحة التي ينتجها النظام مقسوماً على عدد جميع التحليلات/ التحققات التي أنتجت.

(٢) في هذا السياق، يتم تعريف الاسترجاع في نظام معين على أنه عدد التحليلات/التحققات الصحيحة التي ينتجها النظام مقسوماً على عدد جميع التحليلات/ التحققات في مرجعية التقييم.

مرغوب فيه في هذه الأنظمة. جانب آخر من جوانب الأداء هو المتانة ضد المدخلات غير الصحيحة أو الأخطاء الإملائية. بعض من أفضل المحللات تقترح تصحيحاً بديلاً كجزءٍ من التحليل. ويعتبر هذا أمراً ضرورياً للتعامل مع حالات الأخطاء الإملائية الشائعة، مثل أشكال الألف غير المهموزة. وأخيراً، فإن مسألة قابلية الاستخدام تعتمد في الواقع على التطبيق الذي يستخدم فيه المحلل/المولد. في بعض التطبيقات، نجد أن نظاماً ذا تغطية منخفضة لكن بعمق مناسب في الناتج الخارجي، يكون مرغوباً فيه أكثر من نظام ذي تغطية عالية ولكن المخرجات ضحلة أو غير ملائمة.

٥,٢,٢ باما: محلل باكوالتر الصرفي العربي

يستخدم محلل باكوالتر الصرفي العربي (باما)[1] الطريقة المعتمدة على المعجم المتسلسل، حيث تُبنى قواعد التهجئة والتكتيك الصرفي مباشرة في المعجم نفسه بدلاً من تحديدها من حيث القواعد العامة التي تتفاعل لتحقيق المخرجات [88، 23]. يتكون النظام من ثلاثة عناصر هي: المعجم، جداول التوافق والمحرك التحليلي.

المعجم (Lexicon): ينظر إلى الكلمة العربية باعتبارها سلسلة لثلاثة مناطق، وهي منطقة البادئة، ومنطقة الجذع، ومنطقة اللاحقة. ويمكن لمنطقة البادئة واللاحقة أن تكون ملغية. مدخلات المعجم البادئة واللاحقة تغطي كافة المتسلسلات الممكنة من البادئات واللاحقات العربية، على التوالي. لكل مدخل في المعجم، من فئة التوافق الصرفي، يتم تحديد المقابل الإنجليزي وأحياناً بيانات أقسام الكلام (POS). وتتجمع المدخلات المعجمية الجذعية حول معيجيماتها الخاصة التي لم يتم استخدامها في عملية

(١) اختصار Buckwalter Arabic morphological analyzer (Bama) (توضيح من المترجمة).

openkanal.com

التحليل. يوضح الشكل رقم (٥,١) عينة من المدخلات[١]: حيث تمثل القيم الستة الأولى في العمود الأيسر البادئات، والبقية في هذا العمود هي اللواحق، والعمود الأيمن يحتوي على سبعة جذوع تنتمي إلى ثلاثة مُعَيْجِمَات. وتحتوي المدخلات الجذعية أيضاً على مقابلاتها الإنجليزية، مما يسمح للمعجم لأن يعمل كقاموس. ومع ذلك، فإن وجود أشكال مصرفة، مثل النكرة والجمع بين هذه المقابلات يجعلها أقل صلاحية للاستعمال كمترجمات معجمية إنجليزية.

و/wa	Pref-Wa	*and*
بـ/bi	NPref-Bi	*by/with*
وبـ/wabi	NPref-Bi	*and + by/with*
الـ/Al	NPref-Al	*the*
بالـ/biAl	NPref-BiAl	*with/by + the*
وبالـ/wabiAl	NPref-BiAl	*and + with/by the*
ـة/ap	NSuff-ap	*[fem.sg.]*
ـتان/atAni	NSuff-atAn	*two*
ـتين/atayoni	NSuff-tayn	*two*
ـتاه/atAhu	NSuff-atAh	*his/its two*
ـات/At	NSuff-At	*[fem.pl.]*

;; 1_كتب/katab-u_1		
كتب/katab	PV	*write*
كُتب/kotub	IV	*write*
كتب/kutib	PV_Pass	*be written*
كُتب/kotab	IV_Pass_yu	*be written*
;; 1_كتاب/kitAb_1		
كتاب/kitAb	Ndu	*book*
كتب/kutub	N	*books*
;; 1_كتابة/kitAbap_1		
كتاب/kitAb	Nap	*writing*

الشكل رقم (٥,١). بعض من مدخلات قاعدة بيانات باكوالتر المعجمية.

جـــداول التوافـــق (Compatibility Tables): تحدد جداول التوافق الفئات الصرفية التي يسمح أن تتشارك في الظهور. على سبيل المثال، الفئة الصرفية لبادئة حرف العطف و/wa [Pref-Wa] متوافقة مع جميع فئات الاسم الجذعية وفئات الفعل التام الجذعية. ومع ذلك، فإن بادئة حرف العطف [Pref-Wa] غير متوافقة مع جذع الفعل المضارع لأن أحرف المضارعة في محلل باما يجب أن تحتوي على مورفيم بادئة

(١) تم الحفاظ على النقل الكتابي لباكوالتر في الأمثلة المأخوذة من مدخلات معجم باكوالتر (انظر الفصل الثاني).

الفاعل. وبالمثل، فإن جذع كلمة كتاب/ kitAb للمُعَيْجِمة كتاب_١/ kitAb_1 لديها فئة (Ndu)، وهي غير متوافقة مع فئة علامة التأنيث ة/ap [NSuff-ap].. يظهر نفس الجذع، كتاب/ kitAbكأحد الجذوع للمُعَيْجِمة كتابة_١/ kitAbap_1 مع فئة تتطلب لاحقة مع علامة التأنيث. مثل هذه الحالات شائعة جداً وتشكل تحدياً لاستخدام الجذوع كوحدة لفظية حيث يمكنها إضافة غموض غير ضروري.

المحرك التحليلي (Analysis Engine): تعتبر خوارزمية التحليل بسيطة إلى حدٍّ ما، بما أن كل القرارات الصعبة ثُبتت في المعجم وجداول التوافق، بمعنى أن الكلمات العربية تُقسم إلى كل المجموعات الممكنة من سلاسل (Strings) بادئة، وجذعية ولاحقة. ففي التجزئة الصحيحة، تكون السلاسل الثلاث موجودة في المعجم وأيضاً متوافقة من ثلاثة اتجاهات (بادئة–جذع (prefix-stem)، وجذع–لاحقة (stem-suffix)، وبادئ– لاحقة (prefix-suffix)). ينتج باما تحليلات متعددة عبارة عن صفوف (tuples) من التشكيل الكامل، والمُعجمة، وتحليل المورفيم وسمات المورفيم (وتسمى أيضاً سمات باكوالتر لأقسام الكلام، انظر الشكل رقم (٤,٥). على سبيل المثال، كلمة للكتب/llktb سوف يرجع تحليل يحدد تشكيله كـ (lilkutubi) ولمعيجمته (kitAb_1) أما بالنسبة للتحليل المورفيم والسمات المستخدمة فهي كالتالي (li/PREP+Al/DET+kutub/NOUN+i/CASE_DEF_GEN).

حاليا هناك ثلاثة إصدارات من باما، فإصدارة باما ١,٢/١,٠ مشاعة للعموم. أما باما إصدارة ٢,٠ وساما[1] إصدارة ٣,١/٣,٠ فهما متاحتان عن طريق LDC، انظر الملحق د لروابط لهذه المصادر.

---

(١) ساما (SAMA) اختصار (Standard Arabic Morphological Analyzer) وهي في الأساس بإما إصدارة ٣,١/٣,٠

٥,٢,٣ المرجانة (ALMORGEANA)[١]: المحلل والمولد الصرفي العربي المعتمد على المُعَيْجِمَة

المرجانة هو نظام للتحليل الصرفي والتوليدي بُني على قواعد بيانات باما /سما [88، 23][٢]. بخلاف باما، الذي ركز على تحليل لتمثيل سطحي شكلي (surfacy form based representation)، تقوم المرجانة بالتحليل إلى، والتوليد من المستوى الوظيفي (المُعَيْجِمَة والخاصية) من التمثيل. يستعرض الشكل رقم (٥,٢) خصائص مختلفة وقيمها الممكنة[٣]. بهذا المعنى، يقوم معجم المرجانة بتوسيع قواعد بيانات باما الصرفية بمُعَيْجِمَة وخصائص مفتاحية تستخدم في التحليل والتوليد. هذا العمل على المرجانة جعلها قريبة لملحقات نظام باما في الصرف الوظيفي، والمسمى ElixirFM [67] (انظر قسم ٥,٢,٥).

التحليل (**Analysis**): التحليل في المرجانة مشابه لباما. حيث تُقسم الكلمة إلى الشكل الثلاثي سابقة- جذع- لاحقة، وتفحص توافقيتها الثنائية وظهورها المفرد تجاه قاعدة بيانات باما. ويكمن الفرق في خطوة إضافية تستخدم المُعَيْجِمَة والخصائص المهمة المرتبطة بالجذع، وسلسلة اللاحق والسابق لبناء مخرجات الخصائص والمُعَيْجِمَة. على سبيل المثال كلمة (للكتب) (llktb) يسترجع التحليل التالي[٤]:

(٥,١) lilkutubi=[kitAb_1 POS:N l+ Al+ +PL +GEN]=books

---

(١) ARABIC LEXEME-BASED MORPHOLOGICAL GENERATION AND ANALYSIS

(٢) في مقالة سابقة منشورة عن المرجانة ركزت فيها على عنصر التوليد للنظام الذي كان اسمه Aragen [96].

(٣) تجدر الإشارة إلى أن النسخة الحالية للمرجانة لا تعالج تماماً الصرف الوظيفي، إنما لديها مزيج من الخصائص الوظيفية والشكلية.

(٤) المثال المستخدم في هذا الكتاب يعتمد على الإصدارة ٢,٠ من المرجانة.

openkanal.com

نجد هنا أن كلمة (lilkutubi) هو الصيغة المشكلة للكلمة. وبداخل الحواصر المربعة نجد المُعَيْجِمَة الاسمية (kitAb_1) وحرف الجر الزائد (+l) وأل التعريف (+Al) وخاصية (+PL) التي ترمز للجمع وأيضاً خاصية (+GEN) التي تعني حالة جر. تستنبط معظم المعلومات الموجودة في مجموعة الخصائص مباشرة من سمات المورفيم في مخرجات باما لنفس الكلمة كالتالي :

li/PREP+Al/DET+kutub/NOUN+i/CASE_DEF_GEN. إلا أن خاصية (+PL) التي ترمز للجمع لا تظهر في مخرجات باما. لأنها جزء من الإضافة التي عملت في المرجانة لمعالجة قاعدة بيانات باما.

التوليـــد (Generation) : في التوليد، تكون المدخلات عبارة عن المُعَيْجِمَة ومجموعة خصائص. أما المخرجات فتكون كلمةٍ مصرفةٍ ومشكلةٍ بالكامل. على سبيل المثال، [kitAb_1 POS:N l+ Al+ +PL+GEN] يولد كلمة (lilkutubi). عملية التوليد من المُعَيْجِمَة والخصائص مشابهة للتحليل إلا أن المُعَيْجِمَة ومفاتيح الخصائص تستخدم عوضاً عن سلاسل المتواليات (string sequences). أولاً، تُوسع مجموعة الخصائص لتشمل جميع أشكال الخصائص دون المحددة واللازمة، مثل الحالة، والعدد والجنس، إلخ. يليها، اختيار جميع المُعَيْجِمَات ومفاتيح الخصائص في معجم المرجانة التي تتطابق تماماً مع أي مجموعة فرعية من المُعَيْجِمَات ومجموعة الخصائص الموسعة. تُطابق كافة التركيبات من المفاتيح التي تغطي كامل المُعَيْجِمَات ومجموعة الخصائص الموسعة في الشكل الثلاثي (سابقة- جذع- لاحقة). ثم، يُحول كل مفتاح إلى مقابله من متوالية الجذع، أو متوالية السابقة أو متوالية اللاحقة. وتستخدم جداول التوافق نفسها المستخدمة في التحليل لقبول أو رفض الشكل الثلاثي (سابقة- جذع- لاحقة). وأخيراً، تُخرَّج وتُلصق كل الأشكال الثلاثية الفريدة المقبولة. وفي حال لم يُعثر على

openkandi.com

أي شكل سطحي، يُستكشف حل بديل يحاول إعادة التوليد بعد إهمال واحدة من الخصائص المدخلة.

