نظام آلي
للتقطيع والتوسيم
النحوي العربي

نظام آلي
للتقطيع والتوسيم
النحوي العربي

تأليف

د. أفراح عبد العزيز التميمي

الطبعة الأولى

2020 م 1441هـ

نظام آلي للتقطيع والتوسيم النحوي العربي
تاليف: أفراح عبد العزيز التميمي
رقم الإيداع لدى دائرة المكتبة الوطنية: 2020/1/57
ردمك: ISBN 978-9957-74-868-5
الطبعة الأولى 2020م 1441هـ

دار كنوز المعرفة للنشر والتوزيع
وسط البلد - شارع الملك حسين - مقابل بنك الإسكان
عمان - الأردن Amman - Jordan
هاتف Tel: 00962 6 4655877
فاكس Fax: 00962 6 4655875
خلوي Mobile 00962 79 5525 494
www.darkonoz.com
E-mail: info@darkonoz.com
dar_konoz@yahoo.com

تصميم الغلاف والإشراف الفني: محمد أيوب

فهرس المحتويات

فهرس الجداول

فهرس الأشكال

فهرس ملحق الكتاب

قائمة برموز الوسوم المقترحة ومعانيها بالعربية والإنجليزية

الرمز	معناه بالعربية	معناه بالإنجليزية
1	مسند لمتكلم	1st person
2	مسند لمخاطب	2st person
3	مسند لغائب	3st person
A	وسم رئيس: صفة	adjective
A	خاصية: مبني للمعلوم	active voice
AA	صفة مشبهة	participial adjective
ABBREV	اختصار	abbreviation
AC	أفعل التفضيل	comparative adjective
AE	صيغة مبالغة	intensive adjective
AO	صفة مفعول	passive participle
AS	صفة فاعل	active participle
AT	منسوب	attributive form
D	وسم رئيس: ظرف	adverb
D	خاصية: معرف	definite
DIGIT	رقم	digit word
DL	ظرف مكان	place adverb

الرمز	معناه بالعربية	معناه بالإنجليزية
DT	ظرف زمان	time adverb
F	مؤنث	feminine
FOREIGN	كلمة أجنبية	foreign words
I	وسم رئيس: خالفة	interjection
I	خاصية: غير معرف	indefinite
IE	خالفة تعجب	verbal exclamatory
IG	خالفة مدح	*niʿma* interjection
IS	خالفة صوت	imitative word
IV	إخالة	verbal interjection
IX	خالفة ذم	*beʾsa* interjection
L	جمع	plural
M	مذكر	masculine
N	اسم	noun
NA	اسم معنى	abstract noun
NC	اسم ذات	concrete noun
NI	اسم مبهم	undefined noun
NL	اسم مكان	place name
NM	اسم آلة	instrumental noun
NT	اسم زمان	time name
NV	اسم جنس	collective noun

الرمز	معناه بالعربية	معناه بالإنجليزية
P	وسم رئيس: ضمير	pronouns
P	خاصية: مبني للمجهول	passive voice
PD	ضمائر إشارية	demonstrative pronoun
PP	ضمائر شخصية	personal pronouns
PR	ضمائر موصولة	relative pronouns
PUNC	علامة ترقيم	punctuation
R	الرابط الإحالي	clitic pronominal anaphora
RP	أداة	article
S	مفرد	singular
SYMB	رمز	symbols words
U	مثنى	dual
V	فعل	verb
VC	فعل أمر	command verb
VP	حاضر	present verb
VS	فعل ماض	past verb

إن حاجة اللغة العربية إلى مدونة لغوية معالجة آليًّا أمر لا يخفى على كل مشتغل بالعربية مُلمٍّ بمشكلات توسيمها، وبعلاقة هذا التوسيم بالفجوة الرقمية التي نعيشها اليوم، وعجزنا عن الإسهام في صناعة المعرفة المعاصرة أو الاستفادة المرجوة من منجزاتها.

ولكي تكون لدينا مدونة معالجة آليا، فإن الأمر يمر بخطوات علمية هي:

١ـ توفر مدونة لغوية عربية ذات مواصفات علمية بوصفها «قطاعا من المادة اللغوية يتم جمعه مما يقوله المتحدثون السليقيون للغة المدروسة بغرض الوصف والتحليل»[١].

٢ـ معالجة هذه المادة اللغوية الخام على مرحلتين: لغوية وحاسوبية:

أ ـ في مرحلة المعالجة اللغوية تعالج اللغة العربية على كل مستويات التحليل اللغوي؛ ابتداء من المعالجة الصوتية، فالمعالجة الصرفية، فالمعالجة النحوية، فالمعالجة المعجمية، فالمعالجة الدلالية، فالمعالجة التداولية.

ب ـ مرحلة المعالجة الحاسوبية، وفيها يتم التعامل مع المادة المعالجة لغويا طبقا لقواعد اللغة الطبيعية في المرحلة السابقة، بتحويل هذه القواعد إلى لغة رمزية منطقية بها تصير اللغة الطبيعية لغة مشفرة وسيطة يفهمها الحاسوب. وتتوقف كفاءة هذه المرحلة ونجاح إنجازها على شيئين هما: دقة المعالجة اللغوية في المرحلة السابقة من ناحية، وكفاءة من يقوم بالمعالجة الحاسوبية في الإلمام بمعارف ضرورية تتكئ عليها معالجته كالذكاء الاصطناعي وعلم النفس الإدراكي.

ومما سبق يتضح أن المعالجة النحوية لمادة المدونة اللغوية العربية هي في قلب المعالجة اللغوية لمادة هذه المدونة، وعليها تقوم المعالجة الآلية. وفي إطار المعالجة النحوية نجد التحشية بإضافة معلومات ضرورية تجعل المدونة ذات قيمة علمية تخدم

---

(١) من محاضرة علمية بعنوان «المدونات اللغوية بين الواقع والمأمول» للأستاذ الدكتور محمد يوسف حبلص أستاذ علم اللغة أقيمت بمعهد تعليم اللغة العربية بجامعة الإمام محمد بن سعود الإسلامية عام ١٤٣٦ هـ.

المعالجة الحاسوبية، ومن التحشية نجد عملية التوسيم النحوي أو التوسيم **بأقسام الكلام** POS الذي هو معالجة صرفية نحوية، كذلك نجد التحشية **بالتحليل التركيبي** Parsing، وغير ذلك.

وتسمى عملية إضافة المعلومات اللغوية للمدونة اللغوية المحوسبة **تحشية المدونة** Corpus Annotation. ومن التحشيات التي يتم تضمينها في المدونات اللغوية المحوسبة ما يعرف **بـالتوسيم** tagging وهو مصطلح يعني إضافة مستويات من التحشية لبيانات المدونة. أما **الوسم** tag فهو عبارة عن التوسيم نفسه، وقد يتكون من (فونيم أو مورفيم أو كلمة أو شبه جملة أو جملة)[١].

وتتعدد أنواع توسيمات المدونة اللغوية المحوسبة بتنوع المستويات اللغوية المراد الوصول إليها، فالمدونة على المستوى الصوتي يمكن أن توسم كمقاطع صوتية، وعلى المستوى الصرفي يمكن أن توسم بالاعتماد على السوابق واللواحق والجذور، وفي المستوى التركيبي توسم بالتحليل التركيبي والتشجير والأقواس، وفي مستوى تحليل الخطاب يمكن أن توسم المدونة بعرض العلاقات الإحالية، وهكذا[٢].

وأكثر أنواع التوسيم شيوعا في المدونات اللغوية المحوسبة التوسيم بأقسام الكلام[٣]، وهو الهدف الذي يرمي إليها الكتاب. ويعرف التوسيم **النحوي** grammatical أو التوسيم **التركيبي الصرفي** morpho-syntactic **بالتوسيم بأقسام الكلام** POS tagging وهو يختلف عن التوسيم **بالتحليل التركيبي** parsing الذي يستعمل التشجير والأقواس[٤]. ولقد كان

---

(١) Baker, P., Hardie, A., McEnery, A. A Glossary of Corpus Linguistics. Edinburgh University Press, Edinburgh: UK., 2006, pp. 153-154

(٢) McEnery, T, Xiao, R., Tono, Y. Corpus-Based Language Studies, Routledge, USA, 2006, pp. 33-34

(٣) Ibid., p. 34

(٤) McEnery, T. Wilson, A. Corpus Linguistics (An Introduction). Edinburgh University press, Edinburgh: UK, 2011, p.46

شيوع التوسيم بأقسام الكلام في المدونات لأهميته التي يمكن إيجازها في النقاط التالية:

■ يعد مرحلة أولى تسبق عملية تحليل النصوص ومن ثم فإنه يساهم في وسم المدونة على المستوى التركيبي والدلالي والتداولي، كما سيفتح مجالا واسعا للأبحاث اللغوية.

■ لا شك أن كل كلمة في المدونة سيكون لها وسم واحد فقط، وإذا حدث أن كلمة مثل (جنى) وُسمت مرة بفعل ومرة باسم فهذا يعني أن هذه الكلمة لها معان مختلفة أو وظائف في سياقات مختلفة. وجملة مثل: «جنى الجنتين» سيُوسم جزؤها الأول باسم وجزؤها الثاني باسم أيضا. وتوسيم جزئها الأول باسم لا فعل، يشير إلى أن التوسيم بأقسام الكلام سيعين على فهم النصوص لتغير المعنى في الحالين، وسيفيد ذلك في تطبيقات أكثر تعقيدا كالترجمة الآلية مثلا.

■ يحدث أيضا أن يعين التوسيم بأقسام الكلام على إزالة الغموض التركيبي أو الدلالي، فكلمة ذهب في المدونة العربية على سبيل المثال وردت ٦٤ ألف مرة ولا يعلم هل هي المعدن النفيس المعروف أم الفعل. وعندما تكون هناك حاجة لمعلومات خاصة بكلمة، فإن أنواع أقسام الكلام المحيطة بالكلمة هي التي تحدد الاحتمال الصحيح لنوع الكلمة من بين عدد من الاحتمالات الممكنة، وتدعى **الاحتمالات الانتقالية** transition probabilities. فالكلمة ذهب في جملة: ذهب محمد لامع، يمكن أن تكون اسما ويمكن أن تكون فعلا، ولكن جملة القرائن المحتفّة بالسياق الذي وردت فيه الكلمة ـ ومن ضمن هذه القرائن أقسام الكلام لباقي كلمات الجملة ـ هي التي تحدد أنها اسم لا فعل، فقد يسبق ذهب مثلا حرف جر، وبالتالي تكون اسما.

■ تحديد أكثر أقسام الكلام شيوعا في اللغة يمكن بعد توسيم المدونة اللغوية، فقد تتطابق أشكال بعض الكلمات مع اختلاف نوعها أحيانا حتى مع حضور التشكيل، مثل (خَلا) قد تكون حرفا أو اسما أو فعلا، وتوسيمها يقودنا للكشف عن أكثر الأقسام شيوعًا منها في المدونة.

■ تأسيسا على النقطة السابقة، يعين التوسيم بأقسام الكلام على حل بعض الإشكالات التي ترد في غياب التشكيل كما في (من) وهي الاسم (مَنْ) والحرف (مِنْ) والفعل (مَنَّ)، وقد يفيد أيضا في عمليات التشكيل الآلي ويعين عليها فـ(في) مثلا في غياب التشكيل إذا وسمت بحرف ستكون بهذه الصورة (فِـي) وإذا وسمت باسم ستكون (فَـيّ).

■ في مجال معالجة اللغات الطبيعية NLP لا بد أن تتعامل الدراسات الحاسوبية مع حالات غياب التشكيل والأخطاء اللغوية أو المهملات ككتابة الياء مكان الألف المقصورة في مثل: إلي، وهمزات القطع في مثل: ان، وهذا لا يمكن تخطيه بدون وسم المدونات اللغوية على مستويات مختلفة من توسيم أقسام الكلام كما سيُرى.

■ إن وجود مدونة لغوية موسومة بأقسام الكلام يساعد أيضا في تطوير مناهج اللغة وتخطيطها بحيث يتم تضمين الأكثر شيوعا منها في مناهج تعليم اللغات سواء لأهل اللغة أم لغير أهلها.

■ يساعد وجود مدونة لغوية موسومة بأقسام الكلام ـ بطريقة فعالة ـ في بناء المعاجم الحديثة وتنظيمها بالاعتماد على الاستعمال الواقعي للغة.

■ تساعد المدونات الموسومة بأقسام الكلام فيما يعرف **بنمذجة اللغة الإحصائية** Statistical Language Modeling وتعني بناء سلسلة من الكلمات أو الجمل أو العبارات الصالحة لغويا كنماذج لغوية وذلك بالاعتماد على الإحصاء اللغوي من مدونة لغوية. وينتج من النمذجة اللغوية مثلا:

١. معرفة نسبة ورود أنماط الجمل الاسمية أو الفعلية في اللغة.

٢. كما يمكن من خلالها بناء تطبيقات متعددة كالترجمة الآلية واسترجاع المعلومات واستخلاص الأفكار والأشخاص والعلاقات فيما بينها. فلو كان هناك وسم <اسم علم> مع وجود وسم <فعل>في نفس التركيب، لتمكنّا من ربط شبكة من العلاقات ما بين الأفعال

ومن قام بالفعل. وإذا كانت المصطلحات في اللغة عبارة عن **مركبات اسمية** noun phrases تأتي أنماطها في الصور التالية: نمط مفرد (اسم) أو نمط مركب تركيب إضافي (اسم يليه اسم) أو نمط مركب تركيب وصفي (اسم يليه صفة)، فيمكن أن تستخلص الأنماط السابقة من النصوص إذا ما وسمت المدونات.

٣. يمكن الاستـفادة من عملية النمذجة السابقة كذلك في بناء أنطولوجي أو شبكة من المفاهيم في أي مجال معين، كأن تصنف المصطلحات بعد استخلاصها إلى مصطلحات اللغويات والمصطلحات التقنية[١].

ويشكو المشتغلون بحوسبة اللغة والمهتمون بتطوير العربية من عدم وجود مدونة لغوية عربية وافية معالجة حاسوبيا، ويُنتظر من اللغويين القيام بهذه الخطوة كي يتسنى للمشتغلين بحوسبة اللغة القيام بالخطوة التالية وهي المعالجة الحاسوبية للعربية، وهي خطوة لا يمكن الشروع فيها قبل إنجاز مدونة عربية معالجة حاسوبيا. وسوف يترتب على إنجاز هذه المدونة حل المشكلات الجوهرية المبنية على معالجة اللغة حاسوبيا كمشكلة تعليم اللغة العربية، والترجمة الآلية من العربية وإليها، ودفع صناعة المعاجم العربية قدما.

واهتمام الباحثين العرب بالمدونات اهتمام لا يكاد يذكر أمام التقدم الذي أحرزته المدونات في اللغات العالمية الأخرى، حيث مضى أكثر من نصف قرن على إنشاء أول مدونة إنجليزية وعلى نشر دراسة حولها[٢] في حين أن العربية ما زالت تفتقر إلى مدونة عامة مكتملة الأدوات. وثمة أعمال قدمت كمقترحات لوسم المدونات العربية ولكن أغلبها من خارج العالم العربي، ومدعومة من الحكومة الأمريكية خدمة لمصالحها في الشرق

---

(١) Al Qady, M., Kandil, A. Concept Relation Extraction from construction Documents Using Natural Language Processing. Journal of Construction Engineering & Management, Vol. 136, No. 3, USA, 2010, p. 777-780

(٢) مثل مدونة براون Brown لهنري كوسيرا Henry Kucera ودبليو نيلسون W. Nelson التي أنشئت عام ١٩٦٧م ونشرت حولها أول دراسة في نفس العام.

الأوسط[١]، وهي متأثرة بالإنجليزية حيث تستعمل تقسيماتها النحوية أو معتمدة على عمل باكولتر المتأثر بالإنجليزية أيضا، والمعتمدة بياناته على Arabic TreeBank وهو عبارة عن نصوص صحفية فقط ومحدودة الحجم[٢]. وحتى من حاول من أهل العربية، كانت جهوده غير مكتملة حيث لا تعدو أن تكون اقتراحات غير مطبقة[٣]، أو مطبقة ولكنها محدودة في مجال معين[٤]، أو لم تنجز أي نظام يمكن تطبيقه[٥]، أو ليست متخصصة في العربية[٦]، أو تجارية غير متاح العمل عليها[٧]، وهذا يقلل بشكل كبير من فائدتها للأبحاث اللغوية العربية.

---

(١) Habash N., Rambow O., Roth R. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization,Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proc. of the International Conference on Arabic Language Resources and Tools, Cairo: Egypt, 2009

(٢) يذكر مانينق Manning أن «الدقة تنقص بشكل ملحوظ إذا اختلفت طبيعة البيانات، أو اختلف أسلوب الكاتب فيما بين ما درب عليه الموسم وبين البيانات المراد توسيمها». انظر:

Manning, C. D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 2011, pp. 171-189

(٣) الثبيتي، عبد المحسن، وآخرون. طريقة تعتمد على المدونات اللغوية لتجهيز بيانات تدريب واختبار أنظمة الوسوم النحوية. في المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية في دورته الثامنة، مصر: القاهرة، ٢٠١٢

(٤) انظر مثلا:

Sawalha M. Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING. Leeds Uni., UK., 2011

(٥) الحاج، يحيى؛ وآخرون. إعداد وتجهيز نظام إحصائي للتعرف الآلي على المفردات القرآنية: الخصائص والسمات الصرف ـ نحوية وآلية مستحدثة لوسمها. الدورة التاسعة للمؤتمر الدولي لعلوم وهندسة الحاسوب، تونس: الحمامات، ٢٠١٣

(٦) كل ما سبق من أوراق عمل في مجموعها أمثلة.

(٧) مدونة صخر مثالا:

Sakhr. Arabic corpora. 10-9-2017:
http://www.sakhr.com/index.php/en/technology/arabic-resources

إن هذا الكتاب يهدف إلى جمع مدونة متوازنة أقرب للشمول، تتخذ من الإطار النموذجي للمدونة العربية (المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية) KACSTAC إطارا لمدونة من ١٠ آلاف كلمة تقطع وفق قائمة من متغيرات التقطيع التي يعتمد عليها التوسيم النحوي؛ لتقديم نماذج توسيمية نحوية، منطلقة من قواعد النحو العربي، صالحة للتوسيم الآلي، تدرب على جزء منها إحدى خوارزميات تعلم الآلة، ثم يُختبر نظاما التقطيع والتوسيم النحوي على المتبقي منها. ويمكن الاعتماد على المدونة المقطعة والموسمة يدويا في بناء أنظمة للوسم الآلي، كما يمكن بدورهما أن يستعملا في تقطيع وتوسيم مدونات أخرى. وقد تساهم هذه المدونة في تحقيق أهداف متعددة في مجالات لغوية، ولغوية حاسوبية، وأخرى غير لغوية، كحل المشكلات الجوهرية المبنية على معالجة اللغة حاسوبيا، نحو: مشكلة تعليم اللغة العربية، والترجمة الآلية، وصناعة المعاجم، وغيرها.

وحيث لم يقم أي لغوي متخصص في العربية حسب علمي بعملية توسيم نحوي لمدونة عربية منطلقا فيها من نظرية لغوية حديثة لأقسام الكلام، فإن هذا العمل رائد من حيث كونه أول منشور لغوي عربي يتصدى فيه باحث متخصص في اللغة العربية للقيام بهذه المهمة.

الفصل الأول
التوسيم النحوي

تمهيد:

يستعمل اللغويون تقليدياً مصطلح (**مدونة** corpus) للإشارة إلى متن لغوي يمثل الاستعمال الطبيعي للغة ويُعتمد عليه في البحث اللغوي[١]. وتتألف المدونة من نصوص لغوية مكتوبة، أو خطابات منطوقة أو نماذج من المكتوب والمنطوق، وغالباً ما تصمم لتمثل لغة محددة أو تنوعا لغويا[٢].

وفي الستينيات استعمل مصطلح (مدونة) ليشار به إلى متن المادة اللغوية الموجودة بشكل إلكتروني، ويمكن أن تعالج حاسوبياً لأغراض مختلفة كالبحث اللغوي وهندسة اللغة[٣]. ولسرعة تطور الحواسيب فقد ازداد مع تطورها حجم المدونات واختلافها وسهولة الوصول إليها، وفي الوقت نفسه ازداد تطور البرمجيات الخاصة بمعالجة المدونات والوصول إلى المعلومات التي تحويها. لقد أصبحت المدونات الحاسوبية

---

(١) Garside, R., Leech, G., McEnery, T. Corpus Annotation, Routledge, USA, 2013, p.1

(٢) Ibid.

(٣) تعد هندسة اللغة أحد المجالات التي تقع بين اللغة والحوسبة، وتستعمل أدوات مثل القواميس المقروءة آليا والمحللات التركيبية لمعالجة اللغات الطبيعية في تطبيقات مثل التعرف على الكلام وتوليفه والترجمة الآلية؛

Teuber, W. Cermakova, A. Corpus Linguistics: A short introduction, Continuum; UK, 2008, p.53

بصورة سريعة موردا عالمياً للبحث اللغوي خلال الستينيات الماضية[١]. ويرى **ليتش** Leech أن هذه المرحلة ليست مرحلة عرضية وأن سنة ١٩٦١م هي السنة التي يمكن للسانيي المدونات أن يعيدوا النظر إليها كتأريخ بداية المساعي والمساهمات التي عُرفت الآن **بلغويات المدونات** Corpus Linguistics أو باصطلاح أكثر دقة **لغويات المدونات الحاسوبية** Computational Corpus Linguistic[٢]. لقد كان هذا التاريخ هو التاريخ الذي بُدئ العمل فيه على أول مدونة حاسوبية عرفت فيما بعد باسم مدونة **براون** Brown، وانتهى العمل بها وأصبحت جاهزة للنشر عام ١٩٦٤ وبحجم مليون كلمة[٣].

وحيث إن تطور المدونة تزامن مع تطور الحواسيب منذ الستينيات، تبدو مدونة براون الصادرة عام ١٩٦٤م صغيرة بجانب المدونات التي صدرت في التسعينيات وما بعدها. وقد تضمن **بنك الإنجليزية** Bank of English (BoE) الذي أطلق عام ١٩٩١م أكثر من ٣٠٠ مليون كلمة[٤]، واحتوت **المدونة الوطنية البريطانية** British National Corpus (BNC) التي اكتمل بناؤها عام ١٩٩٤م على ١٠٠ مليون كلمة[٥]، وتضم المدونة العربية التي أطلقت عام ٢٠١٢ أكثر من بليون كلمة[٦].

ولا تقاس قيمة المدونة على أية حال باعتبارها أداة للبحث اللغوي بحجمها فقط، حيث إن تنوعها من حيث نوعية نصوصها المكتوبة أو المنطوقة ومواضيعها فضلا

---

(١) Corpus Annotation, 2013, p.1

(٢) Ibid.

(٣) Ibid.

(٤) Kukulska-Hulme, Agnes. Language and Communication: Essential Concepts for User Interface and Documentation designed. Oxford University Press, Oxford: UK, 1999, p.22

(٥) BNC. How the BNC was created. 10-9-2017:
http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=creation

(٦) مدينة الملك عبد العزيز للعلوم والتقنية. المدونة العربية. ٢-٩-٢٠١٦:
http://corpus.kacst.edu.sa/index.jsp

عن توسيمها، يوازي أهمية الحجم بل ويزيد أحيانا[١]. فلقد أثرى التوسيمُ المدوناتِ الحاسوبية ذات النصوص المتنوعة بالمعلومات اللغوية وجعلها منجماً خصباً ومهماً للبحث اللغوي.

وفي هذا الفصل مقدمة موجزة عن التوسيم ودواعيه وما ينبغي أن نلتزم به عند تنفيذه، ثم أنواع التوسيم مع تفصيل للتوسيم اللغوي، يلي ذلك شرح وبيان للتوسيم النحوي وأهميته.

١-١ مفهوم التوسيم بوجه عام:

يُعرّف التوسيم بمعناه العام بأنه إسناد الرموز لكل من المعلومات النصية كاسم الكاتب، وللتحليل اللغوي التفسيري كنوع الكلمة، وذلك لنصوص المدونة. ويستعمل في معناه الضيق ليشار به إلى إضافة التحليلات اللغوية ـ كأقسام الكلام والتحليل التركيبي وغيرها من مستويات التحليل المختلفة بعد تحويلها لرموز ـ إلى نصوص المدونة[٢]. فعملية التوسيم هي إضافة معلومات لغوية تفسيرية مرمّزة لمدونة حاسوبية مكتوبة أو منطوقة، أما التوسيم فيشير إلى المنتج النهائي لهذه العملية الذي تُسند فيه رموز لغوية ذات دلالة إلى المادة اللغوية نفسها الممثلة حاسوبياً[٣]. وقولنا «معلومات لغوية تفسيرية» يشير إلى أن التوسيم منتج لفهم العقل البشري للنص[٤]. فالجملة: «استقبل خادم الحرمين الشريفين» عند توسيمها يكون المنتج النهائي منها بعد إدخالها في موسِّم ستانفورد كالتالي: «استقبل / VBD خادم / NN الحرمين / DTNNPS الشريفين / DTJJ». وهذه الرموز المسندة للكلمات تشير إلى معلومات لغوية زُوّد بها الموسم. فالرمز VBD المسند

---

(١) منقول بتصرف، انظر:

Corpus Annotation, 2013, p.2

(٢) Corpus-Based Language Studies, 2006, p.29

(٣) Corpus Annotation, 2013, p.2

(٤) Ibid.

إلى استقبل يشير إلى أن الكلمة فعل ماض، والرمز NN المسند إلى الكلمة خادم يشير إلى إنها اسم مفرد أو جمع، والرمز DTNNPS يشير إلى أن الحرمين اسم علم مجموع ومعرف بأل، ويشير الرمز DTJJ المسند لكلمة الشريفين إلى أنها صفة معرفة بأل[١].

ويختلف توسيم المدونات في مفهومه الضيق عما يعرف **بالتعليم** mark –up وما يعرف **بالبيانات الواصفة** metadata. فتعليم المدونات يزودنا بمعلومات عما يتعلق بمكونات المدونة وبنائها النصي لكل نص[٢]. والمدونة تستعمل هذا التعليم لتبرز مثلًا بعض الكلمات بحروف مائلة، والعناوين بحجم مختلف، وتحدد بدايات الجمل ونهاياتها وغيرها مما يتعلق بشكل وحدات النص. وفي المدونة المنطوقة يستعمل ليحدد به بداية حديث المتكلم ونهايته مثلاً[٣]. أما الميتاداتا (البيانات الواصفة) فهي تزود بمعلومات عن النص نفسه، كاسم الكاتب، وجنسه، وعمره، وسنة النشر، وغير ذلك مما يتعلق بالنص نفسه[٤]. وهي مفيدة لاستخلاص بعض الخصائص والحقائق المتعلقة باللغة عند جنس معين أو فئة عمرية محددة.

إن عملية توسيم النص هي عملية **لغوية واصفة** metalinguistic فـبدلاً من أن تخبرنا عما يتضمنه النص وما يتألف منه، تقدم لنا معلومات عن لغة هذا النص[٥]. ولا يمكننا أن نقول إن المدونة بعد التوسيم تتضمن معلومات جديدة، فكل ما في الأمر هو أنه قد أضيف إليها معلومات صريحة عن البيانات التي تحتويها المضمنة أصلا فيها، وليست عملية إنشاء أو تحويل للمعلومات[٦]. وبعبارة أخرى، تحديد كلمة بوصفها اسما، لا

---

(١) المثال بتوسيمه مقتبس من مدونة arTenTen12 الموسمة بموسم ستانفورد في موقع سكتش انجن، انظر: sketchengine.co.uk/auth/corpora

(٢) Corpus-Based Language Studies, 2006, p.29

(٣) ماكنري، توني؛ هاردي، أندرو. لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق. ترجمة: سلطان المجيول، جامعة الملك سعود، السعودية: الرياض، ٢٠١٦، صص. ٥٦

(٤) السابق، نفس الصفحة.

(٥) Corpus Annotation, 2012, p.31

(٦) Ibid.

يعني أننا نقلناها إلى الاسمية في هذا العمل، فهي اسم ضمنياً ولكننا أضفنا وسم اسم لها صراحة[١].

ويُستعمل المصطلحان: **التحشية** annotation **والتوسيم** tagging في أدبيات لغويات المدونات ليشار بهما إلى نفس العملية ولكن بفارق بسيط. فعندما تضاف رموز خاصة للكلمات لتدل على خصائص معينة كأقسام الكلام مثلا يقال **توسيم** tagging بدلاً من **تحشية** annotation، وعندما يكون إسناد الوسوم إلى عبارات أو جمل أو فقرات كما في المستويات اللغوية المتعلقة بالتركيب والأسلوب يقال **تحشية** annotation [٢]، وقد يستعملان مترادفين ليشيرا إلى المعنى ذاته[٣].

١-١-١ لماذا نوسم المدونة؟

إن الاستعمال الأساسي للمدونات غير الموسمة هو البحث عن كلمة أو سلسلة من الكلمات، وهذه المهمة هي ما تقوم بها البرامج السياقية التي برمجت وصممت لتسمح بإجراء هذا النوع من البحث اللغوي كالبحث عن المتلازمات اللفظية أو بعض الأبنية التركيبية المحدودة[٤]. وإذا قصرنا البحث اللغوي في المدونات على هذه الطريقة، فكثير من الأبحاث المعتمدة على المدونات في مستويات لغوية مختلفة لن تكون ممكنة. فلنفترض مثلاً أننا نريد إجراء دراسة عن حرف العطف الواو في العربية. وحرف الواو في العربية يتصل بالكلمة كتابيا، والبحث عنه في المدونة سيكون بالبحث عن السلسة الحرفية (و+*)، حيث تشير علامة + مع النجمة إلى أن هناك كلمة تلي الحرف مباشرة

---

(١) Corpus Annotation, 2013. p. 31.

(٢) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٤٤٢؛ Corpus Linguistics (An Introduction), 2011, p.46

(٣) Corpus Based Language ,2006, p.29

(٤) Biber, D, Conrad, S., Repper R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, USA, 2006, p.257

ونحن مضطرون للبحث عن حرف العطف الواو بهذه الصورة لأنه لا يرد إلا على هيئتها. وسيدخل فيها نتائج ليست موضوع البحث كالكلمات المبدوءة بواو من أصل الكلمة لا عطفا، على نحو: ويلي ووفير وغيرها كثير. وسيتحتم علينا التعامل يدوياً مع النتائج مستبعدين العديد من المزايا الخاصة بمنهج البحث المعتمد على المدونات كالموثوقية والدقة والسرعة. ولحل هذه المشكلة تحديداً، ينبغي أن تستعمل مدونة موسمة بأقسام الكلام، تفصل حرف العطف الواو عن الكلمة وتحدد قسمه الكلامي وتسند رمزه للحرف. وبالتالي لن تظهر (ويلي ووفير) في النتائج لاختلاف نوعها، وسيكون بإمكاننا تحديد واو العطف عند تنفيذ عملية البحث.

وهكذا، فإنه من الأهمية بمكان أن نكون قادرين على توسيم المدونة؛ لأن المدونات الموسمة تمكّننا من:

١. استرجاع المعلومات واستخلاصها.

حيث لا تكون المدونات مفيدة إذا لم يكن استخلاص المعرفة والمعلومات منها ممكنا. ولكي نتمكن من ذلك، علينا تزويدها بالمعلومات من خلال التوسيم[١]. فالمدونات الخام في شكلها الهجائي لا تحتوي على معلومات مباشرة. وعلى المستوى النحوي مثلا تعوق المدونات الخام الحالية من التوسيم عمل العديد من التطبيقات والأدوات التي تنفذ عليها، ولو أردنا البحث عن كلمة (في)، ستكون إما حرف جر مجرد، كما في: أدرس في الجامعة، أو حرف جر أسندت له ياء المتكلم، كما في: أثر فيّ، أو اسماً، كما في: أستظل بفيّ الشجر، أو علماً لمؤنث، كما في: أقبلت فيّ، أو فعلاً مسنداً إلى ياء المخاطبة، كما في: فِيِّ الدلو ماء. وهكذا تتعدد المعاني باختلاف نوع الكلمة، ولا يمكن كشفها من خلال شكلها الهجائي خصوصا مع غياب التشكيل. ولكن إذا كانت المدونة موسمة توسيماً نحوياً مثلاً فكل كلمة في مجموع النتائج ستظهر مرفقة بنوعها، وهو أمر ضروري لأي

---

(١) Corpus Annotation, 2013, p.4; Corpus-Based Language Studies, 2006, p.30

صانع معجم يسعى لتحسين معجمه، حيث عيب على المعجميين سابقاً اعتمادهم على المدونات غير الموسمة كمورد أساسي للمعجم[١].

وثمة مثال آخر يوضح فائدة المدونات الموسمة في استرجاع المعلومات واستخلاصها، وهو كلمة (أكل) بوصفها فعلا، حيث يمكن أن تكون فعلاً ماضياً مبنياً للمعلوم، أو فعلاً ماضياً مبنياً للمجهول، وتوسيمها بهذه المعلومات اللغوية يُمكّن **قارئ النصوص الآلي** Speech Synthesize من التميز بينهما، فينطق الكلمة نطقاً سليماً حتى مع غياب التشكيل الذي غالباً ما تفقده النصوص. وهكذا يزود التوسيم القارئ الآلي بالمعلومات التي يحتاجها؛ كي يتمكّن من النطق السليم لها.

ويمكّن التوسيم أيضاً المحللين (بشرًا وآلات) من استرجاع التحليلات التي لا يمكن أن يقوموا بها دون مدونة موسمة. فحتى لو لم تكن تجيد الفرنسية مثلاً، ستكون قادراً على التعامل معها باستعمال مدونة موسمة[٢]. حيث يمكن أن تبحث عن أدوات التعريف في الفرنسية بمعرفة وسمها فقط. وستجد أن الحال ليس كما في العربية إذ تتصرف من حيث الجنس والعدد. وسيزيدك توسيم ما بعدها قدرة على التعامل مع ما يتغير منها كتابيا إذا لحقته حروف متحركة vowel.

٢. إعادة الاستعمال:

ويقصد من ذلك أن يضاف التحليل اللغوي للمدونة من خلال توسيمها الذي يتيح إعادة استعمالها أكثر من مرة وفي الغرض المقصود[٣]. إذ لا يُنفذ في كل مرة يُراد استعمال المدونة لأي غرض. وهذا يوفر الكثير من المال والوقت والجهد، فمتى ما أضيف التوسيم لأي مدونة، ستكون موردا ذا قيمة أكثر من حالها قبل التوسيم. وسيعاد استعمالها للإجابة

---

(١) Corpus Annotation ,2015, p.4

(٢) Corpus –Based language studies, 2006, p.30

(٣) Ibid.

عن أسئلة بحثية مختلفة من قبل العديد من الباحثين[١]. فإذا وسمت نحويا ستجيب مثلا عن ماهية أكثر أقسام الكلام شيوعا، أو ماهية أكثر الأنماط اللغوية استعمالا، أو كيف تبنى الجملة الفعلية، وما إلى ذلك. ولن يعاد توسيمها في كل مرة حتى وإن كان توسيمها لأول مرة لغرض مقصود.

٣. تعددية الاستعمالات:

إذا كان للتوسيم أغراض أو تطبيقات مختلفة فهذا يعني أن المدونة سوف تتعدد استعمالاتها وتزداد بعد توسيمها. ويعزز من ذلك ما ذكر سابقاً (إعادة الاستعمال)، حيث إن هناك أغراضا مختلفة تجعلنا نستعمل المدونة الموسمة أكثر من مرة لمقاصد ربما لم يفكر بها من قام بعملية التوسيم أبداً[٢]. وقد كانت الأمثلة المطروح بعضها سابقاً من هذه التطبيقات. ومن هذه التطبيقات أيضا ما يتعلق بحوسبة اللغة ومعالجتها آليا، كالترجمة الآلية، واستخلاص المصطلحات، وبناء النماذج الحاسوبية للغة واسترجاع المعلومات[٣].

والجدير بالذكر أن بعض مستويات التوسيم تقوم على بيانات التوسيم لمستويات أخرى، وتلك وظيفة أيضا، فالتوسيم النحوي مثلاً يُعد الخطوة الأولى التي تسبق التوسيم بمستويات أصعب كالتوسيم الدلالي والتداولي[٤].

٤. الحصول على بيانات واقعية:

يمثل توسيم المدونات سجلاً تحليلياً واقعياً للغة، وهو سجل مفتوح للدراسة والتمعن والمناقشة[٥]. وبالرغم من أن المدونات أحياناً قد تنتقي نصوصها، أو يستأذن أصحابها المسجلون صوتيا، إلا أن ذلك الإنتاج اللغوي في الحالين عملية غير مُدركة.

---

(١) Corpus Annotation, 2013. P.6

(٢) Ibid.

(٣) Ibid.

(٤) Ibid.

(٥) Corpus –Based-Language Studies,2006, p.30

وبالتالي فإن بيانات توسيمها تعد طبيعية وتمثل الاستعمال الواقعي. كما يمكن أن تكون مورداً مرجعياً لمقارنته ومقابلته مع ما تحويه الموارد الأخرى التي لم تُبن على الاستعمال الحقيقي للغة[١].

٥. الوصول السهل والسريع:

إن استعمال مدونة موسمة لا يتطلب من الباحث عملية جمع البيانات وتوسيمها، فكل المعلومات المتوفرة قد نفذها آخرون، وعليه فقط أن ينتقي منها ما يحقق أهدافه، فضلا عن أن تفاصيل ما نفذه الآخرون مما يتعلق بعملية التوسيم التي ترفق عادة في ملف يسمى ملف التوثيق أو دليل التوسيم، يمكن أن يصل إليها الباحث بسهولة [٢].

٦. النمذجة والتكميم:

ويقصد **بالنمذجة (الاعتيان)** Sampling عملية أخذ عينة للاستدلال بها على المجموع اللغوي، أما **التكميم** Quantification فهو التعامل مع المعلومات اللغوية إحصائياً. وقد وضعت المدونة أساسا لتمثل عينة أو نموذجاً لعموم اللغة، ومن ثم فإن نموذجاً موسما منها يمكن أن يعمم على اللغة[٣]. وتكميم هذه النماذج الموسمة في لغويات المدونات ذو أهمية أكبر من التكميم في الأشكال الأخرى من اللغويات التجريبية؛ لأنه سيزودنا بمعلومات عن اللغة بدلا من نموذج فريد يتخذ للتحليل[٤].

وبرغم ما سبق ـ مما يمكن أن يقدمه لنا توسيم المدونات ـ برزت عدة انتقادات له نلخصها فيما يلي:

١. إن توسيم المدونات يقدم لنا مدونة مبعثرة الكلمات والنصوص، حيث تزعم **هنستن** Hunston أن أكثر عمليات التوسيم تضاف للنصوص نفسها، ومن المهم أن يطلع

---

(١) Corpus Linguistics (An Introduction), 2011, p.130

(٢) Ibid.

(٣) Ibid.

(٤) Ibid.

الباحث على النص الخام دون أن يكون مشوشـاً بمسميات توسيمية[١]. أما **سنكلير** Sinclair فيرى أنها تؤثر في صحة النص واكتماله[٢]. ومع تطور برامج معالجة المدونات، لم يعد هذا الانتقاد محل نظر، فمعظم أدوات توسيم المدونات تسمح لمستعمل المدونة بإخفاء التوسيم لمطالعة النص الأصلي الخام، وينبغي أن يوجه هذا الانتقاد إلى أدوات الاسترجاع والتصفح لا إلى توسيم المدونات[٣].

٢. إن توسيم المدونات يفرض على المستخدمين تحليلا لغويا معينا[٤]. وهذا ليس صحيحا، فرغم طبيعة التوسيم التفسيرية إلا أن مستخدم المدونة لا يجبر على قبوله، ويمكن أن يفرض تفسيراته الخاصة إذا شاء أو أن يتجاهل التوسيم بكل بساطة، وفكرة إمكانية تعدد التفسيرات لا بد أن تكون أمرا مقبولا منذ البدء في عملية توسيم المدونة. فإذا تركنا المدونة دون توسيم لا يعني ذلك عدم وجود عمليات تفسيرية تقع عند تحليل المدونة. بل إن هناك عمليات تفسيرية متعددة تحدث عند استعمال الباحثين للمدونة الخام. إنها موجودة لكنها مخفية وغير معروضة في المدونة، وبالتالي يكون تضمينها في المدونة ميزة لا عيباً[٥].

٣. إن توسيم المدونات يزيد من قيمتها المادية فتقل إمكانية الوصول إليها وترقيتها وتوسيعها[٦]. والواقع أنه ليس بالضرورة ذلك، فكثير من المدونات الموسمة متوفرة للعامة مثل: **مدونة لندن ـ لند للإنجليزية المنطوقة** London-Lund Corpus of Spoken English LLC **ومدونة لانكاستر \IBM للإنجليزية المنطوقة** Lancaster/ IBM Spoken English Corpus SEC، وغيرهما كثير. وأما المدونات الموسمة التي لا تتوفر لعامة

---

(١) Corpus –Based Language Studies, 2006, p.32

(٢) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٢٨٨

(٣) Corpus -Based Language Studies, 2006, P.32

(٤) Ibid.

(٥) Ibid.

(٦) Ibid.

المستعملين فإن سبب ذلك هو مسألة الحقوق الفكرية وليس التوسيم[١]. كما أن هذا الانتقاد يزيحه تماما إمكانية التوسيم الآلي، وتوفير العديد من الأنظمة والبرامج المجانية التي تقوم بعملية التوسيم.

٤. ترى هنستن Hunston أن عملية توسيم المدونات آلياً لا يمكن أن تقدم نتائج تطابق نتائج التوسيم اليدوي ١٠٠٪، ومن جهة أخرى تتشابه الطريقتان في حدوث الأخطاء في كل منهما[٢]. ويرى **سنكلير** Sinclair أن توسيمها يدوياً أو بمساعدة الحاسوب يقلل من اتساق النتائج[٣]. والواقع أن عدم الصحة وعدم الاتساق صفتان بشريتان. وفي هذه الحالة يكمل كل من المحلل البشري، والمحلل الآلي بعضهما الآخر بتقديم دقة واتساق مقبولين[٤]. فضلا عن أن المزايا التي تتيحها المدونات الموسمة تفوق ما قد يلحق بالتوسيم من عدم اتساق.

وعلى أي حال، فإن كل ما سبق من الانتقادات يمحوها مبدأ يعد معيارا للتوسيم، وهو أن الموسم ليس مسؤولاً عن جودة التوسيم إلا إذا نص على ذلك[٥]. إن التوسيم يعني فقط تضمين التحليل اللغوي في المدونة، وهو عمل يؤديه اللغويون منذ قرون مع اختلاف آلياته وفلسفتها[٦].

١-١-٢ معايير التوسيم:

يشير **ليتش** Leech إلى أن قبولنا للتوسيم كوسيلة مفيدة وثرية بالمعلومات ينبغي أن يتوقف على خبرات من أضافوا التوسيم للمدونة، أو مدى فائدة المقترح الذي تبنوه أو

---

(١) Corpus -Based Language Studies, 2006, P.32

(٢) Ibid.

(٣) Ibid.

(٤) Ibid.

(٥) Ibid.

(٦) Ibid.

اعتمدوه[١]. وفي تاريخ التوسيم القصير، لا يوجد بأي وجه من الوجوه ما هو غير مألوف، أو صعب لمصممي المدونات عند إضافة التوسيم لها، في حين أن الآخرين قد يجدونه عملية صعبة ومستحيلة الاستعمال[٢]. ولتجاوز تلك المشكلة يقترح **ليتش** Leech عددا من التعليمات، أو المعايير التي ينبغي أن تطبق على أي مشروع لتوسيم نصوص المدونات، وهي كما يلي[٣]:

١. إمكانية استرجاع النص الخام أو الأصلي بسهولة، أي ينبغي أن تكون المدونة الخام قابلة للاسترداد.

٢. إمكانية فصل التوسيم وعزله عن النص الأصلي وتخزينه في ملف منفصل إذا احتيج إليه.

٣. سهولة وصول مستعمل المدونة إلى ملف التوثيق الذي يتضمن ما يلي من المعلومات:

أ. مقترح التوسيم، وهو ملف يصف ويشرح مقترح التوسيم الذي تقوم عليه عملية التوسيم. وينبغي أن يتوفر مقترح التوسيم لجمهور الباحثين في كتيب يدوي أو ملف إلكتروني منفصل، وذلك للفائدة العملية وعلى افتراض أن الباحثين سيجدونه مفيداً وذا قيمة بدلا من عمل ذلك بأنفسهم، حيث استهلاك الوقت والجهد لإنجازه.

ب. كيف تمت عملية التوسيم، وممن وأين؟

ج. إلى أي حد فحصت هذه المدونة؛ وكم بلغت درجة الدقة للتوسيم؟ وما مدى اتساق أو ثبات تطبيق التوسيم؟ فالتوسيم غالبا ما يتضمن الخطأ، أو عدم الاتساق أو الغموض في بعض عناصره، ولا بد أن تؤخذ بعين الاعتبار جودة عملية التوسيم.

---

(١) Corpus Annotation, 2013, p.6

(٢) Ibid.

(٣) Corpus Annotation, 2013, pp.6-8; Corpus Linguistics (An Introduction), 2011, pp.33-34

٤. لتفادي سوء الاستعمال وسوء التطبيق، من الأفضل أن تعتمد مقترحات التوسيم بقدر الإمكان على نظريات متفق عليها أو محايدة، فمثلاً التوسيم النحوي غالبا ما يتبنى النظريات التقليدية بدلاً عن الاعتماد على نماذج نظرية حديثة.

٥. لا يمكن الادعاء بأن أحد مقترحات التوسيم هو مقترح صحيح تماما، بل إن مقترحات التوسيم تختلف حسب الغرض المراد، وليس ثمة مقترح صالح ومقترح غير صالح. فإذا كان المراد مثلا من توسيم المدونة بأقسام الكلام **التحليل التركيبي** Parsing. لا بد أن يفرّق في وسوم أقسام الكلام بين حروف الجر وحروف العطف. وما يصلح من المقترحات التوسيمية للمدونات المكتوبة قد لا يصلح للمنطوق منها، وما يصلح للغة لا يصلح للغة أخرى.

وبرغم ما سبق، فإنه كثيراً ما يفضل مقترح توسيمي على مقترح توسيمي آخر في المستويات اللغوية المختلفة، فيُتخذ المقترح التوسيمي ذو الأفضلية معياراً لتطوير مقترح توسيمي آخر قاصر[١]. وإذا طبقت نفس المقاييس أو المعايير في كل المقترحات التوسيمية سيكون من السهولة تبادلها بين مختلف الباحثين.

لقد كانت مبادرة EAGLES المحاولة الأولى عالميا لتحديد معايير توسيم المدونات للغات الاتحاد الأوروبي، ومن ذلك تقديمها مجموعة أساسية لخصائص التوسيم النحوي التي تشكل مقياساً ومعيارا لكل اللغات. حيث طبقت على لغات عالمية أخرى خارج الاتحاد الأوروبي كلغات أوروبا الشرقية واللغة العربية[٢].

وهناك العديد من الأسباب التي تجعل المعايير في كثير من الحالات ضرورية، أوجزها فيما يلي[٣]:

---

(١) Corpus Annotation, 2013, p.31

(٢) Corpus Linguistics (An Introduction), 2011, p. 38

(٣) Corpus Annotation, 2013, p. 232

١. تنفذ عملية التوسيم على لغة واحدة من خلال فرق بحثية مختلفة في دول عديدة، وبدون المعايير لن يتشابه أي عمل مع غيره، وستتعدد الأعمال لنفس المهمة على نفس المدونة.

٢. إن عملية توسيم مدونة كبيرة نشاط قيم جداً وحاليا لا يمكن أو من الصعب أن يعاد استعمال الأدوات المطورة لمجموعة بحثية ما من مجموعات أخرى. ووضع معايير للتوسيم يعني أن الأدوات المطورة للتوسيم ولمدونة واحدة سيعاد استعمالها وستتم مشاركتها مع الآخرين، ومن ثم تحفظ الأموال والجهود والأوقات.

٣. تعين المعايير على استثمار أبحاث المدونات الموسمة، فيمكن للمدونة الموسمة باتباع مقياس أو معيار ما أن تستعمل لتطبيقات مختلفة. حيث إن العديد من البرمجيات والمجموعات البحثية لن تضطر لإعادة صوغ مقترح التوسيم ولا الرجوع لتفاصيله، وبالتالي يقدم مقترح التوسيم المعياري كمنتج جاهز يمكن أن يطبق بسهولة.

٤. العديد من المشاريع التي نُفذت أساساً على الإنجليزية كالتوسيم النحوي طُبقت على لغات أخرى، وتقييس التوسيم سيضمن أن تكون المدونات الموسمة من مجموعات بحثية مختلفة في دول عديدة متشابهةً، كما يحدث في المدونات المتوازية للغات مختلفة، وهذا هو قوام وأساس البحث العلمي.

١-٢ أنواع التوسيم:

يمكن أن توسم المدونة يدويا أو آليا بصورة تامة، أو بطريقة شبه آلية بالتفاعل بين الإنسان والآلة. ويلجأ المهتمون إلى التوسيم اليدوي، إذا لم تتوفر أداة للتوسيم، أو عندما تكون دقة الأدوات المتوفرة غير كافية لجعل الوقت المستثمر في التصحيح اليدوي أقل من الوقت المستهلك لو كان التوسيم كله يدوياً. ولأن التوسيم اليدوي يكلف وقتا ومالاً، فهو لا ينفذ إلا في المدونات الصغيرة[١].

---

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٦٠

وفي التوسيم الآلي التام يعمل الحاسوب وحده كموسم باتباع العديد من القواعد والخوارزميات المحددة مسبقاً من قبل مبرمج أو باستعمال خوارزميات تعلم الآلة[١]. وقد يكلف تطوير أداة التوسيم الآلي جهداً ووقتاً ومالاً، ولكنها متى ما اكتملت فإنه يوسم بها أجزاء كبيرة جداً من المدونة باتساق وفي وقت قصير[٢]. إن كثيراً من محاولات التوسيم النحوي نفذت آلياً في لغات متعددة كالإنجليزية والفرنسية، ولم تتجاوز نسبة الخطأ فيها ٣٪ [٣]. وهذا يجعل الباحثين ينظرون إلى هذه الطريقة باعتبارها مصدرا موثوقا ومعتمدا.

وإذا كانت مخرجات عملية التوسيم الآلي غير موثوق بصحتها، كما في بعض **المحللات التركيبية** parsers، أو كانت صحيحة في غالبها، والدقة المطلوبة ليست كافية لاستعمالها كما في المدونة المدربة التي تستعمل في تحسين أداة التوسيم، عندها يكون التدخل البشري مطلوباً، وهو أمر أسرع بكثير من أن تنفذ عملية التوسيم كاملة بطريقة يدوية[٤]. وتزود بعض أدوات التوسيم بواجهة (آلية ـ بشرية) تسمح للمحلل البشري أن يفك حالات اللبس والغموض التي لا تستطيع الآلة فكها. وقد يُنتِج التوسيم الآلي ـ اليدوي، أو شبه الآلي نتائج أكثر صحة من التوسيم الآلي التام، ولكنه أبطأ وأكثر كلفة[٥].

إن أغلب أنواع التوسيم المتوفرة في المدونة الكبيرة نفذت يدوياً أو شبه آليّ، وليس في تلك الطرائق الثلاثة (الآلية ـ شبه الآلية ـ اليدوية) ما تخلو نتائجه من الأخطاء. والتوسيم الآلي بأقسام الكلام في الإنجليزية حقق نسبة عالية من الصحة بلغت ٩٧٪. فالتوسيم اليدوي لا يمكن أن نضمن أن نتائجه ستكون خالية من الأخطاء أيضا، حيث لا يوجد محلل بشري وإن كان لغوياً كاملاً لا يخطئ تماما[٦].

---

(١) Corpus-Based Language studies, 2006, p. 33

(٢) Ibid.

(٣) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٦١-٦٣

(٤) Corpus-Based Language studies, 2006, p. 33

(٥) Ibid.

(٦) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٦١-٦٣

وكما تتباين طرائق إضافة المعلومات للمدونة، تتباين أيضا أنواع المعلومات نفسها التي تزود بها المدونة. فثمة معلومات حول النص نفسه، ومعلومات حول الكتابة الهجائية للنص، ومعلومات حول لغة النص. وفيما يلي موجز لأنواع المعلومات التي تزود بها المدونات مفصلة الحديث عن نوعها الأخير: الهدف.

١-٢-١ المعلومات النصية:

وهي المعلومات التي تتعلق بالنص ككتلة واحدة مثل: اسم الكاتب، وعمره، وجنسه ومجال النص، وموضوعه، وسنه نشره، وغيرها. وهي معلومات لا يمكن تضمينها في اسم الملف الذي يحمل النص، حيث إنها أوسع من الأمثلة المذكورة وأكثر تفصيلاً.

وقد كانت مثل هذه المعلومات توضع سابقا في رأس كامل النص الأصلي. أما الآن فتوضع في ملفات خارجية، أو في قواعد بيانات مستقلة تربط بنصوص المدونة ولا توضع في داخل النص. ثم يستفاد من هذه المعلومات بربطها ببرامج استرجاع المعلومات المنفِّذة لعمليات البحث والفهرسة[١]. فالباحث بعد إضافة مثل هذه المعلومات للمدونة، يستطيع مثلا أن يستعلم آليا عن نصوص الإناث من الفئة العمرية ٢٥ - ٣٠ من خلال متغير (جنس الكاتب) باختيار القيمة (أنثى)، وعن الأديبات منهن من خلال متغير (مجال النص) باختيار القيمة (أدب).

وهناك عدة أنظمة يمكن من خلالها إضافة هذا النوع من المعلومات وعلى رأسها **عداد الكلمات والمكشاف السياقي بنظام أطلس** Word COunt and COncordance on Atlas (COCOA) **ومبادرة ترميز النصوص** (Text Encoding Initiative TEI). ويتألف نظام COCOA من مجموعة من الأقواس ذات الشكل < > تضم داخلها اسم المتغير ثم ما يعبر عنه، على نحو: <Ahmed Ali N> حيث N اسم المتغير لاسم الكاتب وما يليها اسم الكاتب، وهكذا لباقي المعلومات[٢]. أما TEI فهو نظام أكثر شمولًا وتعقيداً

---

(١) Corpus Linguistics (An Introduction), 2011, p. 39

(٢) Ibid.