انظر [97] لمزيد من التفاصيل حول المرجانة وتقييم أدائها. وللمعلومية فإن المرجانة هي المحلل/المولد المستخدم داخل مجموعة أدوات مدى (MADA)، التي سنناقشها بالتفصيل في القسم ٥,٥.

٥,٢,٤ مجيد (MAGEAD)[١]: المحلل والمولد الصرفي للعربية ولهجاتها

يعتبر مجيد محللاً ومولداً صرفياً لعائلة اللغة العربية، التي نعني بها اللغة العربية المعاصرة واللهجات المنطوقة. يربط مجيد (من كلا الجهتين) المُعَيْجِمَة ومجموعة من الخصائص اللغوية بكلمة سطحية عن طريق سلسلة من التحولات. من منظور توليدي، تترجم الخصائص لمورفيمات مجردة، ثم تُرتب وتُمثل على شكل مورفيمات مقيدة.

نوع الخاصية	القيم والتعريفات
أقسام الكلام Part-of-Speech	POS:N *Noun*, POS:PN *Proper Noun*, POS:V *Verb*, POS:AJ *Adjective*, POS:AV *Adverb*, POS:PRO *Pronoun*, POS:P *Preposition*, POS:D *Determiner*, POS:C *Conjunction*, POS:NEG *Negative particle*, POS:NUM *Number*, POS:AB *Abbreviation*, POS:IJ *Interjection*, and POS:PX *Punctuation*

الشكل رقم (٥,٢). خصائص المرجانة وقيمها المختلفة (الإصدارة ٢,٠). خصائص الزوائد مثل حروف العطف والجر اختيارية، ولكن بقية الخصائص إلزامية (مع أنه في بعض الحالات يعتمد على أقسام الكلام، مثال ذلك: الأسماء ليس لها زمن أو صِيغَة البناء للمَجهُول). أما الحالة فيُتعامل معها باستخدام خاصيتين هما: التعريف (definiteness) والملكية (possession).

(١) اختصار عبارة (MORPHOLOGICAL ANALYSIS AND GENERATION FOR ARABIC AND ITS DIALECTS) (توضيح من المترجمة).

نوع الخاصية	القيم والتعريفات
عطف/ربط Conjunction	w+ *'and'*, f+ *'so'*
حرف جر Preposition	b+ *'by, with'*, k+ *'like'*, l+ *'for, to'*
أداة الفعل Verbal Particle	s+ *'will'*, l+ *so as to*
أداة التعريف Definite Article	Al+ *the*
الجنس Gender	+FEM *Feminine*, +MASC *Masculine*
العدد Number	+SG *Singular*, +DU *Dual*, +PL *Plural*
الحالة Case	+NOM *Nominative*, +ACC *Accusative*, +GEN *Genitive*
التعريف Definiteness	+DEF *Definite*, +INDEF *Indefinite*
الملكية possession	+POSS *Construct state*, +NOPOSS *Not construct state*
زمن الفعل Verb Aspect	+PV *Perfective*, +IV *Imperfective*, +CV *Imperative*
صيغة البناء المجهول Voice	+ACT *Active*, +PASS *Passive*
الإعراب Mood	MOOD:I *Indicative*, MOOD:S *Subjunctive*, MOOD:J *Jussive*
الفاعل Subject المفعول به Object التملك Possessive	+S:PerGenNum +O:PerGenNum +P:PerGenNum **Per***son* = *{1,2,3}* **Gen***der* = *{M,F}* **Num***ber* = *{S,D,P}*

تابع الشكل رقم (٥,٢).

تُشبك المورفيمات الصيغية الملفوظة ويضاف لها اللواصق. وتُطبق قواعد منفصلة لإعادة الكتابة المورفوفونيميّة والهجائية. يتبع مجيد في تنفيذ قواعد إعادة الكتابة طريقة [100] في استخدام تمثيل يطلق عليه (multi-tape Finite-state transducer) (FST) ، المماثل

لتطبيقات أخرى معتمدة على FST موجه للصرف العربي [86]. استخدام قواعد لغوية واضحة داخل مجيد ميزها عن غيرها من التطبيقات الأكثر غموضاً مثل باما والمرجانة، التي تُثبت فيها الترميز بفعالية مع شكل الجذع. تجعل هذه الشفافية من مجيد نظاماً أكثر تعقيداً بطرق معينة، ولكنه في ذات الوقت يجعل من السهل توسعته ليستوعب لهجات جديدة. أيضاً يسمح هذا التمييز بين مستويات مختلفة من التمثيل من استخدام نظام مجيد لمجموعة متنوعة من المهام مثل عملية المقابلة من شكل هجائي إلى شكل صوتي. فيما تبقى من هذا القسم، سنناقش عناصر نظام مجيد بمزيد من التفاصيل وباستخدام مثال توضيحي.

المُعَيْجِمَة والخصائص: تمثل تحليلات نظام مجيد الصرفية على شكل مُعَيْجِمَة وخصائص. ويعرف نظام مجيد المُعَيْجِمَة على أنها شكل ثلاثي مكون من جذر، وفئة السلوك الصرفي ((morphological behavior class (MBC) ، ومؤشر للمعنى. فمن خلال النظر للمُعَيْجِمَة بهذا الشكل، يستطيع نظام مجيد أن يكون له تمثيل معتمد على المُعَيْجِمَة وأيضاً يعمل من دون معجم (في حال احتياجها عند التعامل مع لهجة). في الواقع، ولأن المُعَيْجِمَة لها هيكل داخلي، يمكن لنظام مجيد افتراض المُعَيْجِمَة بسرعة من دون الحاجة إلى عمل تخمينات معقدة. على سبيل المثال كلمة (ازدهرت) (Aizdaharat) لديها تحليل المُعَيْجِمَة والخصائص التالية في نظام مجيد:

(٥,٢) Root:zhr MBC:verb-VIII POS:V PER:3 GEN:F NUM:SG ASPECT:PERF

فئة السلوك الصرفي (Morphological Behavior Class): تقوم فئة السلوك الصرفي بمقابلة مجموعة الأزواج اللغوية (الخاصية والقيمة) بمجموعة المورفيمات المجردة. على سبيل المثال، فئة السلوك الصرفي (VERB-VIII) تقابل أزواج الخاصية والقيمة (ASPECT:PERF) لجذر المورفيم المجرد [PAT_PV:VIII]، الذي يتوافق في اللغة العربية

المعاصرة مع مورفيم الجذر المادي V1tV2V3 ، بينما فئة السلوك الصرفي (VERB-VII) تقابل أزواج الخاصية والقيمة (ASPECT:PERF) لجذر المورفيم المجرد [PAT_PV:VII]، الذي يتوافق في اللغة العربية المعاصرة مع مورفيم الجذر المادي V1tV2V3 ، تُعرف فئات السلوك الصرفي باستخدام التمثيل الهرمي مع توارث غير مطَّرِد (non-monotonic inheritance). يسمح التسلسل الهرمي لنظام مجيد بتحديد ولمرة واحدة تلك المقابلات بين الخصائص والمورفيمات لجميع فئات السلوك الصرفي التي تشترك معها. على سبيل المثال ، رأس التسلسل الهرمي من فئة السلوك الصرفي هي كلمة ، وتتشارك جميع الكلمات العربية في تقابلات معينة ، مثل تلك من الخاصية اللغوية (conj:w) للزائدة (+w).. مما يعني أن جميع الكلمات العربية يمكنها أخذ زائدة للعطف. بالمثل ، ضمائر المفعول المتصلة (object pronominal clitics) هي نفسها بالنسبة لجميع الأفعال المتعدية ، بصرف النظر عن شكل وزنها الصيغي. يفترض تصميم نظام مجيد أن التسلسل الهرمي لفئة السلوك الصرفي هي بدائل مستقلة ، بمعنى أنها مستقلة في اللهجة أو اللغة العربية المعاصرة. على الرغم من إضافة خيارات أكثر فإنه سيكون هناك حاجة إلى بعض التعديلات.

المورفيمات (**Morphemes**): لإبقاء التسلسل الهرمي لفئة السلوك الصرفي بديلاً مستقلاً ، يستخدم نظام مجيد تمثيل بديل مستقل للمورفيمات التي يقابلها في التسلسل الهرمي لفئة السلوك الصرفي. ويرمز لهذه المورفيمات بالمورفيمات المجردة ((abstract morphemes (AMs). بعدها تُرتب المورفيمات المجردة بترتيب سطحي لمقابلها من المورفيمات المادية. تحدد ترتيب المورفيمات المجردة بقواعد نحوية مستقلة وخالية من السياق. إذا حاولنا الإنشاء من المثال (٥,٢) ، سنحصل على ما يلي في هذه النقطة :

(٥,٣) [Root:zhr][PAT_PV:VIII][VOC_PV:VIII-act] + [SUBJSUF_PV:3FS]

كما نلاحظ أنه لم يُرتب الجذر والوزن والإعلال بالنسبة لبعضها مع بعض، إنما وضعت ببساطة بعضها بجانب بعض. ترمز علامة "+" إلى ترتيب المورفيمات اللصقية. الآن فقط يمكن ترجمة المورفيمات المجردة إلى مورفيمات مادية (concrete morphemes CMs) التي يتم تسلسلها بترتيب محدد، ليصبح مثالنا كالتالي:

<zhr,V1tV2V3,iaa> +at (٥,٤)

وينتج عن التشابك البسيط للجذر والوزن والإعلال الشكل (iztahar+at). هذا الشكل غير صحيح حيث لم تُطبق أية قواعد صرفية حتى الآن.

القواعد (Rules): لدى نظام مجيد نوعان من القواعد: قواعد مورفوفونيميّة/صوتية تقابل من التمثيل الصرفي إلى التمثيلات الصوتية والهجائية. وقواعد هجائية تعيد كتابة التمثيل الهجائي. وهذه تتضمن على سبيل المثال، قواعد استخدام الشدة. ولمثالنا السابق نحصل على */izdaharat/* على المستوى الصوتي (انظر قسم ٤,٢,٤). باستخدام التهجئة المشكلة للغة العربية المعاصرة المعيارية، يصبح مثالنا (Aizdaharat). وبحذف التشكيل تحول الكلمة إلى صيغتها المألوفة ازدهرت (Azdhrt). لاحظ أنه في وضع التحليل، يخمن نظام مجيد جميع التشكيلات المحتملة (عددها محدود حتى مع التجميع) ويعمل التحليل على ناتج المسار المتعدد المؤتمت (multi-path automaton).

للمزيد حول نظام مجيد انظر [92، 80، 101].

٥,٢,٥ إلكسر إف إم (ELIXIRFM): نظام إلكسر (ELIXIR) العربي للصرف الوظيفي

نظام إلكسر للصرف الوظيفي (ElixirFM) هو تنفيذ عالِ المستوى للصرف

الـوظيفي في العربيـة [67، 102]. اسـتلهم فكرتـه مـن منهجيـة الصـرف الـوظيفي [03]، واعتمدت في البداية على معجم باكوالتر المعاد معالجته [88].

التكتيكـــات الصـــرفية (Morphotactics): بالإضـافة إلى اسـتخدام القواعـد المورفوفونيميّة المتنوعـة في حـدود اللواصـق (مثـل أشـكال التـاء المربوطـة)، واحـدة مـن التجريدات المميزة في نظام (ElixirFM) هو أن أشكال الكلمة يتم ترميزهـا بواسـطة وزن مورفوفونيميّ مصمم بدقة يجمع مع الجذور أو جذوع الكلمـة. هـذه الأوزان في الغالـب ألوفورميـة (allomorphic)، بمعنـى أنهـا ترمـز تـأثير تفاعـل النـوع والجـذر مـع الصـيغة المورفيميـة. علـى سـبيل المثـال، الـوزن والجـذر لكلمـة ميـزان (miyzAn) هـو (wzn + MICAL) بترميز نظام (ElixirFM) أو (wzn + miy2A3) برمز سطحي، في مقابـل الـوزن المورفيمي الذي يدل على أدوات (wzn + mi12A3). وفقاً للتصـميم، يتجنـب هـذا كـل مـن (أ) تعريـف قاعـدة لتحويـل (*miwzAn*) لـ (*miyzAn*) كمـا في مجيـد (انظـر قسـم ٥,٢,٤) و (ب) سـرد الشـكل السـطحي لكـل عنصـر معجمـي كمـا في باما (انظـر قسـم ٥,٢,٢). بالإضافة إلى هذه الحـالات، توجـد بعـض القواعـد لمعالجـة التحـولات العاديـة للجذر والوزن في نظام (ElixirFM)، مثال ذلك إدغام (*t*) في أفعال (Form-VIII) (انظـر قسم ٤,٢,٤). على سبيل المثال، الفعل ازدهرت (*Aizdaharat*) التي اسـتخدمت سـابقاً سيحصل لها التحليل التالي في نظام (ElixirFM):

```
["prosper","flourish"]
Verb [ ] [ ] [ ] [VIII]
izdahar "z h r" IFtaCaL
VP-A-3FS-- izdaharat "z h r" IFtaCaL |<< "at"
```

(٥,٥)

السـطر الأول هـو المقابـل الإنجليـزي. السـطر الثـاني يلخـص المعلومـات المتعلقـة بالمـدخلات المعجميـة للمُعَيْجِمَـة. في هـذه الحالـة، الحواصـر الـثلاث المربعـة بعـد الفعـل "Verb" ستدرج أي جذع يعتمد بشكل معجمي أو استثنائي فعل ماضٍ ومضارع وأمـر،

لكنها في مثالنا فارغة لأن هذه المعلومات يستدل عليها داخلياً من قبل نظام (ElixirFM). أما [VIII] الأخيرة فتمثل بشكل صريح أن الوزن (*IFtaCaL*) ينتمي لفئة (Form VIII) الاشتقاقية. السطر الثالث تبين المادة المعجمية والجذر ووزن المادة المعجمية (الذي صادف أن تكون نفس وزن الكلمة التي حُللت في هذا المثال). السطر الأخير يبين أقسام الكلام (انظر قسم ٥,٤,٥) والشكل الصوتي للكلمة والجذر والوزن واللاحقة. لاحظ أن الوزن المرتبط بالفعل لديه t غير مدغمة.

علم الأصوات والتهجئة (Phonology and Orthography): ميزة أخرى فريدة في نظام (ElixirFM) في أنه يمثل داخلياً وحداته المعجمية بتمثيل صوتي، وتُحول بعد ذلك إلى سلسلة من الأحرف في رموز ArabTEX الموسعة [16]. ويمكن تحويل الرموز بعد ذلك إلى كتابة هجائية أو صوتية. مما يسمح لنظام (ElixirFM) من تجنب تعريف القواعد الهجائية ويقوم في الأساس بفصل علم الأصوات من التهجئة بطريقة مشابهة لنظام مجيد (انظر قسم ٥,٢,٤).

تقطيع النص (Tokenization): وأخيراً، يحتوي التمثيل الخارجي لنظام (ElixirFM) على قرار بسيط لتقطيع النص يتبع طريقة تمثيل بنك بنسلفانيا الشجري للتحاليل النحوية وبنك براغ للتحاليل الشجرية النحوية التبعية للغة العربية (Prague Arabic Dependency Treebank) (انظر قسم ٦,٢). كل وحدة لفظية تستقبل أقسام الكلام الخاصة بها وتحليلها المنفصل. على سبيل المثال، بتكملة تحليل نظام (ElixirFM) لكلمة للكتب (*lilkutubi*) التي ناقشناها سابقاً، يصبح المثال كالتالي:

obeikandi.com

```
["for","to"]
   Prep  []
   li  "l"  "li"

   ["book"]
   Noun  [FuCuL]  []
   kitAb  "k t b"  FiCAL

   li-al-kutubi

   P--------- li  "l"  "li"
   N------P2D  al-kutubi  "k t b"  al >| FuCuL |<< "i"
```

(٥,٦)

كتب قلب نظام (ElixirFM) باستخدام لغة البرمجة الوظيفية هاسكل (Haskell)، في حين تم كتابة واجهات دعم تحرير المعجم، والتفاعلات الأخرى باستخدام لغة بيرل (Perl). انظر ملحق **د** لروابط لنظام (ElixirFM) وواجهته على الويب.

٥,٣ تقطيع النص

الحكمة الشائعة في معالجة اللغة الطبيعية هو أن تقطيع النص للكلمات العربية من خلال التجريد وتخفيض تقليص الاحتمالات الهجائية مفيد للعديد من التطبيقات مثل نمذجة اللغة واسترجاع المعلومات والترجمة الآلية الإحصائية. يقوم تقطيع النص وتقليص الاحتمالات بالحد من التناثر (sparsity) والتعقيد (perplexity)، وخفض عدد الكلمات غير المعروفة (OOV).

obeikandl.com

٥,٣,١ مخططات وتقنيات تقطيع النص

سنميز بين مخططات تقطيع النص وتقنيات تقطيع النص [83]. يعرف المخطط ما هو هدف تقطيع النص، في حين أن التقنية توضح كيفية تنفيذه. وتختلف مخططات تقطيع النص من ناحيتين: ما سيتم تقسيمه (تجزئته)، والشكل الذي تُمثل فيه مختلف الأجزاء المقسمة (المراد تنظيمها). وهناك عدد كبير جدًا من مخططات تقطيع النص الممكنة. ففي سياق استرجاع المعلومات يسمى تقطيع النص في الغالب تجذيعاً stemming [104]. ففي عملية التجذيع تُحذف الزوائد المفصولة والمورفيمات غير الأساسية.

قد تتراوح تقنيات تقطيع النص ما بين تقنية سهلة، مثل استخدام التعابير النمطية الجشعة (greedy regular expression)، إلى استخدام تقنية معقدة تتطلب تحليلاً صرفياً وفكاً للبس (انظر قسم ٥,٥). وبما أن الغموض الصرفي في اللغة العربية متفشٍ، فإنه كلما كان المخطط أكثر تعقيداً كان من الصعب التقطيع بسياق صحيح. وقد تبين أن التقنيات الأكثر تعقيداً كانت الأكثر فائدة [83، 105]، ومع ذلك، تجدر الإشارة إلى أنه في سياقات معينة (الأقل هو الأكثر): على سبيل المثال، الترجمة الآلية الإحصائية القائمة على العبارة (phrase-based SMT) تستفيد فقط من تقطيع النص المعقد مع قليل من البيانات للتدريب حيث يعتبر التناثر مشكلة كبيرة. وكلما أُدخل مزيد من البيانات للتدريب، يبدأ فعلياً تقطيع النص المعقد بالإيذاء مقارنة بالتقطيع الأبسط [83، 105].

٥,٣,٢ تركيب النص (Detokenization)

في سياقات معينة، وعندما تكون اللغة العربية هي اللغة المخرجة، يفضل إنتاج لغة عربية سليمة هجائياً؛ بمعنى أن الكلمات المقطعة والمقلصة احتمالاتها الهجائية من الأفضل تركيبها وإثراؤها (تصحح هجائياً). كمثال على ذلك، من المتوقع وعلى نحو

معقول أن تنتج نظم الترجمة الآلية من الإنجليزية إلى العربية لغة عربية سليمة بغض النظر عن عملية التجهيز المستخدمة لتحسين أداء المترجم الآلي. وأي شيء أقل سيكون مشابه لإنتاج نص بحروف صغيرة بالإنجليزية أو نص غير مشكل أو غير متصل بالفرنسية. قد تكون مهمة تركيب النص ليست سهلة لأن هناك تعديلات صرفية عدة ينبغي تطبيقها في العملية [99، 13، 106]. ومن الواضح أنه كلما كان تقطيع النص أكثر تعقيداً، كان تركيب النص أصعب.

٥,٣,٣ المخططات المختلفة لتقطيع النص

عند الحديث عن تقطيع النص، لابد من التذكير بأنه لا توجد طريقة مثالية لتقطيع النص، فما هو مناسب لنظم استرجاع المعلومات، قد لا يكون مناسباً لنظم الترجمة الإحصائية الآلية. أيضاً ما هو مناسب لنوع معين من الترجمة الإحصائية الآلية قد لا يكون مناسباً لنوع آخر. الاتساق في طريقة التنفيذ أمر مرغوب فيه وغالباً ما توضع قيود على ما يمكن استخدامه من المكونات المختلفة. على سبيل المثال، معظم برامج المحللات التركيبية الجاهزة للاستخدام للغة العربية تستخدم مقطع النص الموجود في بنك بنسلفانيا الشجري للتحاليل النحوية. أما نظام للترجمة الآلية الإحصائية باستخدام محلل أوتوماتيكي عليه التأكد من أن مقطع النص الداخلي متوافق مع المحلل أو على الأقل يعالج هذه المشكلة بطريقة أخرى.