ويستعمل لغة التعليم المعروفة عالمياً SGML التي تستعمل الأقواس المثلثة < > أيضا، وتضع داخلها الوسم المراد الذي يعبر عن أي معلومة نصية، ثم توضع المعلومة النصية، ثم تختم بنفس الوسم مسبوقاً بعلامة (/)، فعلي سبيل المثال:

<creation> 2015 <creation />

حيث يعبر المتغير creation عن سنة التأليف، والرقم ٢٠١٥ هو المعلومة[١].

١-٢-٢ معلومات التهجئة (رسم الكلمات):

قد يبدو في تسمية هذا النوع من المعلومات تناقض؛ لأن رسم الكلمات هو ما يتمثل أمامنا كنص، بينما التوسيم أو إضافة المعلومات تعبر عن النص[٢]. ولكن يمكن أن تكون معلومات التهجئة تفسيرية للنص في تمييزها للوظائف التي تؤديها المعلومات الظاهرة المتنوعة في النص، تلك التي قد تلتصق أحياناً ببعض الكلمات[٣].

وتضمين مثل هذا النوع من المعلومات يفك اللبس الناتج من غموض بعضها، باعتبار أن بعضها له وظيفة وحيدة لا يؤدي غيرها مثل علامة الاقتباس« » والبعض الآخر يؤدي أكثر من وظيفة، وتوسيمه يفك اللبس، كما في العلامات النصية التالية[٤]:

١. في الإنجليزية مثلاً، الأحرف الكبيرة تشير إلى أن الكلمة علمًا، أو أن الكلمة في بداية جملة.

٢. في العربية ولغات أخرى، تشير النقطة (.) إلى نهاية جملة أو إلى اختصار.

٣. في العربية ولغات أخرى، الحروف العريضة تشير إلى عناوين فرعية أو مصطلحات أو أسماء.

---

(١) Corpus-Based Language studies, 2006, p. 35

(٢) Corpus Annotation, 2013, p.14

(٣) Ibid.

(٤) Ibid.

وفي بعض الحالات تُضمّن هذه المعلومات في ترميز سجل التهجئة الخاص بالنص، فتكون مساهمة مفيدة في معالجة المدونة وتوسيمها. وقد كان الرمز (\٠) في مدونة LOB يستعمل للإشارة إلى الاختصارات المتكونة من كلمة واحدة. فالكلمة (in.) اختصار الكلمة inch ترمز هكذا: \0in. وهذا يساعد آلياً في التمييز بين in كحرف جر في نهاية الجملة، وin كاختصار. وحيث إن مثل هذه المعلومات لم تطبق على المدونات بشكل مطّرد، فمن غير المناسب أن يُعتمد عليها في تصميم أدوات التوسيم[١].

١-٢-٣ المعلومات اللغوية:

تنفذ عملية التوسيم (إضافة المعلومات اللغوية) في مستويات لغوية مختلفة بالاستناد إلى طرائق متنوعة. ففي المستوى الصوتي مثلاً توسم المدونة وفقا للحدود المقطعية ويسمى توسيمها صوتياً أو وفقاً للخصائص التطريزية. وفي المستوى الصرفي توسم المدونة بتحديد السوابق واللواحق والجذوع ويسمى توسيماً صرفياً. وفي المستوى النحوي يكون التوسيم بتحديد أقسام الكلام ويسمى توسيماً نحوياً. أو بتحديد المادة المعجمية ويسمى توسيماً معجمياً. أو بتحديد الحقول الدلالية ويسمى توسيماً دلالياً. وفي المستوى التركيبي توسم المدونة بوظائف الكلمات النحوية أو بالطريقة الشجرية أو بالتقويس، وفي التوسيم على مستوى تحليل الخطاب توسم المدونة بالعلاقات الإحالية ويسمى التوسيم تحليل الخطاب، وتوسم بإضافة أفعال الكلام ويسمى التوسيم التداولي، أو بالخصائص الأسلوبية بعرض الأفكار واللغة ويسمى التوسيم الأسلوبي[٢].

وأكثر أنواع التوسيم انتشاراً هو التوسيم النحوي، وأكثرها سرعة في التطور التوسيم التركيبي الإعرابي، وأقلها تطوراً التوسيم التداولي والتوسيم على مستوى تحليل الخطاب[٣]. وفيما يلي عرض مختصر لبعض أنواع التوسيم التي نفذت على بعض المدونات:

---

(١) Corpus Annotation, 2013. p. 14

(٢) Corpus –Based Language Studies, 2006, P.33

(٣) Ibid., p.37

١-٢-٣-١ التوسيم التطريزي Prosodic Tagging:

وهو تمثيل للمنقول من المنطوق إلى المكتوب بالإشارة إلى الطريقة التي نطق بها الكلام، وذلك بتحديد مواضع النبر والتنعيم والوقف والإيقاع[١]. وقد اُختلف فيما إذا كان جزءًا من البيانات الأولية للمدونة أو نوعاً من أنواع التوسيم اللغوي، والحقيقة أنه ومن جهة أخرى يتشابه هذا النوع من التوسيم مع مستويات التوسيم الأخرى حيث إنه ينفذ أساساً من اللغويين، فهو تفسير لغوي صوتي[٢]. وإذا كانت المدونات المنطوقة تحفظ في صورة تسجيل إلكتروني فهي تمثل في صورة مكتوبة لتوسيمها بمثل هذا النوع وغيره من أنواع التوسيم[٣]. وتنفذ عملية التوسيم التطريزي على مقاطع الكلمة بدلاً من تنفيذها على الكلمة كاملة.

ورغم وجود نمو في المدونات المنطوقة إلا أن الموسم منها توسيماً تطريزياً نادر، وذلك لأنه عمل يستهلك وقتاً طويلاً، ويتطلب متخصصين في علم الأصوات على قدر عال من الخبرة والتدريب[٤]. ويستفاد من هذا النوع على قدر محدود مقارنة بالمدونات الأخرى الموسمة بأنواع أخرى. وتشير المراجع الإنجليزية إلى وجود مدونتين فقط وسمتا بهذا النوع وهما: LLC وSEC[٥]. أما عربيا، فقد وسمت به **المدونة القرآنية** Quranic Corpus[٦]. ومن الأعمال التي استفادت من المدونة القرآنية الموسمة تطريزيا، عمل **كلير بريرلي** Claire Brierle، **ومجدي صوالحة** Majdi Sawalha، **وإيرك أتويل** Eric Atwel الذي اعتمد عليها، وطور الباحثون من خلاله برمجية تجمع كل مواضع القلقلة في القرآن الكريم. ومن

---

(١) Corpus Annotation, 2013, P.85

(٢) Ibid.

(٣) Corpus Linguistics (An Introduction), 2011, p.66

(٤) Corpus –Based Language Studies, 2006, p.63; Corpus Annotation, 2013, P.90

(٥) Ibid.

(٦) Sawalha, M. Brierley C., Atwell E. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning. Proceedings of LRE-Rel'2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, 2014

ثم تمكنوا من توليد دليل مربوط **بنظام الألفبائية الصوتية العالمية** International Phonetic Alphabet (IPA) لنطق الكلمات المحتوية على القلقة الكبرى والصغرى والوسطى يستفيد منه الناطق بالعربية وغير الناطق بها[١].

١-٢-٣-٢ التوسيم الصوتي:

إن المدونات المنطوقة بالإضافة إلى كونها منقولة هجائياً بالكتابة، قد تنقل صوتياً بالتوسيم الصوتي. وبرغم محدودية هذا النوع من التوسيم إلا أنه مفيد جداً للباحثين الذين ليس لديهم خبرة في التحليل المعملي للكلام المسجل، حيث ينقل الكلام كما ينطق صوتيا. ومن المدونات الموسمة صوتياً مدونة MARSEC وهي المدونة الإنجليزية الوحيدة الموسمة بهذا النوع من التوسيم، وتعد نسخة من مدونة SEC. ونشر في مجلة العرب العالمية لتقنية المعلومات مقترح للتوسيم الصوتي للعربية قدمه خالد نهار وآخرون[٢]، وفيه أُستعملت حروف لوحة المفاتيح كلها للتوسيم، بخلاف الألفبائية الصوتية العالمية IPA التي تستعمل رموزا مثل ə، ʌ، ... إلخ. فكلمة صيف مثلا نقلت صوتيا هكذا / ssayf/ ، ولوْم هكذا lawm/ / إذ ينطق صوت الياء في الكلمة الأولى إذا سبق بحرف ساكن مفتوح هكذا / ay/ ، وينطق صوت الواو في الكلمة الثانية إذا سبق بحرف ساكن مفتوح هكذا / aw/ . وقد وسم بهذا المقترح مدونة تتألف من ٤٠٠٠ ملف ممثلة في ٥ ساعات مسجلة باللغة العربية المعاصرة لأخبار متلفزة.

١-٢-٣-٣ التوسيم بأقسام الكلام:

يعرف هذا النوع من التوسيم بأسماء مختلفة فيقال التوسيم النحوي أو التوسيم

---

(١) Sawalha, M., Brierley C., Atwell E. Tools for Arabic Natural Language Processing: a case study in qalqalah prosody. In 9th International Conference on Language Resources and Evaluation, Reykjavik; Iceland, 2014, pp. 283-287

(٢) Nahar, Kh., Al-Muhtaseb, H., Al-Khatib, W., Elshafei, M., Alghamdi, M. Arabic Phonemes Transcription using Data Driven Approach. The International Arab Journal of Information Technology, Vol. 12, No. 3, Jordan: Amman, 2015

الصرف تركيبي، وهو أكثر أنواع التوسيم شيوعاً، وسآتي للحديث عنه مفصلاً في الأجزاء التالية من الكتاب حيث الموضوع الهدف.

١-٢-٣-٤ التوسيم المعجمي (Lemmatization):

وهو أحد أنواع التوسيم الذي يقلل من التنوع التصريفي للكلمات بردها إلى موادها lemmas أو وحداتها lexemes المعجمية، كما تظهر في المعجم[١]. فمثلًا المادة العجمية لـ (بيت ـ مبيت ـ بيوت ـ مبات ـ نبيت ـ بات ـ أبيت ـ تبيت ـ يبيت) هي بيت في كلٍّ. ويعد هذا النوع مهما في دراسات المفردات والمعجم، مثل دراسة الانتشار الصرفي لمادة من مواد المعجم. وهو مهم كذلك في تطوير المعاجم والقواميس. وتزيد أهميته في اللغات الغنية تصريفياً؛ لأنه يتمثل في تصريف الأفعال (كل الأزمنة)، وتصريف الأسماء إفرادا وجمعا تذكيرا أو تأنيثا، والأدوات إفرادا وجمعا تذكيرا وتأنيثا. ولأننا في الإنجليزية مثلا نجده يتمثل فقط في تصريف الأفعال لأحد الأزمنة، وفي الأسماء المجموعة، يعد استعماله زيادة وفضولاً، ومن ثم من النادر أن تجد مدونة موسمة في الإنجليزية بهذا النوع، ومن ذلك مدونة SUSANNE[٢]. وفضلا عن تطبيقه على الإنجليزية، طبق على الفرنسية والإسبانية وقدم نتائج عالية[٣]. وطبق أيضا من قبل الربيعة على مدونة الذخيرة النصية الفصحى لجامعة الملك سعود بأداة MADA+TOKAN وحقق دقة بلغت ٨٧٪[٤].

١-٢-٣-٥ التوسيم التركيبي:

هو عملية إضافة المعلومات التركيبية التي تشير إلى البناء التركيبي للنص. وتتم

---

(١) Corpus – Based language studies, 2006, pp. 35-36

(٢) Corpus Linguistics (An Introduction), 2011, p.53

(٣) Corpus – Based Language studies, 2006, pp. 35-36

(٤) Alrabiah, M.Building A Distributional Semantic Model for Traditional Arabic and Investigating its Novel Applications to The Holy Qur'an. (Unpublished doctoral thesis), King Saud University, Riyadh: KSA, 2015

عملية التوسيم للمدونة بعد توسيم المدونة نحوياً POS باستعمال **الأقواس** bracketing، أو باستعمال **شجيرات العلاقات التبعية** treebanks، أو باستعمال المسميات الوظيفية للعناصر مثل فاعل، مفعول ...الخ[١].

وعملية التوسيم التركيبي هي أكثر أنواع التوسيم شيوعاً بعد التوسيم النحوي. وتعد البنوك الشجرية المحللة تركيبيا أكثر فائدة من المدونات الموسمة نحوياً للبحث اللغوي؛ لأنها تزودنا بنوعين من المعلومات اللغوية (النحوية والتركيبية). كما أنه قد يستفاد منها في مجال التعليم حيث تقوم بدور المعلم ويتعلم الطلاب بواسطتها التحليل النحويّ[٢].

وتُنفذ عملية التحليل التركيبي آلياً، ولكن بدقة أقل من عملية التوسيم النحوي، فهي بحاجة إلى التصحيح اليدوي. وهو ما حدث مع البنك الشجري الإنجليزي لبنسلفانيا[٣]. وثمة بنوك شجرية أخرى وسمت توسيماً يدوياً بدون تدخل آلي رغم توفر **المحللات التركيبية** parsers، كالبنك الشجري لليدز ولانكاستر، إذ يمكن استعمال العمل اليدوي أو شبه اليدوي لتدريب المحللات التركيبية الآلية[٤].

ويمكن أن يكون التحليل التركيبي كاملاً، كما يمكن أن يكون **مجملاً أو سطحياً** skeleton. والتحليل التركيبي الكامل يزودنا بتحليل تركيبي مفصل بقدر الإمكان، أما المجمل فهو أقل دقة ويستعمل في الأبنية البسيطة[٥]. ويعد بنك بنسلفانيا الشجري مثالاً للتحليل التركيبي السطحي، فكل الجمل الاسمية وسمت فيه بالرمز(N)، فيما يُميز الإعراب الكامل بين أنواع الجمل الاسمية بخاصتي الإفراد والجمع[٦].

---

(١) Corpus Annotation, 2015, P.34

(٢) Corpus – Based Language studies, 2006, p.36

(٣) Ibid.

(٤) Ibid., p.37

(٥) Ibid.

(٦) حبش، نزار. مقدمة في المعالجة الطبيعية للغة العربية، ترجمة: هند الخليفة، جامعة الملك سعود، الرياض: السعودية، ط١، ٢٠١٤، ص. ١٨٢

وفي العربية يوجد عدد من البنوك الشجرية، **كبنك بنسلفانيا** المعروف **بالبنك الشجري العربي** Arabic Treebank (ATB)، **وبنك براق للتحليل التركيبي العلائقي** Prague Arabic Dependency Treebank (PADT)، **وبنـك كـولومبيا الشجري العربي** Columbia Arabic Treebank (CATiB)[١]. وفي البنك العربي الشجري مثلاً توسم الجملة (من جهة أخرى كشفت مصادر مصرية مطلعة حقيقة الأمر) توسيماً تركيبيا باستعمال الشجيرات كما في شكل (١-١).

حيث تشير S إلى أن (من جهة أخرى كشفت مصادر مصرية مطلعة حقيقة الأمر) بنية جملة يغلق قوسها في نهاية الجملة بعد كلمة (الأمر)، وتشير PP إلى أن كلمة (من) حرف جر متعلق بالعبارة الاسمية (جهة أخرى) الموسمة بـ NP. وتشير VP إلى أن (كشفت) في عبارة فعلية متعلقة بعبارة اسمية هي (مصادر مصرية مطلعة) موسومة بـ NP وهي واقعة موقع الفاعل للفعل (كشفت) ولذلك أضيف لوسمها الرمز SBJ. ثم تلتها العبارة الاسمية (حقيقة الأمر) الموسومة بـ NP-OBJ لوقوعها مفعولا به.

شكل (١-١) التوسيم التركيبي باستعمال الشجيرات في البنك العربي الشجري

```
(S  (PP min من
       (NP jih+ap+K جِهَةٍ
           >uxoraY أُخرَى ))
    (VP ka$af+at  كَشَفَت
       (NP-SBJ maSAdir+u  مَصادِرُ
               miSoriy~+ap+N  مِصْرِيَّةٌ
               muT~aliE+ap+N مُطَّلِعَةٌ ))
       (NP-OBJ Haqiyqata حَقِيقَةُ
               (NP Al->amri الأمرِ )))
```

مِن جِهَةٍ أُخرَى كَشَفَت مَصادِرُ مِصْرِيَّةٌ مُطَّلِعَةٌ حَقِيقَةَ الأمرِ

from another side, well-informed Egyptian sources revealed the truth of the matter

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٩١

١-٢-٣-٦ التوسيم الدلالي:

ويكون التوسيم الدلالي بإسناد الخصائص الدلالية والحقول المعجمية للكلمات في نصوص المدونة[١]. ويوجد نوعان منه: الأول يحدد العلاقات الدلالية بين مكونات الجملة، مثل ما هو موجود في بنك براغ العربي حيث يحدد الفاعل النحوي للفعل والفاعل الدلالي للصفة بنفس الوسم، وهو: -SBJ [٢]، فكلمة (رجل) في الجملتين: (حضر رجل ـ رجل زائر) توسمان بـ -SBJ ويعرف بالإعراب الدلالي. أما النوع الثاني فيحدد الخصائص الدلالية في النص ويصنفها حسب المعاني الدالة عليها، ويفيد هذا النوع في تحليل المحتوى. ومنه **شبكة الكلمات العربية** Arabic WordNet التي تجمع الكلمات في مجموعات من المترادفات تمثل معنى لكلمة فريدة وتشكل المفهوم، إذ تجد فيها مثلا كلمة (زار) مقرونة بـ (تجول ـ جاب ـ دار ـ طاف)[٣] ممثلة لمفهوم الزيارة.

ويشكل التوسيم الدلالي مهمة أكثر تحديا من التوسيم النحوي والتوسيم التركيبي؛ لأنه **يعتمد على المعرفة** الإدراكية knowledeg-base، حيث يتطلب موارد معجمية وأنطولوجية كالمعاجم والمكانز[٤]. وبرغم التحديات التي يواجهها في فك غموض المعاني آلياً، إلا أنه يحقق نجاحات واضحة. ومن أمثلته USAS نظام التحليل الدلالي لمدونة UCREL الذي تتألف مجموعة وسومه من ٢١ قسما أساسيا و٢٣٢ قسماً فرعياً، ويبدأ أولاً بإسناد أقسام الكلام لكل وحدة معجمية باستعمال **نظام التوسيم الآلي بترجيح المكونات** Constituent Likelihood Automatic Word-tagging System (CLAWS)، ثم إدخال مخرجاته في نظام توسيمه الدلالي. وقد بلغت دقته ٩٢٪ بعد تنفيذه على نصوص معاصرة من الإنجليزية[٥].

---

(١) Corpus- Based Language studies, 2006, pp.37-38

(٢) Corpus Linguistics (An Introduction), 2011, p.61

(٣) مقدمة في المعالجة الطبيعية اللغة العربية، ٢٠١٤، ص. ٢٠٥-٢٠٦

(٤) Corpus Linguistics (An Introduction), 2011, pp. 57-58

(٥) Corpus- Based Language studies, 2006, pp.37-38

وكان للحاج محاولة خاصة بالعربية ضمن مشروع من ١٥ مشاركا، لإنشاء معجم دلالي متعدد اللغات، وسمت فيه ٢٥٠ كلمة توسيما دلاليا ويدويا. وقد استعمل المشاركون عملية تصفية حديثة يُطلب فيها من المستخدم تحديد المترادفات في اللغة الهدف، وإزالة المعاني الخاطئة باستعمال مخطط عام للتمثيل الدلالي متعدد اللغات[١].

١-٢-٣-٧ التوسيم الإحالي:

وهو أحد أنواع التوسيم الذي طبق على عدد من المدونات على مستوى تحليل الخطاب. والهدف الأساسي منه تحديد الإحالات في المدونة، كالعلاقات الإحالية بين الضمائر والعبارات الاسمية[٢]. ويمكّن هذا النوع من تتبع كيفية التماسك الشكلي لعناصر النص من خلال الضمائر والتكرار والاستبدال والحذف، وغير ذلك.

ويوجد العديد من المدونات الإنجليزية الموسمة توسيماً إحالياً، وأكبرها مدونة مشروع جامعة Stendhal، ومركز أبحاث زيروكس الأوروبي[٣]. ولا يوجد للإنجليزية مقترح معتمد للتوسيم الإحالي، حيث مازالت مقترحات كثيرة قيد التطوير والتحسين ومنها مقترح ماكنري الذي عدل وطبق على الهندية[٤]. أما عربيا، توجد بعض المحاولات لتوسيم المدونات على هذا المستوى، ومنها محاولة سهى حمدي، وآخرين[٥] التي قدموا فيها أيضا أداة للتوسيم الإحالي تقف على الضمائر وتسمح للموسم البشري أن يختار ما

---

(١) El-Haj, M., Rayson, P., Piao, S. and Wattam, S. Creating and validating multilingual semantic representations for six languages: expert versus non-expert crowds. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Valencia: Spain, 2017, p. 61-71

(٢) Corpus- Based Language studies, 2006, pp.37-38

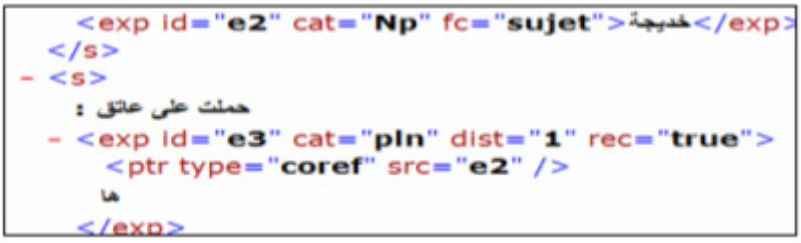
(٣) Ibid., pp.38-39

(٤) Ibid.

(٥) Hammami S., Belguith L., Ben Hamadou A. Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links. The International Arab Journal of Information Technology, Vol. 6, No. 5, Jordan: Amman, 2009

تحيل إليه هذه الضمائر. فالعبارة: ... خديجة. وحملت على عاتقها ...، وسمت إحاليا في مدونتهم كما في شكل (١-٢).

شكل (١-٢) مثال من التوسيم بأداة حمدي وآخرين للتوسيم الإحالي



حيث تشير الخاصية cat إلى نوع الضمير أو القسم الكلامي المحال إليه الضمير. وتشير الخاصية dist إلى المسافة الفاصلة بين المحال والمحال إليه، وهو هنا =١. ويعني أن المحال إليه في الجملة السابقة (خديجة).

١-٢-٣-٨ التوسيم التداولي:

وهو نوع آخر من أنواع التوسيم الذي ينفذ على مستوى تحليل الخطاب[١]. ويركز التوسيم التداولي الحالي على الأفعال الكلامية ومجالات محددة من المحادثات، كالمكالمات الهاتفية، وحوارات الأطباء مع المرضى[٢]. وحتى الآن لم ينجز هذا النوع من التوسيم بطريقة آلية تماماً، حيث لابد فيه من التدخل البشري[٣].

وفي دراسة حديثة تتعلق بالعربية قدم منهج معتمد على القواعد يقف على أنواع متعددة من أفعال الكلام. طور من خلالها نظام خبير لتصنيف ستين نوعا من أفعال الكلام (استفهام، تعجب، نفي ... إلخ)، وتم اختباره على مدونة من ١٥٠٠ جملة وحقق دقة بلغت ٩٢ ,٩٨٪. واُستعمل النظام لتوسيم مدونة (على المستوى

---

(١) Corpus – Based language studies, 2006, p. 40

(٢) Corpus Annotation, 2013, p. 91

(٣) Corpus- Based Language studies, 2006, p. 40

الجملي) بإسناد وسم لكل جملة حسب نوعها الكلامي[١].

١-٢-٣-٩ التوسيم الأسلوبي:

وفيما يركز التوسيم التداولي على أفعال الكلام في المحادثات والحوارات، فإن التوسيم الأسلوبي مرتبط على وجه الخصوص بالخصائص الأسلوبية في الكتابات الأدبية[٢]. ومن ذلك تمثيل أفكار وكلام الناس المعروف بتمثيل **الفكر والكلام** thought and speech، وهو ليس خاصاً بالمشتغلين بعلم الأسلوب، بل حاز اهتمام المتخصصين في علم اللغة التطبيقي وفي علم النفس وفلاسفة اللغة على مدى طويل[٣].

والمدونة الإنجليزية الوحيدة التي وسمت بهذا النوع هي مدونة لانكاستر لتمثيل الكتابة والفكر والكلام، وهي من جزء منطوق وآخر مكتوب، وامتد العمل فيها بين عامي ١٩٩٤ و٢٠٠٣م[٤]. وجاءت الأقسام الرئيسية لمجموعة الوسوم فيها سبعة هي: (صنف مباشر حر، صنف غير مباشر، صنف غير مباشر حر، صنف تمثيل أفعال الكلام/ الفكر/ الكتابة، صنف تمثيل الحالة الداخلية / الصوت، الصنف الإخباري)، مع اختلاف الأقسام الفرعية وفقاً للاختلاف بين اللغة المنطوقة واللغة المكتوبة[٥]. ولأن تركيب النصوص سطحي، فهو ليس مؤشرا كافياً للخصائص الأسلوبية، ومن ثم ليس غريباً أن نقول إن هذه العملية من الصعب تنفيذها آلياً، وإن هذا العمل محلل يدوياً[٦]. وحسب علمي لا يوجد مدونة عربية وسمت بهذا النوع من التوسيم.

---

(١) Sherkawi L., Ghneim N., Al Dakkak O. Arabic Speech Act Recognition Using Boot-strapped Rule Based System. International Journal on Computer and Communications Networks, Computational Intelligence and Data Analytics, Vol. 1, No. 1, Rome: Italy, 2017, p. 1-7

(٢) Corpus- Based Language studies, 2006, p. 41

(٣) Ibid.

(٤) Ibid.

(٥) Ibid.

(٦) Corpus Annotation, 2013, p.99

١-٢-٣-١٠ توسيم الأخطاء (Error Tagging):

وهو نوع لا يعتمد على مستويات لغوية محددة، ولكنه أقرب ما يكون أحد تطبيقاتها[١]. ويرتبط تحديداً بمدونات المتعلمين، ويقصد به إسناد رموز تشير إلى أنواع الأخطاء التي تحدث في مدونة ما[٢]. وهذا يساعد على معرفة أكثر الأخطاء شيوعاً لمتعلمي اللغة الثانية من ناطقين بلغات مختلفة وبخلفيات متعددة ومهارات متفاوتة. ويمكن أن يكشف لنا العلاقة بين الخطأ نفسه، وجنس المتحدثين أو الناطقين بلغة أم محددة. كما يساعد على اكتشاف الخصائص السلوكية التي تحدث في التعلم، كالإفراط في استعمال خصائص لغوية محددة أو التقليل منها.

وتختلف مقترحات التوسيم في هذا النوع حسب أنواع الأخطاء التي تهدف إلى تحديدها، ومن أمثلة هذه المقترحات مقترح كامبرج الذي وسمت به مدونة كامبرج للمتعلمين، وأنواع الأخطاء فيه كانت: «شيء مفقود (M) ـ كلمة أو عبارة غير ضرورية (U) ـ كلمة خطأ اشتقاقياً (D) ـ استعمال خاطئ لكلمة (F) ـ كلمة أو عبارة تحتاج إلى تبديل (R)»[٣].

إن التوسيم بالأخطاء مهمة معملية تستهلك كثيرا من الوقت، كما أنه مع الصعب تطويرها، بسبب قلة المعلومات المتعلقة بأنماط الأخطاء وترددها عند مجموعة من المتعلمين لتعدد لغات المتعلمين الأم وتعدد أجناسهم وأعمارهم ومستوياتهم وغيرها من العوامل الأخرى[٤]. وبرغم ذلك، ثمة العديد من المحاولات لأتمته عملية التوسيم بالأخطاء حيث طورت العديد من الأدوات لاكتشاف أخطاء محددة، كأخطاء التعريف والتنكير[٥].

---

(١) Corpus Annotation, 2013. p. 15

(٢) Corpus - Based Language studies, 2006, p.42

(٣) Ibid.

(٤) Ibid.

(٥) Ibid.

ومن ذلك مجموعة وسوم الفيفي وأتويل المقترحة لتطوير نظام آلي يساعد في توسيم المدونات بأخطاء المتعلمين[١]. وتضمنت ٢٩ نوعا، وقسمت على خمسة مجالات هي الإملاء والصرف والنحو والدلالة والترقيم، فالوسم <OH> مثلا لأخطاء الهمزة وأحد أنواع الأخطاء في مجال الإملاء[٢].

١-٢-٣-١١ التوسيم الموجه لحل المشكلة البحثية (problem –oriented):

ويقصد به التوسيم الذي يوسم الظاهرة اللغوية المتعلقة بالسؤال البحثي فقط. وهذا النوع من التوسيم ضروري ومفيد للأسئلة البحثية التي لا يمكن أن تجيب عنها أنواع التوسيم المتوفرة التي يستفاد منها في مجالات بحثية واسعة، وتنفذ بمقترحات توسيمية متفق عليها بصورة عامة[٣]. وحتى وقتنا الحاضر، تقوم كل أنواع التوسيم بمثل هذا العمل، ماعدا هذا النوع حيث يختلف عنها في أمرين أساسيين هما[٤]:

١. التوسيم الموجه لحل المشكلات البحثية ليس كاملا؛ لأن الظاهرة المتعلقة بسؤال البحث مباشرة هي ما يوسم فقط، وليس المحتوى الكامل للمدونة.

٢. ليس الهدف من مقترح التوسيم الموجه لحل المشكلات البحثية أن يكون معياراً يعتمد عليه، فهدفه الإجابة عن سؤال بحثي محدد. إذ يعتمد على أسئلة فردية بحثية، والمقترحات التوسيمية الناتجة منه خاصة بسؤال البحث، ولا يمكن أن تستعمل لأسئلة بحثية أخرى[٥].

ومن الدراسات التي استعملت هذا النوع، دراسة **هنستون** Hunston التي تسعى

---

(١) Alfaifi, A., Atwell, E. Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. The 8th Saudi Students Conference, London: UK, 2015

(٢) Ibid.

(٣) Corpus- Based language studies, 2006, p.43

(٤) Corpus Linguistics (An Introduction), 2011, P. 69

(٥) Corpus- Based language studies, 2006, p.43

للإجابة عن كيفية كلام الناس عن التطابق والاختلاف[١]. ولا يذكر أن مدونة عربية اعتمدت هذا النوع من التوسيم في أي مستوى لغوي.

وبعد هذه الشروح الموجزة لأنواع التوسيم، يظهر أن مجال التوسيم مجال واسع ومفتوح، ولا يمكن أن يقف عند أنواع محددة. وهذا ما يجعلنا نؤكد أنه رغم عرضنا لأنواع التوسيم في كل المستويات اللغوية تقريباً، إلا أننا سنجد أنواعاً أخرى تظهر في المستقبل القريب.

١-٣ التوسيم النحوي، وأهميته:

يسمى التوسيم النحوي التوسيم بأقسام الكلام، أو التوسيم الصرف ـ نحوي، ويقصد به إسناد رمز لكل وحدة معجمية في النص يشير إلى نوعها (فعل، اسم... إلخ)[٢]. وهو التوسيم الأوسع انتشاراً واستعمالًا حتى اليوم، ومعلوماته هي معلومات أساسية تزيد من إمكانية تحديد البيانات المرغوب في استرجاعها من المدونة. والتوسيم النحوي أساس لمستويات أخرى من التوسيم، مثل التوسيم بالحقول الدلالية، أو التحليل التركيبي[٣]. كما أنه خطوة أولى لفك غموض **المتجانسات هجائيا** homographs، إذ لا يمكن التمييز بين معنى (قاتل) التي قد تعني الأمر بالقتل، أو من قام بالقتل دون وسمها بهذه المعلومات النحوية. وإسناد (فعل) للمعنى الأول، و(صفة فاعل) للمعنى الثاني تفصل بين معنى الفعل ومعنى الصفة. ولذلك، تفيد المدونة الموسم نحوياً (بأقسام الكلام) في مجالات واسعة التطبيقات، ابتداء من فك الغموض في المتجانسات وحتى استعمالات أكثر تعقيداً كحوسبة نظام يحدد أقسام الكلام آلياً.

---

(١) Corpus- Based language studies, 2006, p.43

(٢) Corpus- Based Language studies, 2006, p.34; Corpus Linguistics (An Introduction), 2011, p.46

(٣) Ibid.

١-٣-١ بدايات التوسيم النحوي:

إن أول مشروع لتوسيم المدونات المحوسبة كان التوسيم النحوي لمدونة **براون** Brown تحت إشراف مؤسسي لغويات المدونات المحوسبة Francis وKucera. وقد قام بهذا المشروع طالبا ماجستير في جامعة براون هما **جرينس** Greence **وروبين** Rubin في عام ١٩٧١م، باستعمال مجموعة وسوم تتألف من ٧٧ وسما لأقسام الكلام، وحدث هذا سريعاً بعد إكمال جمع مدونة براون نفسها[١]. ولا تتألف هذه الوسوم النحوية من أقسام الكلام الأساسية فقط (اسم، فعل، ... الخ)، بل تضم أيضا تحديداً للأقسام الفرعية مثل: الصفات المقارنة والعلائقية، والمفرد والجمع من الأسماء ... وهكذا[٢].

وخلاصة هذه التجربة الرائدة التي نفذت من متخصصين لغويين هي أن ٧٧٪ من الكلمات الموسمة كانت صحيحة التوسيم دون إبهام، وذلك باستعمال برنامجهما TAGGIT الذي اتخذ الطريقة **المعتمدة على القواعد** Rule-Based أساساً له، بعد أن وجد الباحثان ضرورة استعمال التوسيم الآلي لأن إكمال المهمة سيكون مملاً ويستغرق وقتا[٣]. واعتمادهما على الطريقة المبنية على القواعد يعني أن عملية التوسيم وفك الغموض في هذا الموسِّم اعتمدت على قواعد شكلية سياقية صممت بالاعتماد على ملاحظة البيانات التي تحدد بعض المعلومات عن الوسم المحتمل، إذ القاعدة هي التي تحدد إمكانية وسم واستحالة آخر[٤].

وظلت جامعة براون في السنوات التالية تضطلع بمهمة توسيم ٢٣٠,٠٠٠ كلمة متبقية غامضة، وذلك بالتحرير اليدوي للمدونة[٥]. وهكذا، أدت تجربة **جرينس**

---

(١) Corpus Linguistics (An Introduction), 2011, p.50

(٢) Corpus Annotation, 2013, p.8

(٣) Ibid.

(٤) Ibid., p.103

(٥) Ibid., p.8

Greence **وروبين** Rubin إلى إنجاز مفيد جداً، وهو مدونة براون الموسمة توسيماً نحوياً والمستعملة من مئات الباحثين في العالم.

ثم تلا ذلك مشروع توسيم **مدونة لانكستر-أوسلو/ بيرقن** /Lancaster-Oslo (Bergen Corpus LOB) النظير البريطاني لمدونة براون، وكان بين عامي ١٩٧٩م - ١٩٨٢م. وفي هذا الوقت كان العمل على برنامج التوسيم **بالطريقة الاحتمالية** Probabilistic [١]. ولحسن الحظ، كانت مدونة براون الموسمة مدخلاً كافياً اعتمدت عليه مدونة LOB في توسيمها الآلي ببرنامج التوسيم CLAWSI [٢]. وقد ازداد معدل نجاح التوسيم النحوي الآلي فيها من٧٧٪ إلى ٩٦,٧٪[٣]. ولأن منهج CLAWSI هو الاعتماد على الطرائق الاحتمالية، يدرج الوسم الأكثر احتمالاً في كل موضع، وإذا فشل في ذلك فإن هذا الموضع هو موضع خطأ[٤]. وهكذا، فشل في ٣,٣٪ من المواضع التي ينبغي وسمها، حيث لم يستطع التنبؤ بالوسم المحتمل، فكان من الضرورة وسمها يدوياً[٥].

وبعد موسم CLAWSI ظهرت العديد من الموسمات النحوية الخاصة باللغة الإنجليزية والمعتمدة على الطرائق الاحتمالية كموسمات Church و DeRose. وظل الخلاف في هذه الفترة على أفضلية أحد المنهجين في التوسيم النحوي (المعتمد على القواعد والاحتمالي)[٦].

إن النماذج الاحتمالية تتطلب مدونة موسمة سابقاً تزودنا بافتراضات أولية للأساسيات التي تدرب عليها برامج التوسيم الاحتمالية، كما في حالة CLAWSI الذي زُوِّد بمعلومات وافرة من مدونة براون. وإن كان كلا من TAGGIT وCLAWSI استعملا

---

(١) Corpus Annotation, 2013. p. 8

(٢) Ibid.

(٣) Ibid.

(٤) Ibid.

(٥) Ibid.

(٦) Ibid., pp.8-9

سياقات محددة جداً لتحديد الوسم الصحيح للكلمة حيث لا يتجاوز كلمتين يساراً ويميناً. كما أن كلا من TAGGIT ووارثه CLAWSI يعملان على اللغة الإنجليزية فقط[١].

ولوقت طويل امتد إلى ما بعد ١٩٨٨م، وُسّم عدد قليل من المدونات في لغات أخرى، كمدونة **لند** Lund السويدية، ومدونة **نيجميجن** Nijmegen الهولندية، والسبب هو عدم وجود مدونات أصلاً[٢]. ومنذ عام ١٩٩٠م ازداد توسيم المدونات في اللغات الأخرى، كالصينية واليابانية والفرنسية والألمانية والأسبانية والهولندية[٣]. وقد ظهر توجه إلى تطوير برنامج لتوسيم المدونات **لا يعتمد على اللغة** independent -language، كما حدث لموسم **زيروكس بارك** Xerox Parc الذي نفذ على الإنجليزية، وأعيد توجيهه ليطبق على الأسبانية[٤].

لقد بدأت مرحلة ازدهار التوسيم النحوي عام ١٩٨٧م تقريباً، حيث طورت العديد من الموسمات النحوية للعديد من اللغات[٥]. أما في العربية، فكانت هناك محاولات من قبل RDI وصخر **وزيروكس** Xerox ولأغراض تجارية. ولكن المحاولات المسجلة في التوسيم النحوي كانت قليلة وفي أوائل الألفية الثانية. وبدأت **بالموسم الهجين** half breed الذي طوره القارح والأنصاري. وهو موسم شبه آلي يستعمل القواعد الصرفية والتقنيات الإحصائية المعتمدة على العربية القديمة في صيغة **نماذج ماركوف الخفية** Hidden Markov Models (HMMs). وقد حقق نتائج وصلت دقتها إلى ٩٠٪[٦]. ثم تفرقت من بعدها المحاولات وكان أبرزها موسم خوجة المعروف باسم **الموسم بأقسام الكلام الآلي**

---

(١) Corpus Annotation, 2013. pp. 8-9

(٢) Ibid.

(٣) Ibid.

(٤) Ibid., pp.151-152

(٥) Ibid.

(٦) El-Kareh, S., Al-Ansary, S. An Arabic Interactive Multi-feature POS Tagger. In Proceedings of the ACIDCA conference, Monastir: Tunisia, 2000, pp. 204-210

Automatic Arabic POS-Tagger (APT)، ويستعمل التقنيات الإحصائية والتقنيات المعتمدة على القواعد معا. ويعد موسم خوجة أول موسم نحوي للعربية استعملت فيه مجموعة وسوم مكونة من ١٣١ وسما مشتقة من مجموعة وسوم BNC. وقد اشتقت خوجة مجموعة وسومها الأولية من قواعد اللغة العربية، وحقق موسمها نتائج وصلت دقتها إلى ٨٦٪[١]. وقد استعمل **فريمان** Freeman منهج تعلم الآلة وطبق موسم بريل لأقسام الكلام على اللغة العربية. واعتمد موسمه على مدونة موسمة بنيت يدويا وتضمنت ٣ آلاف كلمة واستعمل مجموعة وسوم من ١٤٦ وسما[٢]. وطور المعموري **وسيري** Cieri موسما باستعمال المنهج المعتمد على القواعد. واعتمد موسمهما على مخرجات التحشية الآلية من محلل تيم باكولتر الصرفي. وحقق هذا الموسم المطور دقة بلغت ٩٦٪[٣]. وطورت دياب وآخرون موسم **أميرا** AMIRA لأقسام الكلام العربية الذي يستعمل منهج **آلة الدعم الاتجاهي** support vector machine (SVM) وتتألف مجموعة وسومه من ٢٤ وسما نحويا، وحقق نتائج تصل دقتها إلى ٩٦,١٣٪[٤]. وقدم **بانكو** Banko **وموري** Moore موسما معتمدا على نموذج ماركوف الخفي، وقد حقق دقة بلغت ٩٦٪[٥]. وطور حبش

---

(١) APT: Arabic part-of-speech tagger, pp. 20-25

(٢) Freeman, A. Brill›s POS Tagger and a Morphology Parser for Arabic. In: Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France, 2001

(٣) Maamouri M.; Cieri C. Resources for Natural Language Processing at the Linguistic Data Consortium. In Proceedings of the International Symposium on Processing of Arabic, Manouba, Tunisia, 2002, p. 125-146

(٤) Diab, M., Hacioglu, K., Jurafsky, D. Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL, Association for Computational Linguistics, Boston: USA., 2004, pp. 149-152

(٥) Banko, M., R.C. Moore. Part of speech tagging in context. Proceeding of the 20th international conference on Computational Linguistics, Association for Computational Linguistics Morristown, Article No. 556, New Jersey: USA., 2004
DOI: http://portal.acm.org/citation.cfm?id=1220435

ورامبو Ramboo موسم **مدى** MADA الذي يحدد نوع الكلمة بعد تنفيذ ثلاث خطوات هي التحليل وفك الغموض والتوليد، وحقق نتائج تصل دقتها إلى ٩٦٪[١]. وطورت **قوياسا** Guiassa موسما استعمل المنهج الهجين ومنهج التعلم المعتمد على القواعد والمعتمد على الذاكرة، وحقق دقة بلغت ٨٦٪[٢]. وكان قد اعتد ببناء الجملة العربية القليل من الباحثين أمثال الشامسي **وجيسوم** Guessoum حيث طوروا موسما نحويا للنصوص العربية غير المشكلة حاز دقة بلغت ٩٧٪ باستعمال نماذج ماركوف الخفية[٣]. وطور القريني موسما نحويا باستعمال المذهب المعتمد على القواعد يسمى **الموسم الصرف ـ نحوي العربي** (Arabic Morphosyntactic Tagger ATM). وكانت مدخلاته عبارة عن مدونة عربية خام مشكلة جزئيا. ويهدف هذا الموسم إلى إسناد الوسم الصحيح لكل كلمات المدونة بدون استعمال قاموس موسم يدويا أو غير موسم. ويتألف من نوعين من القواعد: قواعد معتمدة على الأنماط وقواعد معجمية وسياقية. وقد حقق هذا النظام دقة بلغت ٩١٪[٤]. وطور محمد **وكوبلر** Kubler منهجين للتوسيم بأقسام الكلام العربية هما: التوسيم لكامل

---

(١) Habash, N., Rambow, O. Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Michigan: USA, 2005, pp. 573– 580

(٢) Tlili-Guiassa, Y. Hybrid method for tagging Arabic text. Journal of Computer Science, Vol. 2, No. 3, NewYork: USA, 2006, p. 245-248

(٣) Al Shamsi, F., Guessoum A. A hidden Markov model-based POS tagger for Arabic. Des Journées internationales d'Analyse statistique des Données Textuelles, Besançon; France, 2006, p. 31-42

(٤) Alqrainy, S., AlSerhan, H. M., Ayesh, A. Pattern-based algorithm for Part-of-Speech tagging Arabic text. Computer Engineering & Systems ICCES International Conference on IEEE, Cairo: Egypt, 2008, pp. 119-124

الكلمة والتوسيم القائم على التقطيع[١]. واستعمل علي **وجاري** Jarray الخوارزمية التكوينية لتطوير موسم نحوي عربي بمجموعة وسوم مقلصة[٢]. وقدم الحاج موسما نحويا آخر يعتد ببناء الجملة العربية ويربط التحليل الصرفي **بنماذج ماركوف الخفية** HMMs، وقد بلغت دقة هذا الموسم ٩٦٪[٣].

١-٣-٢ أهمية التوسيم النحوي:

فيما يلي نظرة عامة على استعمال المدونات الموسمة نحوياً في الدراسات اللغوية التطبيقية بعرض بعض الأعمال التي أُنجزت من خلالها، وبعض الظواهر اللغوية التي تجيب عنها المدونة الموسمة نحوياً. ثم نبذة مختصرة عن معالجة اللغات الطبيعية أو هندسة اللغة المعتمدتين على المدونات الموسمة نحويا.

١-٣-٢-١ استعمال المدونات الموسمة نحويا في الدراسات اللغوية التطبيقية:

إن أكثر الدراسات اللغوية إفادة من المدونات الموسمة نحويا هي الدراسات النحوية والمعجمية. ويرجع ذلك إلى تقديم المدونة الموسمة نحويا بيانات كمية ضخمة للتنوعات اللغوية في اللغة كأكثر أقسام الكلام شيوعا في مجال ما وتوزيع الضمائر في اللغة المكتوبة أو المنطوقة[٤]. وبالإضافة إلى ذلك، يمكن من خلال المدونات الموسمة نحويا اختبار الفرضيات المشتقة من أي نظرية نحوية أو معجمية كالكشف عن المهمل والمستعمل وغيرها.

---

(١) Mohamed, E., & Kübler, S. (2010, May). Arabic Part of Speech Tagging. In LREC.

(٢) Ali, B. B., & Jarray, F. (2013). Genetic approach for Arabic part of speech tagging. arXiv preprint arXiv:1307.3489

(٣) Elhadj, Y. O., Abdelali, A., Bouziane, R., Ammar, A. H. Revisiting Arabic Part of Speech Tagsets. In Computer Systems and Applications (AICCSA), IEEE/ACS 11th International Conference on IEEE, Doha: Qatar, 2014, pp. 793-802

(٤) Corpus-Based language studies, 2006, p.80

وحتى الربع الأخير من القرن العشرين، اعتمدت الدراسات اللغوية على التحليل الكيفي الذي يقدم لنا وصوفا تفصيلية لا يمكننا من خلالها الحكم موضوعيا على أي مسألة بالشيوع أو الندرة أو الشذوذ[١]. ولكن التقدم الذي أحرزته المدونات الموسمة نحوياً بأدواتها، يعني أن التحليل الكمي النحوي سيكون أكثر سهولة وتنفيذاً، وبالتالي سيقدم لنا صوراً ممثلة للاستعمال الحقيقي للنحو ودرجة حدوثه في كل مستوى لغوي داخل اللغة نفسها وفي كل مجال[٢]. وهذا ليس مهما فقط لفهم النحو، بل حتى في دراسة الأنواع المختلفة للتنوعات اللغوية وفي تعليم اللغة[٣].

ومن الدراسات النحوية التي اعتمدت على المدونات الموسمة نحوياً سلسلة دراسات قام بها باحثون في جامعة نوتنقهام، كشفوا فيها عن خصائص لغوية حاضرة في الإنجليزية المنطوقة، ومن جهة أخرى في الإنجليزية المكتوبة، ومنها شيوع بعض الأزمنة الفعلية في اللغة المنطوقة دون المكتوبة[٤]. وفي المقابل ركز **بيبر** Biber على أوجه التشابه بين الشكلين في دراسة استعمل فيها نفس المدونة[٥].

وفي صناعة المعاجم أحدثت المدونات الموسمة نحوياً ثورة فيها. فلا يذكر أن ثمة معاجم صدرت بعد التسعينيات لم تعتمد عليها. فقد استفاد منها صناع قواميس كولينز وأكسفورد ولونقمان وكامبريدج في وجوه مختلفة منها:

١. تـزويد المعجميـن بقـوائم ضخمة من **الهـومونيمات المتـطابقة هجاء** graphic homonyms، **والهومونيمات المتطابقة هجاء ونطقاً** phonotico-graphic homonyms[٦]. ومن ذلك **القاموس المنغولي للكلمات المتجانسات** Mongolian

---

(١) Corpus Linguistics (An Introduction), 2011, p.109

(٢) Ibid.

(٣) Ibid.

(٤) Corpus –Based Language studies, 2006, p.86

(٥) Ibid.

(٦) Ibid., p.81

Homograph Words Information Dictionary المبني على مدونة للمنغولية الحديثة مكونة من ٢٦٠ ألف كلمة موسمة نحويا[١].

٢. تقديم **المتلازمات النحوية** colligation لأي كلمة، وما هو مدى ورود الكلمة مع أحد أقسام الكلام دون غيره، كما في تلازم حروف العطف مع الأسماء والأفعال أو الأدوات[٢]. وفي **قاموس ماكميلان للمتعلمين ذوي المستوى المتقدم** Macmillan English Dictionary for Advanced Learners (MEDAL) والمعتمد على مدونة، أضيفت أكثر المتلازمات النحوية شيوعا للمداخل المعجمية بل وحددت الكلمات أيضا[٣].

٣. توفير قوائم صالحة لأن تكون ملاحق للمعاجم، كأسماء الأعلام وأسماء المدن، والاختصارات. ومن تلك القوائم قوائم الاختصارات الملحقة **بقاموس أكسفورد الإنجليزي** Oxford English Dictionary (OED) [٤].

٤. إعداد المعاجم المتخصصة، كمعاجم الأفعال والحروف والأدوات ومعاجم العبارات الاصطلاحية وغيرها. وقد خصصت أكسفورد مثلا قاموسا للعبارات الاصطلاحية مستندا على مدونتها وهو **قاموس العبارات الاصطلاحية لمتعلمي الإنجليزية** Oxford Idioms Dictionary for learners of English. ولكامبرج قاموس متخصص في الأفعال مبني على مدونتهم وهو **قاموس كامبرج للعبارات الفعلية** Cambridge Phrasal Verbs Dictionary.

---

(١) Hasi, H., Nasun-Urt, N. The Design and Application of Mongolian Homograph Words Information Dictionary. IEEE - Institute of Electrical and Electronics Engineers, Inc., Tianjin: China, 2012

(٢) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٢٤٥؛ Corpus-Based Language studies, 2006, p.148

(٣) Hoey M. What›s in a word?, English Teaching Professional, Issue 27, Hove: UK, 2003

(٤) Oxford University Press. OED-Abbreviations. 2-9-2017: https://public.oed.com/how-to-use-the-oed/abbreviations/

وتتصل معظم الدراسات اللغوية الاجتماعية القائمة على المدونات بالدراسات المعجمية، ومن هذه الدراسات ما قدمه **كيلمر** Kjellmer الذي بحث عن شيوع التذكير والتأنيث في الضمائر وعن الوحدات المعجمية (women-woman-men-man) في مدونة براون الأمريكية ونظيرتها LOB البريطانية، ووجد أن التأنيث أقل تكرارا من التذكير في المدونتين، ولكنه أكثر تكراراً في المدونة البريطانية. وهو ما عزز ملاحظاته حول التحيز الذكوري في الكتابات الإنجليزية الحديثة[١].

وفي مجال التخطيط اللغوي قد تقترح جهات حكومية معينة التدخل في متن اللغة حين يلاحظ شيوع ما في المدونة الموسمة نحويا. وهذا الشيوع قد يكون لكلمة تعد قياسياً خاطئة أو شاذة أو أعجمية. ومثال ذلك ظهور شيوع صيغة معينة في الأسماء المنسوبة رغم شذوذها، ثم تقدم التسهيلات لقبولها. بل إن المدونات الموسمة نحويا في مشروع **المدونة العالمية للإنجليزية** International Corpus of English (ICE) استعملت في هذا المجال لإعادة صياغة نحو الإنجليزية الخاص بناطقين من غير أهلها في أقاليم متعددة كالهند وباكستان والكاميرون والفلبين وسنقافورا، وغيرها[٢].

وتتقارب مع الدراسات اللغوية الاجتماعية، دراسات تحليل الخطاب، حيث تركز على الفكر والثقافة[٣]. وفي دراسة اعتمدت على مدونة BNC الموسمة نحويا كشف الباحث عن مستوى استعمال لفظة بذيئة بأشكال مختلفة في الخطاب المكتوب والخطاب المنطوق، ووجد أن شيوعها في الخطاب المنطوق أكثر من الخطاب المكتوب،

---

(١) Kjellmer, G. 'The lesser man': observations on the role of women in modern English writings. In J. Aarts and W. Meijs (eds) Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora, Rodopi; Amsterdam, 1986, pp. 76-163.

(٢) المجيول، سلطان. التخطيط اللغوي والسياسة: مفاهيم شمولية. حولية كلية الدراسات الإسلامية والعربية بالإسكندرية، المجلد (٢)، العدد (٣١)، ٢٠١٥، صص٤٤١-٤٨٩

(٣) Corpus Based language studies, 2006, p.111

وفسر ذلك بأن اللغة البذيئة تقع في السياق العامي أكثر من السياق الرسمي، وأن مراقبة النصوص المكتوبة هي ما قلل من شيوعها فيها[١].

أما في الدراسات اللغوية النفسية، فإن استعمال المدونات عموماً يركز فيه على مجالين هما: تعلم اللغات ومعالجة اللغة[٢]. ويمكن الإفادة من المدونات الموسمة نحويا في ذلك. حيث يتاح من خلالها دراسة ما أشير إليه سابقا **بالتلازم النحوي** colligation باعتباره ظاهرة لغوية نفسية تعد شكلاً من أشكال **التهيئة المعجمية** lexical priming. فكل كلمة تهيئ عقلياً لحدوث كلمة، أو كلمات أخرى، ومن ثم تكون عملية الفهم أسرع، فتنشأ المتلازمات[٣]. ويمكن بالتالي أن تدرس إمكانية استدعاء بعض الكلمات لنوع معين من أقسام الكلام[٤]. وفي تعلم اللغات تمكن المدونات الموسمة نحوياً من دراسة النمو اللغوي أو التدرج في الاكتساب لفئة معينة من متعلمي اللغة من خلال أنماط محددة من الجمل أو من خلال أقسام الكلام نفسها، كأن يكشف عن أي أقسام الكلام أكثر اكتساباً عند متعلمي اللغة. ومن ذلك أيضا دراسة اعتمدت على مدونة لونقمان للمتعلمين تمكنت من وصف اكتساب متعلمي الإنجليزية من اليابانيين للمورفيمات النحوية وأثبتت الادعاء المطروح في السبعينيات الذي لم يكن مبنيا على المدونات، وكان يرى بأن متعلمي اللغات الثانية يكتسبون المورفيمات النحوية بترتيب زمني مستقل عن مسار النمو اللغوي للتعلم[٥].

ويمكن أن تفيد المدونات الموسمة نحوياً المهتمين بمجال اللغويات التاريخية كثيراً، حتى أنها تكاد تعتمد كليا على المدونات وخصوصاً التاريخية منها، بل إن بعضهم رآها جزءًا من لغويات المدونات[٦]. فمن خلال مدونة **آرتشر** ARCHER التاريخية

---

(١) Corpus -Based Language Studies, 2006, p. 111

(٢) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٣٨٨

(٣) السابق، ص. ٢٧١-٢٧٤

(٤) السابق، نفس الصفحة.

(٥) Corpus –Based Language Studies, 2006, P. 263

(٦) Corpus Linguistics (An Introduction), 2011, p.123

الموسمة نحويا استطاع **بيبر** Biber أن يكشف عن أقسام من الكلام في الإنجليزية اعتراها نقص وبعضها الآخر زاد استعماله خلال ثلاثة قرون ماضية[١]. وتتداخل مع اللغويات التاريخية اللغويات التداولية التاريخية التي زاد الاهتمام بها في الآونة الأخيرة وتقوم كلياً على المدونات[٢]. فمن خلال المدونة الموسمة نحوياً تمكن **تويوتا** Toyota من تتبع أبنية المجهول في الإنجليزية من قديمها إلى حديثها ودراستها من خلال ربط مدونة هلسنكي ومدونة آرتشر ومدونة لندن ـ لند للانجليزية المنطوقة، ومدونة لانكاستر ـ أوسلو/ بيرقن معا. وكان الهدف منها الكشف عن أنماط البناء للمجهول قديما وما اعتراها من تغيير عبر الزمن وحتى وقتنا الحاضر.

وفي الدراسات الأسلوبية يهتم الأسلوبيون بالأعمال الفردية ولذلك نجد أعمالهم التي تنفذ بالاعتماد على المدونات عموماً، إنما تكون في مجال محدد أو لكاتب معين[٣]. ومن الدراسات الأسلوبية التي نفذت على مدونة موسمة نحوياً دراسة تتعلق بإنجليزية السياحة وخصائصها الأسلوبية. وقد حللت هذه الدراسة الخصائص الدلالية لذلك النوع من الإنجليزية الخاص بهذا الغرض بالاعتماد على **مدونة السياحة الإنجليزية** Tourism English Corpus (TEC). واستعملت معها **مدونة فرايبورق** Freiburg-LOB Corpus of British English (FLOB) لإجراء المقارنات. وجاء في النتائج أن للمدونة السياحية خصائص أسلوبية معينة. ورغم أن الأسماء هي الأكثر شيوعا في المدونتين إلا أن المدونة السياحية تزيد في استعمالها للأسماء والصفات وتقل في استعمالها للأفعال والضمائر عن مدونة FLOB. والكشف عن مثل تلك الخصائص يساهم في تخصيص أسلوب لغة السياحة بسمات تمكن من تدريسها للراغبين في تعلم الإنجليزية لأغراض خاصة (السياحة)[٤].

---

(١) Corpus Linguistics (An Introduction), 2011, p. 123

(٢) Corpus –Based Language Studies, 2006, p. 104; Corpus Linguistics (An Introduction), 2011, p.14

(٣) Corpus –Based Language Studies, 2006, P. 114 - 115

(٤) Kang N., Yu Q. Corpus-based Stylistic Analysis of Tourism English. Journal of Language Teaching and Research, Vol. 2, No. 1, Finland, 2011

وفي التسعينيات أصبح استعمال المدونات أمرا مهما في أبحاث وتطبيقات تعليم اللغات، بل وفي تعليم اللغويات نفسها. وظهر ذلك في وصف لغة متعلمي اللغات الأم والثانية والمقارنة بينهما من خلال مدونات المتعلمين[١]. وتساهم مدونات المتعلمين الموسمة نحوياً في وصف أكثر تفصيلاً وعمقاً، بل ويمكن من خلالها إعداد مادة غنية تستعمل في إعداد اختبارات اللغة وتطويرها، وفي المناهج التعليمية للمراحل المختلفة. وقد طور الباحثان **كازوبسكي Kaszubski ووجنوسكا** Wojnowska نظاما **موجها بمدونة** corpus-driven يدعى TestBuilder يمكن من خلاله بناء التدريبات لمتعلمي الإنجليزية القائمة على مدونة خام أو مدونة توسم نحويا بموسم يربط به[٢].

وأختم باستعمال للمدونات الموسمة في الدراسات الجنائية أو القانونية، حيث يمكن للباحث من خلال المدونة الموسمة نحويا أن يحدد مثلا البصمة النحوية لمجموعة نصوص كتبها المتهم، ثم الاستدلال بها قانونياً على صدق ادعاء وشهادة أو عدمها. كأن يكون استعمال المتهم للام التعليل مع الأسماء بهذا التركيب لـ (+أجل+ اسم) بتكرار ذي نسبة عالية، أو بطريقة دائمة في نصوصه، ثم يدعي حقه الفكري لنصوص لم يرد فيها هذا الاستعمال أبداً. ومن الدراسات التايوانية التي طبقت على مدونة موسمه نحويا دراسة **سوزاكوشي** Szakos **ووه** Wang ذات النصوص المنطوقة. وتضمنت المدونة نصوصاً للمحادثات التي جرت في قاعات المحاكم التايوانية بين القضاة والمحكومين في ٣٠ حالة جريمة احتوت على ٧ أنواع من أنواع الجرائم. وركز الباحثان فيها على نماذج شيوع الكلمات وسياقاتها التي من الممكن أن تساعد القضاة في الحصول على الحقيقة والوصول إلى محاكمة عادلة[٣].

---

(١) Corpus Linguistics (An Introduction), 2011, p.121

(٢) Kaszubski, P., Wojnowska A. Corpus-informed exercises for learners of English: The TestBuilder program. In E. Oleksy & B. Lewandowska-Tomaszczyk (Eds.), Research and scholarship in integration processes, Poland–USA–EU, 2003, pp. 337–354

(٣) Corpus –Based Language studies, 2006, P.119

١-٣-٢-٢ المدونات الموسمة نحوياً وهندسة اللغة:

تتعلق هندسة اللغة ببناء أنظمة معالجة اللغات الطبيعية لأداء مهام لغوية على نطاق واسع. ويعد التوسيم أو التحليل النحوي أحد أمثلة استعمال المدونات في مجال هندسة اللغات الطبيعية، حيث إن التقنيات المختلفة التي تقوم بها الأنظمة أو البرامج اللغوية تقدمها المدونات اللغوية، كبرامج الترجمة الآلية، وأنظمة الاسترجاع والتعرف على الكلام والقراءة الآلية وغيرها.

إن بيانات المدونة الموسمة نحوياً تستعمل لتدريب بعض نماذج اللغة التي تكون في الأنظمة. ويحدث التدريب على مدونة خام أو مدونة موسمة يدوياً من قبل اللغويين[١]. ولهذه المدونة الموسمة يدوياً وظيفتان هما[٢]:

١. إتاحة الفرصة للنموذج المطور من قبل برنامج ما أن يكون أكثر دقة. فإذا كان التوسيم يمثل دليلًا قاطعاً على أحد أقسام الكلام التي يريد البرنامج نمذجته، فالمدونة الموسمة نحوياً ستقدم بيانات تدريبية أفضل من المدونة الخام.

٢. يستفاد من التوسيم النحوي كـخيار، لتقييم مثل هذه البرامج. حيث يمكن مقارنة ما نفذ يدوياً بما نفذ آلياً بطريقة سريعة وآلية، وذلك بالسماح للموسم أن يوسم نصا قد وسم سابقاً بطريقة يدوية، بدلاً من الاعتماد البشري، في اختبار وتقدير مخرجات التوسيم النحوي.

وثمة تطبيقات أخرى تستفيد من المدونات الموسمة نحوياً. منها تطبيقات وأنظمة التشكيل الآلي حيث إن تحديد نوع الكلمة يفيد في تشكيلها فكلمة (عيْن) الاسم، تختلف عن (عيّن) الفعل وتمييزها بالوسم يعين في تشكيلها آلياً. فضلا عن المحللات الصرفية والتركيبية الآلية، والمصححات اللغوية، والقارئ الآلي، كلها تستعين بالمدونات الموسمة نحوياً في هندستها.

---

(١) Corpus Linguistics (An Introduction), 2011, P.141

(٢) Ibid.

الفصل الثاني
مكونات بناء نظام التوسيم النحوي الآلي

تمهيد:

يتطلب بناء نظام آلي للتوسيم النحوي المرور بعدد من الخطوات يمكن إيجازها فيما يلي:

١ – تحديد الغرض الذي من أجله سيبنى النظام الآلي للتوسيم النحوي، فحسب الغرض يتحدد نوع الوسوم النحوية وتفاصيلها. إذ إن ثمة أنظمة تبنى لمعالجة أغراض لا تتجاوز الحاجة فيها إلا تعيين الاسم والفعل والحرف، وهناك أنظمة تبنى لحاجات أدق وأعمق تستلزم تحديد خصائص لغوية أخرى كالنوع والعدد أو تحديد نوع الفعل مثلا من حيث زمنه الصرفي (مضارع ـ أمر ـ ماض).

٢ – تصميم وبناء المدونة التي ستستخدم في تدريب نظام التوسيم النحوي واختباره وفقا للأسس المعيارية لبناء المدونات، وحسب مناهجها، وبما يخدم الغرض من بناء نظام التوسيم النحوي.

٣– تحديد معايير المعالجة القبلية لنصوص المدونة التي تعرف بعملية **التفريق** tokenization وفيها يفرق ما بين الكلمات وتفرق الكلمات عن الأرقام والرموز وعلامات الترقيم، فضلا عن المسافات التي تفصل أصلا بين الكلمات في بعض اللغات. والغرض من هذه الخطوة هو تحديد الوحدات التي يجب أن يكون لها وسم ضمن نظام التوسيم النحوي.

٤- تحديد متغيرات تقطيع الكلمة التي تنفذ في عملية تعرف **بالتقطيع** segmentation. وفيها تحدد المورفيمات التي ينبغي فصلها عن الكلمة بعد تفريقها لتتم عملية توسيمها، ومنها مثلا الضمائر المتصلة وحرفا العطف (الفاء والواو) وحرفا الجر (الباء واللام) المتصلات بالكلمة.

٥- تحديد قائمة أو مجموعة الوسوم النحوية التي سوف تطبق على كلمات المدونة بعد تقطيعها، وتخدم الغرض الذي من أجله سوف يبنى نظام التوسيم النحوي. وهي عبارة عن قائمة بالوسوم النحوية والرموز المستعملة التي تضاف داخل النص بعد توسيمه. فالأفعال مثلا قد يكون لها ثلاثة أقسام هي: الماضي والحاضر والأمر، وبالتالي يكون لها ثلاثة وسوم نحوية يرمز لها على التوالي بـ VP – VR – VC حيث تشير V إلى نوع الكلمة الأساسي في الرموز الثلاثة، وتشير P في الرمز الأول إلى الماضي، وR في الرمز الثاني إلى الحاضر، وC في الرمز الثالث إلى الأمر.

٦- تقطيع كلمات المدونة. وتتم في هذه المرحلة عمليتان هما: تفريق الكلمات آليا حسب المعايير المحددة في الخطوة الثالثة أعلاه، ثم عملية التقطيع اليدوي لنصوص المدونة اعتمادا على متغيرات التقطيع التي قررت سابقا في الخطوة الرابعة أعلاه.

٧- توسيم كلمات المدونة يدويا بالوسوم النحوية التي عينت في الخطوة الخامسة أعلاه بعد أن نفذت عليها عملية التقطيع يدويا. ويمكن توسيم الكلمات التي لا تقبل إلا وسما واحد في كل سياقاتها (كأسماء الإشارة والموصولات وغيرها) بصورة آلية لكسب الوقت وتقليل الجهد.

٨- اختيار وتطبيق المنهج الذي سيعتمد عليه بناء نظام التوسيم الآلي. وأول هذه المناهج هو المنهج **المعتمد على القواعد** rule-based. وفي هذا المنهج يتم بناء نظام القواعد الخاص بالتوسيم، ثم يطبق على المدونة قبل توسيمها نحويا. والمنهج الثاني هو

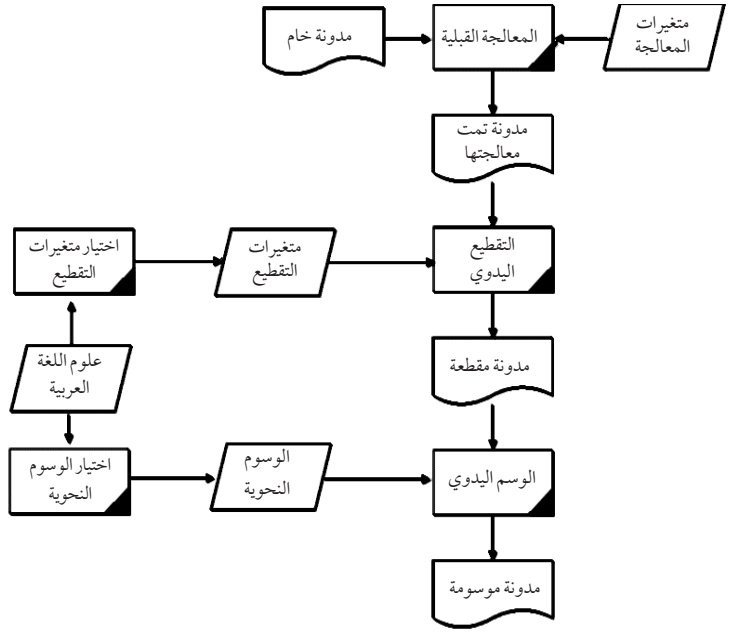
المنهج **المعتمد على تعلم الآلة** machine-learning[١]. وفيه يتم تدريب خوارزمية تعلم الآلة على جزء كبير من المدونة الموسومة يدويا (بين ٧٠٪ إلى ٩٠٪ من حجم المدونة)، ويترك الجزء المتبقي لاختبار نظام التوسيم. وقد يُعتمد على المنهجين معا ويسمى المنهج **بالمنهج الهجين** hybrid.

٩- قياس أداء نظام التوسيم النحوي. وتعتمد طريقة القياس وحجم البيانات المستخدمة فيه على المنهج المتبع. ففي المنهج المبني على القواعد يتم مطابقة نتائج القواعد على المدونة الموسمة يدويا. أما في المنهج المبني على تعلم الآلة، فيستخدم الجزء الذي لم تدرب عليه الخوارزمية من قبل (غير الموسوم يدويا). حيث يتم وسم هذا الجزء آليا ثم تقارن النتائج مع نتائج التوسيم اليدوي لهذا الجزء. وهناك عدة مقاييس تستخدم لقياس أداء النظام الآلي وهي الصحة، والدقة، والاسترجاع، ومقياس ـ ف.

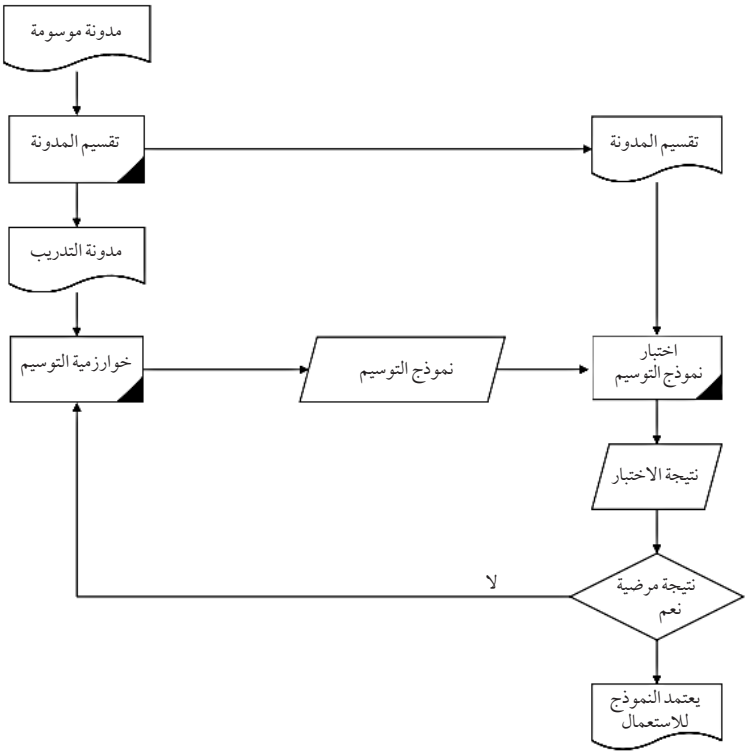
وعادة ما تنفذ الخطوات (١ -٧) من قبل اللغويين، بينما تنفذ الخطوتان (٨ -٩) غالبا بالشراكة بين اللغوي والحاسوبي مع تحمل الحاسوبي للجزء الأكبر منها. ولا شك أن إنجاز مدونة مقطعة وموسومة يدويا (الخطوات ١ -٧)، يعد عملا كبيرا، وخطوات إعدادها تختلف في كل لغة. فمتغيرات التقطيع في الإنجليزية تختلف عن متغيرات التقطيع في العربية وهي أسهل منها بل ويمكن تجاوزها. والوسوم النحوية في اللغة الصينية مثلا تختلف عن العربية، ولكن الجانب الحاسوبي (الخطوتان ٨ -٩) متماثل في الطريقة والمنهج بين جميع اللغات، وقد يكون هو الأسهل من بين جميع خطوات بناء النظام. ويوضح الشكل ٢-١ والشكل ٢-٢ الخطوات المطلوبة لبناء نظام آلي للتوسيم النحوي مع ملاحظة أن ما ينفذ حاسوبيا (٨ -٩) يمكن أن يتم على التقطيع أولا ثم على التوسيم.

---

(١) Mohammad B., Abdulsalam A., Isa B. Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, UK, 2016, p.331



شكل (٢-١) الخطوات اللغوية لبناء نظام آلي للتوسيم النحوي (١-٧)



شكل (٢-٢) الخطوتان الحاسوبيتان (٨-٩) لبناء نظام آلي للتوسيم النحوي (ما ينطبق على التوسيم ينطبق على التقطيع)

وفيما يلي سأوضح ما يتم في هذه الخطوات (١-٧)، وبعض النماذج من الدراسات العربية السابقة في هذا المجال. وحيث إن أحد المعايير الإلزامية التي أشار إليها **ليتش** Leech ضمن مجموعة من معايير التوسيم العامة هو التوثيق، وحيث إنه بدون توثيق كافٍ من الموسمين بعد إنجاز عملية التوسيم ستكون المدونة الموسمة صعبة الاستعمال من قبل الباحثين[١]، ستكون القرارات المتخذة في المبحث الأخير من هذا الفصل بمثابة ملف التوثيق لمقترح التوسيم؛ لضمان قدرة المستعملين مستقبلاً على تطبيق المقترح في اتساق مع مقترحاتهم ومن ثم تطويره.

٢-١ الغرض من التوسيم:

يحدد الغرض من توسيم المدونة نحويا نوع نصوصها، ومتغيرات التقطيع، ومجموعة الوسوم التي ستوسم بها النصوص. فلا يمكن أن نهدف إلى تقديم وسوم عامة للغة ثم نختبرها من خلال مدونة متخصصة في الصحافة أو في الأدب. وتؤثر أيضا الفجوة بين ما هو مرغوب لغوياً، وما هو مطلوب حاسوبياً في تحديد حجم ومكونات مجموعة الوسوم، حيث يرغب اللغوي في الجودة واسترجاع كل المعلومات النحوية المهمة في اللغة، ويرجو الحاسوبي أن يفيد الوسم في عمليات فك الغموض وزيادة دقة التوسيم. فإذا كان الغرض من التوسيم النحوي هو الاسترجاع مثلا، ستكون بعض أقسام الكلام (كالأدوات وتسمى المستبعدات stop words) أقل أهمية من غيرها، وتستبعد قبل المعالجة اللغوية الآلية للبيانات من أجل تحسين البحث، ومن ثم لا حاجة لتصنيفها بدقة، وإنما تصنف تصنيفا أساسيا لتستبعد. ولا تفي بالحاجة الأقسام الثلاثة التقليدية للكلام حين يكون المراد تقديم الأنماط اللغوية الأكثر شيوعا في اللغة إذ من الضروري أن تكون الصفات مثلا محددة بأقسامها للتمكن من تعيين أنماط لغوية دقيقة. وإذا كان الغرض من التوسيم النحوي استخلاص المعاجم اللغوية، فإن الاكتفاء بالأقسام الأساسية دون

---

(١) Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition, 2016, p.241-242

الفرعية، وبالأقسام الفرعية دون الخصائص التصريفية لا يمكن معه استخلاص معاجم للغة. كما أن أنواع التوسيم الأعلى من التوسيم النحوي تختلف حاجتها إلى التفصيل في أقسام الكلام، فحاجة المستوى التركيبي للخصائص التصريفية والتركيبية المضافة لأقسام الكلام في التوسيم التركيبي أكثر من حاجة التوسيم الدلالي. وإن كانت مجموعات الوسوم النحوية المقلصة لا تفيد في المستويات العليا من اللغة، ففي المقابل لا تؤثر مجموعات الوسوم النحوية الموسعة في المستويات الدنيا من التحليل النحوي، حيث يمكن التحكم فيها بتقليصها وتوسيعها متى ما دعت الحاجة.

٢-٢ تصميم وبناء المدونة:

لقد برز اتجاهان واسعان في بناء المدونات هما[١]:

١. منهج المدونة **الراصدة** Monitor، حيث تتوسع المدونة باستمرار لتضم نصوصاً أكثر وأكثر على مر الزمن. وأفضل مثال لذلك **بنك اللغة الإنجليزية** Bank of English (BoE) الذي بدئ العمل به عام ١٩٨٠م، وما زال يزيد توسعا حتى الآن. ويضم ٤٥٠ مليون كلمة في الإنجليزية العامة، و٥٦ مليون كلمة في البيداغوجيا[٢]، والمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية KACSTAC.

٢. منهج المدونة **العينة** Sample، الذي تمثل فيه المدونة اللغة أو مستوى منها كما هي في زمن معين، وفق إطار نموذجي محدد. وهذا النوع من المدونات يهدف إلى تحقيق خاصيتين هما **التمثيل** Representativeness بتمثيل نصوص المدونة للغة أو لهجة معينة، **والتوازن** Balance وذلك بتغطية نصوصها لكل مجالات وأوعية اللغة بطريقة متوازنة تشبه وجودها في مجموع اللغة، من خلال **إطار نموذجي** Sampling frame عبارة عن تحديد للعينات التي تتضمنها المدونة، وكيفية اختيار هذه العينات من مجموع النصوص،

---

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ١١-١٨

(٢) The Bank of English User Guide, 20/12/2018: titania.bham.ac.uk/docs/svenguide.html

وما أنواع النصوص المختارة، وزمنها، وخصائص أخرى، كعدد وطول العينات التي قد تكون الواحدة منها نصاً كاملاً أو ملخص نص.

ويُسمى هذا النوع من المدونات أيضا **المدونات المقتطفة** Snap shot. وأفضل مثال لها مدونة LOB الممثلة للإنجليزية البريطانية المكتوبة عام١٩٦٠م. فقد اتخذت إطاراً نموذجياً مطابقاً للإطار النموذجي لمدونة براون الأمريكية الذي يضم ٥٠٠ عينة موزعة على موضوعات متنوعة في الإنجليزية بطريقة متوازنة، فكل عينة ضمت ٢٠٠٠ كلمة ليكون بذلك متضمناً مليون كلمة(١).

ويهتم اللغويون في اللغويات غالباً بدراسة عموم التنوع اللغوي بدلا من دراسة نص واحد أو نصوص لكاتب واحد. وفي هذه الحالة يكون لدينا خيارات لجمع البيانات في مدونتنا. فإما أن يحلل كل نص مفرد في كل التنوع اللغوي، أو بناء نموذج أصغر من هذا التنوع. والخيار الأول غير ممكن عملياً باستثناء حالات قليلة جداً، كما في اللغات الميتة حيث يمكن إحصاء نصوصها القليلة. أما اللغات الحية كالإنجليزية والعربية فعدد نصوصها يزداد وهو غير محدود نظرياً، لذا فإن تحليل كل نص في هذه اللغات هو مهمة بلا نهاية ومستحيلة(٢).

وثمة منهج يقع في منتصف الطريق بين منهجي المدونات الراصدة والمدونات النموذج، وهو المنهج الذي ينفذ المدونة الراصدة ولكن في إطار نموذجي محدد(٣). ومن الشائع في أبحاث معالجة اللغات الطبيعية أن يكون لدينا مدونة مرجعية خاصة يمكن من خلالها استخلاص المعلومات لتطوير وتحسين البرامج(٤). وفي التوسيم النحوي الآلي تستعمل بيانات التوسيم للمدونة كبيانات تدريب لتوليد الاحتمالات، حيث تعرف بها

---

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ١٩

(٢) Corpus Linguistics (An Introduction), 2011, p. 29

(٣) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، صص. ١١-١٢

(٤) Corpus Linguistics (An Introduction), 2011, p. 139

الاحتمالات الخاصة لكل وسم، كأن يكون نوع الكلمة الأكثر احتمالاً لما بعد الأداتين (لم ـ لن) هو الفعل. وهكذا تستخلص المعلومات لتزود بها **البرامج الآلية المعدة للتوسيم** taggers ثم يطبق على نفس المدونة المنتج النهائي للموسم الآلي، ويقارن بين المدونة الموسمة يدوياً أو الموسمة شبه يدوي بالمدونة الموسمة آلياً بهذا الموسم لتقييمه ومن ثم تحسينه[١]. إن هذا النوع من المدونات يعرف باسم **المدونة القياسية** benchmarking corpus/dataset. وهي مدونات بين المنهجين السابقين حيث إنها ذات حجم محدد، وموسمة يدوياً أو شبه يدوي أو آلياً حسب الغرض المعدة من أجله، ولها استعمالات متعددة ومتنوعة في مجال اللغات الطبيعية[٢].

ويمكن أن تبنى المدونة المرجعية كمدونة خاصة لهذا الهدف تحديداً مثل مدونة البنك الشجري العربي المقتصرة نصوصها على النصوص الصحفية والمحللة صرفيا ونحويا بطريقة يدوية من خلال اختيار الاحتمال الصحيح من القائمة التي يختارها محلل باكولتر لكل وحدة، وطور بالاعتماد عليها موسم منى دياب[٣]، وموسم فان دن بوش[٤]، وموسم ستانفورد[٥]، وغيرها من الموسمات. وقد تُبنى بالاعتماد على مدونة أخرى، واتخاذ **مدونة فرعية** subcorpus منها، كما حدث في مشروع تحسين الوسوم النحوية لمدونة BNC الذي اعتمد على مدونة فرعية منها تضمنت مليوني كلمة[٦].

---

(١) Corpus Linguistics (An Introduction), 2011, p. 139

(٢) Ibid.

(٣) Automatic tagging of Arabic text: From raw text to base phrase chunks, 2004, pp. 149-152

(٤) Van den Bosch, A., Marsi, E., Soudi, A. Memory-based morphological analysis and part-of-speech tagging of Arabic. In Arabic Computational Morphology, Springer Netherlands, 2007, pp. 201-217

(٥) The Stanford Natural Language Processing Group. What POS tag set does the parser use?. 23-3-2016:
http://nlp.stanford.edu/software/parser-arabic-faq.shtml#d

(٦) Corpus-Based Language Studies, 2006, p.350; Corpus Annotation, 2015, p.139

وثمة سؤالان على الباحث أن يجيب عنهما قبل شروعه في تصميم المدونة، وهما:

١- ما الغرض الذي ستعد من أجله المدونة؟

٢- هل توجد مدونة أخرى تقوم بتحقيق هذا الغرض؟

ويحدد الغرض من المدونة من خلال تحديد الباحث لسؤال أو مجموعة من الأسئلة تقود إجاباتها إلى الوقوف على نوع معلومات المدونة وكيفية الحصول عليها. أما الإجابة عن السؤال الثاني فتكون باطلاع الباحث على محتويات وأدوات المدونات المتوفرة والدراسات التي أجريت باستعمالها للتأكد من إمكانيتها للإجابة عن أسئلة بحثه. وإذا لم يجد الباحث المدونة الملائمة فليس أمامه سوى أن يغير أسئلة بحثه ليتمكن من الاستفادة من المدونات المتوفرة، أو أن يصنع مدونته بنفسه[١].

ومن المدونات المرجعية الأخرى التي اعتمد عليها في بناء الموسمات النحوية العربية: المدونة العربية القرآنية التي استعملت لتطوير موسم نحوي عربي سعى مطوروه إلى أن يكون عالي الدقة[٢]، ومدونة شيرين خوجة ذات النصوص الصحيفة المأخوذة من صحيفة الجزيرة السعودية والمستعملة لتطوير موسمها النحوي[٣]. وكلتا المدونتين متخصصة في نوع لغوي واحد: الأولى في عربية القرآن، والثانية في الصحافة السعودية، وهذا يعني أن الموسمين سيحققان دقة أقل عند تطبيقهما على نصوص من نوع آخر غير النصوص التي درب عليهما الموسمان[٤]. ويظهر الجدول (٢-١) أمثلة لبعض المدونات الأخرى المستعملة في بناء وتطوير أنظمة التوسيم النحوي الآلية.

---

(١) الثبيتي، عبد المحسن، تصميم المدونات اللغوية وبناؤها. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، ٢٠١٥، ص. ١٤٩

(٢) Dukes K. Quranic Arabic Corpus. 9-6-2016: http://corpus.quran.com/

(٣) APT: Arabic part-of-speech tagger, 2001, pp. 20-25

(٤) Part-of-speech tagging from 97% to 100%: is it time for some linguistics?, 2011, pp. 171-189

جدول (٢-١) قائمة بالمدونات اللغوية المستعملة في الدراسات العربية المتعلقة بأنظمة التوسيم النحوي

المدونة	النوع	الموسم	المستوى اللغوي	حجمها	
				النصوص	الكلمات
صحيفة الراية[١]	مقالات صحفية	الموسم النحوي المتعدد التقنيات	العربية الحديثة	لم تحدد	لم تحدد
القرآن الكريم وملخصات علمية[٢]	القرآن الكريم وعلوم الحاسوب	الموسم النحوي القائم على التحليل الصرفي	العربية القديمة والحديثة	٢٤٣	لم تحدد
جزء من مدونة العربية المعاصرة[٣]	قصص أطفال	الموسم النحوي المتعدد الطرائق	العربية الحديثة	٢٧	٢١,٩٥٨
مدونة الموقع الرسمي لوزارة التربية الأردنية[٤]	كتب مدرسية	الموسم النحوي القائم على الأوزان الصرفية	العربية الحديثة	لم تحدد	٥٠٠٠

(١) Abuleil, S., Evens, M. Discovering lexical information by tagging Arabic newspaper text. In Proceedings of the Workshop on Computational Approaches to Semitic Languages, pp. 1-7, 1998

(٢) Kanaan K., Al-Shalabi R., Sawalha M. Full automatic Arabic text tagging system. The proceedings of the International Conference on Information Technology and Natural Sciences, Jordan, 2003

(٣) Zribi, C., Aroua, T., Ahmed, M. A Multi-Agent System for POS-Tagging Vocalized Arabic Texts. Int. Arab J. Inf. Technol. 4.4, 2007, p. 322-329

(٤) Pattern-based algorithm for Part-of-Speech tagging Arabic text, 2008, pp. 119-124

المدونة	النوع	الموسم	المستوى اللغوي	حجمها	
				النصوص	الكلمات
القرآن الكريم والأدب العربي[١]	القرآن الكريم والأدب	الموسم النحوي الإحصائي	العربية القديمة	٢	٥٦,٣١٢

والملاحظ أن هذه المدونات في معظمهما تقع ضمن العربية القديمة أو ضمن العربية الحديثة، وتقع في مجال واحد فقط. وما كان منها ما يجمع بين العربيتين، فالعربية القديمة فيه تقع في مجال واحد هو لغة القرآن أو لغة الأدب القديم، والعربية الحديثة فيه تقع في مجال معين كالصحافة والقصص وعلوم الحاسوب. واعتماد الموسم النحوي على بيانات من نوع واحد يعني كما أشرت آنفا أن هذا الموسم الآلي ستقل دقته عند تطبيقه على بيانات من نوع آخر. فلم يرد في القرآن الكريم مثلا كل ما ورد في العربية من الإشاريات، على نحو: تينك وذينك التي ترد في الأدب العربي، وقد نجد عكس ذلك، فهل سيكشف الموسم ذلك إذا لم يدرب على غير بيانات القرآن الكريم؟. ويلاحظ أيضا أن بعض هذه المدونات غير مكتملة المعلومات، إذ لم يحدد إطارها ولا حجمها وإن كان الحجم ليس مهما في حد ذاته حيث التمثيل والتوازن هما قوام المدونة؛ إذ يزيدا من فرص تغطية المجموع اللغوي قيد الدراسة. والجدير بالذكر أن كل تلك المدونات ذات منطلقات حاسوبية ولم يتح الوصول إليها لاستعمالها في بناء مدونات أكبر.

وتمر المدونات اللغوية بمجموعة من الإجراءات تسبق عملية جمع البيانات وتليها[٢]، وهي كما يلي:

---

(١) Elhadj, Y. Statistical Part-of-Speech Tagger for Traditional Arabic Texts. Journal of Computer Science, Vol.5, No. 11, 2009, p. 794-800

(٢) الثبيتي، عبد المحسن. تصميم المدونات اللغوية وبناؤها. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، ٢٠١٥، ص. ١٦١

١- الملكية الفكرية وحقوق النشر:

هناك موضوعان متعلقان بالملكية الفكرية وحقوق النشر عند تصميم المدونة وجمعها، فثمة حقوق لتضمين أي نص في المدونة وحقوق لنشر أي نصوص في المدونة. ولأن انتهاك أي حق من حقوق الملكية الفكرية قد يضيع عمل الباحث، لا بد من اهتمام الباحث بهذه المسألة قبل الشروع في جمع مدونته سواء كان ينوي توزيعها أو حتى الاستفادة منها لأغراضه البحثية الخاصة. ويفضل أن تكون نصوص المدونة غير محمية بقوانين الملكية الفكرية، إذ يتاح للباحث استعمال بعض النصوص بالقدر الذي يشاء، كالنصوص التراثية، أو النصوص التي انتهت حمايتها النظامية بموجب القانون، أو النصوص التي يسمح بعض ملاك الحقوق باستعمالها كاملة كنصوص إصدارات الجهات الرسمية، أو يسمحون باستعمال مقتطفات منها استعمالا عادلا بشروط معينة حسب الحجم والأهمية، كالصحف والمناهج الدراسية[١].

وتنص المادة الثامنة من قوانين الحماية الفكرية بالمملكة العربية السعودية على الاستثناءات التي لا تخضع لقوانين الحماية الفكرية منها نسخ نصوص الأنظمة والأحكام القضائية والاتفاقات الدولية وسائر الوثائق الرسمية، ونسخ المصنفات للاستعمال الخاص، ونسخ المقالات السياسية والاقتصادية والدينية المنشورة في الصحف والمجلات، وما كان معلنا للجمهور كالخطب والمحاضرات والمرافعات القضائية، فضلا عن النسخ التام مع أخذ إذن المؤلف بشرط الإشارة له وبشرط ألا يتسبب بضرر للمؤلف[٢].

٢- تحديد المصادر:

يحكم هذا الإجراء نوع المدونة والمدة الزمنية المراد إنجاز المدونة فيها. فثمة مصادر تقدم نصوصا جاهزة للمعالجة الإلكترونية بصيغ نصية مثل txt وdoc فتوفر الجهد والوقت

---

(١) الثبيتي، عبد المحسن. تصميم المدونات اللغوية وبناؤها. مركز الملك عبد اهلل بن عبد العزيز الدولي لخدمة اللغة العربية، ٢٠١٥، ص. ١٦١

(٢) المملكة العربية السعودية. نظام حماية حقوق المؤلف بالمملكة العربية السعودية. ١٨-١٠-٢٠١٨: https://www.boe.gov.sa/ViewSystemDetails.aspx?lang=ar&SystemID=16&VersionID=24

والمال، وثمة مصادر تتطلب التحويل إلى نصوص قابلة للمعالجة الإلكترونية، كالمصادر المكتوبة يدويا أو المصورة أو ذات الصيغة pdf، فتضطر جامع المدونة إلى تحويلها لملفات ذات صيغ نصية قابلة للمعالجة الآلية عن طريق تطبيقات متعددة كبرامج التعرف الضوئي على الكلام أو تحويل الصيغ ومراجعتها. وحتى يسهل القرار في هذه المرحلة، توضع قائمة بأسماء المصادر وروابطها الإلكترونية وطرق الحصول عليها ومعلوماتها المؤكدة حتى يتمكن جامع المدونة من الحذف والإضافة والاختيار بكل سهولة.

٣- الجمع:

يفضل عند جمع المدونة أن تكون المصادر إلكترونية وجاهزة للاستعمال المباشر. وقد تجمع المواد يدويا بالاستعانة بالقائمة المعدة في الإجراء السابق، كما قد يلجأ لتطبيقات حاسوبية تقوم بالجمع آليا، لكن النصوص ستكون بحاجة للفرز لتحديد أوعيتها وموضوعاتها ومعلوماتها، فضلا عن ضرورة مراجعتها الدقيقة للتأكد من خلوها من الأخطاء، والتكرارات، ومما ليس من نص الموضوع، كروابط المواقع والإعلانات.

٤- إدارة الملفات:

يفضل أن يكون ترميز النصوص موحدا في جميع نصوص المدونة ومتوافقا مع كافة أنظمة التشغيل، وتطبيقات معالجة نصوص المدونات؛ لإعادة استعمال المدونة. فلا يستعمل ترميز **ويندوز** windows الذي لا يتوافق إلا مع أنظمة تشغيل ويندوز، ويستعمل ترميزي UTF8 و UTF16 المقبول من معظم أنظمة التشغيل والتطبيقات. وعند تسمية الملفات يفضل أن يكون مسمى الملفات موحدا وعبارة عن سلسلة من الأحرف الدالة على معيار التصنيف، كأن تبدأ جميع الملفات في الوعاء الواحد بنفس الحرف الدال على الوعاء ثم رقمه السلسلي في مجموعة الملفات. وللتحكم بملفات المدونة بسهولة سيكون من المنطقي أن ترتب المجلدات حسب المعيار المعتمد عليه على أن يضمن كل مجلد الملفات المتعلقة به.

ويُجرى تصنيف المدونة تصنيفا شكليا بناء على النوع والغرض والعدد والتصميم كما بدا في دراسات المدونات اللغوية. وفي داخل كل قسم، هناك اختلافات وتداخلات في التصنيف. وقد نجد مدونات لغوية مصنفة وفقا لأكثر من صنف من هذه الأصناف الأربعة:

١- النوع:

تنقسم المدونات من حيث النوع إلى: **مدونات راصدة** monitor corpora، **ومدونات العينة** sample corpora، **ومدونات الويب** web corpora، **والمدونات المقتطفة** snapshot corpora **والمدونات الفُرصية** opportunistic corpora. ويشير ماكنري وهادري[١] إلى أن كل نوع له خاصية نوعية مميزة في بنائه، فالأول يكون مفتوحا وقابلا للإضافة بهدف التعرف على طبيعة التغير اللغوي للمدونات اللغوية، ويتضمن من التصنيف حسب الغرض النوع التاريخي والتزامني، وقد يكون تقابليا بين لغتين أو أكثر وراصدا في الوقت ذاته. أما المدونة العينة فإن حجمها وزمنها ونوعها ومجالها يتحدد بتحديد سؤال البحث. فمثلا قد تكون مدونة تمثل كتابات طلاب المرحلة الابتدائية في مدينة محددة لغرض الإجابة عن السؤال الآتي: ما حجم تطور الكثافة المعجمية لدى طلاب المرحلة الابتدائية بين السنوات الست؟. أما مدونات الويب، فهي عادة من أسهل المدونات استرجاعا، إلا أن مشكلتها تكمن في حاجتها إلى التنقيح، نظرا لأن البيانات التي تُستورد من المواقع عادة ما تكون **بيانات مشوبة** dirty data. أما المدونات المقتطفة والفرصية فإن الفرق بينهما نابع من منطلق من يجمع كل نوع. فالذي يجمع النوع الأول يكون قاصدا لما يريد جمعه لغرض البحث، أما الفرصية فإن القصد ليس متشكلا عند من يجمع المدونة؛ لأن الدافع إلى الجمع فيها ليس من الأسئلة البحثية أو الغرض من الجمع بل من طبيعة هذه المدونة التي يود الجامع أن يتحين فرص انتقائها لغرض منهجي أو تطبيقي أو بحثي معين[٢]. ولو

---

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ١١-٢٦

(٢) المجيول، سلطان. البحث اللغوي الآلي في المدونة الحاسوبية واللغة العربية. التواصل اللساني، م١٩، ع:١-٢، ٢٠١٨، ص. ٥٧-٦٠

عدنا إلى مثال مدونة الطلاب في المرحلة الابتدائية، ونظرنا إلى طبيعة جمعها من جهة النموذج، والمقتطفة، والفرصية، فإنه من الممكن إعطاء مثال تقريبي يوضح الفروقات بين هذه الأنواع الثلاثة من المدونات من ناحية النوع على النحو الآتي:

• **العينة** sample: توفر شرط التمثيل representativeness والاتزان balanced:

المرحلة الأولى: قياس الكثافة المعجمية بعد كل سنة من سنوات المرحلة الابتدائية.

المرحلة الثانية: التمثيل: جمع نصوص مكتوبة من الطلاب من عينة منتظمة من مدارس المدينة.

المرحلة الثالثة: الاتزان: الحاجة إلى جمع ٦٠ كتابة نصية من طلاب الصف الأول، و٥٠ من الصف الثاني، و٤٠ من الصف الثالث، وهكذا من أجل أن يكون حجم النص لكل مرحلة متزنا.

• **المقتطفة** snapshot:

المرحلة الأولى: قياس الكثافة المعجمية بين سنوات المرحلة الابتدائية.

المرحلة الثانية: جمع عينة نصوص مكتوبة بشكل عشوائي.

• **الفرصية** opportunistic:

المرحلة الأولى: تحين الفرص بعينات توفرت دون قصد التصدي لها.

المرحلة الثانية: قياس الكثافة المعجمية بين طلاب المرحلة الابتدائية.

٢- الغرض:

يمكن تصنيف المدونات اللغوية حسب الغرض[١] إلى **المدونات التاريخية** diachronic corpora **والمدونات المتزامنة** synchronic corpora، **ومدونات المتعلمين**

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٣٤-٣٨، ١٥٤-١٥٨، ١٧٧-١٨٩

learner corpora **والمدونات التقابلية** contrastive corpora، **ومدونة الصور**، وتعرف بـ multimodal corpus [١].

وقد يظن بعض الباحثين أن المقصود بالمدونات التاريخية هي تلك التي تتضمن نصوصا قديمة تعود إلى ما قبل مئات السنين، وذلك غير دقيق؛ حيث إن المدونة التي تتضمن نصوصا تعاقبية زمنيا تُعنى بدراسة التطور أو التغير على اللغة حتى وإن كانت نصوصها ممثِّلةً حقبةً زمنية محددة، على سبيل المثال من ٢٠١٠ إلى يومنا هذا. أما المدونات التزامنية، فإننا لو نظرنا على سبيل المثال إلى مدونة السليطي عن العربية المعاصرة[٢]، فهي تمثل الحقبة المعاصرة. وقد يُوظف هذا النوع من المدونة للدراسات اللغوية التاريخية فيما لو كانت النصوص المعاصرة ممثلة لسنوات معاصرة متعاقبة يُمكن أن يحلل فيها تطور على مستوى أو مجال لغوي معين. أما مدونة المتعلمين، فهي بطبيعة الحال تتضمن نصوصا عربية طبيعية أنتجها متعلمو اللغة، وتختلف في التصميم عن أنواع المدونات الأخرى إذ إن تطبيق التوسيم فيها يكون على مستوى نوعيات الأخطاء. والمدونات التقابلية إما **متشابهة** comparable أو **متوازية** parallel. والفرق الأساسي بينهما هو أن الأول يحتوي على نصوص من اللغة المصدر متقابلة مع نصوص من اللغة الهدف. وهو تقابل ليس من واقع ترجمي حقيقي، بل تقابل يقارب السياقات بين اللغتين وفق أربعة معايير: تقابل الوعاء، وتقابل المجال، وتقابل الفترة الزمنية، وتطابق معايير التصميم والبحث[٣]. أما النوع الثاني فالغرض الأساسي في بنائه يقوم على محاذاة نص من اللغة المصدر مع النص المترجم إليه في اللغة الهدف، وتكون المحاذاة على مستوى الفقرة،

---

(١) Knight, Dawn. (2011). The future of multimodal corpora. Revista Brasileira de Linguística Aplicada, Vol. 11, No., 2, p. 391-415
DOI: https://dx.doi.org/10.1590/S1984-63982011000200006

(٢) البحث اللغوي الآلي في المدونة الحاسوبية واللغة العربية، ٢٠١٨، ص. ٦٠

(٣) السابق، نفس الصفحة.

والجملة، والعبارة، والكلمة إن كان هدف الغرض من بنائها قائما على دقة التصميم[١].

٣- العدد:

تتفرع المدونات تحت هذا النوع إلى مدونات أحادية اللغة، أو ثنائية اللغة، أو ثلاثية اللغة. وينقسم النوعان الأخيران إلى **مدونات متوازية** parallel corpora، **ومدونات متشابهة** comparable corpora، وقد تمت الإشارة لهما في الحديث عن الغرض.

٤- التصميم:

يتجاوز بناء المدونة كثيرا مما هو مذكور في أدبيات لغويات المدونات. وسأشير إلى أشكال التصنيف ـ كون المدونة اللغوية تأخذ أبعادًا تصميمية متشعبة، وذات أغراض مختلفة ـ تحصرها على النحو الآتي:

١. مدونة **الفيرثيون الجدد** neo-Firthian (نصوص خام من دون تعليم وتوسيم وتحشية) في مقابل المدونة الموسمة (توسيم آلي، توسيم شبه آلي، توسيم يدوي على المستوى الصوتي والصرفي والنحوي والدلالي والتطريزي). ويرى الفيرثيون الجدد ومنهم **سنكلير** Sinclair أن تعليم المدونة وتوسيمها وتحشيتها واختبارها دلاليا ليس ذا أهمية؛ لأن تحليل الاستعمال اللغوي الطبيعي هو المحك، إذ إن الحدس اللغوي في معرفة طبيعة التصاحب والكشافات السياقية هو المهم عند معالجة المدونة وتوظيفها في البحث[٢].

٢. مدونة موسمة خالية من البيانات الواصفة في مقابل المدونة الموسمة المحشاة بالبيانات الواصفة. ويقصد **بالبيانات الواصفة** metadata كل المعلومات الإضافية عن مرجع النص الرقمي: اسم المؤلف، وعمره، وجنسيته، إلخ.

---

(١) انظر على سبيل المثال المدونة الإنجليزية-العربية المتوازية English Arabic Parallel Corpus:
Alotaibi, Hind, Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. Arab World English Journal, Vol. 8, No. 3, 2017
DOI: https://dx.doi.org/10.24093/awej/vol8no3.21

(٢) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٢٣٦

٣. مدونة متنوعة من حيث **الوعاء** genre أو **الوسيط** medium، **والمجال** domain، **والموضوع** topic، **والفترات الزمنية** periods of time، **والدول** countries.

٤. مدونة ذات اهتمام في أثناء التصميم **بالسجل اللغوي** register، بمعنى أنها تصنف النصوص إلى نصوص عامة، ونصوص أكاديمية، ونصوص أدبية، ونصوص قانونية، ونصوص للغة الأطفال، إلخ.

٥. مدونات مختلفة التصميم من ناحية بناء أدوات المدونة. وهذه الأدوات هي: البحث المخصص لنوع الكلمة، والبحث عن طريق **المحرف البديل** wild card، وامتداد التتابعات اللفظية، وحجم نافذة الكشافات السياقية، وإحصاءات اختبار الدلالة الإحصائية والغرابة والارتباط، ونوعية الملفات التي يُمكن حفظ نتائج البحث بها.

٦. مدونة مكتوبة في مقابل المدونة المنطوقة: على الرغم من تقدم تصميم المدونات الإنجليزية إلا أن حجم المنطوق غالبا يكون أقل بكثير من المكتوب. على سبيل المثال: ٩٠٪ من مدونة BNC مكتوبة بينما ١٠٪ تشكل المنسوخة من المنطوق.

٧. **مدونة شخصية** personal corpus، وتسمى أيضا بـ do-it-your-self corpus (DIY).

٢-٣ المعالجة القبلية:

تمر النصوص اللغوية في المدونات بمرحلتين **تجهيزيتين preprocessing**: الأولى تفرق فيها الكلمات بعضها عن بعضها، وتفصل عنها علامات الترقيم، والرموز الملتصقة بها، وتحدد فيها أيضا الكلمات المركبة، وتعرف هذه العملية **بالتفريق** tokenization. وتشمل أيضا مهام أخرى تعرف **بالتسوية الهجائية** orthographic normalization كإزالة التشكيل وعلامة الكشيدة، وتوحيد الهمزات. وثمة مقترحات تقلص الاحتمالات الهجائية للكلمات، أي: تحول أشكالاً مختلفة لشكل واحد تطلبها

الإملاء إلى شكل واحد فقط، كشكل الهمزة في حالات الرفع والجر، في نحو: أبنائهم وأبناؤهم، وأشكال الألف المقصورة، في نحو مولاي وإليهم وإلى، فتحول كل همزة متوسطة إلى همزة على السطر، وتحول كل ألف مقصورة إلى (ياء). وقد تراعى السياقات فتحدد شكل الألف المقصورة المناسبة في كل كلمة، فإليهم لا تكون إلي + هم بل إلى + هم، وهكذا[١].

وقد تكون جملة مثل: «مَرّت على المدينة المنورة عاصفةٌ أزالت لوحاتها.»، بعد التفريق، إما:

مرت	على	المدينة المنورة	عاصفة	ازالت	لوحاتها	.

أو:

مرت	على	المدينة	المنورة	عاصفة	ازالت	لوحاتها	.

بحسب القواعد المطبقة في التفريق التي تحددها حاجة نظام التوسيم.

وقد يبدو تفريق الكلمات لأول وهلة عادياً. فالعادة تحدد الكلمة في النص بوضوح من خلال مسافتين سابقة ولاحقة، أو سطر جديد. كما يمكن أن تصاحبها إحدى علامات الترقيم أو الرموز فيسهل تفريقها عنها لإمكانية حصر علامات الترقيم والرموز في قوائم تزود بها أدوات التفريق ولا تحتاج لفك الغموض. وفي اللغات ذات النظام الكتابي الألفبائي كالعربية غالباً ما يحدد رسم الكلمة آليا في التمثيل الكتابي للنص نفسه، وليس هذا في الأنظمة غير الألفبائية كالصينية التي تتطلب معالجات آلية معقدة لتحديد الكلمة[٢]. ومع ذلك، فإن رسم الكلمة في اللغات الحديثة الألفبائية كالإنجليزية والعربية الذي يحدد

---

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٣٨

(٢) Corpus Annotation, 2013, p. 21; Corpus-based language Studies, 2006, p. 35

بالمسافتين القبلية والبعدية ليس بالضرورة أن يكون مثل رسم الوحدة الصرف ـ نحوية لها المراد معالجتها هنا، على نحو: thirty one في الإنجليزية و(أحد عشر) في العربية اللتان تعدان كلمتين في كلٍ، إذا ما نظرنا للمسافات السابقة والتالية، وهما في الحقيقة على وحدة واحدة وليستا على وحدتين. ولذلك، فإن **الكلمة** token هي ما نحتاج إلى تحديده من أجل التوسيم النحوي، وكلمة (أحد عشر) هي كلمتان بالمفهوم الحاسوبي، ولكنهما لا يتطابقان هجائياً مع رسم الوحدة الصرف ـ نحوية الواحدة للعدد المركب: أحد عشر. ولم تتناول أي دراسة سابقة متعلقة بأنظمة التوسيم النحوي العربية المركباتِ على هذا النحو، وأشهرها باكولتر الذي يتعامل معها بالمفهوم الحاسوبي للكلمة الذي لا يصح الاعتماد عليه إذا ما أردنا التعامل مع الكلمات العربية بدقة، وإذا ما كنا ننشد نتائج صحيحة ودقيقة.

٢-٤ متغيرات التقطيع:

وتعرف المرحلة التجهيزية الثانية **بالتقطيع** .segmentation وفيها تحدد المورفيمات الصرفية التي ينبغي فصلها عن الكلمة بعد تفريقها لتتم عملية توسيمها، ومنها مثلا الضمائر المتصلة وحرفا العطف (الفاء والواو). فالجملة السابقة:

بعد التفريق:	العاصفة		التي	مرت	على	المدينة المنورة	أزالت	لوحاتها	.
بعد التقطيع:	ال	عاصفة	التي	مرت	على	المدينة المنورة	أزالت	لوحات	ها	.

ولاحظ أننا نستطيع التحكم بعملية التقطيع حسب ما تقتضيه قوائم الوسوم النحوية في المرحلة التالية. فإن كنا سنخصص وسما لـ ال التعريف، سنقتطعها من الكلمة ـ كما في المثال ـ لأجل وسمها وباعتبارها قسما من أقسام الكلام. وإن كنا سننظر إلى أن تاء التأنيث الساكنة المتصلة بالفعلين الماضيين (مرت ـ أزالت) لاصقة ينتفع بها في أداء معنى الشخص

(الغيبة)[١]، سنبقيها على اتصال بالكلمة ولن نقتطعها باعتبارها وحدة صرف ـ نحوية. وهذا يعني أن كل ما يقتطع لا بد أن يعين له وسم من الوسوم النحوية، فلا يقتطع الموسِّم ال التعريف إذا كان لا يرى أنها قسم من أقسام الكلام العربي.