فيما يلي وصف لبعض المخططات الشائعة لتقطيع النص [83، 105، 107، 99، 13]. وهي ليست مجموعة كاملة، إنما تهدف إلى توضيح التنوع. وانظر للشكل رقم ٥,٣ للحصول على مثال لمقارنة مقطعات النص.

obeikandl.com

- **مُقطع النص البسيط (ST- Simple Tokenization)**: يعتبر مقطع النص البسيط أساس المخططات التجهيزية. حيث يقتصر على تقسيم علامات الترقيم والأرقام من الكلمات. على سبيل المثال، آخر مسافة غير بيضاء للمتوالية الواردة في الجملة في الشكل رقم ٥,٣ "trkyA." قسمت إلى جزأين هما: "trkyA" و ".". مثال آخر على تقسيم الأرقام من الكلمات هو حالة حرف العطف و+ +w الذي يمكن أن يتقدم الأرقام مثلاً عندما توصف قائمة من الأرقام: و١٥ W15. هذا المخطط لا يتطلب فكًّا للبس. وعادة ما تُزال أي علامات تشكيل تظهر في المخطط. ويستخدم هذا المخطط عادة كمدخل لإنتاج مخططات أخرى. حيث يشار أحياناً إلى تقطيع النص هذا بالرمز **D0** (عدم التجريد).
- **تقليص الاحتمالات الهجائية (ON- Orthographic Normalization)**: يعالج تقليص الاحتمالات الهجائية موضوع الإملاء دون المستوى الأمثل في اللغة العربية بجعل الخيارات متسقة. ويكون عادة تقليص الاحتمالات عملية مخفضة (RED) بمعنى أنه يدمج أشكالاً متعددة في شكل واحد، مثل الأشكال المختلفة للألف المهموزة والألف المقصورة تُقلص إلى ألف مجردة وياء منقوطة. أما في تقليص الاحتمالات المدعم (ENR)، فيُحدد شكل السياق المناسب لهذه الحروف [13]. مثال على تقليص الاحتمالات الهجائية يمكن أن يشاهد في إملاء الحرف الأخير للكلمات الأولى والخامسة في المثال المذكور في الشكل رقم (٥,٣) (wsynhý وAlý). أي نوع من تقليص الاحتمالات يمكن تطبيقه من حيث المبدأ على أي مخطط لتقطيع النص.

obeikandl.com

	وسينهى الرئيس جولته بزيارة الى تركيا.						
Input (ST/D0)	wsynhý	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
Gloss	and will finish	the president tour his		with visit	to	Turkey	.
English	The president will finish his tour with a visit to Turkey.						
Scheme							
ON_{Enr}	wsynhy	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
ON_{Red}	wsynhy	Alrŷys	jwlth	bzyArħ	Ăly	trkyA	.
D1	w+ synhy	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
D2	w+ s+ ynhy	Alrŷys	jwlth	b+ zyArħ	Ălý	trkyA	.
D3/S1	w+ s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
S2	w+s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
WA	w+ synhy	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
TB	w+ s+ ynhy	Alrŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
TB_{old}	w+ synhy	Alrŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
MR	w+ s+ y+ nhy	Al+ rŷys	jwl +ħ +h	b+ zyAr +ħ	Ălý	trkyA	.
LEM	Ânhý	rŷys	jwlħ	zyArħ	Ălý	trkyA	.
LEM+TB	w+ s+ Ânhý	rŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
ENX	w+ s+ $Ânhý_{VBP}$ $+S_{3MS}$	$Al+ rŷys_{NN}$	$jwlħ_{NN}$ +h	$b+ zyArħ_{NN}$	$Ălý_{IN}$	$trkyA_{NNP}$	.

الشكل رقم (٥,٣). مثال على مخططات مختلفة لتقطيع النص: يرمز (ST/D0) لتقطيع النص البسيط، و(ON*Enr*) تقليص الاحتمالات الهجائية المدعم و(ON*Red*) تقليص الاحتمالات الهجائية المخفض، و(D1, D2, D3/S1,S2) تمثل الدرجات المختلفة للتجريد، و(WA) تمثل و+ التجريد، و(TB) و(TB*old*) تمثل تقطيع النص الشجري الجديد والقديم على التوالي، و(MR) تمثل المورفيمات، و(LEM) تمثل تعيين المادة المعجمية، و(LEM+TB) تمثل تعيين المادة المعجمية مع تقطيع النص الشجري الجديد، و(ENX) تمثل تقطيعاً للنص مماثلة لـ(D3+LEM+POS) مع علامات لفعل الفاعل.

- **التجريد (D1, D2, D3)**: بدرجات ١ ، ٢ ، ٣ عبارة عن مخططات تقوم بفصل الزوائد. يقوم **D1** بفك فئة زوائد العطف (و+/*w*+ ف+/*f*+) وزائدة الاستجواب النادرة. يقوم **D2** بعمل **D1** بالإضافة إلى فصل فئة الأدوات (ل+/*l*+ ، ك+/*k*+ ، ب+/*b*+ ، س+/*s*+). وأخيراً، يقسم **D3** ما يقوم به **D2** بالإضافة إلى أل التعريف (*Al*+) وجميع الضمائر المتصلة.

- **تجريد حرف العطف و+/*w*+ (WA):** شبيهة بعمل **D1** لكن من دون أخذ ف+/ *f*+ في عين الاعتبار. وقد ذكر أن مقطع النص البسيط مثالي للترجمة الآلية الإحصائية مع مجموعات البيانات الكبيرة جداً [98].
- **مقطع النص من بنك بنسلفانيا الشجري للتحاليل النحوية (TB)**: يستخدم نفس مخطط تقطيع النص المستخدم في البنك الشجري العربي [9]. وهذا شبيه بـ**(D3)** ولكن من دون فصل أل التعريف. وفي نسخة قديمة من **TB** لم يكن يفصل أداة المستقبل س+/*s*+ .
- **يستخدم مخططات (S1) و(S2) بواسطة** [99]. وفي الأساس **(S1) و (S2)** هما مثل **D3**. حيث يقوم **S2** بربط مختلف الزوائد في بداية الكلمة في متوالية واحدة.
- **المورفيمات (MR)**: هذا المخطط يقسم الكلمات إلى جذع ولواصق مورفيمية. وهو مطابق لمقطع النص الأولي المستخدم من قبل [108].
- **المواد المعجمية (LEM)**: يعمل هذا المخطط على تقليص كل كلمة لمادتها المعجمية. ويمكن استخدام المواد المعجمية مع مخططات تقطيع نص أخرى حيث إنها تستخدم في كل وحدة لفظية مقسومة، انظر **LEM+TB** في الشكل رقم (٥,٣).
- **تقطيع النص المشابه للإنجليزية (ENX) المستخدمة في** [105]. ويهدف هذا المخطط إلى تقليل الفوارق بين العربية والإنجليزية. ويقوم بالتجريد مثل **D3** ولكن يستخدم سمات أقسام الكلام والمواد المعجمية عوضاً عن الكلمات المنشأة. خفضت مجموعة سمات أقسام الكلام المستخدمة في نظام (Bies) من مجموعة سمات البنك الشجري العربي (انظر قسم ٥,٤,٢) [9، 109]. بالإضافة إلى أن تصريف الفاعل يشار إليه صراحة كوحدة لفظية مستقلة. ومن الواضح، أن الكثير من الاختلافات الأخرى ممكنة هنا.

openkandil.com

٥,٤ تعيين أقسام الكلام

تعيين أقسام الكلام (Part-of-Speech (POS) tagging)، هي مهمة تعيين سمة صرف- نحوية ملائمة من حيث السياق على كل كلمة في جملة. ومن حيث المبدأ، ينبغي أن تُختار السمات من مجموعة السمات الشاملة والمعرفة بشكل جيد. وبسبب غنى اللغة العربية صرفياً، يمكن أن تكون سمات أقسام الكلام للغة العربية كبيرة جداً. لذا يفضل كثير من الباحثين العاملين في مجال معالجة اللغة العربية العمل على مجموعات أصغر حجماً. فالحجم والتفصيل لمجموعة سمات أقسام الكلام للغة العربية قد تتفاوت بشكل واسع. من جهة، نجد أن تصنيف قواعد اللغة العربية التقليدية لأقسام الكلام عبارة عن تمييز ثلاثي إلى فعل واسم وحرف. هذا التصنيف غير دقيق (عام)، وغالباً لا يستخدم حاسوبياً. من جهة أخرى، نجد أن الشكل الكامل (غير المقطع) لمجموعة سمات بكوالتر والمعتمدة على المورفيمات العربية يمكن أن تصل نظرياً إلى أكثر من ٣٣٠ ألف سمة. ويتفاعل حجم مجموعة السمات أيضاً مع النص من حيث كونه مقطعاً أو لا (وبأي مخطط لتقطيع النص). ومن حيث المبدأ، سمة أقسام الكلام الموضوعة على كلمة غير مقطعة مساوية لربط سمات أقسام الكلام بوحدتها اللفظية.

وعلى الرغم من أن المجموعات الكبيرة هي أكثر اكتمالاً ويمكن أن تخدم الأداء بشكل أفضل للعمليات العليا (تحت ظروف ذهبية)، إلا أنها تميل إلى صعوبة التكهن بها جيداً [110]. أما مجموعة السمات المقلصة فيمكن التنبؤ بها بدقة، وقد أثبتت فائدتها لعدة تطبيقات لمعالجة اللغة الطبيعية [111]. واحدة من مجموعة السمات المقلصة موجودة في برنامج كاتب (CATiB)، ويقال إنها خففت من حمل التوسيم اليدوي للبنك الشجري [112]. ولا توجد مجموعة سمات لأقسام الكلام شاملة ومثالية. فالتطبيقات المختلفة تحتاج إلى مجموعات مختلفة من السمات.

في بقية هذا الجزء، سنستعرض سبع مجموعات لسمات في اللغة العربية مع درجات مختلفة من التفاصيل. وتستخدم مجموعة السمات هذه في مختلف الموارد المتوفرة. وتعرض مقارنة بين مجموعة السمات في مثال واحد في الشكل رقم ٥,٦. وقد ضمن تحليل المرجانة المستخدم في مدى في هذه المقارنة (ارجع لقسم ٥,٥,١) الذي يمكن التفكير فيه باعتباره سمة أخرى لأقسام الكلام.

إن تقنيات تعيين أقسام الكلام والمطورة للغات أخرى، يمكن استخدامها للغة العربية بنفس الطريقة. وقد أجريت الكثير من الأعمال في هذا المضمار [113، 109، 51، 114، 115، 116، 117]. وسنقدم منهجين في القسم ٥,٥.

٥,٤,١ مجموعة سمات باكوالتر

مجموعة سمات بكوالتر التي طورها تيم بكوالتر (Tim Buckwalter) عبارة عن مجموعة سمات شكلية يمكن استخدامها للنص المقطع (tokenized) وغير المقطع (untokenized). سمات النص غير المقطع هي ما أنتجت بواسطة باما (انظر قسم ٥,٢,٢) أما السمات للنص المقطع فتستخدم في بنك بنسلفانيا الشجري للتحاليل النحوية (PATB) (انظر قسم ٦,٢,١). وتستمد السمات المتغيرة للنص المقطع من السمات للنص غير المقطع، ومجموعة كلا الخيارين تستخدم حوالي ٧٠ رمزاً أساسياً من رموز السمات الفرعية (مثل رمز المعرف (DET) ورمز اللاحقة الاسمية (NSUFF) ورمز الصفة (ADJ) ورمز المنصوب (ACC) [82]. انظر الشكل رقم ٥,٤[1]، تجمع هذه

(١) الشكل رقم ٥,٤ يحتوي على بعض المعلمات التي سنحددها هنا:

<PGN> person-gender-number (شخص - جنس - عدد), <GN> gender-number (شخص - عدد)

person (شخص): 1 first, 2 second, 3 third, φ unspecified (غير محدد)

= Gender (جنس): M masculine (مذكر), F feminine (مؤنث), φ unspecified (غير محدد)

openkanal.com

السمات الفرعية لتكوَّن ما مجموعه حوالي ١٧٠ سمة مورفيمية (١٣٥ في PATB إصدارة 1v2.0 و ١٦٩ في PATB إصدارة 3v3.1)، ، مثال على ذلك رمز (NSUFF_FEM_SG) للدلالة على علامة المفرد المؤنث ورمز (CASE_DEF_ACC) للدلالة على حالة النصب. ويتم بناء سمات الكلمات من واحدة أو أكثر من السمات المورفيمية، مثال ذلك DET+ADJ+NSUFF_FEM_SG+CASE_DEF_ACC لكلمة الجميلة (Aljmylħ) في الشكل رقم ٥,٦.

تختلف السمات للنص المقطع وغير المقطع في عدد السمات الفرعية التي يمكن جمعها. على سبيل المثال، في مجموعة سمات بكوالتر في (PATB)، سمتي (CONJ) و(PRON) لا تستخدم في سمات الوحدة اللفظية بل توسم مستقلة بذاتها. قد تصل مجموعة سمات بكوالتر للنص غير المقطع إلى آلاف السمات، بينما قد تصل مجموعة سمات بكوالتر للنص المقطع إلى حوالي ٥٠٠ سمة. وهناك عدة أشكال لمجموعة السمات هذه مستخدمة في إصدارات مختلفة من محلل باما/ساما وفي مختلف الإصدارات من (PATB).