وفي كـثير من اللغات يشكل التـقطيع تحديـاً، حيث يتـطلب ربطاً بالمعجم، واستعمالاً للنماذج الإحصائية كما في العربية والصينية، ولذلك يرتبط فيها التقطيع بما يعرف بالتسوية الهجائية. وهي مهمة تحليلية توليدية يقصد بها إنتاج شكل سطحي صحيح للكلمة بعد عملية التقطيع يهدف إلى الحد من **تناثر البيانات** sparsity الذي يؤدي إلى إنتاج عدد كبير من الأشكال للكلمة الواحدة أو زيادة عالية في **المفردات غير المعروفة** Out of Vocabulary OOV، ومثال ذلك إرجاع التاء المربوطة لشكلها الهجائي السليم بعد فصل الضمير المتصل بها، في مثل: حقيبتها ← حقيبت / ها ← حقيبة / ها[٢].

ويعد الصرف عنصراً أساسياً في عملية التوسيم النحوي؛ لأهميته في التعامل مع شكل الكلمة ووظيفتها وتفاعلها مع الكلمات والوحدات الأخرى[٣]. ويتميز صرف العربية بالمورفيمات **السلسلية** concatenative التي تساهم في تكوين الكلمة بطريقة تسلسلية متلاصقة، حيث تتكون الكلمات من **الجذوع** stems فضلا عن **اللواصق** affixes التي تكون عبارة عن السوابق prefixes **واللواحق** suffixes. فالكلمة: (فسيكتبونها) مثلا، السوابق فيها: (ف ـ س)، والجذع: (يكتب)، واللواحق: (ون ـ ها)[٤]. وتقسيم الكلمة إلى جميع المورفيمات المكونة منها (الجذوع ـ سوابق ـ لواحق) هو ما يعرف بالتحليل الصرفي الشكلي. وتفسير هذه المورفيمات، هو ما يعرف بالتحليل الصرفي

---

(١) أحد المعاني الصرفية التي تؤدى باللواصق. ومنها الشخص ويدخل فيه (التكلم والخطاب والغيبة)، والعدد وفيه (الإفراد والتثنية والجمع)، والنوع ويدخل فيه (التذكير والتأنيث والحياد)، والتعيين، ويدخل فيه: (التعريف والتنكير).

(٢) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ٦٩

(٣) السابق، نفس الصفحة.

(٤) السابق، صص ٧١-٧٢

الوظيفي[١]. فكلمة مثل (الكاتبان) تحلل شكلياً إلى (ال ـ كاتب ـ ان)، ولا يمكن لهذا النوع من التحليل الكشف عن نوع الكلمة من حيث الإفراد والتثنية والتعريف والتنكير، في حين أن التحليل الصرفي الوظيفي يكشف ذلك. وثمة أنظمة تبنى لمعالجة أغراض لا تتجاوز الحاجة فيها إلا تعيين الاسم والفعل والحرف، فلا يلجأ فيها إلا للتحليل الشكلي، وهناك أنظمة تبنى لحاجات أدق وأعمق تستلزم تحديد خصائص لغوية أخرى كالنوع والعدد أو تحديد نوع الفعل مثلا من حيث زمنه الصرفي (مضارع ـ أمر ـ ماض) فيكون التوجه للتحليل الصرفي الوظيفي.

ويشكل الاعتماد على الجذوع باعتبارها وحدة لفظية إشكالية في التوسيم النحوي. فعند توسيم (سيكتب) نحويا قد لا نحتاج لفصل السين إذا خصصنا للفعل الدال على الاستقبال وسماً خاصاً، ولم نخصص للسين وسما مستقلا مع الحروف أو الأدوات. أما لو كانت مجموعة الوسوم لا تصنف الأفعال حسب زمنها الصرفي وتعتبر السين أداة استقبال، فقد نفصل السين عن الجذع (يكتب) ونوسم الجذع باعتباره فعلا دون تخصيصه بزمن. أيضاً، قد لا نقرر في كلمة مُعرّفة بأل فصل ال التعريف عنها باعتبارها سابقة؛ لأن فصلها لن يمكننا من تحديد نوعها من حيث التعريف والتنكير. كما أنه لا يمكننا فصل التاء المربوطة عن الكلمة؛ لأن فصلها لن يمكننا من تحديد نوع الكلمة من حيث التذكير والتأنيث أو الإفراد والجمع. وهكذا، لا يمكن أن نكتفي بالتحليل الصرفي بوصفه خطوة أولى للتوسيم النحوي. إننا نلجأ للتقطيع وليس التحليل لتمثيل الوحدة اللفظية للكلمة بما في ذلك **الجذوع** Stems **والوحدات المعجمية** Lexical Units[٢].

إن الحالة الافتراضية للكلمة من منظور معلوماتي هي أن تكون الكلمة هجائيًا تمثل كلمة فعلية واحدة كما في كتاب وحقيبة وطالب[٣]. وهناك ثلاث حالات أساسية

---

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ٨٤

(٢) السابق، ص. ٦٩

(٣) Corpus Annotation, 2013, p.33

في العربية وفي لغات أخرى أيضا، تعد عدولاً عن العلاقة الافتراضية بين الكلمات الفعلية الصرف ـ نحوية والكلمات الهجائية. هذه الحالات الاستثائية الثلاثة هي[١]:

١. **الكلمة المتعددة** Multiword: وهو أن تقابل أكثر من كلمة ـ هجائياً ـ كلمة فعلية واحدة، مثل الأعداد المركبة تركيباً إضافياً، مثل: أحد عشر، والأعداد المعطوفة، مثل: خمسة وعشرون.

٢. **اللصقيات** Mergers: وهي الكلمات التي تكتب هجاء ككلمة واحدة ويعبر عنها بوحدتين صرف ـ نحويتين أو أكثر. وتتضمن جميع أمثلة هذا النوع في العربية أشكالاً من اللواصق تدمج في الكلمة هجاء وصوتا، على نحو: ﴿أَنُلْزِمُكُمُوهَا﴾ (هود: ٢٨)، حيث تتضمن أكثر من وحدة صرف ـ نحوية وهن: همزة الاستفهام، والفعل المضارع، والضمير (كم) المشبعة حركة الضم فيه واوا، والضمير (ها). وعلى نحو: يسائلونهم، حيث تتضمن وحدتين صرف ـ نحويتين هما: يسائلون، والضمير (هم).

٣. **المركبات** Compound: وهو ما كان كلمة أو أكثر في شكله الهجائي ويعبر عنه بوحدة أو أكثر من الوحدات الصرف ـ نحوية. ومن الصعب وضع حدود لهذا النوع، فالكلمة مثلاً (رأسمال) هي كلمتان أساسا (رأس ومال)، ولها مدخل معجمي مستقل في المعاجم العربية. ومن جهة أخرى، هناك من يكتبها بوصفهما كلمتين مستقلتين هجائياً، وهناك من يفضل استعمال الشرطة بينهما، على نحو: رأس ـ مال للإشارة إلى أنهما من تركيب واحد. وهذا يعني أنه ليس ثمة ضابط لذلك؛ إذ يُعتمد على الأسلوب الكتابي.

والاستثناءات السابقة هي إحدى التحديات التي تواجه الموسم النحوي في توسيمه للمدونات عند تفريق وتقطيع النص لكلمات. فلا بد مسبقاً من تحديد الموسم لأي الكلمات صُنفت من **الكلمات المتعددة** multiword، وأيها صنف من **اللصقية** mergers وأيها صنف من **المركبة** compound؛ وذلك لاتخاذ القرارات بشأن تقطيعها الذي سيُبنى عليه توسيمها نحوياً. حيث قد يضم البعض مثلا للكلمات المتعددة الأفعال

---

(١) Corpus Annotation, 2013. p. 33

المتعدية بحرف جر، أو المصادر الصريحة (أن+ الفعل)، وقد يستثني آخرون العبارات الاصطلاحية في المركبات، وهكذا.

وثمة العديد من النماذج والطرائق التي تقطع بها الكلمات ومنها نموذج التقطيع بإزاء التحليل الصرفي، ويمثله المُقطِّع الأكثر شيوعا وهو مقطع مدى الذي يطبق مرحلتين عند اختيار التحليل الصرفي الملائم من مخرجات محلل (باما)[١]. ولكنه لا يصلح للأغراض التي تستبعد فيها مورفيمات لا يخدم فصلها التوسيم النحوي كالألف والنون في التثنية. وثمة نموذج يفصل فيه الـمُقطِّع عن المحلل الصرفي، ويكشف فيه المقطع المورفيمات ويعين حدودها، غير أنه ما يزال في حاجة لمعلومات متصلة بهذه المورفيمات. ويتميز بقدرته على التعامل مع الكلمات سواء أكانت معروفة للمولد والمحلل الصرفي أم غير معروفة. ويمثله مقطع **أميرا** AMIRA[٢]، ونظام **أسما** ASMA[٣] المعتمدان على تقطيع الزوائد بدون استعمال لأدوات التحليل والتوليد الصرفيين. وهناك نموذج يعتمد على المحلل الصرفي، ويختلف عن نموذج التقطيع بإزاء التحليل الصرفي في أن مخرجاته لا تتضمن أي خصائص صرفية، وتتضمن فقط حدود الكلمة والمورفيمات، ويمثله ما استعمل في المحلل التركيبي الوظيفي والمعجمي للغة العربية التابع لمشروع ParGram[٤]. واقترح

---

(١) Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop, 2005, pp. 573– 580

(٢) Diab, M. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In 2nd International Conference on Arabic Language Resources and Tools, 2009

(٣) Abdul-Mageed, M., Diab, M. , Kubler, S. ASMA: A System for Automatic Segmentation and Morpho-Syntactic Disambiguation of Modern Standard Arabic. Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2013, p. 1-8

(٤) Attia, M.A. Arabic Tokenization System. Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, 2007, p. 65-72.

كوليك ما يمكن أن يستغل من أصناف الكلام المغلقة closed-class في عملية التقطيع[١].

الجدير بالذكر أنه ليس ثمة طريقة معينة ومثالية للتقطيع على مستوى العربية وغيـرها. فما يصلح لمنهج توسيمي معين، قد لا يصلح لمنـهج توسيمي آخر. وما يناسب تطبيقات الترجمة الآلية مثلاً قد لا يناسب تطبيقات الاسترجاع الآلية، وهكذا[٢].

٢-٥ قائمة الوسوم النحوية:

إن إنتاج مدونة موسمة نحوياً عمل قيم يوفر الجهد والوقت بإمكانية إعادة الاستعمال لبيانات المدونة ومشاركتها مع الباحثين الآخرين، فتظل مورداً على مر الزمن. ووضع معايير لعلميات التوسيم وتطبيقها يؤكد أن المدونة الموسمة ستستعمل بإمكانيات أعظم وأوسع. وتظهر أهمية وجود المعايير في المدونات الموسمة على ثلاثة مستويات، هي:

١. المستوى العملي ويتضمن الأسس العملية التي يلتزم بها عند تقديم أي مشروع أو مقترح توسيمي. وقد تحدثت عنها في المبحث الأول من الفصل الأول باعتبارها معايير عامة في تطبيق عمليات التوسيم على المدونات.

٢. شكل التوسيم، وقد عرف على أساس عالمي من خلال **مبادرة ترميز النصوص** Text Encoding Initiative (TEI) التي قدمت تعليمات لاستعمال SGML في توسيم النصوص المقروءة آلياً، وقد أشرت إليها أيضا في الفصل السابق عند حديثي عن التوسيم بالمعلومات النصية[٣].

٣. محتوى التوسيم، وهو ما سأركز عليه في هذا المبحث، ويرتبط بالتعليمات

---

(١) Kulick, S. Exploiting separation of closed-class categories for Arabic tokenization and part-of-speech tagging. ACM Transactions on Asian Language Information Processing (TALIP), Vol. 10, No. 1, 2011

(٢) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٣٧

(٣) Corpus Annotation, 2013, p.231

العامة للتوسيم النحوي خصوصاً التي انفردت بها **المجموعة الاستشارية الخبيرة لمعايير هندسة اللغة** EAGLES، وعُدت كمعيار عام لمحتوى أنظمة التوسيم النحوي[١].

لقد جاءت مبادرة EAGLES بعد طرح المشكلات المرتبطة بتطبيق معايير ثابتة على محتوى التوسيم، ومنها الخلافات اللغوية، حيث إن فرض نظام توسيمي يضر بالبحوث المستقبلية، ففي العربية مثلاً، ليس ثمة حد بين ما إذا كانت «ما» (مصدرية) أو (ظرفية) كما أن الأسس والمعايير لا تصاغ إلا بالاعتماد على أعمال سابقة، والأعمال السابقة حتماً مختلفة، كل يحسب غرضه على مستوى اللغة الواحدة، فكيف على مستوى لغات متنوعة؛ لذا من الصعوبة إلزام الموسمين أو الباحثين بأساس ومعيار موحد[٢].

إن هدف EAGLES هو التوفيق بين المذاهب النظرية المختلفة في اللغة بتوحيد معايير محتوى التوسيم، فقدمت وثيقتين إحداهما في التوسيم النحوي، والأخرى في التوسيم التركيبي، وتأخذ هاتان الوثيقتان شكل التعليمات والتوجيهات، وتترك مجالاً واسعاً لاختلاف الباحثين، بل واختلاف اللغات، إذ نجدها طبقت بنجاح على لغات خارج الاتحاد الأوروبي كالعربية والهندية[٣].

لقد جاءت تعليمات بادرة EAGLES للتوسيم النحوي على ثلاثة مستويات: إجبارية، مستحسنة، اختبارية، وفيما يلي توضيحها[٤]:

١. خصائص أو قيم إجبارية، وهي التي ينبغي أن تضمن في أقسام الكلام، أي الأقسام الأساسية للكلمة، كالفعل، والاسم، ... الخ.

٢. خصائص أو قيم مستحسنة، وهي المقولات النحوية في التوصيفات النحوية

---

(١) Corpus Annotation, 2013, p.235

(٢) Ibid.

(٣) Corpus Linguistics (An Introduction), 2011, p. 38

(٤) Corpus Annotation, 2013, p. 235

التقليدية، كالعدد والجنس والزمن... الخ، بالإضافة إلى فروع الأقسام الأساسية مثل: (عام، علم) للأسماء. وتزيد تلك الخصائص والفروع وتنقص من لغة لأخرى.

٣. امتدادات اختيارية لقائمة من الخصائص والقيم، وهي نوعان:

أ. خصائص أو قيم عامة تضاف لأغراض خاصة، مثل إضافة وسوم دلالية (اسم مكان، اسم زمان) للأسماء.

ب. خصائص أو قيم خاصة بلغات معينة دون غيرها، مثل **الصيغة التعظيمية** honorifics في اللغة الكورية[١].

إن محتوى التوسيم النحوي لا يقف عند هذا الحد، فثمة أمور أخرى ينبغي مراعاتها في مجموعة الوسوم نفسها كمسمياتها وحجمها ومكوناتها.

ـ ما يتعلق بالوسوم ومسمياتها:

الوسوم ببساطة هي قائمة أو مجموعة من الرموز المستعملة في مهمة محددة للتوسيم النحوي، وتمثل **أقسام الكلمة** POS وهي متفاوتة من حيث فائدتها، وقابلية تطبيقها لغوياً؛ إذ إن الحاجة لبعض الالتزامات العملية كالحاجة للدقة والسرعة تظهر فجوة بين ما هو مرغوب لغويا وما هو ممكن عملياً (حاسوبياً)، والتقارب بينهما معتمد على ظروف المشاريع[٢]. وعموماً، يأخذ الموسمون في الحسبان سياق الكلمة، ولكن

---

(١) وهي صيغة تحدد فاعل الفعل بالمقام وليس بالضمائر كما في اللغة العربية. ففي الكورية توجد ٦ مستويات للكلام ولكل مستوى لواحق تلحق قبل نهايات الأفعال، فجملة (هل أكلت؟)، تكتب في أربعة مستويات: 먹냐 -먹어 -먹어요 -먹습니까 (لاحظ أنها جميعا بمعنى واحد، لكن الاختلاف في اختلاف الفاعل في المقام. فالصيغة التعظيمية هي أعلى مستوى من تلك المستويات، ولها لاحقتان هما: 습니다/ㅂ니다، انظر:

Byon, A. S., Teaching Korean Honorifics. the Fifth National Conference on Korean Language Education (The Korean Language in America), Penn.: USA, 2000, pp. 275-289

(٢) Corpus Linguistics, 2013, p.24

إذا صعب التمييز النحوي حتى مع مراعاة السياق، فسيكون من الممكن تركه أو إهماله[١].

ومن المفيد هنا أن نميز بين **الوسم** tag **ومسمى الوسم** label، فالوسم في التوسيم النحوي هو نوع الكلمة الذي يسند إليها يدوياً أو آلياً. وتوجد طرق متعددة لتشكيل هذه الوسوم من خلال لوحة المفاتيح الألفبائية والرقمية، فحروف الجر ممكن أن يكون وسمها: IN أو PREP، وفي تفصيل أدق: NPS، عبارة عن ثلاثة أنواع من المعلومات: الاسم وهو النوع الأساسي ورمزه N، والعلمية مقابل العموم، ورمزها P، والإفراد مقابل الجمع، ورمزه S. وهذا ينعكس على مسمى الوسم الذي سيكون (اسم علم مفرد)[٢].

وعند اختيار مسمى الوسوم ثمة ثلاثة أمور تؤخذ بعين الاعتبار[٣]. وهي:

١. الإيجاز، فالمسميات الموجزة أكثر سهولة في الاستعمال من تلك المطولة.

٢. الوضوح، حيث ما يمكن أن يُفسر بسهولة يكون مألوفاً من غيره، ويتذكر بسهولة، ولذلك PREP أفضل من IN لتسمية وسم حروف الجر.

٣. إمكانية التحليل أو التفكيك، فما يمكن أن يفكك لأجزائه المنطقية هو أفضل من غيره وخصوصاً في المعالجة الآلية، فمثلاً NPS تفكك لما يلي:

N= اسم في مقابل (فعل = V، ضمير = P ...إلخ)

P= علم في مقابل (عام = C).

S= مفرد في مقابل (جمع = P).

ففي عمليات البحث يمثل* N كل الأسماء، و S*N كل المفرد منها، و* NP كل الأعلام من الأسماء مفردها ومجموعها، وهكذا.

---

(١) Corpus Linguistics, 2013. P.24

(٢) Ibid., p.35

(٣) Ibid., p.25

وإذا ما تحققت هذه الاعتبارات كان بالإمكان تحويل مسميات الوسوم آلياً إلى مسميات أخرى، وهذا مفيد جداً في تطوير نظام التوسيم[١]. والجدير بالذكر، أن المدونة يمكن أن توسم بصورة أفقية فتـتطلب الأمور السابقة في مسميات وسومها، أو بطريقة عمودية تأخذ كل كلمة فيها سطراً مستـقلاً، فتـتطلب وسوماً مطولة لا موجزة لسهولة استعمالها[٢].

ـ ما يتعلق بحجم ومكونات مجموعة الوسوم:

إن حجم مجموعة الوسوم هو أقل أهمية، حيث يمكن زيادة أو تقليص عدد الوسوم حسب الهدف الذي يرمي إليه مشروع التوسيم[٣]. وتُعد الأقسام الأساسية للكلام مع الأقسام الفرعية منها صلب مجموعة الوسوم النحوية، ولكن ثمة عناصر هامشية، توسم بإسناد وسوم لها ذات دلالة ومغزى. ففي المدونات المكتوبة، توجد كلمات ليست كلمات في الحقيقة، ولكنها تعالج بوصفها كلمات لأغراض التوسيم النحوي، مثل الكلمات الأجنبية والرموز، وعناوين الإنترنت وغيرها كثير. ومن جهة أخرى في المدونات المنطوقة من المفيد التمييز بين علامات محددة في الكلام، كالتمييز بين علامات التردد وكلمات التعجب بدلاً من تجميعها تحت وسم واحد.

وتؤثر طبيعة اللغة نفسها في حجم الوسوم النحوية، ففي الإنجليزية نجدها ما بين ٣٠ إلى ٢٠٠ وسم، مختلفة حسب الغرض[٤]، وفي الإسبانية تصل حوالي ٥٠٠ وسم[٥]، أما

---

(١) Corpus Linguistics, 2013. pp. 25-26

(٢) Ibid., pp. 26-27

(٣) Ibid.

(٤) Sketch Engine. English part-of-speech tagsets. 2-1-2019: https://www.sketchengine.eu/tagsets/english-part-of-speech-tagset/

(٥) Escartín C. P., Alonso H. M. Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task. In Procesamiento de Lenguaje Natural, Vol. 54, Jaen: Spain, 2015, p. 29-36

في العربية فهي ما بين ٦ إلى ٢٢٠٠ وسم[١]، فكلما كانت اللغة أغنى في صرفها وتعريفها زاد حجم وسومها[٢]، وإن كانت المجموعات الكبيرة هي أكثر اكتمالًا وفائدة للمستويات العليا[٣].

ومما ينبغي التأكيد عليه أن مجموعة الوسوم النحوية لا بد أن يوجد بينها علاقة قابلة للتمثيل كشجرة هيكلية من الملامح والخصائص تُورَّث الخصائص فيها من مستوى لآخر، وهذا ما يعرف بمجموعة الوسوم **المنطقية** Logical[٤]. فلو طبقنا ذلك على التقسيم التقليدي للكلام (اسم ـ فعل ـ حرف)، فسيكون المستوى العام، اسم =N، فعل =V، حرف =P، وسيكون تحت كل قسم خصائص، فالاسم (علم= P وعام=C) ثم العام (معرفة – d، ونكرة–I)، وهكذا يكون مسمى الوسم سلسلة من الأحرف يمثل كل واحد منها فرعا مختلفا من الهيكل الشجري.

وقد استُعملت مجموعات مختلفة من قوائم التوسيم النحوية في الدراسات العربية التي سعت لبناء أنظمة آلية للتوسيم النحوي، وهي كالتالي:

١- القائمة الثلاثية: اسم وفعل وحرف. وهو التقسيم التقليدي للكلام العربي

---

(١) للكلمات العربية عدد كبير من الأنماط التصريفية التي تلتقي مع أقسام الكلام الرئيسة فتكسبها معاني تضيف للقسم الكلامي تفصيلات أخرى كالعدد والنوع والتعريف والحالة الإعرابية والتوكيد والنسب والإعراب. وإذا ما حُسبت احتمالات حدوثها نظريا مع أقسام الكلام فإن أقسام الكلام الموسعة بها قد تصل إلى ٣٣٣ ألف وسم. وقد استعمل منها فعليا في البنك الشجري العربي ٢٢٠٠ وسم في الـ ٢٨٠ ألف كلمة الأولى.

انظر:

Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop, 2005, pp. 573– 580

(٢) Corpus Linguistics, 2013, p. 29

(٣) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص١٤١

(٤) Corpus Linguistics, 2013, p. 29

مع إضافة بعض الخصائص للأسماء والأفعال كالعدد والنوع والشخص. واستعمل هذا التقسيم في دراسة أبو ليل وإيفانز[١]، وكنعان وآخرون[٢]، والقريني وآخرون[٣]، والطعاني والروب[٤]، وهادني وآخرون[٥]. وهو تقسيم عام وغير دقيق، ولا يكفي في أبحاث معالجة اللغة، وإن كان يمكن التنبؤ به بدقة لقلة وسومه.

٢- مجموعات وسوم مقترحة:

أ ـ مجموعة وسوم شيرين خوجة التي تحتوي على ١٧٧ وسما، منها ١٠٣ تخص الأسماء و٥٧ تخص الأفعال و٩ للأدوات و٧ للبواقي ووسم واحد لعلامات الترقيم[٦]. ويوضح الجدولان (٢ -٢- أ) و(٢ - ٢-ب) الخصائص الاسمية والفعلية التي ضمنتها مجموعة الوسوم. ويلاحظ أنها تضيف للأسماء خاصية لا تقبلها إلا الأفعال وهي سمة الشخص، ذلك أن منطلق الدراسة حاسوبي وليس لغويا فوقوع الأخطاء اللغوية متوقع، كما أن نتائج تطبيقها تشير إلى أخطاء وخلط بين الأسماء والصفات وفي الأدوات[٧].

---

(١) Discovering lexical information by tagging Arabic newspaper text, 1998, pp. 1-7

(٢) Full automatic Arabic text tagging system, 2003

(٣) Pattern-based algorithm for Part-of-Speech tagging Arabic text, 2008, pp. 119-124

(٤) Al-Taani, A., Al-Rub, S. A. A Rule-Based Approach for Tagging Non-Vocalized Arabic Words. International Arab Journal of Information Technology (IAJIT), Vol. 6, No. 3, 2009

(٥) Hadni M., Ouatik SA., Lachkar A, Meknassi M. Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. International Journal on Natural Language Computing (IJNLC), Vol. 2, No. 6, 2013, p. 1–15

(٦) APT: Arabic part-of-speech tagger, 2001, pp. 20-25

(٧) Aliwy A. Arabic Morphosyntactic Raw Text Part of Speech Tagging System. Thesis, University of Warsaw: Poland, 2013, p. 103

جدول (٢ -٢- أ) الخصائص الاسمية لمجموعة وسوم خوجة

خصائص الأسماء			
الجنس	مذكر	مؤنث	محايد
العدد	مفرد	مثنى	جمع
الشخص	المتكلم	المخاطب	الغائب
الحالة النحوية	مرفوع	منصوب	مجرور
التعريف	معرفة	نكرة	

جدول (٢ - ٢-ب) الخصائص الفعلية لمجموعة وسوم خوجة

خصائص الأفعال			
الجنس	مذكر	مؤنث	محايد
العدد	مفرد	مثنى	جمع
الشخص	المتكلم	المخاطب	الغائب
الحالة الإعرابية	رفع	نصب	جزم

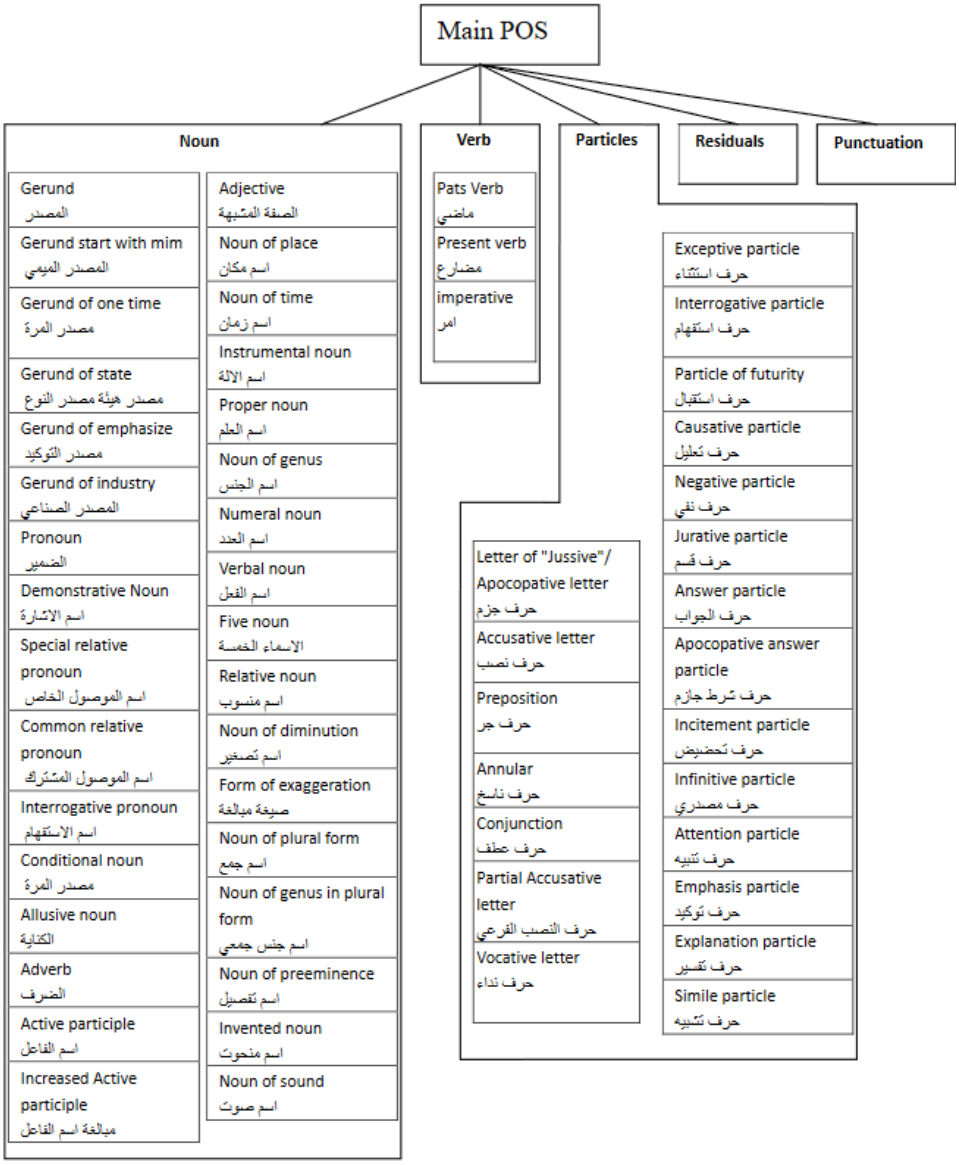
ب ـ مجموعة وسوم صوالحه التي يتألف كل وسم فيها من ٢٢ حرفا[١]. يمثل الحرف الأول القسم الكلامي الأساسي: اسم ـ فعل ـ أداة ـ علامة ترقيم ـ بواقي. وتمثل الثلاثة الأحرف التالية الأقسام الفرعية. وتخصص المجموعة ٣٤ قسما فرعيا للأسماء من حرفين، و٣ أقسام فرعية للفعل من ٣ أحرف، و٢١ قسما فرعيا للأدوات من ٤ أحرف. أما البواقي وعلامات الترقيم فتمثل بخمسة أحرف وستة أحرف على التوالي.

(١) Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING, 2011, p. 122

أما الأحرف التالية فتمثل الخصائص التصريفية التقليدية وهي: النوع (٧)، والعدد (٨)، والشخص(٩)، والصرف(١٠)، والحالة الإعرابية(١١)، وعلامات الحالة الإعرابية(١٢)، والتعريف(١٣)، والبناء(١٤)، والتأكيد(١٥)، واللزوم والتعدي(١٦)، والعاقل وغير العاقل(١٧)، والتصريف(١٨). وأخيرا، توجد ٤ أحرف تمثل المعلومات الصرفية المفيدة في تحليل النصوص العربية وهي التجرد والزيادة (١٩)، وعدد حروف الجذر(٢٠)، جذر الفعل(٢١)، ونهايات الأسماء(٢٢). وتشير الشرطة (-) إلى عدم وجود الخاصية، فتوسم كلمة (إنسان) بالوسم: nq----ms-pafd---htbt-s، حيث تشير n للقسم الأساسي الاسم، وq لنوعه العام، والشرطات الأربع لأقسام لا تمثلها الكلمة، وm للنوع المذكر وs للعدد المفرد ثم شرطة لخاصية الشخص التي لا تتعلق بالأسماء، أما p فتعني أن الاسم لا يتصرف وa حالته الإعرابية المنصوبة وf تعني علامة الفتحة الإعرابية وd تعني أنه معرف، ثم ثلاث شرطات لثلاث خصائص لا تتعلق بالأسماء هي: البناء للمجهول والتأكيد واللزوم والتعدي، وتشير h إلى أن الكلمة تخص العاقل وt إلى أنها اسم ذات وb إلى أنها مزيدة بحرفين وt إلى أنها ذات جذر ثلاثي، أما الشرطة فتقع فيها عدد جذور الفعل ولا ينتمي لها الاسم، وآخر الوسم شرطة لخاصية جذر الفعل ثم خاصية نهايات الأسماء وs تعني أن الاسم صحيح الآخر (انظر الشكل ٢-٣).

ورغم أن هذه المجموعة لا تتجاهل الفروق الاسمية والفعلية إلا أنها تضيف وسوما لا تفيد على هذا المستوى كالوسمين المتعلقين بالجذور. كما أن هناك تداخلا في قسم الأدوات، حيث تخصص لحروف الجر وسما خاصا وتضع لأداة التشبيه وسما آخر، فيصدق على حرف (كـ) وسمان في آن واحد، ولذلك هي غير صالحة للتطبيق عمليا.

شكل (٢ – ٣) مجموعة وسوم صوالحة للحروف الأربعة الأولى فقط (منقول من صوالحه[١])

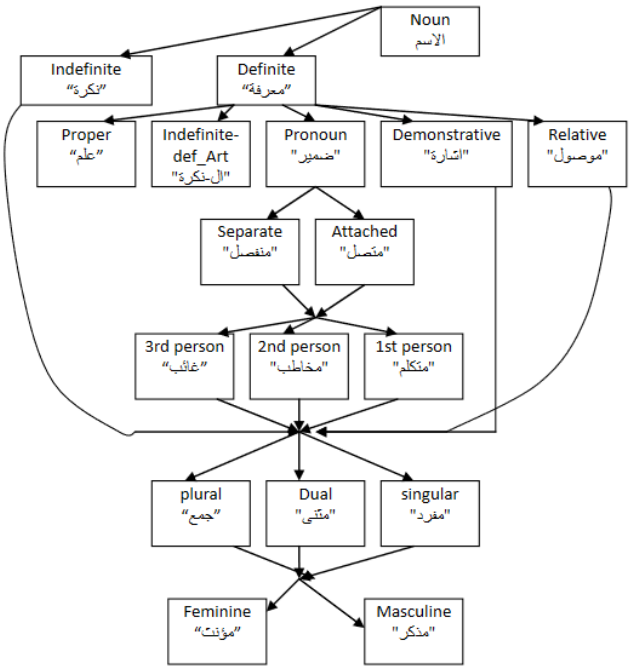


(١) نقلت مجموعة وسوم صوالحه للحروف الأربعة الأولى في صورة مخطط كما في دراسته لتمثيله تقسيم الكلام بصورة أفضل من نقله لجدول، انظر:

Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING, 2011, p. 122

ج ـ مجموعة وسوم الحاج[١]: وهي مجموعة وسوم مخصصة للنصوص العربية القديمة وخصوصا عربية القرآن الكريم. ويبلغ عددها ١٣ وسما، منها ٣ أقسام فرعية للأفعال، و٦ أقسام فرعية للأسماء، و٤ أقسام فرعية للأدوات (انظر سلسلة الأشكال ٢-٤)[٢]. ويلاحظ أن الصفات لم يشر إليها في التقسيم، والأفعال لم تضمن أي خاصية تصريفية. كما أن الأدوات قسمت وظيفيا، وضمنت كل ما ليس اسما أو فعلا كبعض المورفيمات التي تؤدي وظائف تصريفية وليست تقسيمية ولا يمكن أن تستقل بوصفها كلمة، كالجمع.

شكل (٢-٤-أ) الأسماء وتقسيماتها الفرعية في مجموعة وسوم الحاج (منقول من الحاج[٣])



(١) Statistical Part-of-Speech Tagger for Traditional Arabic Texts, 2009, p. 794-800

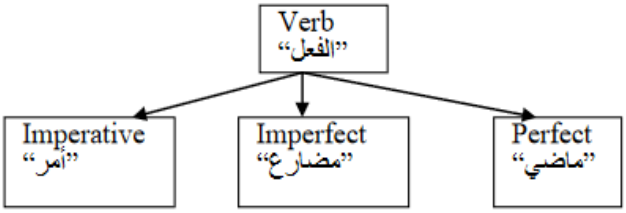
(٢) نقلت مجموعة وسوم الحاج للأسماء في صورة مخطط كما في دراسته لتمثيله تقسيم الأسماء بصورة أفضل من نقله لجدول، انظر:

Ibid.

(٣) Statistical Part-of-Speech Tagger for Traditional Arabic Texts, 2009, p. 794-800

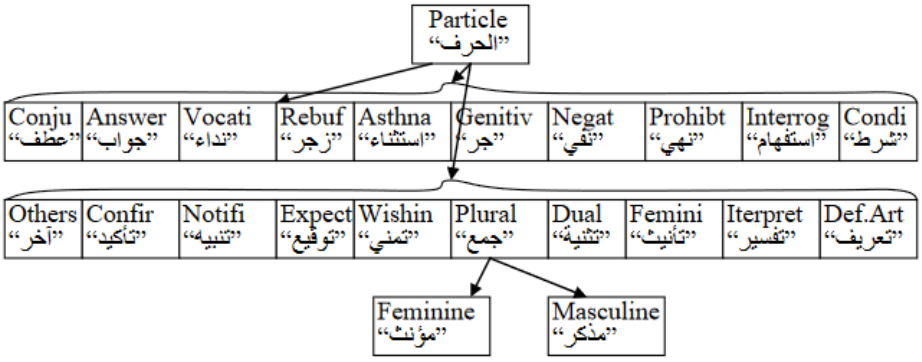
شكل (٢-٤-ب) الأفعال وتقسيماتها الفرعية في مجموعة وسوم الحاج

(منقول من الحاج[١])



شكل (٢-٤-ج) الحروف وتقسيماتها الفرعية في مجموعة وسوم الحاج

(منقول من الحاج[٢])



٣- مجموعة وسوم باكولتر المطورة من قبل تيم باكولتر[٣]. وهي قائمة ضخمة من الوسوم الشكلية التي لا تفرق بين الأقسام والخصائص. وهي ٥٠٠ وسم للنص المقطع و٢٢ ألف وسم للنص غير المقطع. ورغم أنها قد تخدم الباحثين في المستويات اللغوية العليا إلا أنها لا تفرق بين الضمائر والزوائد، وتتعامل مع ياء النسب باعتبارها ضميرا، ويعاب عليها بشكل عام اعتمادها على وسوم اللغة الإنجليزية بما لا يلائم العربية، فتخصص لما لا يعد من أقسام الكلام العربي وسوما خاصة، كتخصيصها

---

(١) Statistical Part-of-Speech Tagger for Traditional Arabic Texts, 2009, p. 794-800

(٢) Ibid.

(٣) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٤٤

للأسماء المكممة[١] quantifier nouns وسما. وقد سعى المهتمون إلى تقليص حجمها في قوائم يسهل التحكم بها، وتعتبر مجموعات الوسوم التالية أحد الأشكال المقلصة لها:

أ. مجموعة وسوم **بيز** Bies: وهي مجموعة طورتها آن بيز ودان بايكل كمجموعة فرعية من مجموعة وسوم باكولتر، واستلهمت من بنك بنسلفانيا الشجري للغة الإنجليزية[٢]. وتحتوي على ٢٤ وسما كما يظهر في الجدول (٢-٣).

ورغم أنها تجريبية وتفتقر لفروقات كثيرة في اللغة العربية كالفروق بين الصفات وبين الأدوات، إلا أنها استعملت في دراسات عديدة، منها: منى دياب وآخرون[٣]، وحبش وأون[٤]، وكوليك وآخرون[٥].

ب. مجموعة وسوم كوليك[٦]: وهي مجموعة طورها سيث كوليك، وتتكون من ٤٣ وسما. وتعد امتدادا لمجموعة وسوم بيز، وكان توسعها في:

---

(١) كلمة أو عبارة تستعمل مع الأسماء للإشارة إلى الكمية، مثل: كثيرا من a lot of - القليل few - بعض some. انظر:

Longman. Longman Dictionary of Contemporary English. 16-1-2019: https://www.ldoceonline.com/dictionary/quantifier

(٢) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٤٤

(٣) Automatic tagging of Arabic text: From raw text to base phrase chunks, 2004, pp. 149-152

(٤) Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop, 2005, pp. 573– 580

(٥) Kulick, S. Gabbard, R., Marcus, M. Parsing the Arabic Treebank: Analysis and Improvements. Proceedings of the Treebanks and Linguistic Theories Conference, Prague: Czech Republic, 2006

(٦) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، صص ١٤٦ – ١٤٧

– إضافة وسم لأفعل التفضيل ADJ_COMP ووسم للأسماء المكممة _NOUN QUANT ووسم للأعداد الوصفية ADJ_NUM ووسم لمشتقات الأفعال DV [١].

– ميزت الإشاريات بـ DEM عن ال التعريف DT.

– أضيفت ٦ وسوم لعلامات الترقيم: النقطة. والفاصلة، والقوسان الدائريان ()، وعلامة الاقتباس « ونقطتا القول: .

– بإضافة وسم لأل التعريف أضيفت وسوم لما يمكن أن يعرف، وهي: - DT+JJ - DT+NN - DT+CD - DT+ ADJ_COMP.

ورغم محاولتها معالجة مشكلات وسوم بيز إلا أنها لا تخلو من الخلط في أدواتها، ويصدق عليها ما على مجموعة وسوم باكولتر في أنها لا تلائم العربية حيث تجعل مثلا للأسماء المكممة وسما، وهو لا يوجد في العربية ومنقول من الإنجليزية.

ج . **مجموعة وسوم إيرتز الموسعة والمقلصة**[٢] Extended ReducedTagSet (ERTS): وهي عبارة عن ٧٢ وسما من مجموعة وسوم باكولتر للنصوص المقطعة، وفيها من مجموعة وسوم بيز بالإضافة إلى ترميز خصائص صرفية إضافية على الأسماء فقط وهي العدد (Du للمثنى وS للجمع ولا شيء للمفرد)، والنوع (F للمؤنث وM للمذكر ولا شيء عند غياب النوع)، والتعريف D وغيابه ø [٣]. وهي مثل مجموعة وسوم باكولتر وسابقاتها الممتدة منها في افتقارها للخصائص الوظيفية في الأدوات حيث لا تميز بينها.

---

(١) في الإنجليزية هي صفات أو أسماء مشتقة من الأفعال، وفي العربية استعمل كوليك هذا الوسم لتوسيم المصادر العاملة عمل الفعل، كالمصدر (صعود) في جملة نحو: «... صعوده الجبل بمشقة». انظر:

Parsing the Arabic Treebank: Analysis and Improvements, 2006, p. 38

(٢) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص١٤٧

(٣) السابق، نفس الصفحة.

جدول (٢-٣) مجموعة وسوم بيز Bies

الأسماء		**الأدوات**	
NN	اسم عام مفرد	CC	حرف عطف
NNS	اسم عام جمع أو مثنى	DT	ضمير إشارة
NNP	اسم علم مفرد	RP	أداة
NNPS	اسم علم جمع أو مثنى	IN	حرف جر ـ أداة ربط ثانوية
PRP	ضمير شخصي	**الأفعال**	
$PRP	ضمير ملكية	VBP	فعل مضارع مبني للمعلوم
WP	ضمير موصول	VBN	فعل مضارع/ ماض مبني للمجهول
JJ	صفة	VBD	فعل ماض مبني للمعلوم
RB	حال أو ظرف	VB	فعل أمر
WRB	حال أو ظرف موصول	**أخرى**	
CD	عدد أصلي	UH	كلمة انفعالية
FW	كلمة أجنبية	PUNC	علامة ترقيم
		NUMERIC_COMMA	الحرف ر يستعمل فاصلة
		NO_FUNC	كلمة لم تحلل

د . **مجموعة وسوم كاتب** CATiB المعدة لمشروع بنك كولومبيا الشجري[١].

وتعد أقصى درجات التقليص لمجموعة وسوم باكولتر حيث لا تحوي سوى ٦ وسوم

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص١٤٧.

(الاسم NOM والعلم PROP والفعل VRB والفعل المبني للمجهول VRB-PASS والأدوات PRT وعلامات الترقيم PNX). من أجل تسريع عملية التحشية اليدوية لبنك كولومبيا. ويلاحظ أنها لا تفرق بين الأسماء والضمائر والصفات ولا بين الأدوات رغم حرصها على المحافظة على الفروق المهمة كما يشير واضعوها.

هـ . مجموعة وسوم **بنك براغ الشجري للعربية** Prague Arabic Dependency Treebank (PADT) [١]. وهي مجموعة الوسوم المستعملة في محلل ElixirFM وكل وسم فيها يتكون من جزئين هما: القسم الكلامي والخاصية. ويبين الجدولان (٢-٤-أ) و(٢-٤-ب) رموز وسوم هذه المجموعة ورموز الخصائص. وهذه المجموعة غير دقيقة ومتداخلة في الأدوات والصفات وتفتقر لفروقات مهمة في الكلام العربي.

جدول (٢-٤-أ) مجموعة وسوم PADT

الوسم	نوعه	الوسم	نوعه
VI	فعل ماض	SR	اسم موصول
VP	فعل مضارع	F	أداة
VC	فعل أمر	FI	أداة استفهام
N	اسم	FN	أداة نفي
A	صفة	C	حرف عطف
D	ظرف	P	حرف جر
Z	علم	I	حرف تعجب
Y	اختصار	G	رمز رسومي
S	ضمير	Q	عدد
SD	اسم إشارة	-	أداة تعريف معزولة

(١) Arabic Morphosyntactic Raw Text Part of Speech Tagging System, 2013, p.35

جدول (٢-٤-ب) خصائص وسوم PADT

الحالة	رفع	نصب	جزم
البناء	معلوم	مجهول	
الشخص	متكلم	مخاطب	غائب
النوع	مذكر	مؤنث	
العدد	مفرد	جمع	مثنى

٢-٦ مناهج التوسيم:

إن تحديد الوسم المناسب للكلمة من خلال النظام يعتمد على المنهج المتبع فيه. وهناك العديد من المناهج التي تُستعمل في أنظمة التوسيم النحوي الآلي متفرقة أو معا، أشهرها: المنهج **المعتمد على القواعد** rule-based، والمنهج المعتمد على **تعلم الآلة** machine-learning[١]. إن المنهج المعتمد على القواعد هو الأسبق ظهورا، ويتخذ من القواعد اللغوية المجموعة يدويا أساسا له في عملية التوسيم. وهي قواعد مستنبطة من القواميس أو المعاجم ترفق من خلالها كل الوسوم الممكنة للكلمات المراد توسيمها. وتستعمل هذه القواعد لتحديد الوسم الصحيح للكلمة حين تتعدد وسومها المحتملة. وتسمى هذه العملية بعملية **فك الغموض** disambiguation. وتكون عن طريق تحليل الخصائص اللغوية للكلمة والكلمة السابقة وكذلك اللاحقة لها فضلا عن جوانب أخرى. فإذا كانت الكلمة السابقة مثلا (إلى)، ستكون الكلمة التالية بلا شك اسما ما لم تكن الكلمة التالية إحدى الكلمات الموجودة في قائمة الأدوات والضمائر نحو: أن ـ من ـ التي، وغيرها. ومن ثمّ ترمز هذه المعلومة كأحد

(١) Btoush MH., Alarabeyyat A., Olab I. Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, UK, 2016, p. 331

القواعد في النظام. وهذا يعني أن عملية التوسيم تتم في النظام على مرحلتين هما[١]:

١- استعمال معجم أو قاموس لإسناد قائمة من أقسام الكلام الممكنة لكل كلمة. فجملة مثل: قاتل في صف العدو، ستكون في هذه المرحلة موسمة مثلا كما يلي:

قاتل	في	صف	العدو
فعل ماض VP فعل أمر VC	حرف جر PP حرف جر متصل بياء المتكلم PP1 اسم نكرة NN فعل أمر VC	فعل ماض VP فعل أمر VC اسم نكرة NN	اسم معرف بأل NND صفة مشبهة معرفة بأل ADJD مثال توضيحي

٢- استعمال قائمة ضخمة من قواعد فك الغموض المكتوبة؛ لترشيح الوسم الملائم للكلمة في سياقها. ومن ثم يظهر المثال السابق بهذه الوسوم:

قاتل	في	صف	العدو
فعل ماض VP	حرف جر PP	اسم نكرة NN	اسم معرف بأل NND مثال توضيحي

ومن الموسمات الآلية التي استعملت هذا المنهج: موسم الصحف العربية[٢]، وموسم كنعان[٣]، والموسم المعتمد على الأوزان الصرفية[٤]، وموسم النصوص غير المشكَّلة[٥].

وفي منهج تعلم الآلة يتعلم نظام التوسيم النحوي كيفية الاستجابة لأحداث معينة

---

(١) Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition, 2016, p.331

(٢) Discovering lexical information by tagging Arabic newspaper text, 1998, pp. 1-7

(٣) Full automatic Arabic text tagging system. The proceedings of the International Conference on Information Technology and Natural Sciences, Jordan, 2003

(٤) Pattern-based algorithm for Part-of-Speech tagging Arabic text, 2008, pp. 119-124

(٥) A Rule-Based Approach for Tagging Non-Vocalized Arabic Words, 2009

بشكل ذاتي دون أن يتم برمجته عليها، إذا لا يغذى بأي معرفة، ويستعاض عن ذلك بإتاحة البيانات له كي يستعملها، ومن ثم يقوم بصياغة منطقه الخاص بنفسه. فبإمكاننا أن نجعل النظام يستخلص السمات المميزة لمئات الآلاف من الكلمات لتصنيفها، وذلك من خلال تدريبه على بيانات معروفة مسبقا ثم تعميم مخرجاتها واستعمالها مع كل المسائل التي لم يتدرب عليها سابقا.

وثمة عدد كبير من خوارزميات تعلم الآلة التي تهدف إلى إيجاد **نماذج** models عبارة عن معادلات رياضية تختزل البيانات وتمثلها بطريقة تسمح **بالتعميم** generalization عند استعمال بيانات جديدة. ويتنوع منهج تعلم الآلة بين تعلم آلة **تحت الإشراف** supervised وتعلم آلة **دون إشراف** unsupervised ويختلفان كثيرا. فتعلم الآلة تحت الإشراف تستعمل فيه مدونة موسمة مسبقا بمجموعة وسوم محددة، وينخفض أداؤه بشكل كبير إذا كانت بيانات الاختبار من مجال مختلف، لكنه يتميز بالدقة العالية. وتندرج تحته خوارزميات **التصنيف** Classification التي تتعلم الآلة بها تصنيف أي بيانات. فرغم أنها نفس الخوارزمية مع أي نوع بيانات إلا أنها تعالج البيانات اللغوية بمنطق تصنيفي مختلف لأنها غذيت ببيانات مختلفة. ومن الموسمات الآلية التي استعملت منهج تعلم الآلة تحت الإشراف: موسم أميرا[١]، والموسم المعتمد على الخصائص الصرفية الوظيفية[٢]، وموسم فان دن بوش[٣]، وموسم مدى[٤]، وموسم

---

(١) pp. 149-152 Automatic tagging of Arabic text: From raw text to base phrase chunks, 2004

(٢) Hajic, J., Smrz, O., Buckwalter, T., & Jin, H. Feature-based tagger of approximations of functional Arabic morphology. In Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, Spain, 2005

(٣) Memory-based morphological analysis and part-of-speech tagging of Arabic. In Arabic Computational Morphology, 2007, pp. 201-217

(٤) Roth, R., Rambow, O., Habash, N., Diab, M., Rudin, C. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, 2008, pp. 117-120

الحاج[١]، والموسم النحوي المحسن من خلال التحليل الصرفي[٢]، وموسم ستانفورد[٣]، وموسم **مداميرا** Madamira [٤]، والموسم النحوي بخاصيتي العدد والنوع[٥].

أما تعلم الآلة دون إشراف فلا تستعمل فيه مدونة موسمة مسبقا وتستقرى مجموعة الوسوم بواسطته، ولديه القدرة على العمل مع أي بيانات، إذ يجمعها ويصنفها إلى مجموعات حسب تشابهها، ثم يقوم بتصنيف بيانات مدونة الاختبار بناء على مدى تشابهها ببعدها عن تلك المجموعات، إلا أنه أقل أداء. وتندرج تحته خوارزميات **التجميع** clus-tering التي تصنف البيانات وتختار مراكز التجميع فيها بطريقة عشوائية وتعاد عملية التجميع فيها أكثر من مرة[٦]. وقد استعمل في العربية مربوطا بتعلم الآلة تحت الإشراف، ومن الموسمات النحوية الآلية التي قامت بذلك موسم العربية المحكية[٧].

---

(١) Statistical Part-of-Speech Tagger for Traditional Arabic Texts, 2009, p. 794-800

(٢) Albared, M., Omar, N., Ab Aziz, M. J. Improving Arabic part-of-speech tagging through morphological analysis. In Intelligent Information and Database Systems, .Springer Berlin Heidelberg, 2011, pp. 317-326

(٣) :The Stanford Natural Language Processing Group. Stanford CoreNLP. 10-12-2017 http://nlp.stanford.edu/software/corenlp.shtml

(٤) Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Roth, R. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014

(٥) Darwish, K., Abdelali, A., Mubarak, H. Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging. In International Conference on Language Resources and Evaluation. 2014

(٦) القنيعير، فارس. تعلم الآلة: مقدمة سريعة. مقال في شبكة الإنترنت (موقع نمذجيات)، ١-١٠-٢٠١٧: https://www.nmthgiat.com/تعلم-الآلة-مقدمة-سريعة/

(٧) Duh K., Kirchhoff K. POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, University of Michigan, Michigan, 2005, p.55

ويعيب المنهج المعتمد على القواعد أنه يستغرق الجهد والوقت والمال؛ إذ يلزمه حزمة ضخمة من القواعد التي لا بد أن يضعها متخصصون لغويون وغالبا ما تتعقد وتتعارض كلما زادت. أما منهج تعلم الآلة فلا يتطلب سوى عينة ممثلة ومتوازنة ذات تنوع وحجم مناسبين للتعلم منها، وإن كان الأول ـ بخلاف الثاني ـ هامش الخطأ فيه أقل إذا ما كتبت قواعده بعناية، ولكن الثاني يستطيع التعلم من أخطائه واكتشاف خصائص لغوية قد لا يعرفها المتخصصون اللغويون أصلا[١].

٢-٧ قياس الأداء:

تتفاوت أساليب قياس أداء نظام التوسيم النحوي الآلي حسب المنهج المتبع فيه، وأيا ما كان المنهج فإن قياس الأداء يتطلب وجود نسختين من المدونة: المدونة في صورتها الخام، والمدونة بعد توسيمها حيث تتم مطابقة نتائج الموسم النحوي المبني على القواعد على المدونة الموسمة يدويا، ويستخدم الجزء الذي لم تدرب عليه الخوارزمية من قبل (غير الموسوم يدويا) في المنهج المبني على تعلم الآلة، فيوسم هذا الجزء آليا (مدونة الاختبار) ثم تقارن النتائج مع نتائج التوسيم اليدوي لهذا الجزء (مدونة التدريب).