٥,٤,٢ مجموعة سمات بكوالتر المختصرة (BIES, KULICK, ERTS)

تعتبر مجموعة سمات بكوالتر غنية جداً لكثير من المشكلات والمنهجيات

Number(عدد): S singular (مفرد) D dual (مثنى) P plural (جمع) 0 unspecified (غير محدد) =

<Mood>: I indicative, S subjunctive, J jussive, SJ subjective/jussive
<Gen>: _MASC masculine, _FEM feminine
<Num>: _SG singular, _DU dual, _PL plural
<Cas>: _NOM nominative, _ACC accusative, _GEN genitive, _ACCGEN accusative/genitive, φ unspecified
<Stt>: _POSS construct/possessor, φ not construct
<Def>: _DEF definite, _INDEF indefinite

Obeikandl.com

الحاسوبية. وقد طورت العديد من مجموعات السمات لتقليص حجمها بشكل قابل للسيطرة. وتعتبر مجموعة سمات كاتب (CATiB tag set) - التي ناقشناها سابقاً- شكل متطرف من التقليص لمجموعة سمات بكوالتر مقارنة مع مجموعة السمات الثلاث التي سنناقشها هنا.

مجموعة سمات بييز (The Bies Tag Set)

طورت مجموعة سمات بييز بواسطة آن بييز (Ann Bies) و دان بايكل (Dan Bikel) كبديل مقلص للسمات العربية وفي مجموعة أصغر من السمات (حوالي أكثر من ٢٠ [109]، تستخدم ٢٤ سمة مختلفة) المستوحاة من مجموعة سمات أقسام الكلام لبنك بنسلفانيا الشجري للتحاليل النحوية للغة الإنجليزية [118]. وعلى الرغم من أن هذه المجموعة كانت تجريبية، إلا أنها استخدمت على نطاق واسع لتعيين أقسام الكلام في اللغة العربية [109، 51، 119]. وتعتبر مجموعة السمات هذه غير وافية لغوياً (linguistically coarse) لأنها تتجاهل فروقاً كثيرة في اللغة العربية، على سبيل المثال: يستخدم رمز (JJ) مع جميع الصفات بغض النظر عن تصريفها (من الواضح، أن السمة الإنجليزية لا يوجد لها تصريف).

الفعلية (Verbs)	
VERB	verb
PSEUDO_VERB	pseudo-verb
PV	perfective verb
PV_PASS	perfective passive verb
PVSUFF_DO:<PGN>	direct object of perfective verb
PVSUFF_SUBJ:<PGN>	subject of perfective verb
IV	imperfective verb
IV_PASS	imperfective passive verb
IVSUFF_DO:<PGN>	imperfective verb direct object
IV<PGN>	imperfective verb prefix
IVSUFF_SUBJ:<PGN> _MOOD:<Mood>	imperfective verb subject and mood suffix
CV	imperative (command) verb
CVSUFF_DO:<PGN>	imperative verb object
CVSUFF_SUBJ:<PGN>	imperative verb subject
الأدوات (Particles)	
PREP	preposition
CONJ	conjunction
SUB_CONJ	subordinating conjunction
PART	particle
CONNEC_PART	connective particle
EMPHATIC_PART	emphatic particle
FOCUS_PART	focus particle
FUT_PART	future particle
INTERROG_PART	interrogative particle
JUS_PART	jussive particle
NEG_PART	negative particle
RC_PART	response conditional particle
RESTRIC_PART	restrictive particle
VERB_PART	verb particle
VOC_PART	vocative particle

الإسمية (Nominals)	
NOUN	noun
NOUN_NUM	nominal/cardinal number
NOUN_QUANT	quantifier noun
NOUN.VN	deverbal noun
NOUN_PROP	proper noun
ADJ	adjective
ADJ_COMP	comparative adjective
ADJ_NUM	adjectival/ordinal number
ADJ.VN	deverbal adjective
ADJ_PROP	proper adjective
ADV	adverb
REL_ADV	relative adverb
INTERROG_ADV	interrogative adverb
PRON	pronoun
PRON_<PGN>	personal pronoun
POSS_PRON_<PGN>	possessive personal pronoun
DEM_PRON_<GN>	demonstrative pronoun
REL_PRON	relative pronoun
INTERROG_PRON	interrogative pronoun
NSUFF<Gen><Num><Cas><Stt>	nominal suffix
CASE<Def><Cas>	nominal suffix
DET	determiner
أخرى (Others)	
PUNC	punctuation
ABBREV	abbreviation
INTERJ	interjection
LATIN	latin script
FOREIGN	foreign word
TYPO	typographical error
PARTIAL	partial word
DIALECT	dialectal word

الشكل رقم (٤,٥). أجزاء مجموعة سمات بكوالتر. انظر الحاشية رقم ٦٧.

بالطبع ، نظراً للطبيعة المعقدة لاتفاق قواعد اللغة العربية (انظر قسم ٦,١,٣)، قد يكون هذا كافياً ما لم يتم استخدام نموذج أفضل من ذلك بكثير. مثال آخر هو استخدام سمات الجمع ليقصد بها كل من الجمع والمثنى. كما تم الإشارة لمجموعة السمات هذه على أنها مجموعة مقلصة من السمات [120] Reduced Tag Set (RTS) وأيضاً مجموعة سمات (PennPOS). وفيما يلي السمات الموجودة في هذه المجموعة :

العبارات الاسمية:

- **الأسماء (Nouns)**: الرمز (NN) يعني اسماً شائعاً مفرداً أو اختصاراً، الرمز (NNS) يعني اسماً شائعاً جمعاً أو مثنى، الرمز (NNP) يعني اسماً علماً مفرداً، الرمز

(NNPS) يعني اسماً علماً جمعاً أو مثنى.

- **الضمائر (Pronouns)**: الرمز (PRP) يعني الضمير الشخصي، الرمز ($PRP) يعني الضمير الشخصي للملكية، الرمز (WP) يعني اسماً موصولاً.
- **أخرى**: الرمز (JJ) يعني صفة، الرمز (RB) يعني حالاً، الرمز (WRB) يعني الحال الموصول أو الظرف، الرمز (CD) يعني عدداً أصلياً، رمز (FW) يعني كلمة أجنبية.
- **الأدوات (Particles)**: الرمز (CC) يعني حرف عطف، الرمز (DT) يعني اسم إشارة، الرمز (RP) يعني أداة، الرمز (IN) حرف جر أو أداة ربط ثانوية.
- **الأفعال (Verbs)**: الرمز (VBP) يعني فعلاً مضارعاً مبنياً للمعلوم، الرمز (VBN) يعني الفعل المضارع/ الماضي المبني للمجهول، الرمز (VBD) يعني فعلاً ماضياً مبنياً للمعلوم، الرمز (VB) يعني فعل أمر.
- **أخرى**: الرمز (UH) يعني كلمة انفعالية، الرمز (PUNC)[1] يعني علامات ترقيم، الرمز (NUMERIC_COMMA) يعني الحرف "ر" يستخدم فاصلة، الرمز (NO_FUNC) يعني كلمة لم يتم تحليلها.

مجموعة سمات كاليك (The Kulick Tag Set)

طورت مجموعة سمات كاليك بواسطة سيث كاليك (Seth Kulick) وتبين أن لها فائدة في تحليل اللغة العربية [119]. تحتوي مجموعة سمات كاليك على حوالي ٤٣ سمة وهي توسع مجموعة سمات بييز. ويمكن تصنيف التوسعات إلى أربعة تصنيفات هي :

(١) على الرغم من أنه في بعض الأحيان يمكن أن تكون علامة الفاصلة، والنقطة والنقطتين مستخدمة نفسها في سمات أقسام الكلام، على سبيل المثال، سمة الفاصلة هي [,].

obeikandi.com

- تعطى علامات الترقيم التالية سمة تتوافق مع شكلها الفعلي، العلامات هي: [،]، [:]،[.]، ["]، [-LRB-][1]، [-RRB-].
- يتم وضع علامة صريحة على الأسماء والصفات التالية: الأسماء المعدودة أو المُكممات (NOUN_QUANT)، صفات المقارنة (ADJ_COMP)، الأعداد الترتيبية/ الوصفية (ADJ_NUM)، ومشتقات الأفعال (DV).
- تميز أداة الإشارة والتعريف بالرمزين (DEM) و(DT) على التوالي.
- ظهور أداة التعريف (DT) توضح في السمة، مثال ذلك: DT+NN وDT+ADJ_COMP و DT+CDو DT+JJ.

مجموعة السمات الموسعة المقلصة

(The Extended Reduced Tag Set) (ERTS)

تعتبر مجموعة السمات الموسعة المقلصة (ERTS) السمات الأساسية المستخدمة في نظام أميرة (انظر قسم ٥,٥,٢). يحتوي ERTS على ٧٢ سمة. وهي مجموعة جزئية من مجموعة بكوالتر الصرفية الكاملة المعرفة على نص مقطع. كما أن ERTS شاملة لمجموعة سمات (Bies/RTS). بالإضافة إلى المعلومات الموجودة في سمات (Bies)، ترمز (ERTS) خصائص صرفية إضافية مثل العدد والجنس والتعريف في العبارات الاسمية فقط. التعريف (أو بعبارة أكثر وضوحاً أداة التعريف) ترمز بخاصية ثنائية بالسمة (D) لظهور الأداة و(φ) لغياب الأداة. يرمز للجنس بالسمة (F) للمؤنث أو (M) للمذكر أو لا شيء عند غياب الجنس. يرمز للعدد بالسمة (Du) للمثنى و(S) للجمع، وفي حال غياب أي وسم فتستخدم للمفرد. على سبيل المثال الأسماء في بييز

(١) يرمز (LRB) إلى الحاصرة المستديرة اليسرى ويرمز (RRB) إلى الحاصرة المستديرة اليمنى.

openkandil.com

(Bies) توسم بالسمة (NN) أو (NNS) للإشارة فقط للعدد، أما سمات الأسماء في (ERTS) فتمثل التعريف والجنس بالإضافة للعدد، مثال ذلك: سمة (DNNM) عبارة عن اسم مفرد مذكر معرف (بمعنى وجود أداة تعريف). وصف كامل لمجموعة (ERTS) موجودة في [111]. وقد تبين أن مجموعة (ERTS) قابلة للعنونة بنفس دقة مجموعة سمات بييز، لكن بإضافة قيمة أكثر كخصائص تعليمية لمهمة حاسوبية عليا، مثل مهمة التقطيع للعبارات الأساسية (Base Phrase Chunking) [111].

٥,٤,٣ مجموعة سمات كاتب لأقسام الكلام (THE CATIB POS TAGSET)

طورت مجموعة سمات كاتب لمشروع بنك كولومبيا الشجري للتحاليل النحوية الشجرية للغة العربية (كاتب) (CATiB) [112, 121]. هناك ست سمات فقط لأقسام الكلام في كاتب. الهدف من البساطة في مجموعة السمات لأقسام الكلام هو لتسريع التحشية البشرية مع المحافظة على الفروق الهامة.

- تستخدم (VRB) مع جميع الأفعال بما فيها الأفعال الناقصة أيضاً تعرف باسم كان وأخواتها.
- تستخدم (VRB-PASS) للأفعال المبنية للمجهول.
- تستخدم (NOM) لجميع العبارات الاسمية مثل الأسماء والصفات والحال واسم الفاعل واسم المفعول والمصدر والضمير (الشخصي، الموصول، الإشارة، الملكية) والعدد (بما فيها الأرقام) وصيغة التعجب.
- يستخدم (PROP) لأسماء العلم.
- يستخدم (PRT) لجميع الأدوات. وهذه مجموعة شاملة تحتوي على العديد من الفئات المغلقة، على سبيل المثال، حروف الجر، والعطف، أدوات النفي، أداة التعريف، الخ.

obeikandl.com

- يستخدم (PNX) لجميع علامات الترقيم.

وتوسعة محددة تلقائياً من إصدار مجموعة سمات كاتب ، أطلق عليها اسم (catibEx) أثبتت فائدتها في التحليل [110]. تقوم التوسعة بربط سلسلة اللاحقة/السابقة المتوافقة للسمة ، مما يزيد حجم مجموعة السمات إلى ٤٤. على سبيل المثال سمة (NOM) للكلمة الكاتبون (AlkAtbwn) تتوسع لـ (Al+NOM+wn). وقد تبين أن مجموعة سمات كاتب يمكن توسعتها بسهولة إلى مجموعة سمات كاليك وبدقة تصل ٩٨,٥٪ (وذلك بوجود شجرة محشاة بالكامل) [112].

٥,٤,٤ مجموعة سمات خوجة (THE KHOJA TAG SET)

طورت مجموعة سمات خوجة بواسطة شيرين خوجة ، وتعتبر واحدة من أوائل مجموعة السمات الكاملة للعربية [122 ، 113]. وهي عبارة عن مجموعة سمات وظيفية على العكس من السمات الصيغية ، إلا أنها لا تعلم حالة البناء (في مقابل حالة المعرفة والنكرة) كما أن ليس لديها تغطية كاملة. على سبيل المثال ، لم تُوضع علامة على حالة أسماء الأعلام والضمائر. وتحتوي مجموعة السمات على ١٧٧ سمة مقسمة كالتالي : ١٠٣ أسماء ، ٥٧ فعلاً ، ٩ أدوات ، ٧مجردات/بواقٍ وعلامة ترقيم واحدة. ويتم تكوين السمات عن طريق ربط علامات حرف أو حرفين في تسلسل محدد تليها سمات محددة. انظر الشكل رقم ٥,٥. على سبيل المثال السمة (NASgMNI) تعني صفة - مفرد - مذكر - نكرة - مرفوعة ، والسمة (VIDu3FJ) تعني فعلاً مضارعاً مجزوماً مسنداً لألف الاثنين (المؤنث). وقد عرفت المجموعة على الكلمات المجردة ، إلا أنه يمكن استخدامه للكلمات بزوائد عن طريق ربط بسيط مع فاصل. على سبيل المثال ، كلمة باسمه (b+Asm+h) يحصل على السمة التالية (PPr_NCSgMGI_NPrPSg3M).

obeikandl.com

• **N** *noun (اسم)*

– **+C** *common (عام)* + **Attribute (سمة):** *number (عدد)-gender (جـــنس)-case (حالـــة)-definiteness (تعريف)*

– **+P** *proper (خاص)*

– **+Pr** *pronoun (ضمير)*

* **+P** *personal (شخصي)* + **Attribute:** *number-person-gender*

* **+R** *relative (موصول)*

· **+S** *specific (محدد)* + **Attribute:** *number-gender*

· **+C** *common (عام)*

* **+D** *demonstrative (إشاري)* + **Attribute:** *number-gender*

– **+Nu** *numerical (عددي)*

* **+Ca** *cardinal* (أساسي) + **Attribute:** *[Sg]-gender (جنس مفرد)*

* **+O** *ordinal (ترتيبي)* + **Attribute:** *[Sg]-gender (جنس مفرد)*

* **+Na** *numerical adjective (صفة عددية)* + **Attribute:** *[Sg]-gender*

– **+A** *adjective (صفة)* + **Attribute:** *number-gender-case-definiteness*

• **V** *verb (فعل)*

– **+P** *perfective (ماض)* + **Attribute:** *number-person-gender*

– **+I** *imperfective (مضارع)* + **Attribute:** *number-person-gender-mood*

– **+Iv** *imperative (أمر)* + **Attribute:** *number-[2]-gender*

• **P** *particle(أداة)*

– **+Pr** *preposition (جر)*, **+A** *adverbial* (ظرف), **+C** *conjunction (عطف)*, **+I** *interjection* (**تعجب**), **+E** *exception (استثناء)*, **+N** *negative* (سلبي), **+A** *answers (إجابات)*, **+X** *explanations* (توضـــيح), **+S** *subordinates (ثانوي)*

• **R** *residual (متبقٍ)*

– **+F** *foreign* (أعجمي), **+M** *mathematical (رياضي)*, **+N** *number (عدد)*, **+D** *day of the week (يـــوم الأسبوع)*, **+my** *month of the year (الشهر في السنة)*, **+A** *abbreviation* (اختصار), **+O** *other (أخرى)*

• **PU** *punctuation (علامة ترقيم)*

• *Attributes (سمات)*

– Gender (جنس): M *masculine* (مذكر), F *feminine (مؤنث)*, N *neuter(محايد)*

– Number (عدد): Sg *singular* (مفرد), Pl *plural* (جمع), Du *dual (مثنى)*

– Person (شخص): 1 *first* (**المخاطب**), 2 *second* (**المخاطب**), 3 *third* (**الغائب**)

– Case (حالة): N *nominative* (مرفوع), A *accusative(منصوب)*, G *genitive(مجرور)*

– Definiteness (تعريف): D *definite* (معرف), I *indefinite(غير معرف)*

– Mood (إعراب): I *indicative*(خبرية), S *subjunctive (منصوب)*, J *jussive (مجزوم)*

الشكل رقم (٥,٥). مجموعة سمات خوجة.

obeikandi.com

٥,٤,٥ مجموعة سمات بادت (THE PADT TAG SET)

طورت مجموعة سمات بادت والمستخدمة في محلل (ElixirFM) (قسم ٥,٢,٥) للاستخدام في بنك براغ للتحاليل الشجرية النحوية التبعية للغة العربية (Prague Arabic DependencyTreebank) [123,114,67]. وقد عرفت مجموعة سمات بادت لنظام (ATB) العربية. كل سمة مكونة من جزأين هما: أقسام الكلام (POS) والخصائص (Features). مكونات أقسام الكلام تتكون من حرفين كالتالي:

- فعل مضارع (VI)، فعل ماضٍ (VP)، فعل أمر (VC).
- اسم (N)، صفة (A)، حال (D)، اسم علم (Z)، اختصار (Y).
- ضمير (S)، اسم إشارة (SD)، اسم موصول (SR).
- أداة (F)، أداة استفهام (FI)، أداة نفي (FN)، حرف عطف (C)، حرف جر (P)، حرف تعجب (I).
- رمز رسومي (G)، عدد (Q)، أداة تعريف معزول (--).

جزء الخصائص من السمة يتكون من متوالية من سبعة أحرف. وكل حرف يقوم بترميز قيمة الخاصية المسندة لموضع الحرف بكفاءة:

- الإعراب (Mood): خبرية (**I**ndicative)، حالة النصب (**S**ubjunctive)، حالة الجزم (**J**ussive) أو D في حال الاشتباه بين حالة S و J.
- صِيغَة البناء المَجهُول (Voice): معلوم (**A**ctive)، مجهول (**P**assive).
- الشخص (**1**:Person) للمتحدث (speaker)، **٢** للمخاطب (addressee)، **٣** للآخرين (others).
- الجنس (Gender): مؤنث (**F**eminine)، مذكر (**M**asculine).

obeikandl.com

- العدد (Number): مفرد (**S**ingular)، مثنى (**D**ual)، جمع (**P**lural).
- الحالة الإعرابية (**1** :(Case= مرفوع (nominative)، **٢** = مجرور (genitive)، **٤** = منصوب (accusative).
- التعريف (Definiteness): نكرة (**I**ndefinite)، معرفة (**D**efinite)، مقلصة (**R**educed)، معقدة (**C**omplex).

على سبيل المثال، سمات أقسام الكلام التالية (VP-A-3MP) تُمثل فعلاً ماضياً مبنياً للمعلوم وفاعلاً لجمع مذكر غائب. لاحظ أن وجود الزائدة القبلية لأداة التعريف يشار إليها فقط من خلال خاصية التعريف، الذي يجمع بينه وبين خاصية الحالة: في بادت (PADT) التعريف يساوي وجود أداة التعريف وحالة التعريف. بينما قيم التعريف المقلصة والمعقدة مساوية لحالة البناء. الفرق هو أن التعريف المعقد لديه أداة التعريف (إضافة غير صحيحة، انظر قسم ٦,١,٣).

٥,٥ حزمتا أدوات

في هذا القسم سنقدم، بشيء من التفصيل، حزمتي أداوت مختلفة إلى حدٍّ ما للمعالجة الحاسوبية للصرف العربي: أداة مدى+توكن (Mada+Tokan) وأداة أميرة (AMIRA). هذه الأدوات متاحة للجميع وقد استخدمت من قبل العديد من المؤسسات البحثية والأكاديمية والتجارية في جميع أنحاء العالم. أيضاً، سنقارن ونقابل بينهما من حيث الوظيفة، والتصميم والأداء في مختلف تطبيقات معالجة اللغة الطبيعية.

٥,٥,١ مدى + توكن (MADA+TOKAN)

المحلل الصرفي ومزيل الغموض للغة العربية، مدى (MADA) ، هي أداة تأخذ نصاً عربياً خاماً، وتضيف عليه أكبر قدر ممكن من المعلومات الصرفية والمعجمية عن طريق إزالة الغموض وفي عملية واحدة، من هذه المعلومات: سمات أقسام الكلام والمُعَيْجِمَة والتشكيل والتحليل الصرفي الكامل [51، 35، 16]. منهجية مدى في العمل تفرق بين مشاكل التحليل الصرفي التي يتولاها محلل المرجانة الصرفي، وبين إزالة الغموض الصرفي. لذا يعتبر مدى نظاماً لإزالة الغموض الصرفي. بمجرد ما يتم اختيار التحليل الصرفي حسب السياق، فإن سمات أقسام الكلام الكاملة والمادة المعجمية والتشكيل تُختار أيضاً (جميعها في خطوة واحدة). تساعد أيضاً معرفة التحليل الصرفي في تقطيع وتجذيع محدد يتولاها جزء توكن (TOKAN) بمجرد ما تنتهي مدى من معالجة النص.

مدى: تقوم مدى بعملها على خطوات. أولاً: تستخدم المرجانة داخلياً لإنتاج قائمة بالتحليلات الممكنة لكل كلمة يواجهها في النص، وفي هذه المرحلة لا تؤخذ سياق الكلمة في عين الاعتبار. بعدها تقوم مدى باستخدام حتى ١٩ خاصية لترتيب قائمة التحليلات. ولكل خاصية، يُستخدم مصنف لخلق تنبؤ عن قيمة هذه الخاصية عن كل كلمة في سياقها. أربعة عشر من الخصائص تستخدم مصنف آلة الدعم الاتجاهي (SVM)، أما بقية الخصائص فتلتقط معلومات مثل الاختلافات الإملائية وإحصاءات (n-gram). وكل توقع للمصنف يتم ترجيحه باستخدام مجموعة ضبط، وتتم مقارنة مجموعة خصائص التنبؤات بقائمة التحليلات الصرفية المحتملة. تلك التحليلات التي تتوافق بشكل وثيق مع مجموعة توقعات الخصائص المرجحة تحصل على درجات أعلى في الترتيب من تلك التي لا تتوافق، أما التحليل الأعلى في

الدرجات فتُوضع علامة عليها على أنها التحليل الصحيح لهذه الكلمة في هذا السياق. وبما أن مدى تختار التحليل الكامل من المرجانة، فإن جميع القرارات المتعلقة بالغموض الصرفي والغموض المعجمي وتقطيع النص والتشكيل وتعيين أقسام الكلام في أي مجموعة سمات أقسام الكلام الممكنة يتم القيام بها دفعة واحدة. وتتميز مدى بدقة تصل إلى ٩٦٪ في تعيين المادة المعجمية وفي الخيارات الصرفية الأساسية (التي تتضمن تقطيع النص ولكن تستثني الحالة النحوية والإعراب والحالة). كما أن لدى مدى دقة تصل إلى ٨٠٪ في التنبؤ بالتشكيل الكامل (بما فيها الحالة النحوية والإعراب). ويمكن الحصول على مقارنة تقييمية وتفصيلية في الأبحاث التالية: [51، 35، 116].