وتستعمل المصفوفة المعروفة **بمصفوفة الإرباك** confusion matrix أو **مصفوفة الخطأ** Error matrix في مجال تعلم الآلة، لوصف أداء **النموذج التصنيفي** أو **المصنفات** classifiers في مجموعة من بيانات الاختبار تكون القيم الصحيحة فيها معلومة. وهذه المصفوفة عبارة عن تخطيط جدولي معين يصور أداء خوارزمية التعلم لآلآلي تحت الإشراف (وتسمى في التعلم الآلي بدون إشراف مصفوفة المطابقة match-ing)، ويقدم معلومات حول **التصنيفات المتوقعة** Predicted class التي يتنبأ بها النظام وحالات **التصنيفات الفعلية** Actual class التي تم تصنيفها من قبل الإنسان. ويمثل كل صف من المصفوفة الأصناف المتوقعة بينما يمثل كل عمود الأصناف الفعلية

(١) Mahafdah R.; Omar N.; Al-Omari O. Arabic part of speech tagging using K-Nearest Neighbour and Naive Bayes classifiers combination. Journal of Computer Science, New York: USA, Vol. 10, No. 10, 2014, p. 1867

(والعكس كذلك). وسميت مصفوفة الإرباك لأنها تسهل رؤية ما إذا كان النظام مرتبكا بين الصنفين، أي: يخطئ في تحديد صنف على أنه الآخر[١]. وعادة يقيم أداء المصنف باستعمال البيانات الواردة في المصفوفة، ويكون حجمها حسب عدد الأصناف.

ولنفترض أننا عرضنا على خوارزمية التصنيف ١٦٥ كلمة (١٠٥ من صنف أ – ٦٠ من صنف ب) وتعرف بصورة صحيحة على ١٠٠ من صنف أ و٥٠ من صنف ب، ستكون المصفوفة بهذا المثال بالشكل التالي:

جدول (٢–٥) مصفوفة الإرباك

العدد الكلي=١٦٥	أ المتوقعة	ب المتوقعة	
صنف أ الفعلي ١٠٥	استجابة صحيحة =١٠٠ True Positive (TP)	رفض خاطئ=٥ False Negative (FN)	ما بين ١٠٥ من الكلمات تعرف النظام على ١٠٠ كلمة فقط
صنف ب الفعلي ٦٠	استجابة خاطئة=١٠ False Positive (FP)	رفض صحيح=٥٠ True Negative (TN)	ما بين ٦٠ كلمة تعرف النظام على ٦٠ كلمة فقط
	ما بين ١١٠ من الكلمات المصنفة ١٠٠ صحيحة و١٠ غير صحيحة	ما بين ٥٥ من الكلمات المصنفة ٥٠ صحيحة و٥ غير صحيحة	

ويبين الجدول (٢–٥) شكل المصفوفة، ومحتوى كل صف وعمود فيها. وتمثل أ وب الصنفين الموجودين، مثلت أ بـ Positive ومثلت ب بـ Negative. أما الاستجابة الصحيحة TP فتمثل عدد الحالات التي توقعها المصنف بشكل صحيح للصنف أ وتنتمي بشكل فعلي للصنف أ. وتمثل الاستجابة الخاطئة TN عدد الحالات التي توقعها المصنف

(١) Powers, David M W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, Australia, Vol. 2, No. 1, 2011, p. 37–63

بشكل صحيح للصنف ب وتنتمي بشكل فعلي للصنف أ. ويمثل الرفض الصحيح FP عدد الحالات التي توقعها المصنف بشكل غير صحيح للصنف ب وتنتمي بشكل فعلي للصنف أ. أما الرفض الخاطئ FN فيمثل عدد الحالات التي توقعها المصنف بشكل غير صحيح للصنف أ وتنتمي بشكل فعلي للصنف ب.

وفيما يلي قائمة بالقيم التي تحسب غالبا من مصفوفة الإرباك للمصنفات الثنائية والمتعددة، بالاعتماد على مصفوفة الإرباك السابقة:

١ – **الصحة** Accuracy: وهو أحد المقاييس المعتمدة والمجمع عليها عند قياس أداء الموسمات النحوية الآلية، ويسمى أيضا trueness. وعلى الرغم من أنه كاف لقياس أداء الأنظمة ذات المجموعات التوسيمية الصغيرة الحجم، إلا أنه غير كاف لقياس المجموعات التوسيمية الغنية بالخصائص التصريفية المرتبطة بالوسوم، وغير كاف أيضا حين تكون من بين نتائج التوسيم كلمات يصدق عليها أكثر من وسم[١].

ويحسب بجمع عدد النتائج التي توصل إليها النظام في كل وسم وصنفت على أنها **استجابات صحيحة** true positive مع عدد النتائج التي نجح النظام في استبعادها true positive، مقسومة على عدد مجموع الاستجابات التي صنفها النظام، أي: /TP+TN TP+TN+FP+FN، وتساوي: ١٠٠+٥٠/ ١٦٥ = ٠,٩١.

٢ – **الدقة** Precision: وهي عدد النتائج التي توصل إليها النظام وصنفت على أنها **استجابات صحيحة** true positives في كل وسم، مقسومة على عدد مجموع النتائج التي صنفها النظام فيه (**الاستجابات الصحيحة** true positives **والاستجابات الخاطئة** false positive). فلو كان تحت وسم حروف العطف ١٠٠ استجابة صحيحة، و١٠ استجابات خاطئة، فالدقة: TP/TP+FP، وتساوي: ١٠٠/ ١٠٠+١٠=٠,٩١.

٣– **الاسترجاع** Recall: هو عدد النتائج التي صنفت على أنها **استجابات صحيحة**

---

(١) Acedánsk, S., Przepiórkowski, A. Towards the Adequate Evaluation of Morphosyntactic Tagger. Proceedings of the 23rd International Conference on Computational Linguistics, Beijing: China, 2010, p. 1

true positives في كل وسم، مقسومة على عدد الاستجابات الصحيحة مع عدد الاستجابات الأخرى التي من المفترض وجودها تحت الوسم ولكن النظام أخطأ ولم يصنفها تحته (**مرفوضات خاطئة** false negative). فلو كان تحت وسم حروف العطف ١٠٠ استجابة صحيحة، ووجد أن هناك ٥ نتائج أخرى صحيحة من المفترض أن تكون تحت هذا الوسم وتجاوزها النظام، فالاسترجاع= TP/TP+FN، أي: ١٠٠/ ١٠٠+٥= ٩٥ ,٠.

٤- **مقياس-ف** f-measure: ويستعمل عند قياس الأداء الكلي، ويكون استعماله ملائما إذا كانت الأخطاء من نوع **الاستجابات الخاطئة** false positives أو **المرفوضات الخاطئة** false negative، ويستعمل كذلك عندما يقاس عدد الأخطاء بالنسبة لعدد **الاستجابات الصحيحة** true positives، وحين تكون **المرفوضات الصحيحة** true negatives غير مهمة. وبدلا من أخذ المتوسط الحسابي للدقة والاسترجاع نحسب من خلال هذا المقياس المتوسط التوافقي، ولا يمكن الحصول على درجته إذا كان أحدهما ذا درجة منخفضة جدا. وللحصول على درجة f عالية لا بد أن يكونا عاليين. وهو يساوي:

٣ * الدقة * الاسترجاع/ الدقة + الاسترجاع

٢-٧-١ الموسمات النحوية الآلية المطبقة على المدونات العربية:

يلاحظ في الموسمات النحوية الآلية المطبقة على المدونات العربية أنها تعطي نتائج متقاربة في الصحة (انظر الجدول ٢-٦) رغم اختلاف منهجية التوسيم المتبعة، واختلاف أنواعها، إلا أن كل موسم منها طبق على نفس نوع المدونات التي درب عليها.

وفيما يلي أنظر في ثلاثة موسمات نحوية آلية طبقت على مدونات عربية مختلفة لم تدرب عليها ومتوفرة في موقع **سكتش إنجن** Sketch Engine، وهي كالتالي:

١. موسم **ستانفورد** Stanford المطبق على مدونة arTenTen.

٢. موسم **مدى** MADA المطبق على **مدونة الذخيرة الفصحى** KSUCCA[١].

(١) King Saud University Corpus of Classical Arabic.

٣. موسم أميرا AMIRA المطبق على مدونة الويب العربية Arabic Web Corpus.

جدول (٢-٦) نتائج الصحة في الموسمات النحوية الآلية المطبقة على المدونات العربية

المنهج	الموسم	الصحة
المنهج المعتمد على القواعد	موسم الصحف العربية (١٩٩٨م)[١]	لم تحدد
	موسم كنعان (٢٠٠٣م)	٪٩٣
	الموسم المعتمد على الأوزان الصرفية (٢٠٠٨)	٪٩١
	موسم النصوص غير المشكَّلة (٢٠٠٩)	٪٩٤
منهج تعلم الآلة	موسم خوجة (٢٠٠١م)	٪٩٠
	موسم أميرا (٢٠٠٤م)	٪٩٥,٤٩
	الموسم المعتمد على الخصائص الصرفية الوظيفية (٢٠٠٥)	٪٩٧,٣٧-٩٥,٢٥
	موسم فان دن بوش (٢٠٠٧م)	٪٩١,٥
	موسم مدى (٢٠٠٨)	٪٩٦
	موسم الحاج (٢٠٠٩)	٪٩٦
	الموسم النحوي المحسن من خلال التحليل الصرفي (٢٠١١)	٪٩٦,٦
	موسم ستانفورد (٢٠١١)	٪٦٩,٥
	الموسم النحوي بخاصيتي العدد والنوع (٢٠١٤)	٪٩٨,١
	موسم مداميرا Madamira (٢٠١٤)	٪٩٥,٩

(١) Discovering lexical information by tagging Arabic newspaper text, 1998, pp. 1-7

ولبيان الحاجة إلى مجموعات توسيمية نحوية تنطلق من العربية ومصادرها نظرت في واقع المجموعات التوسيمية النحوية لكل موسم من الموسمات النحوية الثلاثة (ستانفورد ـ مدى ـ أميرا) خلال تلك المدونات، وفهمت أداءها اللغوي والتقني من خلال أول خمسين كلمة فقط من الكلمات الأكثر تكراراً في كل وسم لكل مجموعة، إذ من الصعب تتبع جميع نتائج التوسيم في هذه المدونات لكبر حجمها الذي يصل أحيانا إلى ١٠٠ ألف كلمة في الوسم الواحد، مع أن بعض الوسوم كوسوم حروف العطف محدودة النتائج ونتائجها أقل من ٥٠ نتيجة.

ويتضمن الأداء اللغوي النظر في صلاحية كل مجموعة وسوم للعربية ثم النظر في محتوى وسومها للتأكد من مطابقة مسمى الوسم لمحتواه، من خلال تطبيقها على المدونات. أما الأداء التقني، فسأنظر في تقييمه إلى جانب **الصحة** Accuracy للمخرجات (الخمسين كلمة الأولى الأكثر تكراراً في كل وسم)، لمعرفة الكلمات الموسمة توسيما صحيحاً وستتحقق من ذلك بالفحص اليدوي، وستنظر في سياقات الكلمات الغامضة، فإن وردت كلمة (في) تحت وسم الأسماء ومن ضمن سياقاتها في النتائج وقوعها حرفا أو وردت تحت الحروف ومن سياقاتها في النتائج ما هو اسم فستعد ذلك خطأ في كل.

أولا: الأداء اللغوي:

١ ـ موسم ستانفورد المطبق على مدونة arTenTen:

ستانفورد نظام موجه بالإشراف للتوسيم النحوي، يعتمد على نماذج تدريبية مختلفة للعديد من اللغات، ومنها العربية. وقد دُرب النموذج الحالي للعربية فيه باستعمال مدونة البنك الشجري العربي ATB القائمة في كل إصداراتها على نصوص صحفية[١].

---

(١) The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus, 2004, p. 102

وطبق موسم ستـانـفورد على جزء من مدونـة arTenTen الذي يتكون من ٧,٤٧٥,٦٢٤,٧٧٩ كلمة. وهي مدونة عربية ضخمة جمعت نصوصها من الإنترنت بعد حذف مكرورا‌تها وتنقيحها باستعمال العديد من الأدوات وتضم ٥٨ بليون كلمة للهجات ومستويات متنوعة من العربية[١].

وتضم مجموعة وسوم ستانفورد ٣٣ وسماً طبقت جميعها على المدونة باستثناء وسم واحد هو وسم الكلمات الأجنبية FW غير أنه قد خصص وسما دون اسم للروابط والأكواد ذات الأحرف اللاتينية. وقد اعتمدت مجموعة وسوم ستانفورد على مجموعة وسوم بيز BIES التجريبية المطورة من مجموعة وسوم باكولتر[٢]. وفي الجدول (٢-٧) قائمة بمجموعة ستانفورد المطبقة على مدونة arTenTen ومقابلها بالعربية والإنجليزية[٣].

وبالتأمل في هذه المجموعة من الوسوم سنجد ما يلي:

١. لا يوجد بين وسوم هذه المجموعة علاقة قابلة للتمثيل كشجرة هيكلية من الملامح والخصائص تورث الخصائص فيها من مستوى لآخر. فليست مجموعة منطقية Logical؛ إذ تخصص مثلا وسوما للأسماء المعرفة DTNN-DTNNS والأعلام المعرفة DTNNP-DTNNPS وتعتبرها وسوما أساسية، ولا تخصص لما تدخل عليه أداة التعريف من الأقسام الأخرى وسوما أخرى كالأعداد الترتيبية ADJ، وأسماء الفاعل والمفعول VN والمصادر VBG.

٢. وفيما تراعي المجموعة خاصية العدد كخاصية التعريف في وسوم الأسماء NNS-NNPS-DTNNPS-DTNNS، إلا أنها لم تكن منضبطة أيضا في ذلك، حيث لم تضفها لباقي الأقسام التي تقبلها، كالأعداد الترتيبية ADJ، وأسماء الفاعل والمفعول VN،

---

(١) sketchengine.co.uk/artenten-arabic-corpus

(٢) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٤٥

(٣) sketchengine.co.uk/stanford-arabic-parser-tagset/

والمصادر VBG، والظروف RB، والصفات JJ، وأفعل التفضيل JJR. فضلا عن أنها لا تعتد بالتثنية ولا بما جمع جمع تكسير.

٣. محاولة نقل الوسوم من الإنجليزية إلى العربية أدى إلى إدراج أقسام أساسية ليس لها وجود في العربية مثل: الموصولات الظرفية WRB والموصولات الضميرية WP وضمائر الملكية PRP$، والأسماء المكممة NOUN.

٤. محاولة نقل الوسوم من الإنجليزية إلى العربية أدى إلى إدراج أقسام فرعية في أقسام أساسية لا تندرج تحتها فليس مكان المصادر VBG أو أسماء الفاعل والمفعول VN قسم الأفعال، فالمصادر أسماء وأسماء الفاعل والمفعول وأفعل التفضيل صفات.

٥. تجاهلت الفروق بين أسماء الفاعل والمفعولين رغم اختلافها في الشكل والوظيفة وإن تطابقا شكلاً في قليل من الأسماء، نحو: مختار، مجتاز... الخ، وخصصت وسما واحد لهما (VN).

٦. تميز مجموعة الوسوم بين الفعل الماضي (VBD) والفعل المضارع (VBP)، ولكنها تخصص وسم (VBN) للفعلين الماضي والمضارع المبنيين للمجهول دون تمييز بينهما. فضلا عن أن البناء للمجهول مبنى تصريفي لا تقسيمي: أي خاصية تصريفية متعلقة بالفعل الماضي والفعل المضارع وتأتي بعد التقسيم الأساسي.

٧. بعض أسماء الوسوم لا يتلاءم مع محتواها، فوسم ADJ يفترض أن يضم الصفات، ولكنه هنا للأعداد الترتيبية.

٨. هناك خلط في وسوم الأدوات حيث تخصص المجموعة للأدوات وسما خاصا RP، ثم تضم عددا من الأدوات مرة أخرى لوسم حروف الجر IN دون معيار واضح.

جدول (٢-٧) قائمة بمجموعة ستانفورد المطبقة على مدونة arTenTen ومقابلها بالعربية والإنجليزية

القسم الرئيس	م	الوسم	معناه بالإنجليزية	معناه بالعربية
أسماء	١	NN	noun, singular or mass	اسم شائع مفرد نكره
	٢	NNS	noun, plural	اسم شائع، مثنى أو جمع نكره
	٣	NNP	Proper noun, singular	اسم علم، مفرد نكره
	٤	NNPS	Proper noun, plural	اسم علم جمع أو مثنى نكره
	٥	NOUN	noun	اسماء مكممه
	٦	DTNN	noun, singular/mass with the determiner al	اسم شائع مفرد معرف بأل
	٧	DTNNS	noun, plural with the determiner al	اسم شائع جمع أو مثنى معرف بأل
	٨	DTNNP	Proper noun, singular with the determiner al	اسم علم مفرد معرف بأل
	٩	DTNNPS	Proper noun, plural with the determiner al	اسم علم جمع ومثنى معرف بأل
أفعال	١٠	VBP	Verb, present tense	فعل مضارع
	١١	VBD	Verb, past tense	فعل ماض
	١٢	VBN	verb, past and present/passive	فعل ماض ـ مضارع/ للمجهول
	١٣	VB	verb, base form	فعل أمر
	١٤	VBG	verb, gerund	مصدر
	١٥	VN	verb, past participle	اسم فاعل/ مفعول

القسم الرئيس	م	الوسم	معناه بالإنجليزية	معناه بالعربية
أدوات	١٦	CC	Coordinating conjunction	حروف عطف
	١٧	DT	determiner	اسم إشارة
	١٨	RP	particle	أداة
	١٩	IN	Preposition or subordinating conjunction	حروف جر وأدوات
ضمائر	٢٠	PRP	Personal pronoun	ضمير شخصي
	٢١	PRP	Possessive pronoun	ضمير شخصي للملكية
	٢٢	WP	Wh-pronoun	ضمير موصول
صفات	٢٣	JJ	adjective	صفة نكرة
	٢٤	ADJ	Ordinal Numbers	أعداد ترتيبية
	٢٥	JJR	adjective, comparative	أفعل تفضيل نكرة
	٢٦	DTJJ	adjective with the determiner al	صفة / معرفة بأل
	٢٧	DTJJR	adjective, comparative with the determiner al	أفعل تفضيل / معرف بأل
ظروف	٢٨	RB	Adverb	ظروف
	٢٩	WRB	Wh-adverb	ظروف موصولة
أخرى	٣٠	CD	Cardinal Number	أرقام أصلية
	٣١	UH	Interjection	انفعاليات
	٣٢	PUNC	Puncituation	علامات الترقيم

٩. بعض الوسوم لا يفهم محتواها حتى بعد الاطلاع عليه، ولا يتوفر ملف توثيقي مع مجموعة الوسوم يشرح المراد من كل وسم، ومن ذلك: NOUN للأسماء المكممة وهو الوسم المنقول من الإنجليزية، ومثله: WRB المقابل للظروف الموصولة، وإن كانت جميع الوسوم بحاجة للتعريف والتوضيح المفصل حتى ما كان معروفا في العربية.

ويظهر الجدول (٢-٨) ملاحظاتي على كل وسم في مجموعة ستانفورد التوسيمية مع عدد النتائج الصحيحة والخاطئة ومجموعها، وذلك بالنظر في النتائج الخمسين الأولى من كل وسم في مدونة arTenTen. ويوضح الجدول أنه فضلا عن تجاهل مجموعة الوسوم الكثير من الفروق في العربية جاءت النتائج مختلطة دون تفريق بسبب الاشتراك اللفظي أحيانا الذي أدى إلى الغموض، فلا يفرق بين تعرَّفَ وتَعْرِفْ وتَعَرُّف فتجدها ومثلها كثير جدا تحت وسم واحد. كما أن الخطأ قد يحدث أيضا بسبب التقطيع، فيقطع أحيانا ما لا ينبغي قطعه، كالفاءات والواوات الأصلية المبتدئة بها بعض الكلمات، كما في: فقد ـ وقى، ويُبقِي ما يستحق القطع مما خُصص له وسم غير مُقتَطَع، كالضمائر المتصلة.

جدول (٢-٨) ملاحظات على كل وسم في مجموعة ستانفورد التوسيمية مع عدد النتائج الصحيحة والخاطئة

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١	NN	تظهر فيه أدوات نحو: عليها ـ مع ـ منها ـ مما، وحروف متفرقة نحو: ف ـ د ـ ق، وأسماء فاعل مثل: خاصّة ـ رئيس، بل ومصادر، نحو: أمر ـ وجود ـ قول، رغم تخصيص المجموعة وسما للمصادر. وعلى أن منذ تأتي اسما إلا أن سياقتها في الوسم تخلط بين وقوعها اسما ووقوعها حرف جر، كما يخلط الوسم بين ما اشترك لفظيا مع وسوم أخرى، وهو كثير هنا، مثل: بين التي ترد فعلا واسما وصفة، ودون التي ترد فعلا، وكتاب التي ترد اسم فاعل.	١٤	٣٦	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٢	NNS	معظم ما صنف هنا كأسماء مجموعة هو بالفعل اسم، ولكن لأن المجموعة تخصص للمصادر ولأسماء الفاعلين والمفعولين وسوما عد وروده هنا خطأ، فتجد: انتخابات (مصدر)، ومتطلبات (اسم مفعول).	٧	٣٤	٥٠
	٣	NNP	لا يوجد ضبط دقيق للأعلام، حيث توجد مثلا ظروف (لما) وأسماء (بيت) وحروف جر (على)، وأفعال (رحمه). كما لا يعتد بالمركب من الأعلام إذ ترد الأجزاء الأولى فقط من الأعلام المركبة، نحو: عبد ـ ابن ـ بني. ولا اعتبار لعلميتها ما دامت مجزوءة. كما أن الأعلام المنقولة من المشتقات، مثل: خالد وسعد ومحمود وسعيد، تختلط النتائج بها ولا تميز.	٢٣	٢٧	٥٠
	٤	NNPS	النتائج لا يستشف منها أي مقصد من الوسم؛ إذ لا توجد أعلام أبدا. وتظهر نتائج غير مفهومة مثل: خلونابحالناوبعدين ـ خلوالنصرمديون، وغالبها ينتهي بـ (ون ـ ين). ويشترك الوسم مع زوجه DTNNPS في وجود كلمة أوكرانيون فقط. وهي خطأ أيضا، وبالتالي فإن جميع النتائج خاطئة.	٠	٢٢	٢٢
	٥	NOUN	لا يتضح من محتوى الوسوم ما الكلمات الفعلية المخصوصة بهذا الوسم، وليس هناك ملف توثيقي يوضح المقصود منه خصوصا وأنه منقول من الإنجليزية. ومع ذلك يمكن أن يحدد الخطأ من خلال وجود بعض الكلمات الخاصة بالوسم نفسه التي لم تقتطع منها الضمائر المتصلة رغم تخصيص وسم لها، مثل: أغلبيتنا، ومن خلال الكلمات التي خصصت لها وسوم أخرى ومع ذلك وردت فيه، وهي غالبا أدوات، مثل: أي، أو كلمات لا معنى لها شبيهة بكلمات تحت هذا الوسم، مثل: اغلبة ـ اغلبيا.	٣٦	١٤	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٦	DTNN	وسم الأسماء غير دقيق، وتختلط نتائجه بكلمات لها وسوم مخصصة، كالصفات: الكثير ـ الملك ـ السيد، والمصادر: الصلاة ـ التعليم ـ النظام.	١١	٣٩	٥٠
	٧	DTNNS	جاءت النتائج خليط من أعلام، مثل: (الإمارات)، ومصادر، مثل: (الاتصالات)، وأسماء فاعلين ومفعولين، مثل: (المسلمون). وخلت من الأسماء سوى من: السنوات.	١	٤٩	٥٠
	٨	DTN-NP	تشترك بعض الأعلام مع أقسام أخرى من الكلام، مثل: الحسن ـ المغرب، وتظهر أجزاء من الأعلام عادة ما تكون الأجزاء الثانية، مثل: العزيز ـ الرحيم ـ الرحمن، أما أجزاؤها الأولى فترد في قسم الأعلام غير المعرف بأل NNP.	١٩	٣١	٥٠
	٩	DTN-NPS	لم يوضح ما إذا كان هذا الوسم خاص بالأعلام التي تأتي بهيئة مثنى كالبحرين وجمع كالإمارات أو التي تأتي أعلاما وتثنى وتجمع كالكوريتان من كوريا. ومعظم نتائج هذا الوسم خاطئة فهي ليست أعلاما مجموعة ولا أعلاما بصورة الجمع أو المثنى. إنها أسماء مجموعة جمع مؤنث ومذكر سالمين، نحو: الآيات ـ الكشميريين.	٤	١٥	١٩
أفعال	١٠	VBP	يضم الوسم خطأ بعض الأدوات، مثل: أن التي تسبق الأفعال المضارعة. بالإضافة إلى: لكن عل ـ لعل (أخوات إنّ)، وأداة لا السابقة تحديدا لكلمة: بُدّ. وتختلط الأفعال المتشاركة في شكلها بأقسام أخرى من الكلام ببعضها، إذ لو كان الفعل صحيحا، فالنتيجة ليست دقيقة وبالتالي عدت خطأ، على نحو: (تعرف) التي تختلط بتعرّفْ الأمر وتُعرف المبنية للمجهول وتعرُّف المصدر.	٢٠	٣٠	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١١	VBD	تظهر نتائج لكلمات من حرفين فقط (تح ـ صل) وليست من الأفعال الماضية المضعفة العين واللام. وهي عبارة عن كلمات فعليه فصل عنها الحرف الأول باعتباره أداة. في الأولى الفاء، وفي الثانية الواو.	٤٦	٤	٥٠
	١٢	VBN	يبدو لأول وهلة أن هذا الوسم دقيق، فأفعال مثل: (علم ـ قتل ـ نشر ـ يعتقد) الظاهرة في النتائج هي بالفعل أفعال مبنية للمجهول، ولكن بالعودة للسياقات، يظهر الخلط في كل ما كان شكله في البناء للمعلوم والمجهول واحدا، وكل ما كان شكله مطابقا للمصدر منه، كما في الأمثلة السابقة. كما تختلط النتائج أيضا بالحروف المضافة للضمائر نحو: مني وهي ليست كذلك بل حرف جر متصل بياء المتكلم (منِّي). وقد نجح النظام في تحديد كل ما جاء فيه البناء للمجهول بصيغة تختلف عن المعلوم والمصدر منه، نحو: أجريت ـ يراد.	١٧	٣٣	٥٠
	١٣	VB	يضم الوسم خطأً أسماء من الأسماء الخمسة: أخاه ـ أخيه ـ أباه ـ أمه ـ أختي، كما يضم أيضا أفعل التفضيل المبدوءة بالهمزة أيضا (أعلاه ـ أصلح ـ اعرض)، وأسماء أخرى مبدوءة بهمزة أيضا: ابني ـ امرئ ـ أجره. وهناك من الخطأ ما هو نتيجة التقطيع الخاطئ حيث فصلت الحروف الأولى كالفاء باعتبارها حرف عطف، في مثل: اق ـ قه ـ از. بل إن (تذكر) اختلطت فيها المصادر بالأفعال الماضية والمضارعة والأمر لتطابق الصيغة في كل. فالتشابه في شكل أفعال الأمر مع هذا النوع من الكلمات أثر على نتائج هذا الوسم.	١٢	٣٨	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١٤	VBG	يخلط المصادر بالمشتقات من الفاعل والمفعول، فترد مثلا فيه: قادما ـ القائلين ـ المكلفة. وتختلط بعض المصادر مثل: عطف ـ حدث بالأفعال الماضية. ويلاحظ أيضا من كلمتين مصدر هما: القائ ـ اعطائ أن المقطع لا يعيد الهمزة لشكلها الأصلي بعد فصل الضمائر المتصلة عنها فجاءت هاتين الكلمتين غير المعروفتين، ومع ذلك صحيحتان.	٣٥	١٥	٥٠
	١٥	VN	ورغم خلط الوسم لأسماء الفاعل بالمفعولين أساسا إلا أنه دقيق في تعيينها، فحتى النتيجة الخاطئة (معناها)، لم يأت خطؤها بسبب صيغتها بل لاتصال الضمير بها والنظام يخصص له وسما.	٤٩	١	٥٠
	١٦	CC	عدد حروف العطف في العربية تسعة هي: و ـ ثم ـ ف ـ حتى ـ لكن ـ بل ـ أو ـ أم ـ حتى، ويغيب في هذا الوسم حرف العطف (حتى)، وتضم ما ليس منها: (كما وأما). وهما أداتان لا تأتيان بأي وجه حرف عطف، ولذا فإن ورودهما هنا في هذا الوسم خطأ.	٧	٢	٩
	١٧	DT	أسماء الإشارة من الوسوم المغلقة ولكن النتائج هنا متداخلة مع كلمات أخرى شبيهة بها مثل: هذاه ـ هذيان، أو لم تفصل فيها أسماء الإشارة عما يليها وهذا الأغلب، على نحو: هذاالموضوع ـ هذاالرجل ـ هذاالوطن.	١٩	٣١	٥٠
	١٨	RP	تغيب بعض الأدوات رغم عدم تخصيص وسم لها كأداة الاستفهام والشرط من، وما ورد هنا من مَنْ جاءت حرف جر. وعلى أن (كأن) في (كأنما وكأنكما) من أخوات إن إلا أنه لا يرد سواها هنا، ووردت متصلة بما بعدها وهو يفصل ما بعدها في كلمات ومواضع أخرى.	٢١	٦	٢٧

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١٩	IN	يختلط هذا الوسم بالظروف فتجد فيه مثلا: إذ ـ إذا ـ لما، وتجد حروفا متفرقة: ا ـ ن. كما أنه يورد حروف جر لم تفصل عما يليها مثل: فيما، ويضيف ما يصنف من ضمن الأسماء كبينما وريثما.	١٦	١٥	٣١
	٢٠	PRP	هذا الوسم يظهر ٣٠ نتيجة رغم أن الضمائر الشخصية محدودة في العربية. فيظهر ما ليس له علاقة بالضمائر، كحرفي ة ـ ظ، والفعل يلائم، وحرف a، وبعض الكلمات المتصلة بالضمائر من لهجات عربية، مثل: جينا ـ اتقولوا. فضلا عن الأخطاء في تعيين الضمائر، كاعتبار التاء المربوطة هاء الغيبة والكاف الجارة كاف الخطاب ومعظم نتائج ألف الاثنين الألف في كلمة لا وال التعريف.	١٤	١٦	٣٠
	٢١	$PRP	لأن المجموعة تعتبر الضمائر المتصلة مباني تقسيمية، فالضمائر المتصلة في الوسم أعلاه هي نفسها الضمائر المتصلة في هذا الوسم، فياء المتكلم واحدة في الوسمين، ومع ذلك مرة تأتي هنا ومرة تأتي هناك. وإذا كنا سننقل الوسم كما هو من الإنجليزية، فالمفترض أن تكون هنا متصلة بالحروف والأسماء لتعطي معنى الملكية، وفي الوسم السابق متصلة بالأفعال لتؤدي خاصية الشخص Pearson. وثمة أخطاء في تعيين ضمير ياء المتكلم هنا، حيث جاءت كل النتائج ياء مقطوعة من حرف الجر في، وياء في مثل كلمة حقيبتي لا تفصل ولا تدرج لا في هذا الوسم. ولا في الوسم PRP. وكذا الحال في هاء الغائب حيث صنفت التاءات المربوطة على أنها هاء الغائب داخل هذا الوسم.	٧	٢	٩

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
صفات	٢٢	WP	الموصولات بالعربية محدودة كالضمائر. وقد نجح النظام في تحديد ما كان منها غير مشترك في شكله الهجائي، أي: الذين ـ اللذان ـ اللتان ـ التي ـ الذي. أما: من فقد جاءت نتائجها مختلطة مع مَن (حرف الجر وأداة الاستفهام وأداة الشرط). واختلطت ما بالماءات الأخرى في العربية: كالاستفهام وما الزائدة. كما أن الوسم يرد فيه ما ليس له علاقة بالموصولات، مثل: ماذا ـ مهما ـ م. أما أي الموصولة فقد أخفق وكل ما حدده من نتائجها خاطئ.	١٢	٦	١٨
	٢٣	JJ	يحدث خلط واسع في نتائجه حيث يخطئ في جعل الأسماء المنسوبة (سياسية ـ عسكرية) والمصادر الصناعية: (ديموقراطية ـ وطنية) صفات، وأصاب حين أورد أفعل التفضيل (أخرى) ومشتقات الفاعل (مختلف ـ مختلفة) والمفعول (موجود ـ موجودة) لأنها صفات. ولكنه قد خصص لهما وسمين VN -JJR فعددت ورودهما هنا خطأ وجاءت معظم النتائج خاطئة وتبقى ما هو صفة مشبهة ولم يخصص لها وسم، كجديد وكثير.	٢٠	٣٠	٥٠
	٢٤	ADJ	تظهر في نتائج الوسم علامة التعجب المصاحبة للنقطة مثل: ..!ـ...!ـ...!!ـ.!، ويعتبر كلمة الأولياء جمع للعدد الترتيبي (أول). كما يضم الوسم العلم: ثامر!	٤٢	٨	٥٠
	٢٥	JJR	أخطأ هذا الوسم تماما في كلمة واحدة فقط، هي: اقى. أما باقي الأخطاء فهي بسبب اختلاط النتائج في كل كلمة بكلمات أخرى لها وسوم مخصصة، مثل: أقصر، أطيب، المتشابهة شكلا بصيغة التعجب والفعل المضارع.	١١	٣٩	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٢٦	DTJJ	يحدث خلط واسع في نتائجه حيث يخطئ في جعل الأسماء المنسوبة (العسكرية ـ الاقتصادية) والمصادر الصناعية: (الديموقراطية ـ الوطنية) صفات، وأصاب حين أورد أفعل التفضيل (الأخيرة) ومشتقات الفاعل (القادمة ـ المتحدة) لأنهما صفات. ولكنه قد خصص لهما وسمان فعددت ورودهما خطأ وجاءت معظم النتائج خاطئة وتبقى ما هو صفة مشبهة ولم يخصص لها وسم، كالعظيم والجديد والكبير.	٥	٤٥	٥٠
	٢٧	DTJJR	من أكثر الوسوم صحة، ويحدث الخطأ هنا في ثلاث كلمات هي: الاصطلاح ـ الاصحاح ـ الاصبع.	٤٧	٣	٥٠
	٢٨	RB	معظم ما يرد هنا ليس له علاقة بالظروف، مثل: فقط ـ إذن ـ سيما ـ طالما، وأكثر ما يرد في هذا الوسم ما اتصل بالإشارة (ذا أو ذاك) مثل: وقتذاك ـ حينذاك ـ هكذا، بالإضافة إلى: هناك ـ ثمة. كما توجد أخطاء هنا أيضا سببها التقطيع، مثل: عدم فصل إذ الظرفية عن اسم الإشارة وعن الأسماء الأخرى، فنجد مثل: إذّاك ـ عندئذ ـ حينئذ.	١	٤٩	٥٠
	٢٩	WRB	ليس في العربية ظروف موصولة، ولا يعلم ما الكلمات المفترض أن يوسم بها هذا الوسم. والاضطراب ملاحظ في نتائجه، حيث يضم أدوات استفهام ومثل: أين ـ كيف، وشرط: أينما ـ كيفما، وكلمات غير معروفة، مثل: متينافيما ـ لكاهما ـ كينما ـ كيمفما ـ كادما ـ اينةوما. وتختلط نتائجه بعلامة الترقيم النقطة (.)، والفاصلة العلوية مع النقطة (.')، وعلامة الاقتباس (")، ولها وسم خاص PUNC.	٠	٣٣	٣٣

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
أخرى	٣٠	CD	تختلط بنتائجه علامات ترقيم مخصص لها وسم مستقل مثل: "ـ(-)ـ!!ـ، وترد كذلك حقول فارغة. ويخلط أيضا بين الأعداد الرقمية والمكتوبة ومن الأفضل والأدق التفريق بينهما لتقديم نتائج استرجاع مفيدة لتطبيقات مختلفة.	٣٣	١٤	٥٠
	٣١	UH	ليس ثمة أساس واضح لمحتوى الوسم. والظاهر أنه خصص لما هو عليه في الإنجليزية. ولم يظهر في نتائج الوسم منها إلا: نِعْم وتختلط بنعم الجوابية، بالإضافة إلى لا ـ كلا ـ اللهم، وكلها خطأ.	٠	٤	٤
	٣٢	PUNC	جاءت جميعها صحيحة ولكن هناك مثلها يرد في وسم آخر هو الأعداد الأصلية CD.	١٩	٠	١٩

٢ ـ موسم مدى MADA المطبق على مدونة KSUCCA:

مدى هي أداة أو تطبيق يستعمل على نص عربي خام ويضيف إليه معلومات صرفية ومعجمية في عملية واحدة[١]. ومن تلك المعلومات التي تضيفها: نوع الكلمة ـ التشكيل ـ تحديد المادة المعجمية ـ تقطيع النص[٢]. ولتحديد نوع الكلمة فإنه ينفذ ثلاث خطوات هي: التحليل ـ فك الغموض ـ التوليد. حيث يستعمل محلل المرجانة الصرفي دون النظر في سياق الكلمات، فتظهر قائمة من التحليلات لكل كلمة، ثم يستعمل عليها ١٩ خاصية. وترتب هذه النتائج، والنتيجة الأعلى هي التحليل الصحيح للكلمة في سياق ما[٣]. وأخيراً، يستفاد من المرجانة في عملية التوليد بعد تقطيع الكلمات، حيث يعاد إنشاء الكلمة بعد فصل الزوائد عنها. فكلمة (عزيمته) مثلاً يفصل الضمير المتصل (الهاء)، وتولد التاء

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٥٣

(٢) السابق، نفس الصفحة.

(٣) السابق.

المربوطة فتصبح بعد التقطيع (عزيمة+ هـ)[١]. وهكذا، يعتمد في كل معلوماته على محلل المرجانة الصرفي المبنى في أساسه الأول على بيانات محلل باكولتر الصرفي[٢]. وتضم مجموعة وسوم مدى ٣٥ وسماً متفرعاً من ثمانية أقسام أساسية هي: الاسم والصفة والفعل والأداة والظرف والضمير والانفعاليات، وأخرى.

وقد طبق موسم مدى على مدونة الذخيرة الفصحى لجماعة الملك سعود KSUCCA الخاصة باللغة العربية ما بين القرن السابع والقرن الحادي عشر الميلادي. وتغطي مجالات الدين واللغة والأدب والعلوم وعلم الاجتماع والتراجم في تلك الفترة، وتضم ما يقارب الخمسين مليون كلمة[٣]. ويوضح الجدول (٢-٩) قائمة بمجموعة وسوم مدى المستعملة في وسم مدونة KSUCCA، ومسمياتها بالإنجليزية ومقابلاتها العربية[٤].

جدول (٢-٩) قائمة بمجموعة مدى MADA المطبقة على مدونة KSUCCA ومقابلها بالعربية والإنجليزية

القسم الرئيس	م	الوسم	معناه بالإنجليزية	معناه بالعربية
أسماء	١	noun	Noun	اسم جامد
	٢	noun_num	Number Noun	اسم عدد
	٣	noun_quant	Quantifier Noun	اسم يدل على جزء
	٤	noun_prop	Proper Nouns	اسم علم

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ١٥٦

(٢) السابق، نفس الصفحة.

(٣) Sketch Engine. KSUCCA: King Saud University Corpus of Classical Arabic. 2-8-2017: /https://www.sketchengine.co.uk/corpus-of-classical-arabic-ksucca

(٤) Sketch Engine. Arabic MADA system tagset. 10-9-2017: https://www.sketchengine.co.uk/arabic-mada-system-tagset/;

جامعة الملك سعود. مدونة الذخيرة الفصحى. ٢٣-٣-٢٠١٦: /http://ksucorpus.ksu.edu.sa

صفات	٥	adj	Adjectives	صفة
	٦	adj_comp	Comparative Adjectives	افعل التفضيل
	٧	adj_num	Number Adjectives	صفة عددية
ظروف	٨	adv	Adverbs	ظروف واحوال
	٩	adv_interrog	Interrogative adverb	موصولات استفهامية
	١٠	adv_rel	Relative adverbs	موصولات ظرفية
ضمائر	١١	pron	Personal pronoun	ضمائر
	١٢	pron_dem	Demonstrative	اسم إشارة
	١٣	pron_exclam	exclamatory pronoun	اسم تعجب
	١٤	pron_interrog	Interrogative pronouns	اسم استفهام
	١٥	pron_rel	Relative pronouns	اسم موصول
أفعال	١٦	verb	Verb	فعل
	١٧	verb_pseudo	Pseudo Verb	حروف مشبهة بالفعل
أدوات	١٨	part	Particle	أدوات أخرى
	١٩	part_det	Determiner particle	حرف تعريف
	٢٠	part_focus	Focus Particle	حرف تفصيل
	٢١	part_fut	Future particle	حرف استقبال
	٢٢	part_interrog	Interrogative Particle	حرف استفهام
	٢٣	part_neg	Negative particle	حرف نفي
	٢٤	part_restrict	Restriction Particle	أداة استثناء
	٢٥	part_verb	Verb Particle	حروف تدخل على الفعل
	٢٦	part_voc	Vocative particle	حرف نداء
	٢٧	prep	Preposition	حرف جر
	٢٨	conj	Conjunctions	حرف عطف
	٢٩	conj_sub	Subordinating conjunction	حرف مصدر

انفعاليات	Interjections	interj	٣٠	انفعاليات
اختصارات	Abbreviations	abbrev	٣١	أخرى
علامات ترقيم	Punctuation	punc	٣٢	
حروف لاتينية	Foreign/Latin	latin	٣٣	
أعداد رقمية	Digital Numbers	digit	٣٤	
كلمات غير محللة	Unanalyzed words	NULL	٣٥	

وبالتأمل في هذه المجموعة من الوسوم سنجد ما يلي:

١. مدى يشبه ستانفورد في تجاهله فروقاً في العربية، فلا تفرق المجموعة بين أنواع الأفعال من حيث الأزمنة الصرفية، إذ كل الأفعال الماضية والمضارعة والأمر ترد تحت وسم واحد هو verb. ولكن هذا الوسم قابل للتوسيع بإضافة خاصية الزمن والشخص، خصوصا وأن النظام لا يفصل الضمائر المتصلة ولا يخصص لها وسما.

٢. لا تفرق المجموعة بين أنواع الضمائر الشخصية، فتضعها تحت وسم واحد هو pron، وتعزل عنها ضمائر النصب لتضعها في وسم part مدرجة مع الأدوات. ومهما اختلف النحاة في ضمائر النصب المنفصلة (إياك ـ إياهم) فإنهم لا ينزلونها منزلة الأدوات أبداً[١].

٣. تخصص مدى لأفعل التفضيل وسماً من ضمن الصفات هو Adj-comp، لكنها لا تفرق بين أسماء الفاعلين والمفعولين والصفة المشبهة وتجعلها تحت وسم واحد تسميه الصفة Adj رغم اختلافها شكلاً ووظيفة. وتستبعد صيغة المبالغة من الصفات وتجعلها من الأسماء.

٤. تصنيف الأدوات يتطلب دقة، إذ إن إهمال بعضها يؤدي إلى الخلط. ومجموعة وسوم مدى تركز على بعض الأدوات وتهمل غيرها، حيث خصصت مثلاً وسماً لأدوات

(١) ابن عقيل، عبد الله بن عبد الرحمن العقيلي الهمداني المصري. شرح ابن عقيل على ألفية ابن مالك. دار التراث، القاهرة: مصر، ١٩٨٠، ص. ٣٩

النفي part_neg وأهملت أدوات النهي، كما أن ثمة خلطا في تخصيصها للأدوات التي تدخل على الفعل وسماً part_verb، ثم تخصيصها وسماً آخر لأداة لا تدخل إلا على الأفعال كحرف الاستقبال part_fut.

٥. تقع المجموعة في خطأ تقسيمي نتيجة النقل من الإنجليزية، حيث تدرج أسماء الاستفهام وأسماء التعجب من ضمن مجموعة وسوم الضمائر. وإذا كانت الإنجليزية تستعمل فيها أدوات الاستفهام كضمائر موصولة، فليس في العربية ذلك. فالضمائر في العربية هي: ضمائر شخصية وضمائر إشارة وضمائر موصولة، وأسماء الاستفهام هي عند اللغويين أسماء وحروف أو أدوات، أما ما يخص التعجب فليس في العربية أسماء تعجب بل توجد صيغتان هما: ما أفعله ـ وأفعل به وصيغ أخرى سماعية، وجميعها تعرف بالخوالف عند أهل اللغة[١].

٦. محاولة نقل الوسوم من الإنجليزية إلى العربية أدى إلى إدراج أقسام ليس لها وجود في العربية مثل: الموصولات الاستفهامية adv_interrog والموصولات الظرفية adv_rel والأسماء الدالة على جزء noun، رغم محاولة تعريب التوسيم بغير ما يقابله في الإنجليزية وتكييفه مع العربية.

٧. تخصص مدى وسماً للحروف المشبهة بالفعل مع الأفعال، وهي إن وأخواتها، ولكنها تهمل كان وأخواتها، وغيرها مما اختلف اللغويون في كونه فعلاً أو أداة[٢].

٨. المجموعة تخصص وسما لأل التعريف part_det ولكنها لا تقتطعها.

ويظهر الجدول (٢-١٠) ملاحظات على كل وسم في مجموعة مدى التوسيمية مع عدد النتائج الصحيحة والخاطئة ومجموعها، وذلك بالنظر في النتائج الخمسين الأولى من كل وسم في مدونة KSUCCA. ويوضح الجدول أن الملاحظات على مجموعة الوسوم النحوية هذه شبيهة بالملاحظات على مجموعة الوسوم النحوية السابقة، وفضلا عن تجاهل مجموعة وسوم مدى الكثير من الفروق في العربية ومجيء النتائج مختلطة

---

(١) اللغة العربية، معناها ومبناها، ١٩٩٤، ص.١١٠-١١٣

(٢) السابق، ص. ٣٢١

دون تفريق بسبب الاشتراك اللفظي فإن أبرز ما يلاحظ هنا هو عدم اقتطاع ما لا يختلف في أنه قسم من أقسام الكلام الأساسية كحروف الجر والعطف، وتخصيصه وسما لـ(أل التعريف) وهو لا يقتطعها.