Arabic	Gloss	Buckwalter/PATB	CATiB	Bies	Kulick	ERTS	Khoja	PADT	ALMORGEANA/MADA
خمسون xams+uwna	fifty	NOUN_NUM+ NSUFF_MASC_PL_NOM	NOM	CD	CD	CD	NNuCaPlM	QL------1I	POS:NUM +MASC +PL +NOM
الف Âalf+a	thousand	NOUN_NUM+ CASE_DEF_ACC	NOM	CD	CD	CD	NNuCaSgM	QM-----S4R	POS:NUM +DEF +ACC
سائح sAŷiH+ī	tourist	NOUN+ CASE_INDEF_GEN	NOM	NN	NN	NNM	NCSgMGI	N------S2I	POS:N +INDEF +GEN
زاروا zAr+uwA	visited	PV+ PVSUFF_SUBJ:3MP	VRB	VBD	VBD	VBD	VPPI3M	VP-A-3MP--	POS:V +PV +S:3MS
مدينة madiyn+aħ+a	city	NOUN+ NSUFF_FEM_SG+ CASE_DEF_ACC	NOM	NN	NN	NNF	NCSgFAI	N------S4R	POS:N +FEM +SG +DEF +ACC
نا+ +nA	our	POSS_PRON_1P	NOM	PRP$	PRP$	PRP$	NPrPPl1	S----1-P2-	+P:1P
الجميلة Al+jamiyl+aħ+a	beautiful	DET+ADJ+ NSUFF_FEM_SG+ CASE_DEF_ACC	NOM	JJ	DT+JJ	DJJF	NASgFAD	A-----FS4D	POS:AJ Al+ +FEM +SG +DEF +ACC
في fiy	in	PREP	PRT	IN	IN	IN	PPr	P---------	POS:P
ايلول Âay.luwl+a	September	NOUN_PROP+ CASE_INDEF_GEN	PROP	NNP	NNP	NNPM	Rmy	N------S2I	POS:PN +INDEF +GEN
الماضي Al+mADiy	past	DET+ADJ	NOM	JJ	DT+JJ	DJJM	NASgMGD	A-----MS2D	POS:AJ Al+
. .	.	PUNC	PNX	PUNC	.	PUNC	PU	G---------	POS:PX

الشكل رقم (٥,٦). مقارنة لعدة مجموعات لسمات أجزاء الكلام للجملة (خمسون ألف سائح زاروا مدينتنا الجميلة في ايلول الماضي) (xmswn Alf sAŷH zArwA mdyntnA Aljmylħ fy Aylwl AlmADy).

تتميز مدى بمرونة عالية وخيارات واسعة في تهيئة التشغيل. وابتداء من الإصدارة ٢,٠ ، قامت مدى بتطبيق الأوزان على كل من الخصائص التسع عشرة المستخدمة لتحسين الدقة ، وتُحدد هذه الأوزان بناء على مجموعة ضبط وتُحسن لأغراض مختلفة مثل : تقطيع النص ، أو التشكيل ، أو تعيين أقسام الكلام. وتأتي مجموعات الأوزان هذه مع الحزمة ، وينبغي أن تُختار من قبل المستخدم بناء على كيفية استخدام مدى. ومع ذلك ، يمكن أيضاً للمستخدمين اختيار تعيين هذه الأوزان مباشرة بأنفسهم. الخيار الافتراضي لمدى ، هو محاولة ترتيب التحليل الكامل من حيث صحتها إجمالاً. فعن طريق اختيار خاصية بديلة ومجموعة أوزان ، من الممكن جعل مدى تركز بشكل أكثر تحديداً للحصول على جانب معين من التحليل بشكل صحيح. على سبيل المثال ، يمكن للمستخدمين تحقيق تحسن مطلق بنسبة ٠,٤٪ في دقة تعيين أقسام الكلام إذا استخدموا مجموعة أوزان تُضبط لتعيين أقسام الكلام ، بدلاً من المجموعة الافتراضية. ومع ذلك ، قد تعاني دقة مخرجات مدى الأخرى (على سبيل المثال التنبؤ بالمُعَيْجِمَة). كما يتضمن مدى إجراء للتراجع الصرفي الذي يمكن تشغيلهأو إيقافه من قبل المستخدم.

توكن : يعتبر توكن مُقطعاً عاماً للنص العربي الذي يوفر مصدراً سهل الاستخدام لتقطيع نص مدى الغامض إلى مجموعة كبيرة من الاحتمالات [83 ، 97]. تقوم مدى بتحديد ما إذا كان هناك حرف جرٍ أو عطفٍ كزائدة في الكلمة العربية ، لكن عملية التقطيع الفعلية للزوائد بما فيها مختلف التكتيكات النحوية والتسوية الإملائية تقوم بها توكن. ويمكن استخدام مخطط التقطيع كمعلمة (parameter) في تعلم الآلة لمختلف التطبيقات مثل ، الترجمة الآلية والتعرف على أسماء الأعلام (Named Entity Recognition).

يدخل إلى توكن ملف مدى- لفك الغموض ووصف لمخطط تقطيع النص الذي يحدد النص المقطع الهدف. خذ في عين الاعتبار المواصفات التالية :

"w+ f+ b+ k+ l+ s+ Al+ REST + / + POS +P: +O: -DIAC"

يقوم هذا المخطط بفصل حروف العطف والجر والأدوات الفعلية وأداة التعريف والضمائر المتصلة وإضافة سمات أقسام الكلام الأساسية لشكل الكلمة. أيضاً يبين المخطط أن التشكيل يُنتج. وبتحليل الكلمة "وسيكاتبها" (wasayukAtibuhA) سيتم تقطيعها إلى (wa+ sa+ yukAtibu/V +hA). أما مخطط بسيط مثل (w+ f+ REST) فسينتج (w+ sykAtbhA). اطلع على [83، 105] للحصول على وصف مفصل لمخططات عدة تم تبنيها واتباعها منذ نشرها. يحتوي توكن على مجموعة كبيرة من خصائص أخرى تسمح للمستخدم القيام بعمليات مختلفة لتقليص الاحتمالات الهجائية أو التحكم بكيفية ترتيب المخرجات وعرضها لأنها قد تناسب الاحتياجات المختلفة لأنظمة مختلفة. جميع مخططات تقطيع النص المعروضة في الشكل رقم ٥,٣ تدعمها توكن.

يستخدم توكن التوليد الصرفي داخلياً، (بواسطة المرجانة) لإعادة إنشاء الكلمة بمجرد ما يتم فصل الزوائد المختلفة. هذه الطريقة في التوليد الرجعي تسمح لنا بتعديل المحتوى الصرفي في الكلمة بسهولة بما فيها على سبيل المثال، حذف/تعيين خصائص محددة لكلمة افتراضياً. يضمن هذا أن شكل الكلمة المولدة تم تقليصها واتساقها مع تكرارات ظهور هذه الكلمة. على سبيل المثال، فصل الضمير المتصل من كلمة فيها تاء مربوطة (ة h) سوف يبقي التاء المربوطة بشكل كلمتها الداخلي (كتاء عادية ت t). وفي توكن يتم إنشاء التاء المربوطة حسب مقتضى الحال. على سبيل المثال، كلمة "جولته" (jwlth) تُقطع إلى جولة+ه (jwlh+h) وليس جولت+ه (jwlt+h) ولا يعتبر هذا إملائياً صحيحاً.

مدى+توكن لتطبيقات معالجة اللغة الطبيعية: استخدم مدى+توكن من قبل العديد من المعاهد البحثية والأكاديمية والتجارية في جميع أنحاء العالم. وسنذكر هنا بعض الأمثلة على استخدامها. في سياق الترجمة الآلية من العربية للإنجليزية، قام [83] و[105] بالبحث في استخدام مختلف مخططات المعالجة القبلية وتوليفتها. وقد اتُبعت نتائج بحثهم من قبل مجموعات مختلفة من الباحثين العاملين في مجال الترجمة الآلية من اللغة العربية للغة الإنجليزية [124، 125، 126]. أما [127] فقد بحث في استخدام التشكيل المنتج بواسطة مدى في الترجمة الآلية. أيضاً قام [107] بتحسين محاذاة الكلمة آليا للترجمة الآلية من العربية للإنجليزية باستخدام مزيج من مختلف مخططات تقطيع النص المنتجة بواسطة مدى+توكن. اطلع على [97] لمزيد من التفاصيل حول التمثيلات المختلفة للصرف العربي في الترجمة الآلية. وقد استخدم [99] مدى في سياق الترجمة من الإنجليزية للعربية. كما استخدمت مدى أيضاً لإنتاج خصائص لنظم التعرف على أسماء الأعلام [128، 129].

٥,٥,٢ أميرة (AMIRA)

تتكون أميرة من مجموعة من الأدوات التي بنيت خلفاً لنظام (Asvmt) التي طورتها جامعة ستانفورد [١٠٩] وشرحت بالتفصيل في [117]. وتحتوي مجموعة الأدوات على مقطع للنص ومعنون لأقسام الكلام ومقطع للعبارات الأساسية (base phrase chunker-BPC)، والمعروف أيضاً باسم المحلل التركيبي السطحي (shallow syntactic parser). سيكون تركيزنا في هذا الجزء على (Amira-Tok) و(Amira-Pos). تعتمد التقنية في أميرة على التعليم الموجه/تحت الإشراف (Supervised Learning) مع عدم وجود اعتماد واضح على معرفة عميقة بالصرف، ومن ثَم، على عكس مدى، فهو يعتمد على البيانات السطحية ليتعلم التعميمات. وبشكل عام، تستخدم الأدوات

إطاراً موحداً يحيل كل مشكلة في المكونات إلى مشكلة تصنيف. والتقنية المستخدمة الكامنة وراء الأدوات هي آلة الدعم الاتجاهي (SVM) في إطار تسلسل النمذجة.

تركز أميرة- توك (**AMIRA-TOK**) في الأساس على تقطيع الزوائد. ولا تعتمد أدوات أميرة على التحليل الصرفي أو أدوات التوليد في أي من عملياتها. ومن ثَم، تتعلم أميرة- توك تعميم تقطيع الزوائد من تجزئة الزوائد المستخدمة في بنك بنسلفانيا الشجري للتحاليل النحوية (PATB) مباشرة دون الاعتماد على القواعد بشكل واضح.

تقوم أميرة- توك بتجزئة مجموعة الزوائد التالية: زوائد حروف العطف (و+ w+، ف+ f+) زوائد حروف الجر (ك+ k+، ل+ l+، ب+ b+)، زوائد علامات المستقبل (س+ s+)، زوائد أدوات الفعل (ل+ l+)، زائدة أداة التعريف (أل+ Al+) والضمائر المتصلة في الكلمة التي تدل على ضمائر الملكية والمفعول به.