جدول (٢-١٠) ملاحظات على كل وسم في مجموعة مدى التوسيمية مع عدد النتائج الصحيحة والخاطئة

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١	noun	تختلط بعض النتائج بالأجزاء الأولى من الأعلام المركبة، مثل: ابن ـ أبي، وتظهر نتائج خاطئة من الأفعال، مثل: حدثنا ـ رضي. كما تظهر نتائج تختلط فيها الأفعال بالأسماء، مثل: دون ـ مثل، والأسماء بالصفات، مثل: كتاب ـ الحديث، وبسبب الاشتراك اللفظي.	٢٧	٢٣	٥٠
	٢	noun_num	هذا الوسم نتائجه جيدة، وما ورد هنا من خطأ بسبب عدم فصل ما يتصل بالأسماء من حروف الجر والعطف، فتجد على نحو: بألف ـ وسبعين. فضلا عن ظهور الأجزاء الأولى من الأرقام المركبة مثل: أحد عشر، لعدم تعامل النظام معها.	٣٨	١٢	٥٠
أسماء	٣	noun_quant	يفهم من محتوى الوسوم ما الكلمات الفعلية المخصوصة بهذا الوسم، ولكن ليس هناك ملف توثيقي يؤكد المقصود منه خصوصا وأنه منقول من الإنجليزية. ومع ذلك يمكن أن يحدد الخطأ من خلال وجود بعض الكلمات الخاصة بالوسم نفسه التي لم تقتطع منها حروف الجر والعطف رغم تخصيص وسم لها، مثل: بكل ـ وكل، ومن خلال الكلمات التي خصصت لها وسوم أخرى ومع ذلك وردت فيه، وهي إما أفعال وأعلام، مثل: قيس، أو أسماء مثل: الآية ـ آية، أو أفعل التفضيل، مثل: أكثر.	٢٤	٢٦	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
صفات	٤	noun_prop	لا يوجد ضبط دقيق للأعلام. فلا يعتد النظام بالمركب من الأعلام ومن ثم ترد الأجزاء الأولى فقط من الأعلام المركبة، نحو: بن ـ أبي ـ أبو، فلا اعتبار لعلميتها ما دامت مجزوءة. كما أن الأعلام المنقولة من المشتقات، مثل: حميد وسعد وسعيد، تختلط النتائج بها ولا تميز.	٢٥	٢٥	٥٠
	٥	adj	يخصص هذا الوسم للصفات من صيغ الفاعل والمفعول والصفة المشبهة، ولكنه يخلط بينها والأسماء المنسوبة، فنجد: الأعرابي ـ البخاري. ونجد أيضا ما جاء من الأعلام على صيغتها، نحو: المثنى ـ حسان. وفي نتائج كلمتي (آخر ـ الآخر) تختلط صيغة فاعل بأفعل التفضيل لتشابههما شكلا، وقد وردت أخرى هنا وهي صيغة تفضيل مؤنثة من آخَر.	٣٠	٢٠	٥٠
	٦	adj_comp	معظم الأخطاء هنا هي بسبب الاشتراك اللفظي بين أفعل التفضيل وصيغ الماضي والمضارع والأمر وبعض الأعلام، فتأتي نتائج مختلطة بهم جميعا، مثل: أسلم ـ أظلم.	١٤	٣٦	٥٠
	٧	adj_num	هذا الوسم خصص لما جاء من الأعداد على صيغة فاعل، ولكنه يدخل ما ليس منها فيها، مثل: أولو. ويدرج الأجزاء الأولى أو الثانية من الأعداد المركبة، مثل: الحادي ـ عشر ـ عشرة. فضلا عن عدم فصل الحروف المتصلة بها.	٣٢	١٨	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
ظروف	٨	adv	كل ما خصص تقريبا لهذا الوسم خصص خطأ وهو عبارة عن أسماء اتصلت بالظرف (إذ)، مثل: حينئذ ـ عندئذ ـ يومئذ، أو أسما لم تتصل بالظرف، مثل: طالما ـ سيما، أو أسماء إشارة، مثل: هناك ـ هنا. والنتيجة الصحيحة الوحيدة هي: إذا.	١	٤٩	٥٠
	٩	adv_interrog	ليس في العربية موصولات استفهامية، ولذلك كل النتائج هنا خاطئة؛ لوجود وسوم خاصة بالنتائج الظاهرة. فكان من المفترض أن تكون بعض نتائج هذا الوسم، مثل: أين ـ متى ـ كيف ـ لماذا ـ أنى في وسم أسماء الاستفهام pron_interrog.	٠	٤٥	٤٥
	١٠	adv_rel	ليس في العربية موصولات ظرفية ولذلك كل النتائج هنا خاطئة؛ لوجود وسوم خاصة بالنتائج الظاهرة. فكان من المفترض أن تكون بعض نتائج هذا الوسم، وهي: متى ـ حيث ـ أين ـ حيثما ـ أينما ـ كلما ـ لما، مخصصة بالوسم adv لأنها ظروف مع إذ ـ إذا.	٠	٥٠	٥٠
	١١	pron	تلتصق بالضمائر أدوات كحروف الجر والعطف فتأتي هنا مقترنة بها أحيانا، نحو: فأنتما ـ وهو، وهو ما لا يعد نتيجة صحيحة لأن هذه الكلمات عبارة عن كلمتين لهما وسمان مختلفان. ويلاحظ أن الضمير (أنت) لم تضم نتائجه إلى (انت) الغائبة عنها همزة القطع، ولا يفرق بينه والضمير (أنتِ). ومثل الكثير من الكلمات العربية يحدث الاشتراك اللفظي بين الضمائر والأدوات والأفعال والأسماء، كما في: هم ـ أنا ـ نحن ـ هما.	١١	٣٩	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
ضمائر	١٢	pron_dem	تلتصق بأسماء الإشارة أدوات كحروف الجر والعطف فتأتي هنا مقترنة بها أحيانا، نحو: فذلك ـ وهذا، وهو ما لا يعد نتيجة صحيحة لأن هذه الكلمات عبارة عن كلمتين لهما وسمان مختلفان.	١٠	٤٠	٥٠
	١٣	pron_exclam	يخصص هذا الوسم لما التعجبية ولكن النتائج تظهر ما النافية (ما ـ وما ـ فما)، وما التعجبية يضعها مع الأسماء الموصولة.	٠	٤	٤
	١٤	pron_interrog	تغيب أسماء استفهام كثيرة عن هذا الوسم، مثل: كيف ـ لماذا ـ أين ـ متى ـ أيان ـ من ـ أنى. كما أنه من بين النتائج الخاطئة أسماء استفهام اقترنت بالواو والفاء العاطفتين وحروف الجر، نحو: فبم ـ ومن ـ ولمن، أو اشتركت لفظيا مع كلمات أخرى مثل: ما ـ أيا.	٩	٤١	٥٠
	١٥	pron_rel	الأسماء الموصولة في العربية محدودة، ولكن يحدث الاشتراك اللفظي في (من ـ ما) الموصولتين مع الأدوات، فتظهر نتائج ما ومن مختلطة بمن وما الاستفهاميتين وأدوات أخرى. ويضم هنا للأسماء الموصولة ما ليس منها، نحو: مهما ـ الآية.	١١	٣٩	٥٠
أفعال	١٦	verb	أخطاء هذا الوسم هي بسبب اقتران الفعل بكلمة أخرى من نوع الحرف لم تقتطع منه، نحو: وقال ـ وسلم ـ فكان، والاشتراك اللفظي مع المصادر، في مثل: بشر ـ ذكر، والأفعال، في مثل: يزيد.	٣٦	١٤	٥٠
	١٧	verb_pseudo	النظام لا يقتطع الفاء ولا الواو مما تتصل به ولذلك كل الأخطاء هنا جاءت بسبب ذلك، فيما عدا: (لا ـ ولا ـ فلا) التي ترد مع الحروف المشبهة بالفعل (إن وأخواتها) وهي ليست منها.	٢١	٢٩	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	١٨	part	معظم ما خصص به هذا الوسم هو ضمائر النصب المتصلة، والقليل مما ورد هنا هو أدوات لها وسوم محددة في المجموعة. فمثلا: كأنما من الحروف المشبهة بالفعل verb_pseudo.	٠	٥٠	٥٠
	١٩	part_det	تضع مجموعة الوسوم لال التعريف وسما ولكنها لا تخصه بها حيث تضمنت النتائج: أل ـ ال التي تعود على الاسم: أل التعريف، وتضمنت أيضا الاسم آل الذي يسبق الأعلام، مثل: آل عمران.	٠	٢٣	٢٣
	٢٠	part_focus	يخلط أمَا الاستفتاحية بأمّا التفصيلية، أما (إما) فتظهر بنتائج صحيحة. ويظهر في النتائج أيضا كلمات تشابه كلمتي إما وأما شكلا، نحو: أمأ. وفي نتيجة واحدة جاءت (امأ) وهي في الأصل كلمة امرأً ولكن اقتطعت الراء منها من النص نفسه.	٢	١٣	١٥
	٢١	part_fut	يستبعد سين الاستقبال الملتصقة بالفعل المضارع، ولا يدرج سوى الأداة سوف. وبسبب الاشتراك اللفظي بين سوف الأداة وسوّف الفعل جاءت نتائج سوف مختلطة.	٠	٣	٣
	٢٢	part_inter-rog	يتوقع أن تكون هنا نتيجتان فقط هما: أ ـ هل، لأنهما حرفا الاستفهام الوحيدان في العربية. ولكن تظهر هنا ٢٨ نتيجة جميعها خاطئة وتضمنتها (أ ـ هل). حيث لم ينجح النظام في تحديد (همزة الاستفهام) لأنه لا يقتطعها، ولم يميز (هل الاستفهامية) عن (هل الفعل). أما باقي النتائج الأخرى فهي خاطئة بسبب التقطيع، نحو: فهل ـ وهل، أو بسبب التصنيف الخاطئ، مثل: هلا ـ كا ـ با.	٠	٢٨	٢٨

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٢٣	part_neg	لا يفرق الوسم بين ليس الأداة وليس الفعل فتجد نتائجها مختلطة، فترد هنا وترد مع الأفعال بنفس إشكالية الخلط. ويضم الوسم أي أداة نفي مسبوقة بحرف التاء حتى وإن تكونت من كلمة غير معروفة، نحو: تلن ـ تما ـ تلم ـ تبلا، فضلا عن نتائج أخرى لم تقتع الأدوات فيها عن أدوات أخرى، مثل: أفلم ـ فلن ـ ولا ـ فلا. كما يضم أدوات ليست للنفي، بل خصصت لها وسوم، كأداة الاستثناء (إلا) على اعتبار أنها إن + لا ولكن النظام لا يميزها.	٥	٤٥	٥٠
	٢٤	part_re-strict	يسمى هذا الوسم بأداة استثناء ولكنه يضمن أداة الحصر: إنما، ويستبعد أدوات تعد استثناء مثل: خلا ـ عدا ـ حاشا وهي أفعال عند العرب، ولكنه يضمن (سوى) وهي مثل (غير) من حيث إنهما يعدان أسماء عند العرب، ولكنك تجدها مع الأسماء أيضا دون فارق يذكر. وتظهر في النتائج ألا أن+لا المدغمة على أنها إلا خطأ.	٦	٣١	٣٧
	٢٥	part_verb	هذا الوسم للأدوات التي تدخل على الأفعال، ولكنه خصص فقط لحرف التحقيق والتشكيك (قد). وعزل عنها أدوات الاستقبال part_fut. وقد ورد بصيغه المختلفة دون فصل لما اتصل به عنه، فنرد: وقد ـ فقد ـ لقد ـ فلقد، فضلا عن الفعل تقد. ولشبه قد الأداة بالفعل قدّ فقد جاءت نتائج قد مختلطة، وبالتالي لم تظهر أي نتيجة صحيحة تحت هذا الوسم.	٠	٧	٧

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٢٦	part_voc	هذا الوسم هو لأداة النداء، ولكن ما يرد فيه هو تركيب نداء ما عرف بأل، فتجد: أيها ـ أيتها بكل ما يتصل بها. ويغيب عن هذا الوسم أدوات النداء: أيا ـ أي ـ وا ـ هيا ـ الهمزة. ولا يرد هنا سوى أداة النداء: يا.	١	٢٧	٢٨
	٢٧	prep	الأخطاء في هذا الوسم يقع بعضها من إدخال ما ليس حرف جر في مجموع النتائج، نحو: إلا، وتأتي أيضا من عدم فصل حروف الجر عن الأسماء الموصولة المدغمة فيها، نحو: ممن ـ مما، ومن عدم قطع الحروف المتصلة بها، كالفاء والواو، في مثل: وعلى، فمن. كما أن المشترك اللفظي في مثل ما بين في ـ فيه ـ فيها الحروف ومثيلاتها الأسماء أدى لاختلاط النتائج والوقوع في الخطأ.	٣٦	١٤	٥٠
	٢٨	conj	أحرف العطف تسعة وهي: الواو ـ الفاء ـ ثم ـ حتى ـ أم ـ أو ـ لا ـ بل ـ لكن. والنظام لا يقتطع الفاء والواو العاطفتين من الكلمات التي اتصلت بها، ولذا فإن ما يأتي من النتائج هنا لهذين الحرفين، إما واو عاطفة منفصلة من المصدر المكتوب عما بعدها كما في: و« أبو مرزوق»، أو واوات منفصلة سواء أكانت ترقيما أو رمزا أو اختصارا. أما ما يتعلق بالفاء العاطفة، فهي لا ترد هنا أبدا وما يرد من الفاءات إنما هو فاءات لاختصارات أو ترقيمات. أما الحروف: بل ـ أم ـ لكن فهي تشترك لفظيا مع الأفعال والأدوات. وأما حتى فقد استبعدت من هذا الوسم.	٢	٤٧	٤٩

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٢٩	conj_sub	الوسم مخصص للحروف المصدرية، ولكنه يدخل ما ليس منها، كالظروف، مثل: إذ ـ إذا، والأدوات، مثل: لولا ـ حتى، كما أنه لا يفرق بين أن وإن ولو وما المصدريات عن مثيلاتهم من أقسام الكلام الأخرى، ولذلك جاءت جميع النتائج خاطئة.	٠	٥٠	٥٠
	٣٠	interj	يقصد بهذا الوسم أسماء الأفعال والأصوات (الخوالف) ولكنه يخلط بها بعض الأدوات مثل: نعم ـ كلا ـ بلى، والأفعال، مثل: تأوه، والاختصارات، مثل: (ا.ه) لكلمة انتهى، وصيغة فاعل لـ: واه، لأن النظام لا يقتطع حروف العطف.	٧	٤٣	٥٠
	٣١	abbrev	هذا الوسم خصص للاختصارات ولكنه لم يحدد هل المقصود بها الحروف التي ترمز لشيء معين، نحو: ص. (الصفحة)، أو ص. (الترقيم)، أو المنحوتات، نحو: صلعم ـ إلخ، أو أنها جزء من اختصار مركب، نحو: آي. بي. إم (IBM). ومع ذلك جاءت معظم النتائج حروفا متفرقة مقطوعة أو من مواد أو جذور للكلمات العربية وليست اختصارات، نحو: ة ـ ا ـ ظ ـ ط، وكلمات غير معروفة، نحو: وط ـ وس ـ وص، أو معروفة، نحو: وإن ـ إيه ـ أم ـ أرَ، أفّ.	٢	٤٨	٥٠
	٣٢	punc	يجمع هذا الوسم علامات الترقيم والرموز الرياضية. ويغيب عن نتائج هذا الوسم الشكل العربي للفاصلة الاعتيادية (،)، والفاصلة المنقوطة (؛) وعلامة الاستفهام (؟). والملاحظ فصل شرطتي الجملة الاعتراضية عن بعضهما، فتأتي نتائجهما متفرقة مع تطابقهما شكليا.	٢١	٠	٢١

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٣٣	latin	كل النتائج هنا بحروف لاتينية.	٥٠	٠	٥٠
	٣٤	digit	كل النتائج عبارة عن أعداد رقمية.	٥٠	٠	٥٠
	٣٥	NULL	توجد كلمات معروفة من المفترض أن تكون مخصصة بوسوم موجودة في هذه المجموعة. وهي غالبا أعلام، نحو: دريد ـ الفرزدق ـ سيبويه، واختصار، نحو: ثنا ـ نا، ونتيجة واحدة لا يمكن تحليلها، وهي حرف: ئ، رغم وجود مثيلاتها خطأ من ضمن الاختصارات.	١	٤٩	٥٠

٣ ـ موسم أميرا AMIRA المطبق على AWC:

أميرا AMIRA تطبيق يضم مجموعة من الأدوات هي: مقطع النص AmiraTok وموسم لأقسام الكلام Amira ومقطع للعبارات الأساسية المعروف باسم **المحلل التركيبي السطحي** Shallow syntactic parser[١]. ولمستعمل أميرا الحرية في إدخال نص خام أو نص مقطع بنظام يتوافق مع أحد مخططات أو أنظمة للتقطيع الذي يعرفها Amira–tok. فإذا كان النص مقطعاً يتم التوسيم على الشكل السطحي، أما لوكان خاماً فإن أميرا تطبق مقطعها على النص ثم تنفذ عملية التوسيم[٢]. وتعتمد أميرا مجموعة وسوم ERTS الموسعة والمقلصة التي تضم ٧٢ وسما وتعتمد هي الأخرى على مجموعة باكولتر الكاملة المعرفة على نص مقطع. وقد استعملت ٢٣ وسماً من هذه المجموعة، وتعد الوسوم الأساسية بالنسبة لها[٣].

---

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ٨١

(٢) السابق، نفس الصفحة.

(٣) sketchengine. :14-10-2017. Sketch Engine. POS tag set for Modern Standard Arabic /co.uk/pos-tag-set-for-modern-standard-arabic

وطبقت مجموعة وسوم أميرا النحوية على مدونة الويب العربية AWC التي جمع **سيرج شاروف** Serge Sharoff نصوصها من الإنترنت في عام ٢٠٠٩، ونقحت وحذفت مكروراتها وضمت ١٧٥ مليون كلمة تقريباً[١]. ويظهر في الجدول (٢-١١) قائمة لمجموعة وسوم أميرا التي وسمت بها مدونة الويب العربية المتوفرة على موقع سكتش انجن.

وبالتأمل في هذه المجموعة من الوسوم سنجد تجاهل مجموعة وسوم أميرا أيضا لفروق مهمة في العربية، وعدم مراعاتها لخصائص العربية وتظهر فيما يلي:

١. فيما تعتد المجموعة بخاصية العدد في الأسماء إفرادا وجمعا، إلا أنها تتجاهل المثنى الذي يختلف عن الجمع شكلاً ووظيفة. ولأن الجمع في العربية ليس كله سالماً في العربية، نجد جمع التكسير هنا يضمن مع المفرد حيث لا تعتد هذه المجموعة إلا بالجمع السالم.

٢. خصص وسم VBN للأفعال الماضية والمضارعة المبنية للمجهول ولم يفصل بينهما، بينما فصل بينهما في صيغة المعلوم؛ وهذا ما يجعل مجموعة الوسوم بعيدة عن التقسيم المنطقي، فالبناء للمجهول والمعلوم خاصية تصريفية تلي تقسيم الفعل من حيث الزمن الصرفي ولا يعتد بها في التقسيم الأساسي.

٣. يخصص للصفات وسم واحد فقط هو: . JJ والصفة باب واسع في العربية يضم اسم الفاعل، واسم المفعول وصيغ المبالغة وأفعال التفضيل والصفة المشبهة، وجميعها تختلف شكلاً ووظيفة.

٤. يخصص لحروف الجر والعطف وسمان هما: IN-CC، وباقي الأدوات يخصص لها وسم واحد RP. والعربية تزخر بأدوات كثيرة تؤدي وظائف خاصة يمكن أن تقسم الأدوات بناء عليها كالنفي، والنهي، والتوكيد، والاستفهام، والشرط وغير ذلك كثير.

---

(١) Sketch Engine. Arabic corpus (arWaC). 14-10-2017: https://www.sketchengine. /co.uk/arabic-web-corpus-wac

٥. بسبب إسقاط وسوم الإنجليزية على العربية، يظهر هنا أيضا وسم الموصولات الظرفية RB الذي ليس لها أساس في العربية رغم تخصيص وسم للظروف WRB.

٦. لم يُخصص للأرقام أو الأعداد إلا وسم فريد هو CD وهو وسم يُعنى بالأعداد الأصلية، لذا فقد تضمن الرقمية والمكتوبة منها، وهذا ليس مفيدا عند الاستفادة من نتائج الاسترجاع في التطبيقات المختلفة، كالترجمة الآلية وغيرها.

ويظهر الجدول (٢-١٢) ملاحظات على كل وسم في مجموعة أميرا التوسيمية مع عدد النتائج الصحيحة والخاطئة ومجموعها، وذلك بالنظر في النتائج الخمسين الأولى من كل وسم في مدونة AWC. ويوضح الجدول أن الملاحظات على مجموعة الوسوم النحوية هذه شبيهة بالملاحظات على مجموعتي الوسوم النحوية السابقة، وفضلا عن تجاهل مجموعة وسوم أميرا الكثير من الفروق في العربية ومجيء النتائج مختلطة دون تفريق بسبب الاشتراك اللفظي، فإن أبرز ما يلاحظ هنا هو عدم اقتطاع ما لا خلاف عليه في أنه قسم من أقسام الكلام الأساسية، كسين الاستقبال. وكذلك اطراد أخطاء كثيرة في كل وسم تقريبا، كورود كلمات ثنائية غير معروفة في الوسوم: PRP- CC- UH-VB- NUMCOMMA VERB-NOFUNC، وتكرار ورود حروف متفرقة معينة تحت أكثر من وسم، مثل: M١- ٠أ.

ويبدو أن المقطع يحافظ على الهمزات ولا يقوم **بتوحيدها** normalizatiom على عكس مُقطِّع ستانفورد ومقطع مدى، وهذا منع الموسم من الوقوع في بعض الأخطاء، كالخلط بين الفعل الأمر اسمع والفعل المضارع أسمع. ويلاحظ أن النص الأصلي للمدونة الذي طبقت عليه مجموعة وسوم أميرا يتأثر بالمُقطِّع في الكلمات المنتهية بضميري هاء الغائب والغائبة، نحو: (بها ـ انه ـ له ـ منه ـ أ ـ أي) فترد في النص الأصلي مقطوعة بعلامة + بهذا الشكل (ب+ها ـ ان+هـ ـ ل+هـ ـ من+هـ) فلا يتعرف عليها الموسم ويتعامل مع الكلمة بوصفها كتلة واحدة، ومن ثم ترد متناثرة بين الوسوم ومتفرقة.

جدول (٢-١١) قائمة بمجموعة مدى AMIRA المطبقة على مدونة AWC ومقابلها بالعربية والإنجليزية

القسم الرئيس	م	الوسم	معناه بالإنجليزية	معناه بالعربية
أسماء	١	NN	noun singular or mass	اسم مفرد
	٢	NNP	Proper noun, singular	علم مفرد
	٣	NNS	noun, plural	اسم (جمع ـ مثنى) سالم
	٤	NNPS	Proper noun, plural	علم (جمع ـ مثنى) سالم
ضمائر	٥	PRP	personal pronoun	ضمير شخصي
	٦	DT	demonstrative pronoun	اسم إشارة
	٧	WP	relative pronoun	اسم موصول
صفة	٨	JJ	adjective	صفة
ظروف	٩	RB	adverb	الظروف والأحوال
	١٠	WRB	wh-adverb	موصول ظرفية
أدوات	١١	CC	coordinating conjunction	حروف عطف
	١٢	IN	preposition	حروف جر
	١٣	RP	particle	أدوات
أفعال	١٤	VERB	Verb, based form	أفعال جامدة
	١٥	VB	Verb, based form	فعل أمر
	١٦	VBP	Verb, non-3rd person singular present	فعل مضارع
	١٧	VBD	Verb, past tense	فعل ماض
	١٨	VBN	Verb, past participle	أفعال مبنية على المجهول

القسم الرئيس	م	الوسم	معناه بالإنجليزية	معناه بالعربية
أخرى	١٩	UH	interjection	علامات انفعالية
	٢٠	PUNC	punctuation	علامات ترقيم
	٢١	NOFUNC	without function	كلمات غير محللة
	٢٢	CD	cardinal number	أرقام أصلية
	٢٣	NUM-COMMA	remove all non-numeric characters and convert "," to "." and vise versa	الفاصلة ر

جدول (٢-١٢) الملاحظات على كل وسم في مجموعة أميرا التوسيمية مع عدد النتائج الصحيحة والخاطئة

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
أسماء	١	NN	الأخطاء هنا معظمها تحدث بسبب الاشتراك اللفظي بين الأسماء والأقسام الأخرى من الكلام، كما في: كل ـ سبب ـ عمل. وترد أيضا القليل من الصفات، مثل: الرئيس ـ الشيخ، وكلمات متصلة بضمير الغائب: ب+ها ـ ب+هـ ـ علي+هـ.	٣٢	١٨	٥٠
	٢	NNP	يتضمن هذا الوسم أخطاء مطردة، حيث ترد كلمات كثيرة متصلة بهاء الغائب ـ الغائبة على نحو: علي+هـ ـ قول+هـ ـ لأن+هـ ـ نفس+هـ ـ سبحان+هـ ـ في+ها، كما ترد أفعال ماضية على نحو: صلى ـ وسلم، وأجزاء أولى من أعلام مركبة، نحو: عبد ـ أبو، وأسماء، نحو: ابن ـ بن. ويحدث الاشتراك اللفظي فتتداخل الأعلام مع الأسماء والصفات والأفعال، في مثل: حسن ـ عمر ـ خالد.	٢٠	٣٠	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٣	NNS	نتائج الوسم لا تعطي نتائج فعلية لما هو مجموع من الكلمات حيث لا تعتد بما جمع جمع تكسير وتكتفي بالجمع السالم وبالتأنيث. كما تظهر مع النتائج الصفات المجموعة جمعا سالما، مثل: المؤسسات ـ المسؤولين.	٣٤	١٦	٥٠
	٤	NNPS	لم يوضح ما إذا كان هذا الوسم خاص بالأعلام التي تأتي بهيئة مثنى كالبحرين وجمع كالإمارات أو التي تأتي أعلاما وتثنى وتجمع كالكوريتان من كوريا. ومعظم نتائج هذا الوسم خاطئة فهي ليست أعلاما مجموعة ولا أعلاما بصورة الجمع أو المثنى. إنها أسماء مجموعة جمع مؤنث ومذكر سالمين، نحو: الآيات ـ الكشميريين.	٤	٣٦	٤٠
ضمائر	٥	PRP	تضع المجموعة الضمائر الشخصية تحت وسم واحد ولا تفرق بين المتصل والمنفصل منها، وتعزل عنها ضمائر النصب فتجدها متناثرة في كل أقسام الكلام. ونتائج هذا الوسم جاءت بعلامات ترقيم وأفعال وأدوات وكلمات غير معروفة وحروف متفرقة، مثل: [ـ هام ـ هل ـ هذ ـ ة. والملاحظ أن النظام أحيانا حين يفصل الضمير المتصل بالكلمة، يأتي بالكلمة مع ما اتصلت به ويسمها كاملة بوسم الضمير، فتجد: أن+نا، وهي أننا المكونة من أداة وضمير وكلاهما قسمان من الكلام خصص لهما وسمان، وفي الغالب يسمها باعتبارها ضمير المتكلم أنا. وتجد أيضا الضمائر المتصلة كياء المتكلم مثلا ليست في النتائج ياء متكلم إطلاقا، وإنما ياء نسب أو غيرها، وياء المتكلم لا تقتطع أبدا، وكذا الحال في كاف الخطاب إذ لا تقتطع ونتائجها كلها هي حرف الجر.	٩	٤١	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
	٦	DT	لم يتضمن الوسم اسم الإشارة (ذا) فقط، بل تضمن (ذو ـ ذوي ـ ذي)، وهذه الكلمات من الأسماء الخمسة بمعنى صاحب، وإن كانت الأخيرة تستعمل للإشارة أيضا. كما تضمن نفس الوسم حروفاً متطابقة متصلة مع بعضها، مثل: (ذذ ـ ذذذ)، وكلمات ليس لها صلة بضمائر الإشارة إلا في التشابه الحرفي، على نحو: (هلك ـ فلك) المشابهة في أحرفها لـ (ذلك)، و(هذيان) التي تشبه (هذان ـ هذين). كما يحدث الاشتراك اللفظي في أسماء الإشارة العامية في مثل: هَذِيل وهُذيل. واللافت هنا أنه يقتطع ال التعريف ويجعلها من أسماء الإشارة.	٢٤	٢٦	٥٠
	٧	WP	يضمن هذا الوسم الخاص بالأسماء الموصولة بعض الأدوات والظروف، مثل: أين ـ متى ـ مهما ـ ماذا. وتأتي النتائج بكلمات عديدة غير معروفة، مثل: مو ـ اللا ـ مي ـ مذا. ولم يكن الوسم دقيقا في (من ـ ما) الموصلتين بسبب الاشتراك اللفظي بينها وأقسام أخرى من الكلام.	١٨	٣٢	٥٠
صفة	٨	JJ	رغم وجود مشتركات لفظية في النتائج إلا أن النظام يستطيع تمييز بعضها، فهو يفرق بين العام الصفة والعام الاسم وأول الصفة وأوَّلَ الفعل وإن أخفق في أكثر وأكبر اللتان امتزجت نتائجهما بالأفعال.	٢٨	٢٢	٥٠
ظروف	٩	RB	ورغم وجود الظروف في العربية لكن لم أتمكن من الكشف عما خصص به هذا الوسم من كلمات لكثرة الأخطاء في نتائجه. فمعظم النتائج الخمسين الأولى إما أسماء وأعلام، نحو: فوق ـ أمس ـ هند ـ هناء، أو أدوات، نحو: ثم ـ معك، أو ضمائر شخصية، نحو: أنا، أو إشاريات، نحو: هنا ـ هناك، أو بين الأسماء والأفعال، نحو: همس، أو صفات، نحو: هنيئا. وإن وردت بعض الظروف فإنها ترد هنا ضمن كلمة ولم تقتطع، نحو إذ في: عندئذ.	٠	٥٠	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
أدوات	١٠	WRB	يظهر في نتائج هذا الوسم ظروف يفترض أن تكون في وسم الظروف وهي: حيث ـ أين ـ متى. ومعظم النتائج عبارة عن كلمات تتصل بها (ما)، مثل: كيفما ـ طالما ـ ريثما ـ مهما، عندما ـ بعدما، أو كلمات تنتهي بـ (هما) وغالبها تكون مقتطعة بعلامة الزائد+ ولكنها وسمت كاملة بهذا الوسم، نحو: عن+هما ـ عند+هما ـ بعد+هما.	٥	٤٥	٥٠
	١١	CC	تغيب عن هذا الوسم ثم العاطفة ويجعلها من ضمن الظروف RB. كما تغيب أيضا حتى ولكن العاطفتان. وما ورد من حروف العطف كان الاشتراك اللفظي سببا في خلط نتائجه، حيث إن (أم) تشترك مع الأفعال والأسماء، وكذلك الحال في (بل ـ ف ـ و). وتطغى على النتائج الكلمات ذات الحرفين سواء أكانت معروفة مثل: به ـ عم ـ قم وبعض الأسماء الخمسة مثل: أخ ـ أخو ـ أب، أو غير معروفة، نحو: أأ ـ مو ـ با.	٢	٤٨	٥٠
	١٢	IN	يحدث الاشتراك اللفظي في حروف جر مثل: في ـ من ـ عن فتختلط في الحرف بفي الاسم، ومن الحرف بمن الاسم والفعل، وكذا الحال في عن. أما حروف الجر: الباء والكاف واللام، فيقتطعها النظام أحيانا وأحيانا لا يقتطعها فتبقى ملتصقة بالكلمة ولا يعتد بها. ويظهر في هذا الوسم كلمات خصصت لها وسوم معينة ولا يختلف في أنها ليست من حروف الجر، مثل: أن ـ بعد ـ كما ـ ضد ـ لذلك، وغيرها كثير.	٩	٤١	٥٠
	١٣	RP	تكثر في نتائج هذا الوسم كلمات غير معروفة وليس لها صلة بالأدوات، مثل: لêـ مْ ـ قم ـ لى ـ لد، وكلمات خصصت لها وسوم في نفس المجموعة، مثل: ما الموصولة، ومن الجارة والموصولة، وأين ومتى الظرفان، وليس الفعل الناقص، وكلمات لم تقطع، مثل: لك ـ أيها، وكلمات متصلة بهاء الغائب والغائبة، مثل: ل+ها ـ ل+ك. فضلا عن الخلط الحاصل في الأدوات بسبب المشترك اللفظي، كما في: هل ـ كيف.	٢١	٢٩	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
أفعال	١٤	VERB	يقابل هذا الوسم في العربية الأفعال الجامدة، ورغم محدودية الأفعال الجامدة وإمكانية حصرها، إلا أنه لم يرد تحت VERB في الخمسين كلمة الأولى الأكثر تكراراً أي نتيجة تعد فعلا ناقصا.	٠	٥٠	٥٠
	١٥	VB	تأتي في النتائج أفعال ليست أفعال أمر وتشترك لفظيا مع الماضي والمضارع، مثل: كذبوا ـ أحب ـ قل، وكلمات غير معروفة، مثل: أأ ـ Mا، وكلمات لا صلة لها بالأفعال، نحو: مش ـ آه ـ زي ـ ، وأسماء أو أدوات متصلة بالضمائر، مثل: ل+هـ ـ من+هـ ـ أن+ك	١٥	٣٥	٥٠
	١٦	VBP	تظهر في النتائج أفعال مقترنة بأداة الاستقبال ولم تفصل عنها، مثل: سيأتي، وأفعال جامدة خصص لها وسم VERB، مثل: يزال ـ يصبح. ولوجود وسم خاص بما بني للمجهول فكثير من الأفعال الموجودة تحت هذا الوسم تأتي بمثل الصيغة في المعلوم والمجهول، نحو: يَعد ويُعد ـ يَقوم ويُقوّم، وقد تأتي فعلا معلوما ومجهولا واسما، نحو: تكون وتُكوّن وتكوُّن، فيحدث الخلط.	٢٣	٢٧	٥٠
	١٧	VBD	تظهر في النتائج أفعال جامدة خصص لها وسم VERB، مثل: أصبح ـ صار. ولوجود وسم خاص بما بني للمجهول فكثير من الأفعال الموجودة تحت هذا الوسم تأتي بمثل الصيغة في المعلوم والمجهول، نحو: أكَّدَ وأكِّد ـ بَلَّغ وبُلِّغ، وقد تأتي فعلا معلوما ومجهولا واسما، نحو: طَلَب وطُلِبَ وطَلَب، فيحدث الخلط.	٢٠	٣٠	٥٠
	١٨	VBN	يحدث الخلط في هذا الوسم بنفس الأسلوب الذي يحدث فيه الخلط في الوسمين VBD.- VBP فكثير من الصيغ المبنية للمجهول تتطابق مع صيغها عند بنائها للمعلوم. ومن ثم لا يفرق بينها، فتجد (يذكر) مثلا تحت وسم VBN تارة، وتارة تحت وسم VBP تارة أخرى.	١٩	٣١	٥٠

القسم الرئيس	م	الوسم	الملاحظات	الصحيحة	الخاطئة	المجموع
أخرى	١٩	UH	يقابل هذا الوسم أسماء الفعل والصوت في العربية غير أنه يخصص به أدوات النداء، نحو: يا ـ هيا ـ وا، وأحرف الجواب، مثل: لا ـ نعم ـ كلا. وترد الكثير من النتائج كلمات غير معروفة مكونة من حرفين، نحو: تا ـ ما ـ را، أو كلمات متصلة بضمائر، نحو: كلا+هما ـ طبع+هما ـ كل+هم.	٣	٤٧	٥٠
	٢٠	PUNC	يجمع هذا الوسم علامات الترقيم والرموز الرياضية والاختصارات. ويغيب عن نتائج هذا الوسم الشكل العربي للفاصلة الاعتيادية (،)، والفاصلة المنقوطة (؛) وعلامة الاستفهام (؟). كما تظهر في النتائج رموز التشكيل: ً . والملاحظ فصل شرطتي الجملة الاعتراضية عن بعضهما، فتأتي نتائجهما متفرقة مع تطابقهما شكليا.	١٦	٣٤	٥٠
	٢١	NOFUNC	خصص هذا الوسم للكلمات غير المحللة، ومعظمها كلمات فصل المقطع ضمير الغيبة عنها بعلامة +، ورغم ذلك تجدها متناثرة بين الوسوم، وإن كان وجودها هنا بشكل كبير جدا وكأن الوسم مخصص لها، نحو: أن+ها ـ على+ها ـ إلي+ها. كما توجد كلمات عربية أخرى قابلة للتحليل، مثل: إله ـ إليها ـ اسمها.	٠	٥٠	٥٠
	٢٢	CD	تختلط الأرقام الأصلية بالأرقام المكتوبة، ومن الأفضل والأدق عزل الأعداد الرقمية عن المكتوبة لتقديم نتائج استرجاع مفيدة لتطبيقات مختلفة.	٣	٤٧	٥٠
	٢٣	NUMCOMMA	لا يمكن أن يستشف من النتائج أن المقصود بهذا الوسم هو نفس المقصود به في الإنجليزية، حيث لم يتضمن أي نتيجة فيها حرف ر مستعملا كالفاصلة. وإنما تضمن تسع نتائج مثلها متناثر في بقية الوسوم في المجموعة، وهي: MI ـ أ٠ ـ تث ـ يا ـ ته ـ PM ـ آ ـ ١.	٠	٩	٩

ثانيا: الأداء التقني:

رأينا كيف أن مجموعات الوسوم لا تنطلق من العربية ولا تتلاءم معها، ولكي أؤكد أن الأنظمة القائمة الحالية يعتريها النـقص من جانب آخر نتيجة تطبيقها على مدونات لم تتدرب على نوعها من قبل، حاولت كشف جانب الصحة فيها Accuracy فقط؛ لأن المعلومات الصحيحة والافتراضية في عملية التوسيم الأصلية التي اعتمد عليها النظام غير معلومة، وبالتالي لن نستطيع بدونها قياس الأداء بواسطة المقاييس الأخرى.

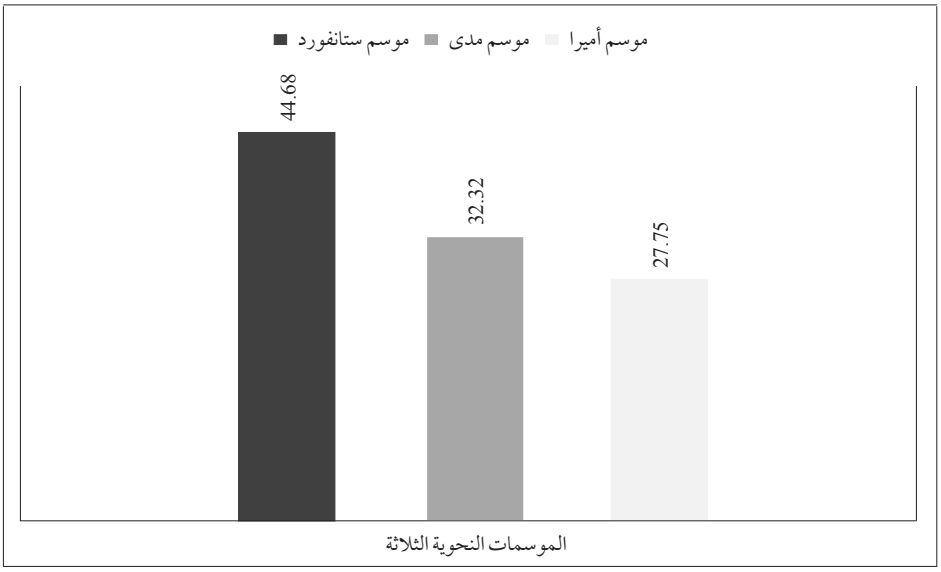
وللكشف عن نسبة **الصحة** Accuracy في الموسمات النحوية الثلاثة تقسم عدد الكلمات الموسمة توسيما صحيحاً على مجموع المخرجات، ثم يضرب الناتج في ١٠٠، وتلك النسبة تعد مؤشراً مفيداً لجودة التوسيم في كل موسم، فإذا صحح ما كشف خطؤه؛ أفاد ذلك في تطوير الموسمات النحوية الآلية.

ويبين الجدول (٢-١٣) نسبة الصحة بعد حسابها في كل موسم، مع بيان عدد النتائج الخاطئة في كل موسم. ورغم أن المقارنة لا تستقيم إلا بتطبيق الموسمات على نفس المدونة، يمثل الشكل (٢-٥) نسبة الصحة في كل موسم بالمقارنة مع الموسمين الآخرين، وهذا يُمكِّن من تفسير بعض الجوانب المتعلقة بمجموعات الوسوم.

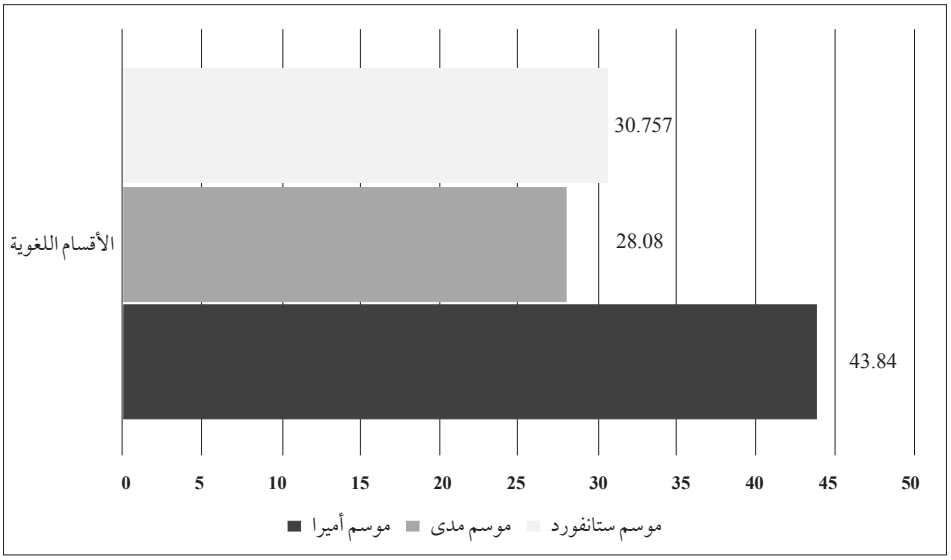
جدول (٢-١٣) نسبة الصحة Accuracy في كل موسم

اسم الموسم	مجموع النتائج الصحيحة والخاطئة	مجموع النتائج الصحيحة	مجموع النتائج الخاطئة	الصحة
ستانفورد	1271	568	691	44.68%
مدى	1460	472	988	32.32%
أميرا	1099	305	794	27.75%

شكل (٢-٥) نسبة الصحة في كل موسم بالمقارنة مع غيره بالنسبة للكلمات الخمسين الأكثر شيوعا في كل وسم داخل كل مجموعة



شكل (٢-٦) مقارنة الصحة بين الأقسام اللغوية في كل موسم



ويلاحظ من الجدول (٢-١٣) تفوق موسم ستانفورد في تطبيقه على مدونة arTenTen على موسم مدى وموسم أميرا. وقد أشير سلفاً إلى أن مدونة arTenTen تضم نصوصاً متنوعة من العربية الفصحى القديمة والحديثة ولهجات عربية مختلفة، أما مدونة KSUCCA فهي محدودة حيث لا تضم إلا نصوصاً من العربية الفصحى القديمة وبالتالي فإن نوعية نصوص المدونة تؤثر في نتائج وأداء أي موسم، وفي ذلك إشارة إلى أن تفوق موسم ستانفورد في تطبيقه على مدونة arTenTen كان لتدريب موسم ستانفورد على نصوص البنك العربي الشجري المتخصصة في النصوص الصحفية. وهي نوع النصوص التي تشكل أكبر مجال في نصوص مدونة arTenTen التي طبق عليها فحسنت من الأداء[١].

ومع أن موسم مدى يعتمد على المحلل الصرفي (المرجانة) تحليلاً وتوليداً، واعتماد موسم أميرا اعتماد سطحي على الصرف[٢]، إلا أن موسم مدى وإن تقدم بصورة مجملة على أميرا، فهو لم يتقدم عليه في أقسام الكلام اللغوية (أي أقسام الكلام (الاسم ـ الفعل ـ الضمير ...إلخ) بدون علامات الترقيم وما شابهها مما لا يعد كلاما)، وقد يكون لتخصيصه وسوماً للكلمات غير المحللة unanalyzed، وللكلمات اللاتينية latin، والاختصارات abbrev، أثر في الحد من اختلاط نتائجها بغيرها من الوسوم، ومن ثم ضبط الوسوم الأخرى، وهذا ما لم يكن في موسم أميرا (انظر الشكل ٢-٦).

إن هذه الموسمات لم تنجح في تطبيقها على هذه المدونات؛ لأنها لم تراع قواعد العربية، ولم تُدرّب على نصوص متنوعة تغطي عصورا للعربية قديمة وحديثة. فالتوسيم

---

(١) Sketch Engine. Non open source. 7-6-2018: https://the.sketchengine.co.uk/corpus/wordlist?corpname=preloaded/artenten12_stanford;wlmaxitems=1000;wlattr=doc.urldomain;wlminfreq=1;include_nonwords=1;wlsort=f;wlnums=docf

(٢) Automatic tagging of Arabic text: From raw text to base phrase chunks, 2004, pp. 149-152

النحوي الآلي القائم حالياً في العربية تتخلله مشكلات مختلفة على عدة مستويات هي:

١. أن ثمة اضطرابا في تقسيم الكلام عند النحاة أنفسهم قديمهم وحديثهم.

٢. أن معظم مجموعات الوسوم المطبقة، هي محاولات لاستيعاب العربية في إطار الإنجليزية؛ إذ اعتمدت على مجموعة وسوم باكولتر غير الملائمة للعربية، ومن ثم كثُرت فيها الأخطاء اللغوية.

٣. ما يوافق التقسيم العربي التقليدي للكلام منها غير صالح لمعالجة النصوص حاسوبيا لافتقاره الدقة ولاستبعاده بعض الأقسام التي تؤدي إلى خلط بعض الأقسام ببعضها.

٤. بعض المجموعات ضخمة جدا، وإن كانت صالحة نظريا فهي غير قابلة للتطبيق عمليا.

٥. لا توفر هذه المجموعات ملفا توثيقيا أو دليلا يُشرح فيه كل ما يتعلق بالوسوم.

٦. بعض الأقسام الكلامية الموجودة في هذه الوسوم لا وجود له في العربية، مثل: الأسماء المكممة. وقد وسم بها في مدى ما يعد جزءا من كلّ من الأسماء، نحو: نصف وربع وبضع، ومن الصفات، نحو: قليل وكثير.

٧. لم تشر كل مجموعات الوسوم النحوية إلى جموع التكسير واكتفت بجمعي المذكر والمؤنث السالمين.

ووجود مجموعة من الوسوم النحوية تنطلق من العربية ولا تستل من لغة أخرى يضبط عملية التوسيم. بالإضافة إلى أن توفر مدونة مكونة من نصوص حديثة وقديمة، ومن عدة أوعية مختلفة موسمة يدويا، يمكِّن الموسم من التعرف بدقة على الوسوم عندما يطبق على مجال من مجالاتها.

الفصل الثالث
نظام التوسيم النحوي الآلي المقترح وتطبيقاته

٣-١ تحديد الغرض من النظام التوسيمي النحوي المقترح:

يعد التوسيم النحوي أحد مكونات أنظمة معالجة اللغات الطبيعية لأي لغة وإحدى أدواتها. ولكن اهتمام الباحثين العرب بالمدونات اللغوية لا يكاد يذكر أمام التقدم الذي أحرزته في اللغات الأخرى وعلى رأسها الإنجليزية. فقد مضى أكثر من نصف قرن على إنشاء أول مدونة إنجليزية موسمة نحويا، وتوالت بعدها الجهود في اللغات الأخرى، ولكن العربية ورغم ما يبذل ما زالت تفتقر لمدونة موسمة نحويا بوسوم منطلقة من علوم اللغة العربية المستقرة والمدروسة على امتداد تاريخ النحو العربي.

ولقد كانت أولى خطوات بناء نظام توسيم نحوي آلي التي أشرت إليها هي تحديد الغرض من التوسيم النحوي. فبحسب الغرض يتحدد نوع الوسوم النحوية وتفاصيلها، إذ إن ثمة أنظمة تبنى لمعالجة أغراض لا تتجاوز الحاجة فيها إلا تعيين الاسم والفعل والحرف، وهناك أنظمة تبنى لحاجات أدق وأعمق تستلزم تحديد خصائص لغوية أخرى كالنوع والعدد أو تحديد نوع الفعل مثلا من حيث زمنه الصرفي (مضارع ـ أمر ـ ماض). والنظام هنا سيكون للتوسيم النحوي وليس للتوسيم الصرفي النحوي، ولذلك ستكون عمليات التقطيع فيها مبنية على ما يحتاجه الوسم النحوي وليس على ما تتألف منه الوحدة النحوية.

ولعدم وجود مدونة عربية ممتدة عبر أزمنة العربية وأمكنتها وموضوعاتها وموسمة نحويا من قبل لغويين متخصصين بما يتوافق مع نحو العربية، فإنه لا يوجد حاليا نظام آلي للتوسيم النحوي العربي يلبي حاجة الدراسات العربية المعتمدة على المدونات ويقبله المتخصصون في العربية، بحيث يكون مبنيا على المعرفة بنحو العربية. ومن ثم جاء الغرض من التوسيم هنا، وهو: بناء نظام آلي للتوسيم النحوي ينطلق من نحو العربية ويؤسس على معرفة بعلومها، ويكون خاصا بالمستوى الفصيح من العربية على امتداد عصورها وأمكنتها وموضوعاتها، حتى يتمكن الباحث من استعماله على نصوص مختلفة ومتنوعة دون أن تتأثر دقته؛ إذ إن دقة الوسوم النحوية تتناقص إذا طبقت على مدونات ذات طبيعة مختلفة عن طبيعة ما درب عليه الموسم النحوي كأن يختلف زمانها ومكانها وموضوعاتها. وسأضع تقسيمات تمام حسان السباعية تحت النظر لمعرفة مدى ملاءمتها وفائدتها العملية رغم ما وجه لها من انتقادات سأعمل ببعضها.

٣-٢ تصميم وبناء مدونة النظام:

يهدف الكتاب إلى تقديم مقترح للوسوم النحوية يطبق على العربية الفصحى بامتداد عصورها وتنوع أوعية نصوصها. والمدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية المعروفة بالمدونة العربية هي المدونة الملائمة لهذا العمل؛ وذلك لتنوع أوعية نصوصها (انظر الجدول ٣-١) الممتدة زمنيا من العصر الجاهلي إلى العصر الحديث (انظر الجدول ٣-٢)، والمتضمنة نصوصا منشورة في مختلف مناطق العالم، ولتخصصها بمستوى لغوي واحد هو اللغة العربية الفصيحة. ولكن بحجمها الذي يفوق البليون كلمة يستحيل أن تتم معه عملية التوسيم يدوياً أو حتى شبه يدوي، ولذا فإنه ليس أمامي إلا استخلاص مدونة فرعية قوامها مليون كلمة من المدونة العربية أسترشد فيها بإطار المدونة العربية النموذجي لتكون نموذجاً ومرجعاً يتخذ في استخلاص المعلومات وتطوير وتحسين الأدوات والبرامج المصممة للعربية.

جدول (٣-١) الإطار النموذجي للمدونة اللغوية العربية

الأوعية	عدد النصوص	عدد الكلمات	النسبة
الإصدارات الرسمية	684	7,216,593	0.61%
الإنترنت	35,534	18,484,346	1.56%
الدوريات المحكمة	3,173	27,494,533	2.33%
الرسائل الجامعية	1,504	36,296,019	3.07%
الصحف	1,067,404	446,170,582	37.73%
الكتب	1,424	167,706,638	14.18%
المجلات	187,724	151,987,925	12.85%
المخطوطات المحققة	2,451	301,566,866	25.50%
المناهج الدراسية	1,237	15,176,879	1.28%
وكالات الأنباء	22,050	10,415,252	0.88%
المجموع	1,323,185	1,182,515,633	100.00%

وبالاعتماد على الإطار النموذجي للمدونة اللغوية العربية الموضح في الجدول (٣-١)، ووفقا للغرض الذي بنيت من أجله المدونة، ستستخلص المدونة الفرعية بنفس خصائص المدونة الأم (المدونة العربية) أي بالنسب والتناسب من الأوعية وبمراعاة الخط الزمني والجغرافي، كما يظهر في الجدول (٣-٢). وحيث إن استعمال النسب والتناسب يتطلب معلومات متكاملة عن التوزيع داخل المجالات والأوعية وهي غير متوفرة في المدونة العربية لتعلقها بمسائل حقوق الملكية الفكرية، فضلا عن تعذر تقطيع وتوسيم مليون كلمة من قبل شخص واحد في فترة محدودة؛ سأكتفي بمدونة من ١٠ آلاف كلمة أقوم بجمعها وفق الإطار النموذجي للمدونة اللغوية العربية وطبيعة نصوصها، وسأوثق

خلالها كل التجارب والتحديات والتساؤلات التي ستكون دليلا في توسيم المليون كلمة (انظر الجدول ٣-٣).

راعيت أن تكون نصوص المدونة نصوصا لا تخضع لقانون الحماية الفكرية، أو نصوصا تضمن في المدونة بطريقة مستثناة من نظام الحماية الفكرية، فلا تضمن في وعاء الكتب مثلا مؤلفات منشورة كاملة، بل نصوصا مقتطعة لا تتجاوز عددا محدودا من الكلمات.

لقد قمت بفهرسة محتويات المدونة في ملف اكسل واحد تضمن ١٠ أوراق عمل، خصصت كل ورقة عمل لوعاء واحد في المدونة، وتضمنت كل ورقة عمل داخل ملف الفهرسة ٨ أعمدة، هي: (عنوان النص ـ المؤلف ـ الرابط الإلكتروني إن وجد ـ عدد الكلمات في كل ملف نصي ـ اسم النص ـ الزمن ـ الموضوع ـ البلد). فإذا كان عدد الملفات النصية المقرر في وعاء الكتب ٢٨ ملفا نصيا، فستكون صفوف هذه الأعمدة ٢٨ (انظر الشكل ٣-١).

وبعد أن تشكل الإطار العام للمدونة شرعت في عملية الجمع. وقد سهل عملية الجمع اعتمادي على مصادر إلكترونية تقتطع منها النصوص، ثم أقوم بمراجعتها إملائيا والتأكد من عدد كلماتها المطابق للمعايير التي قررتها فيما يتعلق بحجم النص. أما من حيث الصعوبات فإن أكثر ما أعاق عملية جمع المدونة هو قلة المحتوى العربي في بعض البلاد العربية كموريتانيا، أو عدم سهولة الوصول إليه، كان يكون من مواقع محجوبة أو بنسخ غير قابلة للمعالجة الآلية، كالنسخ المحمية أو نسخ النصوص ذات الصيغة pdf.

لقد حرصت على أن يكون ترميز النصوص متوافقا مع أنظمة تشغيل متعددة، ومع تطبيقات معالجة النصوص العربية، فوحّدت ترميز الملفات النصية ليكون UTF-8 بدلا من الترميز الافتراضي ANSI للبرنامج المستعمل في **المفكرة** NotePad.

جدول (٣-٢) الإطار الزمني للمدونة اللغوية العربية

الفترة الزمنية	عدد النصوص	عدد الكلمات	النسبة
0-600	41	259,925	0.02%
601-700	24	679,991	0.06%
701-800	56	1,706,032	0.14%
801-900	148	12,993,528	1.10%
901-1000	138	11,372,816	0.96%
1001-1100	173	12,095,781	1.02%
1101-1200	138	17,531,348	1.48%
1201-1300	314	37,886,100	3.20%
1301-1400	549	62,448,895	5.28%
1401-1500	405	46,138,884	3.90%
1501-1600	126	23,169,794	1.96%
1601-1700	117	17,541,585	1.48%
1701-1800	65	14,693,982	1.24%
1801-1900	306	36,139,232	3.06%
1981-1990	185	27,730,883	2.35%
1901-1980	1,946	41,311,818	3.49%
2001-2010	904,902	563,472,048	47.65%
1991-2000	32,105	83,635,704	7.07%
2011-2020	381,447	171,707,287	14.52%
المجموع	1,323,185	1,182,515,633	100.00%

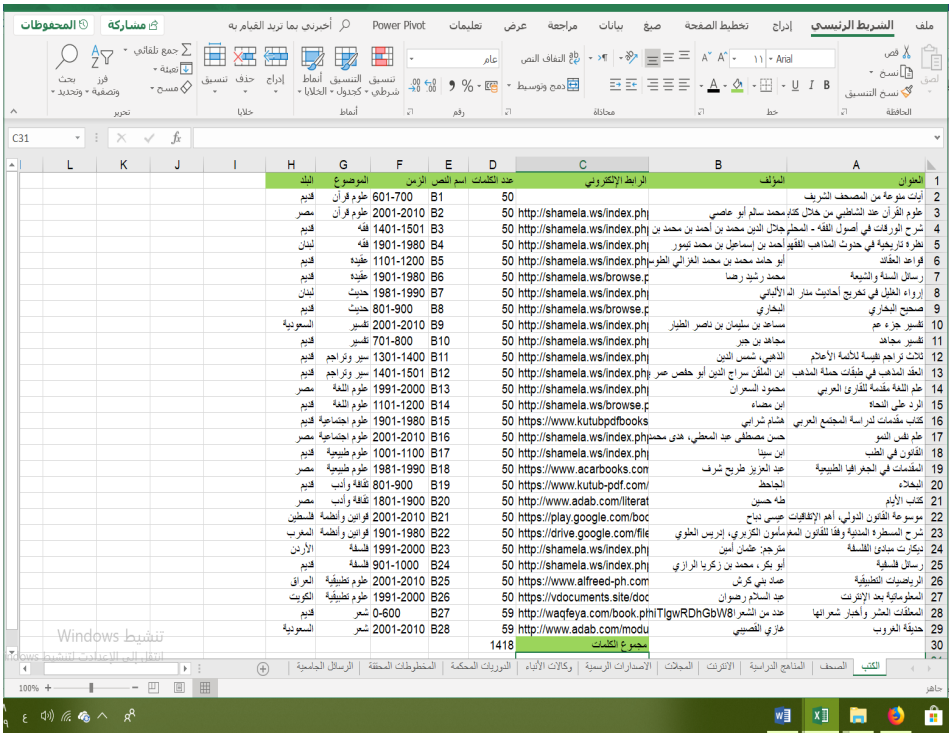
جدول (٣-٣) عدد الكلمات وتكرارها النسبي من كل وعاء

الأوعية	عدد الكلمات	النسبة	لمليون كلمة	لعشرة آلاف كلمة
الصحف	446,170,582	37.730629	377,306	3773
المخطوطات المحققة	301,566,866	25.502146	255,021	2550
الكتب	167,706,638	14.182192	141,822	1418
المجلات	151,987,925	12.852932	128,529	1285
الرسائل الجامعية	36,296,019	3.06939	30,694	307
الدوريات المحكمة	27,494,533	2.325088	23,251	233
الإنترنت	18,484,346	1.563138	15,631	156
المناهج الدراسية	15,176,879	1.28344	12,834	128
وكالات الأنباء	10,415,252	0.880771	8,808	88
الإصدارات الرسمية	7,216,593	0.610275	6,103	61
المجموع	1,182,515,633	100	1,000,000	10,000

وضعت المدونة في مجلد أسميته (المدونة الخام)، وقسمته لعشرة مجلدات سمتها بأسماء الأوعية العشرة للمدونة، وهي: (الصحف ـ المخطوطات المحققة ـ الكتب ـ المجلات ـ الرسائل الجامعية ـ الدوريات المحكمة ـ الإنترنت ـ المناهج الدراسية ـ وكالات الأنباء ـ الإصدارات الرسمية). وتضمن كل وعاء ملفات نصية بصيغة txt بلغ عددها ١٨٨ ملفا، يختلف عددها داخل الوعاء الواحد حسب وزنه في المدونة (انظر الشكل ٣-٣). وحيث إن الحيز الأكبر كان لوعاء الصحف جاءت الملفات النصية فيه ١١٤ ملفا نصيا، أما الأقل فهو للإصدارات الرسمية إذ تضمن ملفا نصيا واحدا، فيما توزعت باقي الملفات على باقي الأوعية. وقد راعيت أن تكون تسمية ملفات كل وعاء تبدأ بحرف يشير للوعاء نفسه ويليه ترتيبه في سلسلة الملفات، (انظر الشكل ٣-٢). ويبين

الجدول (٣-٤) عدد الملفات النصية في كل وعاء، وكذلك عدد الكلمات في مجموع الملفات النصية داخل كل وعاء مع رموز الأوعية التي ضُمنت داخل ملفاتها.