وينظر إلى حل أميرة- توك على القيام بمعاملة تقطيع النص العربي على أنه مشكلة تجزئة على مستوى الحرف. مما يسمح باستخدام حلول (IOB) للمجزئات النحوية التي تستخدم عادة في مستوى العبارة وعلى مستوى الكلمة الفرعية. هنا، يتم وضع حواش لكل حرف (بما فيها علامات الترقيم) كالتالي: داخل الجزء (I)، خارج الجزء (O)، بداية الجزء (B)، ومن هنا جاءت تسميته (IOB). لسمة (I) و(B)هناك خمس فئات محتملة: زوائد حروف العطف يرمز لها بفئة (Prefix 1)، وحروف الجر يرمز لها بفئة (Prefix 2)، وأداة التعريف يرمز لها بفئة (Prefix 3)، والضمائر المتصلة يرمز لها (Word, Suffix). وهذا يؤدي إلى ما مجموعه ١١ فئة في البيانات كالتالي:

O, B-PRE1, I-PRE2, B-PRE2, I-PRE2, B-PRE3, I-PRE3, B-WORD, I-WORD, B-SUFF, I-SUFF

ومن خلال تعلم كيفية تعيين تسميات هذه الوسوم، تتعلم (Amira-Tok) كيف تجزئ الكلمات.

openkandil.com

لا تقوم أداة أميرة-توك بإنتاج كلمات مجذعة لا تعتبر كلمات عربية صحيحة. كما أن الأداة تطبق بعض الإرشادات لعكس تأثير التكتيك الصرفي مثل فقدان الألف في أل التعريف عند وجودها في سياق زائدة حرف الجر ل+ +l. ومعظم ترميمات التكتيك الصرفي يتم تطبيقها بشكل محدد. غير أن التكتيكات الصرفية غير المحددة مثل تلك التي تشمل علامة التأنيث الاسمية (التاء المربوطة) والألف المقصورة يتم تحديدها تلقائياً من خلال طبقة أخرى من التعلم لمشكلة تصنيف حروف نهاية الكلمة. بالنسبة للتاء المربوطة: التاء في آخر الجذع يمكنها أن تبقى ت t أو تحول إلى ة ħ، أما بالنسبة للألف المقصورة: الألف في نهاية الجذع يمكنها أن تبقى ألفاً ا A أو تحول إلى ى ȳ.

وعلى الرغم من أن تقطيع النص في أميرة يعمل على فصل الزوائد ويقلص الجذع، إلا أن واجهة الأداة تسمح لعدد محدود من المتغيرات، التي تشمل مستوى تجزئة الزوائد وما إذا كان التقطيع يستدل به بالفراغات (مما يغير عدد المقاطع) أو بعلامة زائد فقط (مما يحتفظ بعدد المقاطع). على سبيل المثال، كلمة "وللبلاد" (wllblAd) يمكن أن تقطع إلى الشكل التالي من بين طرق أخرى: بطريقة أميرة- توك الداخلية تصبح (w+ l+ Al+ blAd)، وبطريقة حروف العطف فقط تصبح (w+ llblAd)، وبطريقة حروف الجر تصبح (w+l+ AlblAd)، وبطريقة أل التعريف تصبح (wl+Al+ blAd)، وبطريقة اللواحق فقط تصبح (wllblAd)، وبطريقة كل السوابق واللواحق تصبح (wll+ blAd).

ويعتبر أداء أميرة- توك عالياً في مقياس (F-score) حيث يصل إلى ٩٩,٢٪.

أميرة- أقسام الكلام (AMIRA-POS): تستخدم أميرة في الأساس مجموعة سمات (ERTS POS) الذي يفترض أن النص مقطع حسب الزوائد. كما أن تعيين أقسام الكلام في أميرة يعتمد على طريقة التصنيف المعتمد على آلة الدعم الاتجاهي

(SVM) وباستخدام المتتابعات الحرفية (character n-grams) كخصائص في النماذج المتسلسلة.

لدى المستخدم المرونة اللازمة لإدخال النص الخام أو المقطع في نظام يتوافق مع أحد المخططات التي تعرفها أميرة- توك. وبناء على ذلك، يمكن للمستخدم أن يطلب تعيين سمات أقسام الكلام على الشكل السطحي. أما داخلياً، وفي حالة ما إذا كانت المدخلات خاماً، فإن أميرة- أقسام الكلام تشغل أميرة- توك على النص الخام ثم تنفذ عملية تعيين أقسام الكلام. عندها تُمثل المخرجات كنص مقطع أو معنون بأقسام الكلام أو بدون تقطيع حيث تُعين سمات أقسام الكلام على سطح الكلمات. في هذه الحالة الأخيرة، وتُلحق سمات (ERTS tag) بسمات زوائد أقسام الكلام لتشكيل وسومٍ أكثر تعقيداً، حينها يمكن للمستخدم اختيار إما أن يَسِمَ باستخدام ERTS أو RTS (انظر قسم ٥,٤,٢).

ومن المثير للاهتمام، أن دقة معنون (ERTS) هي ٩٦,١٣٪، ودقة معنون (RTS) هي ٩٦,١٥٪. ويشير هذا إلى أن اختيار المعلومات التي تدرج في مجموعة سمات ERTS يعكس الانقسام الطبيعي في الفضاء النحوي. وقد أثبتت مجموعة سمات (ERTS) الثرية أنها تُحسن جودة المعالجة النهائية مثل مهمة تجزئة العبارات الأساسية BPC [111, 120].

أميرة لتطبيقات معالجة اللغة الطبيعية: استخدمت عدة مجموعات أميرة بنجاح في سياق الترجمة الآلية للنص، وتحديداً لتحسين المحاذاة وإعادة الترتيب في سياق الترجمة الآلية الإحصائية [130]، وأيضاً لتحديد النص الصعب للغة المصدر [131]. وعلاوة على ذلك، تم استخدام حزمة أميرة في سياق الترجمة الآلية للكلام [132]. وقد تم بحث استخدام حزمة أميرة لأغراض استرجاع المعلومات عبر اللغات في عمل

بواسطة [104]. وقد استخدمت أميرة لإنتاج سمات أقسام الكلام وخصائص BPC لنظام التعرف على أسماء الأعلام (NER) [133, 129].

٥,٥,٣ مقارنة مدى+توكن مع أميرة

في هذا القسم، سنقارن ونقابل برنامج مدى+توكن وأميرة من ناحية تصميمهما ووظائفهما وأدائهما.

التصميم: أما بالنسبة لتصميم البرنامجين، فقد يكون من المساعد تأطير الأدوات المختلفة من حيث استعمالها الأساسي في حزمتين: حزمة مدى وحزمة أميرة. فمن ضمن حزمة مدى، هناك خطوة واضحة للتحليل الصرفي يتولاها نظام المرجانة. والثاني، الذي يعتبر عنصراً أساسياً في حزمة مدى، هو نظام مدى، الذي يقوم بفك غموض التحليلات التي تنتجها المحللات الصرفية. وأخيراً فإن جزء توكن يستفيد من قوة التوليد الصرفي للمرجانة لتقطيع التحليل غير الغامض عن طريق إعادة توليدها. وفي حزمة أميرة، فإن المكونين يركزان على التقطيع أميرة- توك (Amira-Tok) وتعيين أقسام الكلام (أميرة- أقسام الكلام (Amira-Pos)).

أما ما يتعلق في تصميمهما، فإن أميرة- توك وأميرة- أقسام الكلام مختلفتان عن حزمة مدى من حيث إنهما تتخذان منهجاً مكوناً من خطوتين لتعيين أقسام الكلام هما: التقطيع ثم التوسيم. وبالمقارنة فإن مدى تتخذ منهجاً مختلفاً يقوم على تفكيك المشكلة إلى ثلاث خطوات هي: (تحليل، فك غموض، توليد)، وهذا النهج متعامد مع طريقة أميرة في التجزئة. على الرغم من وجود ثلاث خطوات في مدى، فإن قرار التقطيع وتعيين أقسام الكلام تتم معاً بخطوة واحدة. إحدى الطرق لتمييز هذه الأدوات يكمن في عمق المعرفة اللغوية المطلوبة. ويعتبر نظام أميرة سطحياً لأنه يركز على

الصرف الشكلي (وتحديداً الزوائد) المكتسبة من البيانات المحشاة، بينما لدى مدى إمكانية للوصول بشكل أعمق إلى صرف وظيفي منمذج معجمياً. وثمة فرق آخر بين حزمة مدى الحالية وحزمة أميرة تكمن في أن الأول قد لا ينتج أي تحليل لكلمة معينة إذا لم تكن موجودة في الأدوات الصرفية الأساسية (على الرغم من أنه عادة ما يستخدم تحليلاً متراجعاً في مثل هذه الحالات) في حين أن حزمة أميرة دائماً ما تنتج تقطيعاً افتراضياً وسمات لأقسام الكلام عن كل كلمة في النص.

ومن حيث احتياجاتها التدريبية، فإن حزمة مدى تتوقع وجود كل من محلل صرفي وبيانات تدريبية للتعليم الموجه، في حين أن حزمة أميرة تحتاج فقط إلى بيانات تدريب محشاة. ويمكن إنشاء بيانات التدريب من خلال عدة طرق، بما في ذلك استخدام التحليل الصرفي يليها توسيم بواسطة شخص ما، ولكن هذا ليس شرطاً لحزمة أميرة. تضع هذه المتطلبات المختلفة ولكنها متماثلة حدوداً مشابهة على نوع التوسعات التي يمكن القيام بها في أي نهج. على سبيل المثال، يتطلب الذهاب إلى اللهجة العربية وجود بعض المحللات/ المولدات الصرفية للهجة في مدى، ولكن ليس لأميرة. ومع ذلك، فكلاهما بحاجة إلى بعضٍ من البيانات المحشاة للتدريب عليها.

الوظيفة: من حيث الأداء الوظيفي، سنأخذ في عين الاعتبار خمسة تطبيقات هي: تقطيع النص، والتشكيل، وتعيين أقسام الكلام، وتعيين المادة المعجمية، ومجزئ العبارات الأساسية. بالنسبة لمجزئ العبارات الأساسية موجود فقط في حزمة أميرة، لكنها في الواقع وحدة منفصلة يمكن استخدامها بشكل مستقل مع حزمة مدى. يتم التعامل مع التطبيقات الأربعة الأخرى جميعاً في مدى كجزء من عملية فك الغموض الصرفي. ولا تتعامل أميرة مع عملية تعيين المادة المعجمية أو التشكيل. أما بالنسبة لتقطيع النص وتعيين أقسام الكلام، فبما أن مدى أعمق من أميرة، فإن هناك مجموعة

واسعة من مخططات التقطيع الممكنة وسمات أقسام الكلام التي يستطيع مدى إخراجها. على الرغم من أن نتائج المقارنة السابقة أظهرت أن أميرة أكثر محدودية من مدى، إلا أن أميرة تستطيع معالجة تقطيع النص وسمات أقسام الكلام الأكثر استخداماً. وعلى الباحثين المهتمين باستكشاف عدد كبير من مجموعات مختلفة من تقطيع النص كخصائص في أنظمتهم النظر في مدى. أما الباحثون المهتمون فقط في مقارنات محدودة أو تطبيقات محددة، تحتوي على عملية تقطيع النص وسمات أقسام الكلام المدعومة من أميرة، فإن عليهم النظر في أميرة.

الأداء: من الصعب مقارنة أداء حزمتي أميرة ومدى. ففي محاولات سابقة من قبل [٥١] تبين أنه من الممكن الحصول على أداء مماثل في المهام المشتركة مثل: تقطيع النص بطريقة (PATB) وسمات أقسام الكلام. ويمكن أن تكون أميرة أسرع بكثير من مدى، ومع ذلك، فإن مدى تتطلب تشغيلها مرة واحدة فقط ويمكن أن تنتج عدداً أكبر بكثير من تقطيع النص وسمات أقسام الكلام (بالإضافة إلى مخرجات أخرى غير مدعومة من قبل أميرة) عن طريق تشغيل خطوة توكن السريع (fast Tokan step).

openkandi.com