شكل (٣-١) فهرس محتويات مدونة النظام لوعاء الكتب

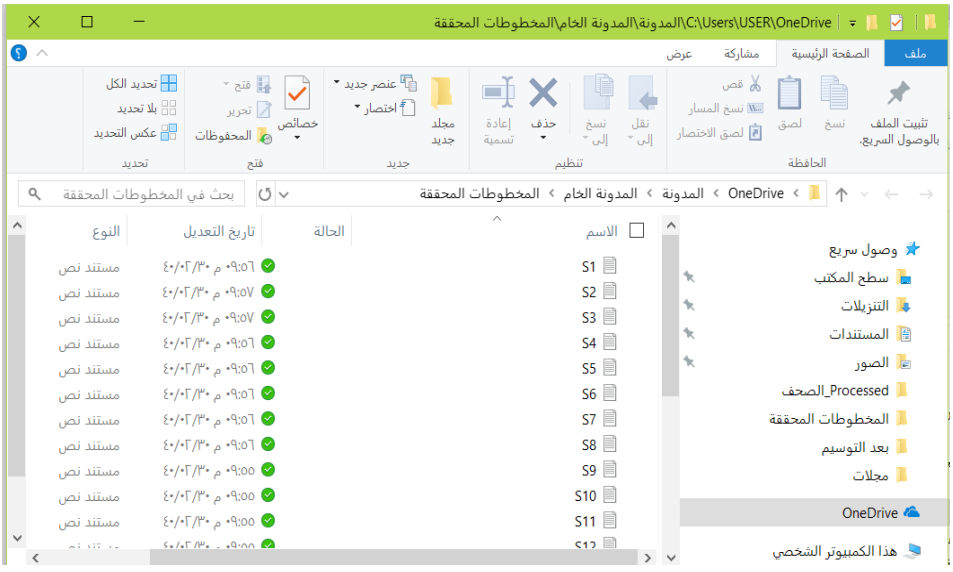


جدول (٣-٤) عدد الملفات النصية وعدد الكلمات في مجموع الأوعية مع رموزها

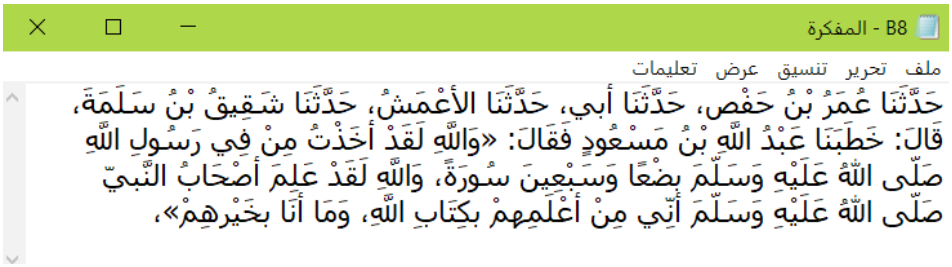
الأوعية	رمز الوعاء	عدد الملفات النصية	عدد الكلمات
الصحف	PR	114	3773
المخطوطات المحققة	S	14	2550
الكتب	B	28	1418
المجلات	M	12	1285
الرسائل الجامعية	R	3	307

الأوعية	رمز الوعاء	عدد الملفات النصية	عدد الكلمات
الدوريات المحكمة	A	4	233
الإنترنت	W	3	156
المناهج الدراسية	C	6	128
وكالات الأنباء	P	3	88
الإصدارات الرسمية	K	1	61
المجموع		188	10,000

شكل (٣-٢) أسماء ملفات محتوى وعاء المخطوطات المحققة



شكل (٣-٣) عينة من ملف نصي خام من ملفات المدونة في وعاء الكتب



٣-٣ المعالجة القبلية (تحديد متغيرات التفريق):

لقد ميزت في الفصل السابق بين **التفريق** tokenization **والتقطيع** segmentation، فالأول يتعلق بفصل الكلمات بعضها عن بعضها، وفصل العلامات والرموز عنها ويدخل فيه **التسوية الهجائية** Orthographic Normalization التي تزال فيها غالبا الكشيدة وعلامات التشكيل، وتسوى فيها الهمزات والياءات أحيانا، أما الثاني فيتعلق بفصل ما يستدعيه الوسم النحوي عن الكلمة من سوابق ولواحق بصرف النظر عن تركيب الكلمة.

وفيما يتعلق بالتفريق (المعالجة القبلية)، عالجت نص المدونة وفق ما يلي من الخطوات:

١- تفريق الكلمات باعتبار المسافة السابقة واللاحقة لها، فكل **سلسلة حرفية** String في النص يسبقها فراغ ويتلوها فراغ تعد كلمة. وجملة مثل: (محمد بن سلمان: مثلث الشر: إيران، والإخوان، وجماعات التطرف)، تفرق على صورتها الكتابية: (محمد/ بن/ سلمان:/ مثلث/ الشر:/ إيران،/ والإخوان،/ وجماعات/ التطرف)/ . وقد استعنت بمعالج الكلمات WORD لتعديل المسافات المضاعفة لمسافة واحدة، ووضع كل كلمة في سطر بطريقة آلية عبر خاصية البحث والاستبدال، حيث تبحث عن المسافتين وأكثر وتحولها إلى مسافة واحدة؛ حتى لا يعتبرهما النظام حيزا لكلمة، وبعد تقليص المسافات تُحوّل كل مسافة إلى سطر جديد بالبحث عن المسافة الواحدة ثم تحويلها إلى سطر جديد باستعمال الرمز p^ الذي يعني في معالج الكلمات WORD سطرا جديدا.

٢- فصل الرموز بأشكالها (@#$٪^&*><×÷) وعلامات الترقيم بأنواعها (. ! ؟ , ": ؛ ، - ( )[ ] ~) والأرقام المفردة والمركبة (١٢٣٤٥...- ٤٥ ٤٨ ...) عن الكلمات التي تتصل بها وفيما بينها، فالجملة: (نسبة الطلاب الناجحين في الرياضيات، والعلوم،

والحاسوب لا تتجاوز: ٧٠٪.)، تفرق على الصورة التالية: (نسبة/ الطلاب/ الناجحين/ في/ الرياضيات/ ، / والعلوم/ ،/ والحاسوب/ لا/ تتجاوز/ :/ ٧٠ / ٪ / .). وهنا استعملت برنامج المشذب العربي[١] لتفريق الكلمات عن الأرقام وعلامات الترقيم والرموز آليا دون تدخل مني.

٣- التسوية الهجائية، وهي على مرحلتين:

١- التسوية قبل التقطيع: أزيلت فيها علامات التشكيل، والكشيدة عن الكلمات، وسوّيت الأرقام بتحويلها من العربية للهندية. ونفذت في نفس المرحلة التي فصلت فيها الكلمات عن الأرقام والرموز بالاستعانة ببرنامج المشذب العربي. ولم تسوَّ أو توحّد الأشكال الكتابية للهمزات والياءات؛ لأن الغرض من التوسيم هنا يركز على المستوى النحوي فقط وليس الصرفي، وتسويتها تزيد من غموض الكلمة، فكلمة مثل (استعلم) تختلف عن (أستعلم) من حيث النوع، ولو سويت الهمزتان وحوِّلتا إلى همزة وصل تطابقت الكلمتان، ولن يتمكن النظام من التفريق بينهما.

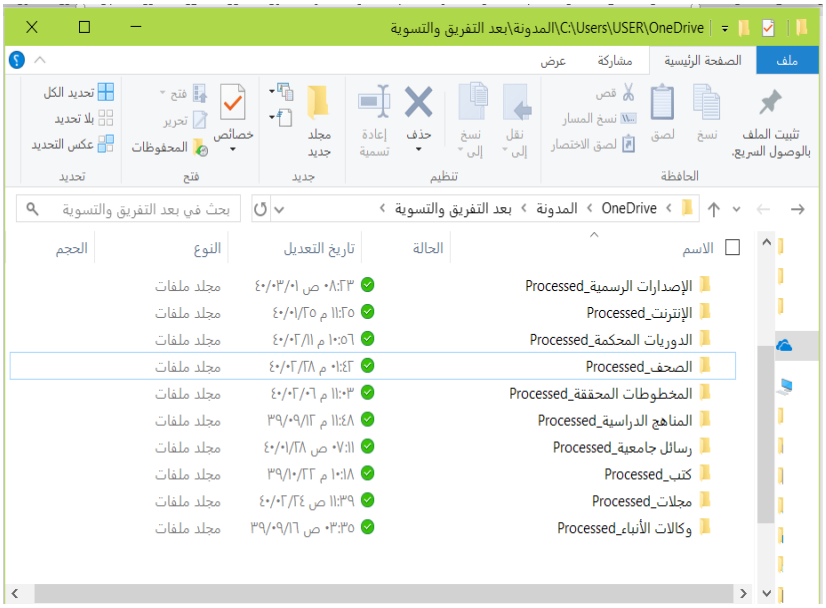
ب ـ التسوية بعد التقطيع: وستناقش بعد تحديد متغيرات التقطيع.

وبعد انتهاء إجراءات التفريق أصبحت لدي نصوص خام فرقت كلماتها ووضعت كل كلمة مفروقة في سطر مستقل مع بقائها في ملفات مستقلة الوعاء والموضوع. ومن ثم أنشأت مجلدا ثانيا أسميته (بعد التفريق والتسوية)، وجمعت فيه كامل ملفات المدونة الخام بعد تفريقها وتسويتها الأولى ونقلتها لجداول أكسل مُسلسلة (انظر الشكل ٣-٤ والشكل ٣-٥).

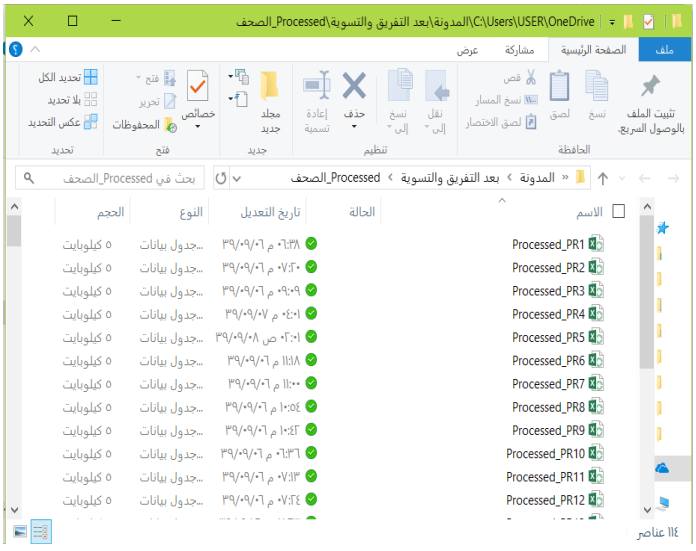
---

(١) أداة للمعالجة القبلية (preprocessing) لنصوص المدونات اللغوية العربية أو أنظمة التنقيب في النصوص العربية.

شكل (٣-٤) مجلدات المدونة بعد التفريق مصنفة حسب الأوعية (تصنيف المدونة الخام)



شكل (٣-٥) الملفات النصية المفرقة الكلمات لمجلد الصحف



٣-٤ تحديد متغيرات التقطيع:

بعد الانتهاء من المعالجة القبلية (التفريق) التي تدخل فيها عملية التسوية الهجائية قبل التقطيع، بدأ العمل يدويا على فصل متغيرات التقطيع. وقد حددت متغيرات التقطيع التي تقتضيها عملية التوسيم النحوي، ولا توجد إلا في النوع الثاني من الكلمات التي تُعد عدولاً عن العلاقة الافتراضية بين الكلمات الفعلية الصرف ـ نحوية والكلمات الهجائية، وهي **اللصقيات** Mergers. وحيث إن الخصائص الشكلية للكلمات اللصقية ليست معبرة على الدوام عن معان وظيفية، لم أفصل التاء في طلحة وحقيبة، فهي ليست للتأنيث وفصلها لن يمكننا من تحديد نوع الكلمة. وفي المقابل يوجد لواصق تفيد في تحديد وسوم أكثر دقة[١] لم أفصلها. وأوسع هذه اللواصق معاني هي الضمائر الشخصية المتصلة ويستفاد منها في تحديد خصائص تصريفية مختلفة للكلمة، كالشخص (المتكلم ـ المخاطب ـ الغائب)، والعدد (الإفراد ـ التثنية ـ الجمع)، والنوع (التذكير ـ التأنيث)[٢]. وأضيقها مجالًا (أل التعريف) وتفيد في تحديد المعرفة والنكرة، والتعريف «كظاهرة عامة... أوسع من أن يقتصر على دلالة (أل) بمفردها»؛ إذ إن عدم اقتران الكلمة بأل لا يعني أنها نكرة[٣]. بالإضافة إلى حروف المضارعة التي لن نتمكن بفصلها من تحديد نوع الكلمة، ولواصق أخرى كحرفي التوكيد (النونين، الخفيفة والثقيلة: نْ، نّ) وياء النسب وألف الإطلاق، وألف التنوين التي تزيد تعقيد عمليتي التقطيع والتوسيم. وقد عمدت إلى تلك القرارات؛ لأن التوسيم النحوي يقضي بأن يكون كل متغير من متغيرات التقطيع عبارة عن مبنى تقسيمي، وما سبق مما قررت عدم فصله هو مبان تصريفية لا تقسيمية. ولذلك قررت فصل اللواصق التي تعد مباني تقسيمية وهي متميزة دائما بالصدارة في الكلمة. وفي الجدول (٣-٥) تفصيل لهذه اللواصق (متغيرات التقطيع) التي قررت فصلها قبل البدء بعملية التوسيم لهذا النوع من الكلمات.

أنشأت مجلدا ثالثا بإزاء مجلد (المدونة الخام)، ومجلد المدونة (بعد التفريق

---

(١) مقدمة في المعالجة الطبيعية للغة العربية، ٢٠١٤، ص. ٨١

(٢) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ١٥٩

(٣) السابق، ص. ١٥٩

والتسوية) أسميته (بعد التقطيع)، واستعملت فيه ملفات المدونة في مجلد التفريق والتسوية بعد إضافة عمود آخر بمحاذاة عمود الكلمات (بعد التفريق) وسمته (بعد التقطيع)، وعرضت فيه الكلمات المفرقة نفسها التي تحتاج للتقطيع مقطعة، فكلمة (والعشرون) المفروقة عن النقطتين الرأسيتين (:)، قطعت كما في الشكل (٣-٦) إلى: (و العشرون).

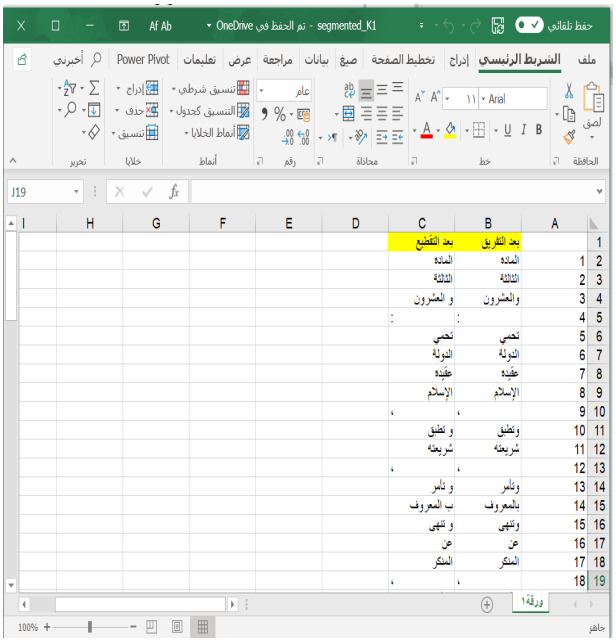
جدول (٣-٥) متغيرات التقطيع التي بني عليها التوسيم النحوي

المتغير	مثاله
همزة الاستفهام والتسوية (أ)	أيرضيك ذلك؟ـ«وسواء عليهم أأنذرتهم أم لم تنذرهم»
حرف الجر والقسم (ب)	ببيتك ـ بالله
تاء القسم (ت)	تالله
حرف الاستقبال (س)	سيزور
عن المدغمة نونها في ميم من وما (ع)	عمّن ـ عمّا ـ عم
فاء الشرطية والسببية والاستئنافية والعاطفة (ف)	إذا مررت فسلمْ عليهم ـ ولا يؤذن لهم فيعتذرون ـ فتعالى الله عما يشركون ـ دخل محمد فخالد
حرف الجر (ك)	كالقمر
حرف الجر ولام الأمر ولام التأكيد (ل)	للمنزل ـ لينته عنه ـ إن في ذلك لعبرة
من المدغمة نونها في ميم من وما (م)	ممن ـ مما
هاء التنبيه (ها ـ هـ)	أيها ـ هأنت
حرف العطف والمعية والحال والقسم (و)	كبر وعظم ـ سرت والليل ـ مضيت وأنا سعيد ـ والله
من الجارة المدغمة ميمها في ميم الاستفهام (م)	ممّ
في الجارة المتصلة بمن (في)	فيمن
حروف الجر المتصلة بما الاستفهامية المحذوفة الألف (ب ـ في ـ ل ـ على ـ إلى ـ حتى)	بم ـ فيم ـ لم ـ علام ـ إلام ـ حتام
إذ الظرفية المتصلة ببعض الكلمات	حينئذ ـ عندئذ

٣-٤-١ التسوية بعد التقطيع:

أشرت في المبحث السابق إلى أن نص المدونة الخام عولج قبليا على ثلاثة مراحل. وكانت التسوية الهجائية قد أتت على مرحلتين تحدثت عن المرحلة الأولى فيها في سياقها. أما المرحلة الثانية البعدية من التسوية الهجائية فقد ظهرت الحاجة إليها في هذه المرحلة بسبب طبيعة الكلمات العربية التي يتغير شكلها الكتابي بعد فصلها. فكلمة مثل: ممّن، أصبحت بعد التقطيع (م من) وكان لا بد من إعادة (م) لشكلها الطبيعي قبل إدغامها في كلمة (من) لتكون كلمتين منفصلتين ومستقلتين، على النحو التالي: (مِنْ مَنْ)، وكلمات أخرى أيضا، كالكلمات المعرفة بأل التي تتصل بها الأداة (ل)، نحو: للأيام وحذفت ألفها.

شكل (٣-٦) مثال من ملفات مرحلة تقطيع نصوص المدونة الخام بعد تفريقها



٣-٥ تحديد الوسوم النحوية:

أسعى في مجموعة الوسوم النحوية المقترح تطبيقها على مدونة النظام إلى مراعاتها

معايير EAGLES في شكلها ومحتواها. وقد اعتمدت في مجموعة الوسوم المقترحة هنا على تقسيم تمام حسان السباعي لأقسام الكلام؛ حيث يسير هذا التقسيم وفق معطيات المنهج الوصفي وتطبيقا له، كما أن تقسيمه تقسيم منطقي قابل للتمثيل الهيكلي ويمكن من خلاله استقراء جوانب الكلمات، واستيعابها بحيث لا تفلت أي منها من أيّ قسم من أقسام الكلمة، فضلا عن إمكانية تكييفه مع تعليمات مبادرة EAGLES الساعية لتوحيد معايير محتوى التوسيم النحوي. وليس هذا غريبا بحكم تأخره التأريخي، واختلاف انتماء لسانه المتأثر بأستاذه **فيرث** Firth إذ لا يمكن تكييف التقسيم الثلاثي مع تعليمات مبادرة EAGLES لأنه لا يميز بين الأسماء والضمائر مثلا. وحيث إن تقسيمه أيضا لم ينل حظه من التطبيق عمليا كما حدث للتقسيم الثلاثي التقليدي رأيت الاستناد إليه.

لقد خصص تمام حسان في كتابه (اللغة العربية، معناها ومبناها) فصلا لأقسام الكلام، وجعلها سبعة أقسام، هي: الاسم والصفة، والفعل، والضمير، والخالفة، والظرف، والأداة. وقد جعل هذا التقسيم على أساس المبنى وأساس المعنى معا. وضم المبنى فوارق صورية هي: (الإعراب، والرتبة، والصيغة، والجدول، والالصاق والتضام والرسم الإملائي)، أما المعنى فقد ضم فوارق معنوية هي: (التسمية، والحدث، والزمن، والتعليق، والمعنى الجملي)، وليس هذه الكتاب مجال التفصيل فيها.

وقد عملت على هذا التقسيم بوصفه تقسيما أساسيا بصورة تامة أما في التقسيمات الفرعية فستبني على ما بنى عليه، ولكن مع تغيير في بعض ما تستدعيه الآلة ابتداء ولا يخل بتقسيمه بل يضيف إليه. ومن ثم فإن مجموعة الوسوم النحوية في هذا الكتاب أتت على مستويين حددتهما EAGLES ويمكن توسيعها وتقليصها:

١- الأقسام الأساسية (القيم الإجبارية)، وهي محددات نحوية عبارة عن مبان تقسيمية تأتي في صدر السلسلة التحليلية في الوسم، وهي سبعة: الأسماء ـ الأفعال ـ الصفات ـ الضمائر ـ الظروف ـ الأدوات ـ الخوالف، بالإضافة إلى التقسيمات غير

اللغوية التي يستلزمها التحليل النصي الحاسوبي، وهي: علامات الترقيم ـ الاختصارات ـ الكلمات الأجنبية ـ الرموز (انظر الجدول ٣-٦).

٢- الأقسام الفرعية (القيم المستحسنة)، وهي محددات دلالية عبارة عن معان للمباني التقسيمية، وسمات دلالية من خلال المباني التصريفية. فالمحددات الدلالية في الأسماء: اسم ذات، واسم معنى، واسم مكان، واسم زمان، واسم آلة، واسم جنس، واسم مبهم. وفي الأفعال: ماض، وحاضر، وأمر. وفي الصفات: صيغة فاعل، وصيغة مفعول، والصفة المشبهة، وأفعل التفضيل، وصيغة المبالغة. وفي الضمائر: الضمائر الشخصية، والإشارة، والموصولات. وفي الظروف: ظرف زمان، وظرف مكان. وفي الخوالف: خالفة تعجب، ومدح، وذم، وإخالة، وصوت. وحيث إن الأدوات في العربية متعددة، وتتضمن عند تمام حسان ما ليس من حروف المعاني، وليس هناك دراسات تكفي لاتخاذ القرارات بشأن ما إذا كان بعضها محولا أو أصليا، عمدت إلى إبقائها دون التقسيم الذي جاء به تمام حسان كتقسيم فرعي. ويمكن العمل عليها خارج إطار هذا الكتاب كدراسة متممة وتوسيعها لمحددات دلالية أكثر وأدق (انظر الجدول ٣-٧).

جدول (٣-٦) أقسام الكلام الأساسية ووسومها

القسم الكلامي الأساسي	وسمه
الاسم	N
الفعل	V
الصفة	A
الضمير	P
الأداة	RP
الخالفة	I
الظروف	D

أما السمات الدلالية فهي معاني المباني التصريفية. وتتعدد معاني المباني التصريفية في الكلمة العربية لكني ارتكزت على بعض المعاني أو الخصائص التصريفية التي أرى ضرورتها للمضي في عملية توسيم واضحة المنهج للقارئ. هذه الخصائص التصريفية هي:

١- التعريف بأل، وتخص: (الأسماء).

٢- الشخص، وتخص: (الأفعال).

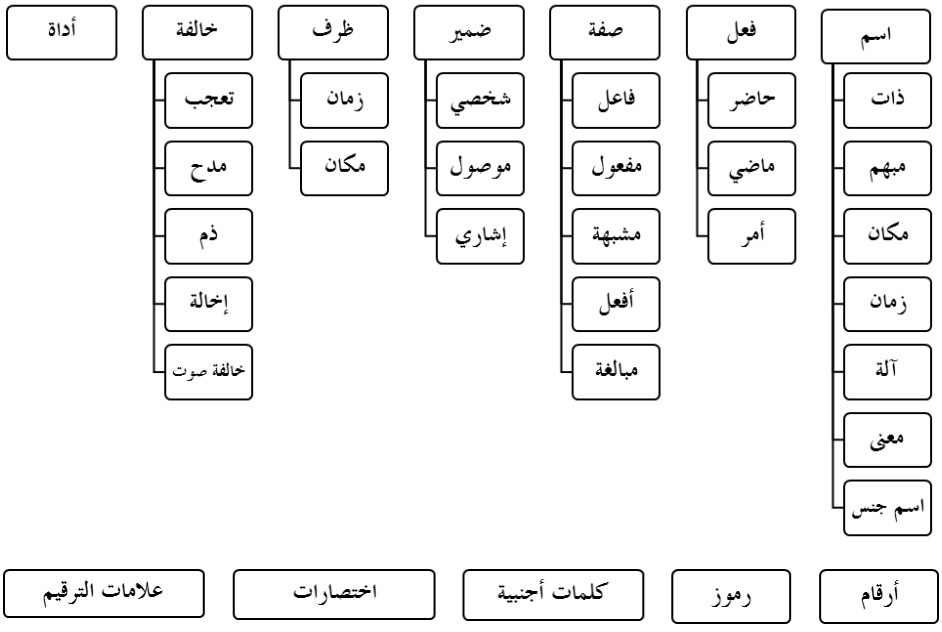
٣- العدد والجنس، وتخصان: (الأسماء والأفعال والصفات والضمائر).

٤- البناء للمجهول، وتخص: (الأفعال).

٥- النسب، وتخص: (الأسماء).

٦- الربط، وتخص: (الأسماء والأفعال والصفات).

شكل (٣-٧) أقسام الكلام الفرعية لمجموعة الوسوم النحوية المقترحة



كما يمكن أن يضاف إليها أكثر مما اكتفيت به، **كالحالة الإعرابية** case/mood في الأسماء والأفعال والصفات، **والجهة** aspect في الأفعال. وسآتي بالتفصيل على ما ارتكزت عليه من خصائص فيما يلي من المباحث. ومن خلال الشكل (٣-٧) والجدول (٣-٨)، يتضح التقسيم الأساسي والفرعي والخصائص التي اعتمدتها بالاستناد إلى تقسيم تمام حسان السباعي للكلام.

جدول (٣-٧) وسوم الأقسام الفرعية

القسم الكلامي الأساسي	وسمه	القسم الكلامي الفرعي	وسمه
الاسم	N	اسم ذات	NC
		اسم معنى	NA
		اسم مكان	NL
		اسم زمان	NT
		اسم آلة	NM
		اسم جنس	NV
		اسم مبهم	NI
الفعل	V	فعل ماض	VS
		حاضر	VP
		فعل أمر	VC
الصفة	A	صفة فاعل	AS
		صفة مفعول	AO
		صيغة مبالغة	AE
		صفة مشبهة	AA
		أفعل التفضيل	AC

القسم الكلامي الأساسي	وسمه	القسم الكلامي الفرعي	وسمه
الضمير	P	ضمائر شخصية	PP
		ضمائر موصولة	PR
		ضمائر إشارية	PD
الخالفة	I	إخالة	IV
		خالفة صوت	IS
		خالفة تعجب	IE
		خالفة ذم	IX
		خالفة مدح	IG
الظرف	D	ظرف زمان	DT
		ظرف مكان	DL
الأداة	RP		

أولا: الأقسام الرئيسة والفرعية:

١ – الاسم:

وهو كل ما دل على مسمى خال من المعنى الزمني[١]، ويدخل فيه:

أ ـ اسم الذات أو الجثة أو المعين: وهو ما دل على مسمى معين، مثل الأعلام نحو: محمد ـ صالح ـ عبد الله، والأعراض المختلفة، كالألوان، نحو: أحمر ـ أخضر، والأجسام، نحو: شمس ـ قمر ...إلخ.

ب ـ اسم الحدث أو المعنى: وهو ما دل على الحدث أو نوعه أو عدده، وتدخل فيه المصادر وأسماء المصادر وأسماء المرة وأسماء الهيئة.

---

(١) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ٩٠

ج ـ اسم جنس: ويتضمن اسم الجنس الجمعي نحو: عرب وترك، واسم الجمع، نحو: إبل ونساء.

د ـ الميميات، وهي المشتقات المبدوءة بميم زائدة، وهي:

- اسم الزمان، نحو: مولد ومنطلق.
- اسم المكان، نحو: مجلس ومخرن.
- اسم الآلة، نحو: مفتاح وغسالة.

هـ ـ اسم مبهم: ويدخل فيه كل ما لا يدل على معين، ويحتاج إلى وصف أو تمييز أو إضافة أو غيرها لبيان مقصوده، نحو: الجهات (شمال ـ يمين)، والأوقات (يوم ـ سنة)، والموازين (طن ـ كيلو)، والمقاييس (بوصة ـ متر)، والمكاييل (أونصة ـ جرام)، والأعداد (خمسة ـ أحد عشر)، ونحوها مثل: كل ـ غير ـ جميع ـ كلتاهما ـ كلاهما ـ مثل ـ نحو.

٢- الفعل:

وهي كل كلمة تدل على الحدث والزمن معا[١]، ويدخل فيه:

أ ـ الفعل الماضي، ويميز بقبوله تاء الفاعل وتاء التأنيث: نحو: أكل ـ ركضتُ ـ سمعتْ.

ب ـ الفعل المضارع، ويميز ببدئه بأحد أحرف المضارعة، وقبوله لام الأمر ونوني التوكيد والنسوة، وقبول مجيئه مقترنا بسين الاستقبال، وسوف ولم ولن، نحو: أسمع ـ ليتأمل ـ يكتبنّ ـ يقرأنَ ـ سيسافر ـ .

ج ـ الفعل الأمر، ويتميز بقبول نوني التوكيد والنسوة فقط من بين كل ما سبق، نحو: اكتب ـ ادرسَنْ ـ اقرأنّ.

٣- الصفة:

وهي كل كلمة تدل على موصوف بالحدث[٢]، وتتضمن الصيغ التالية:

---

(١) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ١٠٤

(٢) السابق، ص. ٩٨

أ ـ صيغة فاعل، وتدل على وصف الفاعل بالحدث، نحو: صادق ـ رائدون ـ مستبِدّ.

ب ـ صيغة مفعول، وتدل على وصف المفعول بالحدث، نحو: مشهور ـ مفتَقد ـ مسرورون.

ج ـ صيغة المبالغة، وتدل على وصف الفاعل بالحدث مبالغةً، نحو: كذاب ـ حدادون.

د ـ صيغة أفعل، وتدل على وصف الفاعل بالحدث تفضيلا، نحو: أكبر ـ فضلى ـ أفاضل.

هـ ـ الصفة المشبهة، وتدل على وصف الفاعل بالحدث على سبيل الدوام، نحو: فرِح ـ حزين ـ عظماء.

٤ – الضمير:

وهو كل كلمة تتضمن مطلق الحاضر أو الغائب وتستعمل عند الاستغناء عن إعادة ذكر الاسم الظاهر[١]، وبالتالي تحمل معنى وظيفيا لا معجميا، وتتضمن ما يلي:

أ ـ الضمائر الشخصية: ولا تتضمن إلا ضمائر الرفع المنفصلة، نحو: هم ـ هن ـ هي ... إلخ. وضمائر النصب المنفصلة، نحو: إياك ـ إياكم ـ إياكن ... إلخ.

ب ـ ضمائر الموصول[٢]: وهي: جميع الموصولات: الذي ـ التي ـ اللائي ... إلخ.

ج ـ ضمائر الإشارة: وهي: هذه ـ هذا ـ هاتان ... إلخ

٥ – الظرف:

وهي كلمات لا صيغة لها، وتدل على معنى صرفي عام هو الظرفية المكانية أو الزمانية[٣]، وتشمل:

---

(١) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ١٠٨

(٢) على مذهب البصريين، لأن الكوفيين يضيفون (أل) للموصولات إذا اتصلت بالصفات، انظر: الزجاجي، أبو القاسم، اللامات في اللغة. تحقيق: مازن المبارك، دار الفكر، دمشق: سوريا، ط٢، ١٩٨٥، صص. ٤٠-٤٥

(٣) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ١١٩

أ ـ ظروف الزمان: ومنها: إذا ـ لما...إلخ.

ب ـ ظروف المكان: ومنها: أين ـ متى ـ حيث ... إلخ.

وتشترك مع الأدوات كثيرا، إلا أن الأدوات تتميز عنها بالصدارة في الجملة. فليست متى في جملة مثل: متى تأتينا؟ ظرفا، وهي ظرف في: آتيك متى تصحو.

٦- الخالفة:

هي كل كلمة تفصح عن موقف انفعالي أو تأثري[١]، وهي أربعة أنواع:

أ ـ خالفة الإخالة (اسم الفعل)، نحو: هيهات ـ آمين ...إلخ.

ب ـ خالفة الصوت (اسم الصوت)، نحو: كخ ـ هش... إلخ.

ج ـ خالفة التعجب (صيغة التعجب)، نحو: أكرِم بـ ـ وأكرَمَ في: ما أكرم.

د ـ خالفتا المدح والذم (فعلي المدح والذم)، وهما: نعم ـ بئس.

٧- الأداة:

وهي كل كلمة تحمل معنى وظيفيا عاما وهو التعليق الذي يتناول التركيب الكلامي ككل وليس الأجزاء التحليلية له[٢]. وهي عند تمام حسان على قسمين لم أعتمدهما ـ كما أسلفت ـ لتضمنها عند تمام حسان ما ليس من حروف المعاني، وحاجتها لدراسة متممة تكفي لاتخاذ القرارات في تقسيماتها الفرعية. وهما:

أ ـ أداة أصلية: والمقصود بها الأدوات ذات المعاني التي استعملت أصلا في معانيها ولم تحول عن غيرها، مثل حروف الجر، وحروف العطف، وأدوات النفي، وأدوات النهي، والنواسخ، وغيرها.

---

(١) اللغة العربية، معناها ومبناها، ١٩٩٤، ص. ١١٣

(٢) الساقي، فاضل. أقسام الكلام العربي من حيث الشكل والوظيفة. القاهرة: مصر، مكتبة الخانجي، ١٩٧٧، ص. ٢٦٤

ب ـ أداة محولة: وهي الأدوات ذات المعاني التي استعملت محولة من معانيها وتعرف بصدارتها، وتكون:

- ظرفية، في مثل: متى هلَّ العيد أزرك ـ أينما سكنت أسكن معك.
- اسمية، في مثل: كيفما تريد أفعل ـ كم ترى قد جمعت.
- فعلية، في مثل: كان وأخواتها ـ كاد وأخواتها.
- ضميرية، في مثل: من يجدّ يجد ـ ما على الرسول إلا البلاغ.

ثانيا: الخصائص التصريفية:

وهي المقولات التي تساعد في تحديد هوية الكلمة، وضبط العلاقات التركيبية بين الكلمات، وتعبر عن المعاني النحوية المتصلة بالكلمة سواء أكانت لواصق أو صيغا[١].

١ - التعريف بأل:

وهو خاص بالأسماء والصفات فقط، فالفرق بين النكرة من الأسماء والصفات في حالة التنكير، وحالتها في التعريف هو وجود (أل) في حالة التعريف، إلا أن التعريف عموما لا يكون بأل وحدها، فقد يحدث بالإضافة والتخصيص ويمكن أن تستقل هاتان الخاصيتان الصرفيتان في مستوى أكثر دقة من مستويات التوسيم النحوي.

٢ - الشخص:

وتختص به الأفعال والضمائر، ويعبر عنه بضمائر الرفع المتصلة في الفعل الماضي، وهي: التاء المتحركة تْ ـ تَ ـ تُ ـ تِ، ونا الدالة على الفاعلين، وواو الجماعة، وألف الاثنين، وياء المخاطبة، ونون النسوة، وقد يستتر الضمير، في مثل: ذهب. وفي الفعل المضارع يعبر عنه بحروف المضارعة، وهي: أ ـ ن ـ ي ـ ت. أما فعل الأمر فلا يأتي إلا لشخص واحد هو

---

(١) عبد الواحد، عبد الحميد. الكلمة في اللسانيات الحديثة. كلية الآداب والعلوم الإنسانية وحدة بحث اللسانيات والنظم المعرفية المتصلة بها، صفاقس: تونس، ٢٠٠٧، صص. ٤٢-٤٣

المخاطب ولا يحتاج للواصق. كما تعبر عنه ضمائر الرفع المنفصلة، نحو: هم ـ هن ...إلخ، وضمائر النصب المنفصلة، نحو: إياك ـ إياكم ...إلخ. ويتعدد الشخص فيكون:

- متكلم.
- مخاطب.
- غائب.

٣- العدد:

وتختص بالأفعال من خلال الضمائر المتصلة، فمع الماضي يظهر العدد في التكلم (تُ ـ نا)، وفي الخطاب (تَ ـ تُم ـ تنَ ـ تما) وفي الغيبة (الاستتار ـ ـوا ـ ـا). أما في المضارع فيظهر العدد في التكلم بحرفي المضارعة (أ ـ ن)، وفي الخطاب بالاستتار وياء المخاطبة وألف الاثنين وواو الجماعة ونون النسوة (ـ ان ـ ا ـ ون ـ ن)، وفي الغيبة بالاستتار وألف الاثنين وواو الجماعة ونون النسوة (ـان ـ ا ـ ون ـ ن)، وفي الأمر يظهر العدد في الخطاب بالاستتار وياء المخاطبة وألف الاثنين وواو الجماعة ونون النسوة (ان ـ ا ـ ون ـ ن) أيضا. كما يختص العدد بالصفات والأسماء من خلال اللواصق (ان ـ ين ـ ون)، والصيغ (جموع التكسير). فيكون:

- مفردا.
- مثنى.
- جمعا.

٤- الجنس:

وتختص به الأسماء والصفات والأفعال. فأما في الأسماء والصفات المفردة فيعبر عنه بتاء التأنيث والألف المقصورة والهمزة الممدودة في التأنيث، وبعدمها في التذكير، وفي الأسماء المجموعة بالألف والتاء للمؤنث، وعلامات أخرى للمذكر. أما في الأفعال فيعبر عنه بتاء التأنيث ونون النسوة وواو الجماعة وألف الاثنين وياء المخاطبة. ويكون جنس الكلمة:

- مؤنثا، نحو: حقيبة ـ حسنى ـ أكلت.

- مذكرا، نحو: مركب ـ قادم ـ قدموا.

٥- البناء للمجهول:

وتختص به الأفعال الماضية والمضارعة، أما الأمر فلا يبنى للمجهول. ويعرف المبني للمجهول بصيغته ففي الماضي يكسر ما قبل آخره ويضم كل متحرك قبله، نحو: أكرم> أُكرِم، وفي المضارع يضم أَوله ويفتح ما قبل آخره، نحو: يكرم> يُكرِم. أَما الأَلف التي قبل الحرف الأخير فتقلب ياءً في الماضي، وأَلفاً في المضارع.

٦- الربط:

ويكون بالضمير المتصل العائد على مذكور في جميع هذه الأقسام الكلامية، وتختص به الأسماء والأفعال والأدوات والإخالة وخالفة التعجب. ويدخل فيه:

- نوع الربط إن كان لمتكلم أو مخاطب أو غائب.
- جنسه إن كان مذكرا أو مؤنثا.
- عدده إن كان مفردا أو مثنى أو جمعا.

أما من حيث تسمية الوسوم فسيراعي الإيجاز والوضوح وقابلية التفكيك، بحيث يكون مسمى الوسم سلسلة من الأحرف يمثل كل واحد منها فرعا مختلفا من الهيكل الشجري، فيمثل كل وسم بحرف يشير للقسم الأساسي يكون أول حرف من المقابل الإنجليزي له، وحرف للقسم الفرعي يكون غالبا أول حرف من المقابل الإنجليزي له، ثم شرطة تحتية تليها رموز الخصائص المتعلقة القسم الكلامي على التوالي، ومن ثم يكون وسم اسم الزمان مثلا المعرف بأل، والمفرد المذكر منه: NT_DSM، حيث ترمز N: noun (اسم) و T: time (زمان) وD: definite (معرف بأل) وS: singular (مفرد) و M: masculine (مذكر)، كما توضح سلسلة الجداول (٣-٩). وقد تعاملت مع **الكلمة المتعددة** multiwords التي تكون فيها الكلمة الفعلية الواحدة أكثر من كلمة على النحو الكتابي من أسماء الدول والمدن والأعلام الشخصية والأعداد المركبة إضافيا والمعطوفة، نحو: الولايات المتحدة

ورأس الخيمة وعبد الله وأحد عشر وتسعة وعشرون باستعمال **الوسوم المزدوجة** ditto tag التي يضاف فيها نفس الوسم لكل كلمات الكلمة المتعددة، وتوضع في نهاية الوسم لكل الكلمات فيها رقمين بعد شرطة تحتية، فيمثل الرقم الأول ترتيب الكلمة داخل الكلمة المتعددة، ويمثل الثاني عدد الكلمات فيها. وبالتالي توسم كلمة عبد الله مثلا باعتبارها كلمة متعددة في مجموعة الوسوم الأساسية المقترحة: عبد N_١٢/ الله N_٢٢.

جدول (٣-٨) الأقسام الرئيسة والفرعية وما يتعلق بها من خصائص تصريفية

(تشير + إلى وجود الخاصية والفراغ إلى عدمها)

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		معرف بأل	غير معرف بأل	مجهول	معلوم	متكلم	مخاطب	غائب	مفرد	مثنى	جمع	مذكر	مؤنث	منسوب	النوع			العدد			الجنس	
															متكلم	مخاطب	غائب	مفرد	مثنى	جمع	مذكر	مؤنث
الاسم	اسم ذات	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم معنى	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم آلة	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم مكان	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم زمان	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم مبهم	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم جنس	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
الفعل	مضارع			+			+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+
	ماضي			+			+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+
	أمر						+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		غير معرف بأل	معرف بأل	مجهول	معلوم	متكلم	مخاطب	غائب	مفرد	مثنى	جمع	مذكر	مؤنث	منسوب	النوع			العدد			الجنس	
															متكلم	مخاطب	غائب	مفرد	مثنى	جمع	مذكر	مؤنث
الصفة	صفة فاعل	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صفة مفعول	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	أفعل تفضيل	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صفة مشبهة	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صيغة مبالغة	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
الضمائر	شخصي						+	+	+	+	+	+	+										
	إشارة								+	+	+	+	+										
	موصول								+	+	+	+	+										
الظروف	مكان																						
	زمان																						
النحو الف	إخالة														+	+	+	+	+	+	+	+	+
	خالفة صوت																						
	تعجب														+	+	+	+	+	+	+	+	+
	ذم																						
	مدح																						
الأدوات															+	+	+	+	+	+	+	+	+

جدول (٣-٩-أ) تسمية وسوم الاسم

القسم الأساسي	القسم الفرعي	الخصائص التصريفية																							
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط											
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم 1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس			
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF			
الاسم N	اسم ذات C	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم معنى A	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم آلة M	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم مكان L	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم زمان T	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم مبهم I	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	اسم جنس V	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

جدول (٣-٩- ب) تسمية وسوم الفعل

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																							
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط											
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم 1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس			
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF			
الفعل V	مضارع P			+		+	+		+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+
	ماضي S			+		+	+		+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+
	أمر C					+			+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+

جدول (٣-٩-ج) تسمية وسوم الصفة

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم 1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس	
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF
الصفة A	صفة فاعل S	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صفة مفعول O	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	أفعل تفضيل C	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صفة مشبهة A	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	صيغة مبالغة E	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

جدول (٣-٩-د) تسمية وسوم الضمائر

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم 1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس	
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF
الضمائر P	شخصي P					+	+	+	+	+	+	+	+										
	إشارة D								+	+	+	+	+										
	موصول R								+	+	+	+	+										

جدول (٣-٩-هـ) تسمية وسوم الظروف

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس	
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF
الظروف D	مكان L																						
	زمان T																						

جدول(٣-٩- و) تسمية وسوم الخوالف

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	النوع			العدد			الجنس	
															متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF
الخوالف I	إخالة V														+	+	+	+	+	+	+	+	
	خالفة صوت S																						
	تعجب E														+	+	+	+	+	+	+	+	
	ذم X																						
	مدح G																						

جدول (٣-٩-ز) تسمية وسوم الأدوات

القسم الأساسي	القسم الفرعي	الخاصية التصريفية																					
		التعريف		البناء		الشخص			العدد			الجنس		النسب	الربط									
															النوع			العدد			الجنس		
		معرف D	غير معرف I	معلوم A	مجهول P	متكلم 1	مخاطب 2	غائب 3	مفرد S	مثنى U	جمع L	مذكر M	مؤنث F	منسوب AT	متكلم R1	مخاطب R2	غائب R3	مفرد RS	مثنى RU	جمع RL	مذكر RM	مؤنث RF
الأدوات R															+	+	+	+	+	+	+	+	+

الفصل الرابع
تطبيق المنهج وقياس الأداء

تمهيد:

انتهيت بنهاية الفصل الثالث من جمع مدونة مكونة من ١٠ آلاف كلمة أصبحت بعد تفريقها ١١,٥٩٩ كلمة، وبعد تقطيعها حوالي ١٣,٦٠٦ كلمة، منها ٤٩ كلمة متعددة من كلمتين و٩ كلمات متعددة من ٣ كلمات، ووسمت يدويا بأقسام الكلام على ثلاثة مستويات:

١- مستوى الوسوم الرئيسة. وقد ظهرت جميعها في مدونة النظام، وهي ١٢ قسما: الاسم والصفة والفعل والضمير والظرف والخالفة والأداة وعلامات الترقيم والرموز والاختصارات والكلمات الأجنبية والأرقام.

٢- مستوى الوسوم الفرعية. وقد ظهر منها في مدونة النظام ٢٧ وسما، ولكنها ٣٠ قسما:

- الاسم، ويتفرع إلى ٧ أقسام، هي: اسم الذات، واسم المعنى، واسم المكان، واسم الزمان، واسم الآلة، واسم الجنس، والاسم المبهم.

- الصفة، وتتفرع إلى ٥ أقسام، هي: صفة الفاعل، وصفة المفعول، والصفة المشبهة، وصفة المبالغة، وأفعل التفضيل.

- الفعل، ويتفرع إلى ٣ أقسام، هي: الماضي، والمضارع، والأمر.

- الضمير ويتفرع إلى ٣ أقسام، هي: الضمير الشخصي، وضمير الإشارة، والضمير الموصول.

- الظروف، ويتفرع إلى قسمين، هما: ظرف الزمان وظرف المكان.

- الخالفة، وتتفرع إلى ٤ أقسام، هي: الإخالة، وخالفة الصوت، وخالفة المدح، وخالفة الذم، وخالفة التعجب.

- الأداة.

- علامات الترقيم.

- الرموز.

- الاختصارات.

- الكلمات الأجنبية.

- الأرقام.

٣- مستوى الوسوم الموسعة بالخصائص التصريفية (التعريف ـ الشخص ـ العدد ـ الجنس ـ البناء ـ النسب ـ الرابط): وقد بلغ عدد المستعمل منها في المدونة ٣٩٧ وسم، وقد يظهر عدد أكبر من هذا العدد من الوسوم الموسعة إذا وسعت مدونة النظام، وطبقت عليها مجموعة الوسوم المقترحة في هذا المستوى.

وللتقييم مزية أساسية تتشكل في أهمية العلاقة بين اللغويات واللغويات الحاسوبية منذ التسعينات الميلادية في القرن العشرين. وفي إشارة **جونسون**[١] Johnson لتلك الأهمية من الناحية التقييمية، فإن العلاقة الوثيقة بين المنظور اللغوي والمنظور التقني في معالجة اللغة الطبيعية تتمحور في الجوانب العلمية النوعية والإحصائية. حيث إن مجال معالجة

---

(١) Johnson, M. How relevant is linguistics to computational linguistics?. Linguistic Issues in Language Technology - LiLT, USA, Vol.6, No.7, 2011, p. 1-23

اللغة الطبيعية يؤطر مناهجه وتطبيقاته بصورة موسعة بالاعتماد على تعلم الآلة والإحصاء، على عكس الجانب اللغوي الذي يركز نوعيا على الجوانب اللغوية النظرية والتطبيقية.

ومن المهم في سياق التقييم اللغوي والتقني الكشف عن طبيعة مدونة النظام، فالنتائج اللغوية سند للنتائج التقنية والعكس كذلك. ويعتمد التقييم اللغوي على اتجاهين، هما:

الأول: النظر في أكثر الكلمات تكرارا قبل التقطيع وبعد التقطيع وبعد التوسيم من حيث الكلمة الفعلية نفسها، وتفسير ذلك لغويا على الوسوم بمستوياتها الثلاثة (الرئيسة، والفرعية والموسعة بالخصائص التصريفية).

الثاني: قياس الأنماط التـتابعية من الناحية النحوية المعجمية (النحو المعجمي lexical grammar)، وذلك بالمقارنة بين مستويات الوسوم الثلاثة في الأنماط التتابعية العشرة الأكثر تكرارا في مدونة النظام على تتابع كلمتين وثلاث كلمات وأربع كلمات n-grams[١]، كما عند الفيرثيين الجدد وسنكلير[٢].

وأما ما يتعلق بالتقييم التقني، فإن التقييم يمر بمرحلتين، هما:

١- تدريب خوارزمية تعلم الآلة على ٨٠٪ من مدونة النظام (مدونة التدريب).

٢- قياس أداء نظام التوسيم النحوي الآلي، باستخدام الجزء الذي لم تدرب عليه الخوارزمية من قبل (مدونة الاختبار). ويكون التقييم بمقاييس الأداء المتبعة في قياس أنظمة التوسيم النحوي الآلية، وهي أربعة مقاييس: **الصحة** Accuracy **والدقة** Precision **والاسترجاع** Recall **ومقياس ف** F-measure.

---

(١) تستعمل المتتابعات اللفظية n-grams في مجالات التجهيز الإحصائي للغات الطبيعية. وهي مجموعة غير محددة العدد من الكلمات باعتبارها كلمات فعلية متلازمة. انظر: قاموس مصطلحات المعلوماتية واللغويات الحسابية، ٢٠٠٣، ص. ٢٤٤

(٢) المجيول، سلطان. مناهج التهيئة المعجمية في تعليم العربية لغير الناطقين بها. المؤتمر الدولي الثاني اتجاهات حديثة في تعليم العربية لغة ثانية، معهد اللغويات العربية بجامعة الملك سعود، مجلد٢، الرياض: المملكة العربية السعودية، ٢٠١٦، ص. ٦٠١-٦٠٢-٦٠٣

٤-١ نتائج التقييم اللغوي:

يعد التحليل الإحصائي لبيانات المدونة اللغوية مفتاحا أساسيا لاستكشاف المدونة ومعرفة خصائصها اللغوية. ويفيد تحليل كلمات المدونة اللغوية في معرفة انتظامها التوزيعي داخل المدونة، فتوزيع الكلمات في أي لغة يتبع نمطا بيانيا يمكن التنبؤ به حيث توجد علاقة إحصائية بين حجم النص وتكرارات الكلمات فيه تشير إلى تناسب النص اللغوي وانسجامه، وتعرف هذه العلاقة **بقانون زيف** Zipf›s Law [١]، وسنرى فيما سيأتي إن كانت مدونة النظام تتسم بذلك. فقد توصل العالم اللغوي **جورج زيف** George Zipf بعد بحثه في نصوص من الإنجليزية والألمانية والصينية إلى أن مجموعة قليلة من الكلمات في اللغة تظهر في النصوص بتكرار عال، وأن معظم الكلمات الباقية لا تستخدم إلا نادرا[٢]. ووجد أن احتمالية ظهور الكلمات الأكثر شيوعا مرتبطة بترتيبها، وأن عدد الكلمات المستعملة في النصوص المكتوبة والمنطوقة له علاقة بتكراراتها حيث يستعملها الكتاب والمتحدثون أكثر من غيرها. فوجد في العينات النصية أن الكلمة الأولى الأكثر شيوعا تظهر مرة واحدة كمعدل عام بين كل عشر كلمات. وأن الكلمة الثانية الأكثر شيوعا تظهر مرة واحدة بين كل عشرين كلمة، أي أنها تظهر بعدد مرات يساوي نصف عدد مرات ظهور الكلمة الأولى. وأن الكلمة الثالثة الأكثر شيوعا تظهر بين كل ثلاثين كلمة، أي أنها تظهر بعدد مرات يساوي ثلث عدد مرات ظهور الكلمة الأولى. وهكذا حتى الكلمة ذات الرتبة المائة التي تظهر مرة واحدة في كل ألف كلمة من المجموع اللغوي. وبعد ترتيبه للكلمات ضرب القيمة الرقمية **لرتبة** Rank كل كلمة في عدد مرات **تكراراتها** Frequencies وحصل على **الناتج**

---

(١) David M. W. Powers. Applications and Explanations of Zipf›s Law. In D. M. W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL, 1998, p.151

(٢) Rousseau, R., Goerge Kingsley Zipf: life, ideas, his law and informetrics. Glottometrics, Germany, Vol. 3, 2002, p. 11

Product الذي وجده ثابتا في مجمل قوائم نصوصه. وبالتالي فإن قانون زيف الرياضي المتعلق بالكلمات الأكثر تكرارا هو: R*F= C، حيث R هي الرتبة، وF تكرارات الكلمة وC هو الناتج الثابت، وسأنظر في هذا القانون على نصوص مدونة النظام.

ويوضح الجدول (٤-١-أ) الكلمات الخمسين الأكثر تكرارا في مدونة النظام قبل تقطيعها، وهي تعادل ٢٦,٧٪ من كامل المدونة. كما يوضح الجدول (٤-١-ب) الكلمات الخمسين الأكثر تكرارا في مدونة النظام بعد تقطيعها، وهي تعادل ٣٩,٩٪ من كامل المدونة. وبالمقارنة بين الجدولين يلاحظ أن الكلمات الأكثر تكرارا متشابهة في الجدولين، ولكن تكراراتها في الجدول (٤-١-أ) أقل بكثير من تكراراتها في الجدول الآخر بسبب المعالجة اللغوية التي تقتضي فصل ما يعد في نظرية تمام حسان كلمة يجب تصنيفها وإن كان لا يعد كذلك حاسوبيا، حيث تظهر الواو والفاء والباء واللام والكاف والسين منفصلة باعتبارها **كلمات فعلية** tokens. فكلمة: (الله) وردت من غير تقطيع ٦٩ مرة بينما وردت بعد فصل ما يعد عند تمام حسان كلمة فعلية ٨١ مرة. وأما الكلمات التي وردت بنفس التكرار في الجدولين، فهي كلمات لم ترد مع لواصق، وفي ذلك إشارة إلى قلة تلازمها مع متغيرات التقطيع في مدونة النظام، مثل كلمة: بن. وبالمقارنة بين الجدولين أيضا، يلاحظ أن متغيرات التقطيع ترد أكثر من الكلمات المستقلة بنفسها. حيث ظهرت الواو والباء بتكرار عال في الجدول (٤-١-ب) وظهرتا في الجدول (٤-١-أ) ملتصقات بكلمات أكثر تكرار أيضا. ويبين الجدولان أيضا أن الواو والباء تأتيان أكثر من حروف المعاني التي لا تتصل إلا بروابط إحالية، نحو: (في وعلى وإلى)، وأن اتصال حروف الجر بالكلمات الفعلية من متغيرات التقطيع أكثر من اتصالها بالروابط الإحالية، حيث تتأخر كثيرا تكرارات كلمات مثل: (به-له) عن الباء واللام. ومع تقطيع المدونة يلاحظ تراجع بعض الأدوات التي كانت تتردد كثيرا وتأتي في الخمسين كلمة الأكثر تكرارا في المدونة قبل التقطيع، ومن تلك الكلمات: حرف العطف (ثم) الذي حلت محلة متغيرات التقطيع التي تطغى تكراراتها على كامل المدونة.

جدول (٤-١-أ) الكلمات الخمسون الأكثر تكرارا في مدونة النظام قبل التقطيع (٣٠٩١ كلمة فعلية)

م	الكلمة	التكرار النسبي المئوي[١]	م	الكلمة	التكرار النسبي المئوي
١	،	3.99	٢٦	كان	0.25
٢	في	2.55	٢٧	لم	0.24
٣	.	2.51	٢٨	"	0.24
٤	من	2.29	٢٩	غير	0.24
٥	على	1.22	٣٠	به	0.23
٦	أن	1.06	٣١	حتى	0.22
٧	:	0.89	٣٢	فيه	0.22
٨	إلى	0.84	٣٣	ولا	0.21
٩	عن	0.61	٣٤	؟	0.21
١٠	الله	0.60	٣٥	مع	0.20
١١	ما	0.60	٣٦	وقد	0.19
١٢	التي	0.49	٣٧	كما	0.19
١٣	أو	0.39	٣٨	قال	0.19
١٤	عليه	0.39	٣٩	فإن	0.19
١٥	)	0.39	٤٠	خلال	0.19

(١) التكرار النسبي هو تكرار الفئة مقسوما على مجموع التكرارات لكل الفئات، أما التكرار النسبي المئوي فهو عبارة عن التكرار النسبي مضروبا في مائة.

انظر:

Statistics In Corpus Linguistics (A practical Guide),2018, p. 43

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١٦	)	0.39	٤١	إن	0.18
١٧	لا	0.38	٤٢	هو	0.17
١٨	هذه	0.38	٤٣	أي	0.17
١٩	هذا	0.38	٤٤	ثم	0.17
٢٠	بن	0.36	٤٥	–	0.17
٢١	الذي	0.31	٤٦	بعد	0.16
٢٢	له	0.30	٤٧	وكان	0.16
٢٣	كل	0.30	٤٨	وهو	0.15
٢٤	بين	0.26	٤٩	قد	0.15
٢٥	ذلك	0.26	٥٠	فيها	0.14

جدول (٤-١-ب) الكلمات الخمسون الأكثر تكرارا في مدونة النظام بعد التقطيع (٥٤٢١ كلمة فعلية)

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١	و	8.24	٢٦	)	0.33
٢	،	3.40	٢٧	(	0.33
٣	ب	2.57	٢٨	عليه	0.33
٤	في	2.31	٢٩	أو	0.32
٥	ل	2.25	٣٠	"	0.32
٦	من	2.19	٣١	هو	0.31

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
٧	.	2.13	٣٢	بن	0.30
٨	ف	1.15	٣٣	قال	0.30
٩	على	1.08	٣٤	الذي	0.27
١٠	ما	1.05	٣٥	لم	0.26
١١	أن	1.00	٣٦	له	0.25
١٢	:	0.76	٣٧	بين	0.24
١٣	إلى	0.74	٣٨	غير	0.24
١٤	الله	0.60	٣٩	هي	0.21
١٥	عن	0.55	٤٠	)	0.20
١٦	لا	0.54	٤١	به	0.19
١٧	التي	0.46	٤٢	حتى	0.18
١٨	قد	0.44	٤٣	فيه	0.18
١٩	ذلك	0.43	٤٤	مع	0.18
٢٠	هذا	0.40	٤٥	؟	0.18
٢١	هذه	0.39	٤٦	س	0.17
٢٢	ك	0.38	٤٧	بعد	0.16
٢٣	إن	0.36	٤٨	خلال	0.15
٢٤	كان	0.35	٤٩	–	0.15
٢٥	كل	0.34	٥٠	أي	0.15

ويثبت الجدول (٤-١-أ) والجدول (٤-١-ب) أيضا ما قال به اللغوي **جورج**

زيف George Zipf [١] حول أثر القلة الأكثر تكرارا على الكثرة الأقل تكرارا. والمقصود من القلة الأكثر تكرارا في المدونة اللغوية الكلمات الوظيفية، وهي الضمائر والأدوات والظروف والخوالف وبعض الأسماء المبهمة وعلامات الترقيم والرموز والاختصارات، حيث إنها من أكثر الكلمات تكرارا رغم قلتها، على عكس الكلمات ذات المحتوى التي قد لا يأتي عليها العد، وقد لا ترد إلا مرة واحدة[٢]. وعلى الرغم من ذلك، نجد أن ثمة ثلاث كلمات محتوى من ضمن الكلمات الأكثر تكرارا في مدونة النظام، وهي: لفظ الجلالة (الله)، والاسم (بن) والفعل (قال). ويُعزى سبب ذلك على التوالي إلى المحتوى اللغوي العربي الثقافي الإسلامي المعظم للخالق، وفي الأسلوب العربي الشائع وخصوصا في التراث في إيراد الأعلام منسوبة لآبائها، وفي الاهتمام في الثقافة العربية بالنقل القولي النطقي دون النقل النصي.

وفي الجداول (٤-٢-أ) و(٤-٢-ب) و(٤-٢-ج) تظهر الكلمات الخمسون الأكثر تكرارا في مدونة النظام بعد توسيمها مفصولة عن وسمها بشرطة مائلة (/). وهذا يعني أن ثمة تمييزا سيحدث على مستوى الكلمة الواحدة، فالباء في مدونة النظام مثلا ليست كلها أدوات، فقد يرد منها اختصارات ولذلك نراها في قائمة الكلمات في هذه الجداول بنسبة تكرار أقل من الجدول (٤-١-ب)، والحال نفسه في (ما) التي جاءت ضميرا باعتبار التقسيم الأساسي، وموصولة بحسب التقسيم الفرعي أكثر من كونها أداة في مدونة النظام، ولكنها

---

(١) David M. W. Powers. Applications and Explanations of Zipf›s Law. In D. M. W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL, 1998, p.151

(٢) الكلمات الوظيفية هي كلمات تعبر عن العلاقات النحوية بين الكلمات في إطار الجملة، وتشير إلى العلاقات البنيوية التي تربط بين الكلمات، فلا تقوم بمعنى بنفسها في الجملة بل تحتاج إلى غيرها، وهي الضمائر والأدوات والظروف والخوالف والأسماء المبهمة. أما كلمات المحتوى فهي كلمات ذات دلالات معنوية مستقلة. ومنذ اقتراحهما لأول مرة في عام ١٩٥٢ من قبل فرايز Fries كان هذا التمييز مؤثرًا بشكل كبير في الدراسات اللغوية. انظر:
Fries, Charles Carpenter. The Structure of English. Harcourt Brace, New York: USA, 1952

في الحالين ترد في قائمة الكلمات الخمسين الأكثر تكرارا في الجداول (٤-٢-أ) و(٤-٢-ب) و(٤-٢-ج). وهذا التمييز الحاصل هنا يعود بنا إلى أهمية التوسيم النحوي من حيث الاستعانة به على فهم النصوص وإزالة الغموض التركيبي والدلالي لتغير المعنى في الحالين. ويلاحظ أيضا تصدر الأدوات قائمة الوسوم في كل مجموعات الوسوم إذ تبلغ النصف تقريبا من الخمسين في كل مجموعة، حيث إنها قسم أساسي لم ينظر إلى فروعه بحكم تناوله بهذا الحال في هذا المقترح. كما تتصدر خاصيتي الإفراد والتذكير والتأنيث من مجموع الخصائص التصريفية (انظر الجدول ٤-٢-ج).

جدول (٤-٢-أ) الكلمات الخمسون الأكثر تكرارا في مدونة النظام بعد التوسيم بالوسوم الرئيسة (٥٤٢١ كلمة فعلية)

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١	و\RP	8.19	٢٦	كل\N	0.33
٢	،\PUNC	3.38	٢٧	(\PUNC	0.33
٣	ب\RP	2.56	٢٨	)\PUNC	0.33
٤	في\RP	2.30	٢٩	عليه\RP	0.33
٥	ل\RP	2.24	٣٠	أو\RP	0.32
٦	.\PUNC	2.10	٣١	"\PUNC	0.32
٧	من\RP	2.05	٣٢	هو\P	0.31
٨	ف\RP	1.15	٣٣	بن\N	0.30
٩	على\RP	1.08	٣٤	قال\V	0.29
١٠	أن\RP	0.99	٣٥	الذي\P	0.27
١١	:\PUNC	0.76	٣٦	لم\RP	0.26
١٢	إلى\RP	0.74	٣٧	له\RP	0.25

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١٣	ما\P	0.70	٣٨	بين\N	0.24
١٤	الله\N	0.60	٣٩	غير\N	0.24
١٥	عن\RP	0.55	٤٠	هي\P	0.21
١٦	لا\RP	0.54	٤١	به\RP	0.20
١٧	التي\P	0.46	٤٢	حتى\RP	0.18
١٨	قد\RP	0.44	٤٣	مع\N	0.18
١٩	ذلك\P	0.43	٤٤	فيه\RP	0.18
٢٠	هذا\P	0.40	٤٥	؟\PUNC	0.18
٢١	هذه\P	0.39	٤٦	س\RP	0.17
٢٢	ك\RP	0.38	٤٧	بعد\N	0.16
٢٣	إن\RP	0.36	٤٨	خلال\N	0.16
٢٤	كان\RP	0.35	٤٩	-\PUNC	0.15
٢٥	ما\RP	0.34	٥٠	أي\RP	0.15

جدول (٤-٢-ب) الكلمات الخمسون الأكثر تكرارا في مدونة النظام بعد التوسيم بمجموعة الوسوم الفرعية (٥٤٢١ كلمة فعلية)

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١	و/ RP	8.09	٢٦	)/ PUNC	0.33
٢	،/ PUNC	3.28	٢٧	(/ PUNC	0.33
٣	ب/ RP	2.53	٢٨	عليه/ RP	0.33
٤	في/ RP	2.25	٢٩	كل/ NI	0.33

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
٥	ل/ RP	2.22	٣٠	أو/ RP	0.32
٦	PUNC /.	2.06	٣١	PUNC /"	0.32
٧	من/ RP	2.00	٣٢	هو/ PP	0.31
٨	ف/ RP	1.13	٣٣	بن/ NC	0.30
٩	على/ RP	1.05	٣٤	قال/ VS	0.29
١٠	أن/ RP	0.95	٣٥	الذي/ PR	0.27
١١	PUNC /:	0.75	٣٦	لم/ RP	0.26
١٢	إلى/ RP	0.71	٣٧	له/ RP	0.25
١٣	ما/ PR	0.67	٣٨	بين/ NI	0.24
١٤	الله/ NC	0.58	٣٩	غير/ NI	0.23
١٥	عن/ RP	0.55	٤٠	هي/ PP	0.21
١٦	لا/ RP	0.53	٤١	به/ RP	0.20
١٧	التي/ PR	0.44	٤٢	مع/ NI	0.18
١٨	قد/ RP	0.43	٤٣	فيه/ RP	0.18
١٩	ذلك/ PD	0.41	٤٤	حتى/ RP	0.18
٢٠	هذه/ PD	0.39	٤٥	؟/ PUNC	0.18
٢١	هذا/ PD	0.39	٤٦	س/ RP	0.17
٢٢	ك/ RP	0.37	٤٧	خلال/ NI	0.16
٢٣	إن/ RP	0.35	٤٨	بعد/ NI	0.16
٢٤	كان/ RP	0.33	٤٩	أي/ RP	0.15
٢٥	ما/ RP	0.34	٥٠	–/ PUNC	0.15

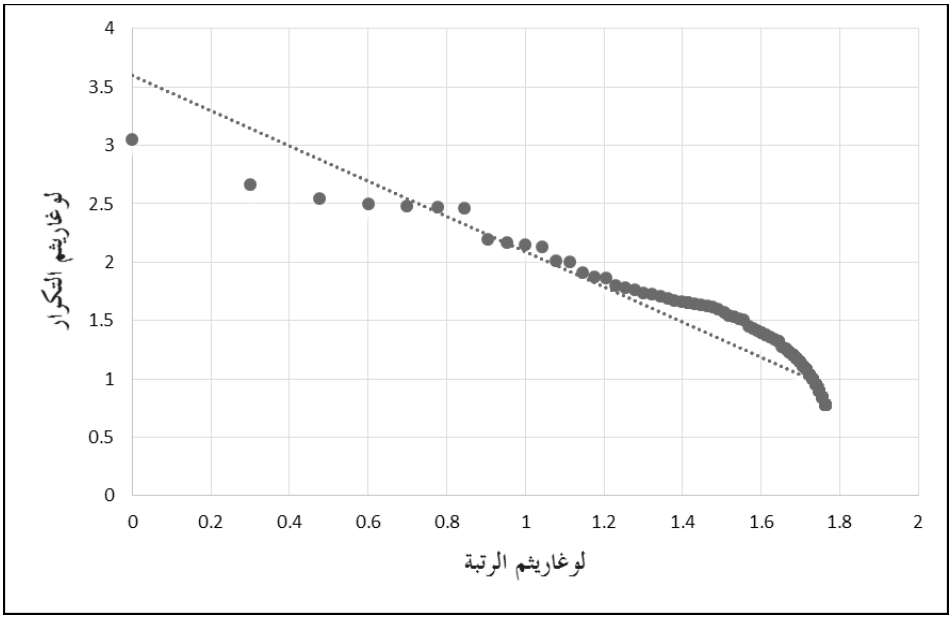
جدول (٤-٢-ج) الكلمات الخمسون الأكثر تكرارا في مدونة النظام بعد التوسيم بمجموعة الوسوم الموسعة (٥٤٢١ كلمة فعلية)

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
١	و/ RP	8.09	٢٦	كل/ NI_ISM	0.33
٢	،/ PUNC	3.28	٢٧	أو/ RP	0.32
٣	ب/ RP	2.53	٢٨	(/ PUNC	0.31
٤	في/ RP	2.25	٢٩	)/ PUNC	0.31
٥	ل/ RP	2.22	٣٠	"/ PUNC	0.31
٦	./ PUNC	2.06	٣١	هو/ PP_3SM	0.30
٧	من/ RP	2.00	٣٢	بن/ NC_ISM	0.30
٨	ف/ RP	1.13	٣٣	قال/ VS_A3SM	0.28
٩	على/ RP	1.05	٣٤	الذي/ PR_SM	0.27
١٠	أن/ RP	0.95	٣٥	التي/ PR_SF	0.25
١١	:/ PUNC	0.75	٣٦	لم/ RP	0.25
١٢	إلى/ RP	0.71	٣٧	له/ RP_R3RSRM	0.25
١٣	ما/ PR_SM	0.65	٣٨	بين/ NI_ISM	0.23
١٤	عن/ RP	0.55	٣٩	غير/ NI_ISM	0.23
١٥	لا/ RP	0.53	٤٠	به/ RP_R3RSRM	0.19
١٦	الله/ NC_ISM	0.53	٤١	التي/ PR_LF	0.18
١٧	قد/ RP	0.43	٤٢	حتى/ RP	0.18
١٨	ذلك/ PD_SM	0.41	٤٣	مع/ NI_ISM	0.18
١٩	هذا/ PD_SM	0.39	٤٤	فيه/ RP_R3RSRM	0.17
٢٠	ك/ RP	0.37	٤٥	؟/ PUNC	0.17

م	الكلمة	التكرار النسبي المئوي	م	الكلمة	التكرار النسبي المئوي
٢١	هذه/ PD_SF	0.36	٤٦	س/ RP	0.17
٢٢	إن/ RP	0.35	٤٧	خلال/ NI_ISM	0.15
٢٣	كان/ RP	0.33	٤٨	أي/ RP	0.15
٢٤	عليه/ RP_R3RSRM	0.33	٤٩	بعد/ NI_ISM	0.15
٢٥	ما/ RP	0.33	٥٠	–/ PUNC	0.15

ويوضح الشكل (٤-١) أن العلاقة بين التكرار والرتبة لكلمات المدونة علاقة خطية بعد استبعاد الكلمات ذات التكرارات الأقل من ٦؛ لأنها لا تعد من ضمن الكلمات الأكثر تكرارا حسب ما يرى زيف. ويلاحظ بأن كل كلمات المدونة الأكثر تكرارا منحرفة تنازليا بشكل طفيف بين تكراراتها ورتبها عدا الانحراف البين عن خط زيف بسبب الكلمة الأولى (و). ووجود هذا الانحراف عن هذا الميل قد يكون بسبب قلة كلمات المدونة حيث يحتاج قانون زيف عينات ضخمة من النصوص.

شكل (٤-١) العلاقة بين التكرار والرتبة لكلمات المدونة



وبعد حساب التكرار للكلمات الفعلية بعد إجراء عملية التقطيع اسُتخلصت نتائج متغيرات التقطيع الأكثر تكرارا وجميعها يقع من ضمن الكلمات الخمسين الأكثر تكرارا في مدونة النظام. وكانت الكلمة الوظيفية: (الواو)، والكلمة الوظيفية: (الباء)، والكلمة الوظيفية: (اللام)، والكلمة الوظيفية العاطفة: (الفاء)، ثم الكلمة الوظيفية: (كاف التشبيه)، فالكلمة الوظيفية الدالة على المستقبل للأفعال الحاضرة: (السين)، انظر الجدول (٤-٣).

جدول (٤-٣) متغيرات التقطيع المستعملة في المدونة وتكراراتها

م	المتغير	التكرار النسبي المئوي	م	المتغير	التكرار النسبي المئوي
١	و	8.19	٤	ف	1.15
٢	ب	2.53	٥	ك	0.37
٣	ل	2.24	٦	س	0.17

ويعرض الجدول (٤-٤) تكرارات الوسوم الرئيسة المطبقة في هذا المقترح، ويلاحظ أن كل الوسوم الرئيسة المقترحة قد ظهرت في مدونة النظام. وقد يبدو أن بعض وسوم الكلمات ذات المحتوى أكثر من وسوم الكلمات الوظيفية، فيظهر مثلا أن الأسماء وردت أكثر من الأدوات ـ رغم أن الأدوات عادة ما تكون من ضمن الكلمات الفعلية الأكثر تكرارا ـ وتليها الفعل ثم الصفة ثم بقية الوسوم. والواقع أن جزءا من القسم الرئيس (الاسم) هو أسماء مبهمة منها ما هو كلمات وظيفية، نحو: (بين ـ غير ـ خلال ـ مع ـ كل)، وجميعها ورد في القائمة الخمسين الأكثر تكرارا في مدونة النظام ويمثل مع الأدوات تكرارا نسبيا قدره ٠,٣٤ (انظر الجدول ٤-٥)، وأن الأدوات هي جزء من الكلمات الوظيفية وليست جميعها، وبالتالي فإن وسوم الكلمات الوظيفية بمجموعها (الأدوات مع الضمائر والظروف والخوالف وعلامات الترقيم والاختصارات والرموز والأرقام) أكثر تكرارا من وسوم الأسماء، وهذا ما يقره زيف في قانونه، إذ كل الجمل في العربية تعتمد في تلخيص العلاقة بين أجزائها على الكلمات الوظيفية.

جدول (٤-٤) الوسوم الرئيسة المستعملة في المدونة وتكراراتها

م	الوسم الرئيس	القسم الرئيس	التكرار النسبي المئوي
١	N	اسم	36.87
٢	RP	أداة	30.16
٣	V	فعل	9.81
٤	A	صفة	9.30
٥	PUNC	علامة ترقيم	8.48
٦	P	ضمير	3.72
٧	DIGIT	رقم	0.85
٨	D	ظرف	0.29
٩	ABBREV	اختصار	0.20
١٠	FOREIGN	أجنبي	0.18
١١	SYMB	رمز	0.10
١٢	I	خالفة	0.04
المجموع		100	

ويبين الجدول (٤-٥) أنه عند تقسيم الوسوم الرئيسية إلى وسوم فرعية أصبحت الوسوم الوظيفية (الأدوات) أكثر تكرارا من الأسماء التي كانت أكثر تكرارا في جدول الوسوم الرئيسة لمدونة النظام. كما يبين الجدول (٤-٥) أن ثمة وسوما فرعية مقترحة لم تستعمل في نصوص المدونة، وهي: خالفة التعجب IE وخالفة الصوت IS وخالفة الذم IX وخالفة المدح IG التي تندرج تحت الوسم الرئيس (خوالف I)، ولعل ذلك بسبب طبيعة أجناس النصوص حيث لا تعنى النصوص المكتوبة بالتخاطب المفعم بالانفعالات والأصوات، إذ يرى مصطفى جمال الدين أن مثل خوالف الأصوات «لا تدخل في طبيعة مفردات اللغة باعتبارها واسطة نقل الأفكار من ذهن إلى ذهن، ولا تدخل في الجمل العربية للقيام بوظيفة

الرابط أو المرتبط فيها، إلا على سبيل الحكاية»[١]. فضلا عن أن بعض خوالف الأصوات ليس في اللغة من الحروف ما يدونها، كما أن صغر حجم المدونة له دور في ذلك.

وبين الجدول في ملحق الكتاب قائمة الوسوم الموسعة بالخصائص التصريفية المستعملة في المدونة مع تكراراتها النسبية المئوية، وقد حُسبت منها التكرارات النسبية المئوية للخصائص التصريفية. وتظهر في الجدول (٤-٦) جميع الخصائص التصريفية المستعملة في مدونة النظام مرتبة حسب تكراراتها. وحيث إن الخصائص التصريفية للكلمات العربية تختلف باختلاف قسم الكلمة، قد تتأثر تكرارات الخصائص بتكرارات الوسوم الرئيسة. ولكن الذي بدا واضحا أن صيغة الإفراد مع التذكير ومع التأنيث في كل الأقسام الكلامية تتكرر أكثر من صيغة الجمع معهما حتى في الروابط الإحالية (انظر الجدول ٤-٧)، وأن التثنية نادرة جدا مقارنة بهما؛ وذلك لكثرة الأفراد وقلة الجماعات والمثنيات، والحاجة هي ما يتحكم في الاستعمال، ويلاحظ أيضا زيادة صيغ الغائب على صيغ المخاطب والمتكلم في الضمائر والأفعال، ويعود ذلك إلى أن العرب يستعملون الألفاظ المؤنثة أو المذكرة جمعا كانت أو مفردة ليعبروا بها عن المعاني، ثم إذا أرادوا الحديث عن هذه المعاني مرة أخرى عادوا إليها بصيغة الغائب. ومن الأمثلة على ذلك في نصوص المدونة: «فإن مبدأ الحياة هو القلب والروح»، «...المعصِرات: السماء، وهذا إن حُمل على التفسير على المعنى ...»، كما أن معظم النصوص في مدونة النظام هي النصوص الوصفية الصحفية والإخبارية التي تتجنب استعمال ضمير المتكلم، ولا توجه الخطاب من المتكلم إلى المتلقي، وهذا يؤيد أيضا ما يراه زيف حول كثرة ما تقل حروفه. وبدا أيضا أن البناء للمعلوم يزيد عن البناء للمجهول في الأفعال، حيث إن الأصل هو البناء للمعلوم، فلا يكون الفعل دون فاعل، وأما البناء للمجهول فلتحقيق أغراض عارضة. كما ظهرت الجموع المؤنثة أكثر من الجموع المذكرة، وذلك لأن هذا النوع يدخل فيه جمع المؤنث السالم وجمع التكسير الذي يعامل في بعض صيغه كالمؤنث.

---

(١) جمال الدين، مصطفى. رأي في تقسيم الكلمة. مجلة تراثنا، العدد ٦، مؤسسة آل البيت، قم: إيران، ١٩٨٦، ص١٢٣

جدول (٤-٥) الوسوم الفرعية المستعملة في المدونة وتكراراتها

م	الوسم الفرعي	القسم الفرعي	التكرار النسبي المئوي
١	RP	أداة	30.17
٢	NA	اسم معنى	16.98
٣	NC	اسم ذات	12.70
٤	PUNC	علامة ترقيم	8.48
٥	VP	فعل مضارع	4.91
٦	VS	فعل ماض	4.74
٧	NI	اسم مبهم	4.73
٨	AS	صفة فاعل	4.42
٩	AA	صفة مشبهة	2.07
١٠	AO	صفة مفعول	1.72
١١	PR	ضمير موصول	1.64
١٢	PD	ضمير إشارة	1.38
١٣	NV	اسم جنس	1.23
١٤	NL	اسم مكان	1.00
١٥	AC	أفعل تفضيل	0.87
١٦	DIGIT	رقم	0.85
١٧	PP	ضمير شخصي	0.70
١٨	AE	صفة مبالغة	0.22
١٩	ABBREV	اختصار	0.20
٢٠	FOREIGN	أجنبي	0.18

م	الوسم الفرعي	القسم الفرعي	التكرار النسبي المئوي
٢١	VC	فعل أمر	0.16
٢٢	DT	ظرف زمان	0.15
٢٣	NM	اسم آلة	0.15
٢٤	DL	ظرف مكان	0.14
٢٥	SYMB	رمز	0.10
٢٦	NT	اسم زمان	0.08
٢٧	IV	إخالة	0.04
المجموع		100	

جدول (٤-٦) تكرارات الخصائص التصريفية في مجموعة الوسوم المقترحة

الخاصية بالعربية	وسمها	تكرارها النسبي المئوي	القسم الكلامي الذي وردت فيه
غير معرف بأل ـ مفرد ـ مذكر	ISM	25.17	أسماء وصفات
معرف بأل ـ مفرد ـ مذكر	DSM	14.27	
رابط إحالي (*) تشير لخصائص الرابط	*R*R*R	10.58	أسماء وصفات وأفعال وأدوات وخوالف
معرف بأل ـ مفرد ـ مؤنث	DSF	7.69	أسماء وصفات
غير معرف بأل ـ مفرد ـ مؤنث	ISF	7.06	
معلوم ـ غائب ـ مفرد ـ مذكر	A3SM	6.90	أفعال
منسوب	AT	4.47	أسماء وصفات
معرف بأل ـ جمع ـ مؤنث	DLF	4.42	
غير معرف بأل ـ جمع ـ مؤنث	ILF	3.69	

الخاصية بالعربية	وسمها	تكرارها النسبي المئوي	القسم الكلامي الذي وردت فيه
مفرد ـ مذكر	SM	2.72	ضمائر
معرف بأل ـ جمع ـ مذكر	DLM	2.12	أسماء وصفات
معلوم ـ غائب ـ مفرد ـ مؤنث	A3SF	1.94	أفعال
معلوم ـ غائب ـ جمع ـ مذكر	A3LM	1.07	
مفرد ـ مؤنث	SF	1.01	ضمائر
معلوم ـ غائب ـ جمع ـ مؤنث	A3LF	0.79	أفعال
معلوم ـ متكلم ـ جمع ـ مذكر	A1LM	0.76	
مجهول ـ غائب ـ مفرد ـ مذكر	P3SM	0.70	
غير معرف بأل ـ جمع ـ مذكر	ILM	0.70	أسماء وصفات
معلوم ـ متكلم ـ مفرد ـ مذكر	A1SM	0.66	أفعال
غائب ـ مفرد ـ مذكر	3SM	0.44	ضمائر
جمع ـ مؤنث	LF	0.34	
معلوم ـ مخاطب ـ مفرد ـ مذكر	A2SM	0.31	أفعال
مجهول ـ غائب ـ مفرد ـ مؤنث	P3SF	0.28	
مخاطب ـ مفرد ـ مذكر	2SM	0.22	أفعال وضمائر
غير معرف بأل ـ مثنى ـ مذكر	IUM	0.21	أسماء وصفات
غير معرف بأل ـ مثنى ـ مذكر	IUF	0.18	
مجهول ـ غائب ـ جمع ـ مؤنث	P3LF	0.18	أفعال
معرف بأل ـ مثنى ـ مذكر	DUM	0.18	أسماء وصفات

الخاصية بالعربية	وسمها	تكرارها النسبي المئوي	القسم الكلامي الذي وردت فيه
غائب ـ مفرد ـ مؤنث	3SF	0.17	ضمائر
غائب ـ جمع ـ مؤنث	3LF	0.15	
جمع ـ مذكر	LM	0.13	
متكلم ـ مفرد ـ مذكر	1SM	0.10	أفعال
معرف بأل ـ مثنى ـ مؤنث	DUF	0.10	أسماء وصفات
معلوم ـ غائب ـ مثنى ـ مذكر	A3UM	0.06	أفعال
غائب ـ جمع ـ مذكر	3LM	0.06	ضمائر
متكلم ـ جمع ـ مذكر	1LM	0.04	
مجهول ـ غائب ـ جمع ـ مذكر	P3LM	0.04	أفعال
معلوم ـ مخاطب ـ جمع ـ مذكر	A2LM	0.03	
معلوم ـ مخاطب ـ مفرد ـ مؤنث	A2SF	0.01	أفعال وضمائر
مجهول ـ متكلم ـ مفرد ـ مذكر	P1SM	0.01	
مخاطب ـ مفرد ـ مؤنث	2SF	0.01	أفعال وضمائر
غائب ـ مثنى ـ مذكر	3UM	0.01	
مخاطب ـ جمع ـ مذكر	2LM	0.01	
غائب ـ مثنى ـ مؤنث	3UF	0.01	أفعال
المجموع		100	

وفيما يتعلق بالاتجاه الثاني من التقييم اللغوي لمدونة النظام، تكشف سلسلة الجداول (٤–٨) التالية الأنماط اللغوية العشرة الأكثر تكرارا في المدونة في مجموعات الوسوم الثلاثة (الرئيسة والفرعية والموسعة) على تتابع كلمتين 2-grams وثلاث كلمات

3-grams وأربع كلمات 4-grams. وتشير توزيعات الأنماط في جدول (٤-٨-أ) إلى أن الوسوم الرئيسة لكلمات المحتوى فيها أكثر من الوسوم للكلمات الوظيفية. ولا يرد في الأنماط العشرة الأول في الوسوم الأساسية سواء على تتابع وسمين أو ثلاثة أو أربع أي تركيب معروف للجملة الفعلية: (فعل ـ اسم / فعل ـ اسم ـ اسم / فعل ـ اسم ـ صفة) بينما يرد تركيب الجملة الفعلية للفعل اللازم المسند لضمائر مستترة، وهو: (فعل ـ أداة ـ اسم). وتغيب تماما في كل التتابعات بعض الأقسام الرئيسة كالضمائر والظروف والخوالف، وهذا بسبب تراجعها التكراري في مدونة النظام.

جدول (٤-٧) الروابط الإحالية المستعملة في مجموعة الوسوم المقترحة

وسم الرابط الإحالي	الرابط الإحالي بالعربية	التكرار النسبي المئوي
R3RSRM	رابط محال لغائب مفرد مذكر	5.37
R3RSRF	رابط محال لغائب مفرد مؤنث	2.51
R3RLRM	رابط محال لغائب جمع مذكر	0.86
R1RSRM	رابط محال لمتكلم مفرد مذكر	0.60
R1RLRM	رابط محال لمتكلم جمع مذكر	0.51
R2RSRM	رابط محال لمخاطب مفرد مذكر	0.25
R2RLRM	رابط محال لمخاطب جمع مذكر	0.15
R3RLRF	رابط محال لغائب جمع مؤنث	0.15
R3RURM	رابط محال لغائب مثنى مذكر	0.13
R2RSRF	رابط محال لمخاطب مفرد مؤنث	0.03
R3RURF	رابط محال لغائب مفرد مؤنث	0.01
المجموع		10.58

جدول (٤-٨-أ) الأنماط اللغوية للوسوم الرئيسة في مدونة النظام

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي	النمط اللغوي 4-grams	التكرار النسبي المئوي
أداة اسم	14.79	اسم أداة اسم	6.94	اسم اسم أداة اسم	2.90
اسم اسم	13.30	أداة اسم اسم	5.68	أداة اسم أداة اسم	2.70
اسم أداة	11.57	اسم اسم اسم	4.59	اسم أداة اسم اسم	2.59
أداة فعل	5.17	أداة اسم أداة	4.48	اسم أداة اسم أداة	2.32
أداة أداة	5.01	اسم اسم أداة	4.44	أداة اسم اسم أداة	2.05
ترقيم أداة	4.78	اسم ترقيم أداة	2.78	أداة اسم اسم اسم	1.77
اسم ترقيم	4.58	صفة أداة اسم	2.30	اسم اسم اسم اسم	1.61
اسم صفة	4.29	فعل أداة اسم	2.23	اسم اسم اسم أداة	1.48
فعل أداة	3.98	أداة فعل أداة	2.23	اسم صفة أداة اسم	1.24
صفة أداة	3.90	اسم صفة أداة	1.96	أداة فعل أداة اسم	1.21
الكلمات الوظيفية: ١٠		الكلمات الوظيفية: ١٢		الكلمات الوظيفية: ٢٧	
كلمات المحتوى: ١٠		كلمات المحتوى: ١٨		كلمات المحتوى: ١٣	

ويُلاحظ من جدول (٤-٨-ب) أن العشرة الأولى من أنماط الوسوم الفرعية بالمقارنة مع أنماط الوسوم الرئيسة تغيب عنها الصفة تماما فضلا عن الضمير والظرف والخالفة، كما تغيب الجمل الفعلية تماما بكل تراكيبها، مع ورود الفعل بقلة في ثلاثة أنماط فقط وغياب فعل الأمر. وليس من الغريب أن تظهر في الأنماط من بين الأسماء أسماء المعنى وأسماء الذوات دون غيرها؛ حيث إنها من الأسماء الأكثر تكرارا في مدونة النظام.

جدول (٤-٨-ب) الأنماط اللغوية للوسوم الفرعية في مدونة النظام

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي	النمط اللغوي 4-grams	التكرار النسبي المئوي
أداة اسم_معنى	8.19	اسم_معنى أداة اسم_معنى	2.56	أداة اسم_معنى أداة اسم_معنى	1.25
اسم_معنى أداة	5.69	أداة اسم_معنى أداة	2.51	اسم_معنى أداة اسم_معنى أداة	0.93
أداة أداة	5.01	أداة اسم_معنى اسم_معنى	1.76	أداة اسم_معنى اسم_معنى أداة	0.73
علامة_ترقيم أداة	4.78	علامة_ترقيم أداة أداة	1.66	اسم_معنى أداة اسم_معنى اسم_معنى	0.56
اسم_ذات أداة	4.12	أداة اسم_ذات أداة	1.40	اسم_ذات أداة اسم_ذات أداة	0.53
أداة اسم_ذات	4.11	اسم_ذات علامة_ترقيم أداة	1.26	اسم_معنى اسم_معنى أداة اسم_معنى	0.52
اسم_معنى اسم_معنى	3.21	أداة فعل_مضارع أداة	1.25	أداة اسم_معنى علامة_ترقيم أداة	0.50
أداة فعل_مضارع	2.76	اسم_معنى اسم_معنى أداة	1.23	أداة اسم_ذات أداة اسم_ذات	0.46
اسم_ذات اسم_ذات	2.74	اسم_معنى علامة_ترقيم أداة	1.15	اسم_ذات علامة_ ترقيم أداة أداة	0.42

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي	النمط اللغوي 4-grams	التكرار النسبي المئوي
أداة فعل_ماض	2.36	اسم_ذات أداة اسم_ذات	1.12	اسم_معنى علامة_ ترقيم أداة أداة	0.41
الكلمات الوظيفية: ١٠		الكلمات الوظيفية: ١٧		الكلمات الوظيفية: ٢١	
كلمات المحتوى: ١٠		كلمات المحتوى: ١٣		كلمات المحتوى: ١٩	

جدول (٤-٨-ج) الأنماط اللغوية للوسوم الموسعة بالخصائص التصريفية في مدونة النظام

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي
ترقيم أداة	4.69	ترقيم أداة أداة	1.48
أداة أداة	3.66	أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة	0.80
أداة اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر	2.05	أداة أداة أداة	0.56
أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر	1.77	ترقيم ترقيم أداة	0.43
اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة	1.65	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر ترقيم أداة	0.43
أداة اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر	1.18	ترقيم أداة فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر	0.41

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي
أداة فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر	1.18	أداة اسم_معنى_معرف^بأل ـ مفرد ـ مؤنث أداة	0.37
أداة فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر	1.18	اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر	0.37
اسم_ذات_معرف^بأل ـ مفرد ـ مذكر أداة	0.94	أداة فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر أداة	0.37
أداة اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث	0.93	ترقيم أداة أداة	0.37
الكلمات الوظيفية: ٨		الكلمات الوظيفية: ٧	
كلمات المحتوى: ١٢		كلمات المحتوى: ٢١	

النمط اللغوي 4-grams	التكرار النسبي المئوي
ترقيم أداة أداة أداة	0.22
أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر	0.19
اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر أداة	0.16
أداة اسم_معنى_معرف^بأل ـ مفرد ـ مذكر ترقيم أداة	0.15
ترقيم ترقيم أداة أداة	0.13
ترقيم أداة أداة فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر	0.12

النمط اللغوي 2-grams	التكرار النسبي المئوي	النمط اللغوي 3-grams	التكرار النسبي المئوي
فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر	أداة_رابط^إحالي^غائب ـ مفرد ـ مذكر أداة	0.12
اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر ترقيم أداة		0.12
ترقيم أداة	أداة فعل_ماض _معلوم ـ غائب ـ مفرد ـ مذكر		0.11
الكلمات الوظيفية: ٢٥			
كلمات المحتوى: ١٥			

وبالنظر في الجدولين (٤-٨-أ) و(٤-٨-ب)، ومقارنتهما بالجدول (٤-٨-ج) يُلاحظ أنه مع توسيع الوسوم يزيد استعمال الكلمات الوظيفية في جميع التتابعات اللفظية، وذلك بسبب تفريع الأسماء التي يتصدر عددها في المدونة قائمة الكلمات، حتى أن النمط الأول الأكثر تكرارا في كل تتابع يرد دون أي كلمة محتوى، وترد أنماط أخرى دون كلمات محتوى على ترتيبات مختلفة. ففي التتابع على كلمتين ٨ كلمات محتوى و١٢ كلمة وظيفية، وفي التتابع على ٣ كلمات ٧ كلمات محتوى و٢١ كلمة وظيفية، وفي التتابع على ٤ كلمات ١٥ كلمة محتوى و٢٥ كلمة وظيفية. وقد استمر عدم ورود الأقسام الأخرى من الأسماء وغياب الصفة والضمير والظرف والخالفة مع زيادة في ورود الأفعال الماضية والمضارعة. والملاحظ فيما يخص الأفعال في الجداول الثلاثة أن ورود الأفعال مسبوقة بالأدوات أكثر من ورودها دونها، حيث تتنزل بعض الأدوات منزلة الجزء من الفعل كالسين مثلا[1]،

(١) المالكي، أبو محمد بدر الدين حسن بن قاسم بن عبد الله بن عليّ. الجنى الداني في حروف المعاني. تحقيق: د فخر الدين قباوة ـ الأستاذ محمد نديم فاضل، دار الكتب العلمية، بيروت: لبنان، ط١، ١٩٩٢، ص.٢٦

ويستلزم تركيب الجملة إذا ما ورد الفعل في داخل النص وليس أوله حتى وإن كان أول الجملة أن يكون مستأنفا أو معطوفا. فلا تبدأ الجملة داخل النص بفعل دون استئناف بفاء أو واو.

ويظهر الجدول (٤-٩) كل المتتابعات اللفظية الواردة في المدونة على تتابع ٣ كلمات للوسوم الرئيسة فقط، أما الجدول (٤-١٠) فيوضح نسبة ظهورها في المتتابعات اللفظية.

جدول (٤-٩) الأنماط للوسوم الرئيسة على الـ 3-grams

اسم أداة اسم	أداة اسم ضمير	ضمير أداة فعل	صفة اسم ضمير	اختصار اسم ترقيم	صفة ضمير صفة	صفة ترقيم ظرف	اسم اختصار اختصار
أداة اسم اسم	أداة أداة صفة	رقم اسم اسم	اسم ترقيم ظرف	فعل ضمير صفة	ضمير صفة فعل	ترقيم اسم رمز	ترقيم ترقيم ضمير
اسم اسم اسم	أداة فعل صفة	فعل اسم ضمير	أداة أداة ترقيم	اسم ضمير صفة	اختصار ترقيم ترقيم	ضمير ترقيم ضمير	اسم فعل ظرف
أداة اسم أداة	أداة صفة ترقيم	أداة فعل ضمير	ترقيم صفة اسم	اسم ضمير ضمير	ترقيم أجنبي ترقيم	رقم أداة رقم	فعل أجنبي أجنبي
اسم اسم أداة	اسم ترقيم فعل	ضمير اسم صفة	أداة ظرف أداة	ظرف أداة أداة	رقم صفة أداة	فعل ظرف ضمير	اسم أداة أجنبي
اسم ترقيم أداة	فعل ترقيم أداة	فعل ضمير اسم	ترقيم اسم فعل	رقم اسم ترقيم	أداة ظرف ضمير	رقم اسم فعل	ترقيم اختصار رمز
صفة أداة اسم	ضمير اسم أداة	فعل ترقيم ترقيم	فعل ضمير فعل	صفة ظرف فعل	فعل رقم اسم	فعل اختصار رقم	أجنبي اسم اسم
فعل أداة اسم	صفة أداة فعل	ضمير اسم ضمير	ترقيم أداة ظرف	اسم أجنبي أداة	صفة ظرف أداة	أجنبي أجنبي صفة	أجنبي فعل اسم

أداة فعل أداة	اسم اسم ضمير	صفة ترقيم ترقيم	ضمير فعل ضمير	اسم ظرف أداة	أداة اسم رمز	صفة اسم رمز	فعل ضمير ضمير
اسم صفة أداة	صفة أداة أداة	صفة صفة صفة	رقم ترقيم فعل	صفة فعل ضمير	ضمير ضمير صفة	صفة ترقيم رقم	أجنبي صفة صفة
اسم اسم ترقيم	فعل أداة فعل	اسم اسم رقم	فعل فعل أداة	صفة رقم اسم	اختصار فعل اسم	رقم أداة صفة	ترقيم اختصار ترقيم
أداة اسم صفة	ترقيم اسم اسم	أداة اسم رقم	فعل ضمير أداة	رقم أداة فعل	أداة ظرف اسم	رقم رمز اسم	رمز أداة فعل
أداة فعل اسم	أداة أداة ضمير	ضمير أداة أداة	ترقيم اسم ضمير	ضمير رقم اسم	أداة رقم رمز	صفة اسم ظرف	اسم فعل رقم
أداة أداة اسم	صفة صفة أداة	ترقيم ترقيم ترقيم	صفة صفة فعل	ضمير ضمير أداة	اختصار أداة فعل	ظرف ترقيم رقم	أداة صفة ظرف
ترقيم أداة أداة	ترقيم أداة ضمير	اسم ترقيم رقم	أداة رقم اسم	اسم اسم ظرف	ضمير صفة ضمير	ترقيم اسم رقم	فعل فعل ترقيم
أداة أداة فعل	اسم صفة صفة	فعل اسم فعل	اسم أداة ظرف	خالفة فعل اسم	أداة ضمير رقم	ترقيم فعل رقم	رمز ترقيم أداة
أداة اسم ترقيم	أداة ضمير أداة	ترقيم ترقيم فعل	ضمير أداة ضمير	أجنبي أجنبي أجنبي	ضمير ضمير اسم	ضمير أداة رقم	رقم اختصار أداة
فعل اسم اسم	فعل اسم ترقيم	أداة ترقيم اسم	ضمير ترقيم فعل	ترقيم رقم رقم	اسم رمز أداة	رمز رقم ترقيم	اسم أجنبي رقم
اسم أداة فعل	صفة اسم ترقيم	ضمير فعل ترقيم	ضمير صفة صفة	ضمير ضمير ضمير	رمز أداة أداة	اختصار ترقيم فعل	اسم فعل اختصار

اسم اسم صفة	ترقيم اسم ترقيم	فعل ترقيم اسم	رقم أداة اسم	أداة اسم ظرف	اسم اختصار رقم	فعل اسم خالفة	اسم اختصار ترقيم
فعل اسم أداة	فعل صفة اسم	ترقيم ترقيم اسم	صفة اسم رقم	اختصار ترقيم أداة	ترقيم رقم اسم	أداة ظرف ترقيم	ظرف اسم أداة
اسم أداة أداة	فعل صفة أداة	ظرف فعل اسم	أداة اسم اختصار	اسم صفة ظرف	رقم رمز أداة	فعل فعل ظرف	أداة أجنبي اسم
ترقيم أداة اسم	اسم ضمير اسم	أداة صفة فعل	ترقيم ظرف فعل	ترقيم رقم اختصار	ضمير أداة ترقيم	فعل رقم رمز	رمز اسم صفة
ترقيم أداة فعل	اسم صفة فعل	اسم رقم اسم	أداة ترقيم فعل	رقم رقم ترقيم	اختصار رقم اختصار	اسم رقم فعل	فعل أداة اختصار
أداة صفة أداة	صفة اسم صفة	ضمير فعل صفة	صفة أداة ترقيم	أداة ترقيم ترقيم	ترقيم خالفة فعل	أجنبي أداة فعل	أداة رقم صفة
اسم أداة صفة	ضمير اسم اسم	ترقيم صفة ترقيم	اسم أجنبي أجنبي	فعل صفة فعل	ضمير ترقيم صفة	اختصار ترقيم اختصار	أجنبي رقم أداة
أداة صفة اسم	فعل أداة ضمير	رقم ترقيم أداة	اسم ترقيم ضمير	ترقيم اختصار اسم	فعل ترقيم صفة	أداة أداة ظرف	اختصار رمز اسم
أداة أداة أداة	ترقيم أداة صفة	رقم ترقيم ترقيم	رقم اختصار ترقيم	اسم رقم رمز	فعل اختصار ترقيم	اسم اختصار أداة	اختصار اختصار أداة
اسم صفة اسم	أداة صفة صفة	ضمير صفة أداة	صفة ضمير أداة	أجنبي ترقيم اسم	رمز اسم ترقيم	ترقيم ترقيم خالفة	ترقيم أداة رمز
صفة ترقيم أداة	ترقيم رقم ترقيم	اسم رقم أداة	اسم فعل فعل	ضمير ترقيم اسم	ضمير صفة ترقيم	ضمير اختصار ترقيم	صفة فعل أجنبي

صفة اسم اسم	ضمير أداة اسم	ترقيم صفة أداة	اسم ترقيم اختصار	رقم اسم صفة	فعل ضمير ترقيم	صفة صفة ظرف	رمز فعل اسم
صفة اسم أداة	ترقيم فعل أداة	ترقيم فعل صفة	اسم اسم أجنبي	ترقيم ظرف أداة	أداة فعل اختصار	ظرف اسم اسم	اسم خالفة اسم
ضمير فعل أداة	صفة أداة ضمير	أداة صفة ضمير	أجنبي أداة اسم	ظرف أداة فعل	اسم رقم رقم	أداة خالفة أداة	فعل ترقيم رقم
اسم صفة ترقيم	صفة فعل اسم	أداة فعل فعل	رقم أداة أداة	اسم رقم صفة	صفة رقم ترقيم	رقم رقم اسم	اختصار رقم صفة
فعل أداة أداة	صفة اسم فعل	رقم ترقيم رقم	ترقيم فعل ضمير	اسم ترقيم أجنبي	اسم خالفة فعل	أداة فعل رقم	اختصار صفة فعل
أداة ضمير اسم	ترقيم اسم أداة	فعل أداة ترقيم	رقم اسم أداة	اسم أداة رقم	اسم اختصار فعل	أجنبي ضمير اسم	ترقيم ضمير أداة
اسم فعل اسم	اسم صفة ضمير	ضمير أداة صفة	ظرف فعل أداة	ضمير اسم رقم	رقم فعل اسم	صفة ضمير رقم	رقم اسم رقم
اسم أداة ضمير	صفة ضمير فعل	ترقيم صفة صفة	ترقيم ضمير اسم	ظرف ضمير اسم	أجنبي أداة صفة	صفة صفة رقم	ترقيم أداة رقم
صفة أداة صفة	صفة ترقيم اسم	فعل صفة صفة	صفة ترقيم صفة	اختصار ترقيم اسم	اسم ترقيم خالفة	صفة ضمير ترقيم	ترقيم ضمير ضمير
اسم ضمير فعل	صفة صفة اسم	ترقيم ترقيم رقم	فعل اسم رقم	رقم صفة اسم	اسم أجنبي فعل	خالفة اسم أداة	رمز رقم رمز
اسم فعل أداة	صفة ترقيم فعل	ترقيم ترقيم صفة	اسم فعل ضمير	اختصار رقم أداة	رقم اختصار اختصار	رقم رمز فعل	أداة ضمير اختصار
اسم ترقيم اسم	أداة ترقيم أداة	أداة ضمير ضمير	صفة فعل صفة	رقم ترقيم اسم	فعل أداة رقم	فعل فعل صفة	أجنبي أداة ضمير

أداة اسم فعل	فعل ترقيم فعل	ضمير ترقيم أداة	ترقيم أداة ترقيم	رمز رقم اسم	اسم اسم خالفة	رقم ترقيم أجنبي	اختصار ترقيم رقم
أداة ضمير فعل	صفة فعل أداة	اسم فعل ترقيم	صفة ترقيم ضمير	رقم أداة ضمير	صفة أداة ظرف	ظرف أداة ضمير	ظرف فعل صفة
ضمير فعل اسم	ترقيم فعل ترقيم	ضمير صفة اسم	أداة اسم أجنبي	أداة ترقيم رقم	فعل أداة خالفة	فعل صفة رقم	صفة رقم أداة
فعل اسم صفة	ترقيم اسم صفة	اسم ضمير ترقيم	اسم رقم اختصار	صفة صفة ضمير	ضمير فعل رقم	اسم رمز ترقيم	أداة اختصار صفة
ترقيم فعل اسم	صفة صفة ترقيم	ترقيم ضمير فعل	اسم صفة رقم	أداة ترقيم صفة	فعل ترقيم ضمير	فعل رقم صفة	ترقيم صفة ضمير
أداة فعل ترقيم	اسم ضمير أداة	فعل فعل اسم	اسم ظرف فعل	ظرف ضمير أداة	ظرف فعل ترقيم	ترقيم رمز رقم	رقم رقم أداة
اسم اسم فعل	اسم ترقيم صفة	صفة ضمير اسم	أداة ظرف فعل	فعل اسم أجنبي	رقم صفة ترقيم	فعل ظرف فعل	فعل أداة ظرف
ترقيم ترقيم أداة	اسم أداة ترقيم	أداة ضمير ترقيم	أجنبي أجنبي أداة	رقم ترقيم ضمير	اختصار أداة	ترقيم أجنبي ضمير	أداة رمز أداة
اسم ترقيم ترقيم	أداة ضمير صفة	ظرف أداة اسم	رقم رمز رقم	ضمير فعل فعل	اختصار اختصار فعل	اسم ظرف ضمير	اسم رمز اسم
فعل أداة صفة	اسم فعل صفة	ضمير ضمير فعل	فعل صفة ضمير	رقم اختصار رقم	صفة ترقيم رمز	خالفة أداة اسم	صفة ضمير ضمير
أداة اسم ضمير	ضمير اسم ترقيم	فعل صفة ترقيم	صفة أداة رقم	رمز أداة اسم	اسم اسم اختصار	أجنبي أجنبي ترقيم	ضمير اسم ظرف
أداة أداة صفة	ضمير اسم فعل	عدد الأنماط:	٤٢٥				

جدول (٤-١٠) توزيعات الأنماط للوسوم الرئيسة على الـ 3-grams

توزيع الأنماط على مستوى الـ 3n-grams	التكرارات
اسم	17.18
أداة	15.37
ترقيم	13.18
فعل	12.08
صفة	11.29
ضمير	9.49
رقم	8.16
اختصار	3.76
ظرف	3.37
أجنبي	2.82
رمز	2.35
خالفة	0.94
مجموع توزيع الأنماط على ثلاث كلمات n-grams	**التكرارات النسبية**
الأنماط للوسوم الرئيسة للكلمات الوظيفية	57.57
الأنماط للوسوم الرئيسة للكلمات ذات المحتوى	42.43

ويعود تقدم الكلمات الوظيفية على الكلمات ذات المحتوى هنا إلى طبيعة فرز التكرار لـ n-grams على ثلاث كلمات متتابعة. فلو كان لدينا جملة من تسعة عشر كلمة فعلية، على نحو:

وإذا رجعنا إلى الشعر نفسه ل التعرف على مكانة الطبيعة فيه؛ تبين أن الشعراء لا يستغنون عنها

سنلحظ كما في الجدول (٤-١١) استمرارية ازدياد الكلمات الوظيفية على الكلمات ذات المحتوى، ولو قسنا ذلك على نص يزيد على المليون الكلمة بدلا من تسع عشرة كلمة، فإن رجحان الازدياد لصالح الكلمات الوظيفية سيزيد نسبة أكثر وأكثر، وهذا هو تفسير **قانون زيف** Zipf›s Law للمدونات اللغوية. وبالتالي فإن التركيز عليها في الدراسات اللغوية الحاسوبية سيحل كثيرا من المشكلات المتعلقة بالأنظمة اللغوية الحاسوبية، حيث ستكون الكلمات القلة الأكثر تكرارا (الكلمات الوظيفية) هي المسؤولة عن الصحة والدقة ومشكلات أخرى بنسبة أكثر من الكلمات الأكثر الأقل تكرارا (كلمات المحتوى)، خصوصا وأنها كلمات ذات قوائم مغلقة.

جدول (٤-١١) طبيعة فرز التكرار لـ n-grams على ثلاث كلمات متتابعة

ثلاث كلمات متتابعة	التكرار	تكرار الكلمات الوظيفية	تكرار الكلمات ذات المحتوى
و إذا رجعنا	١	٢	١
إذا رجعنا إلى	١	٢	١
رجعنا إلى الشعر	١	١	٢
إلى الشعر نفسه	١	٢	١
الشعر نفسه ل	١	٢	١
نفسه ل التعرف	١	٢	١
ل التعرف على	١	٢	١
التعرف على مكانة	١	١	٢
على مكانة الطبيعة	١	١	٢
مكانة الطبيعة فيه	١	١	٢
الطبيعة فيه؛	١	٢	١

ثلاث كلمات متتابعة	التكرار	تكرار الكلمات الوظيفية	تكرار الكلمات ذات المحتوى
فيه؛ تبين	١	٢	١
؛ تبين أن	١	٢	١
تبين أن الشعراء	١	١	٢
أن الشعراء لا	١	٢	١
الشعراء لا يستغنون	١	١	٢
لا يستغنون عنها	١	٢	١
المجموع	١٧	٢٨	٢٣

وهكذا، حاولت استكشاف بعض خصائص المدونة ومعرفة طبيعتها اللغوية من خلال النظر في تكرارات الكلمات فيها قبل التقطيع وبعد التقطيع وبعد التوسيم، ثم الوقوف على الأنماط التتابعية من الناحية النحوية المعجمية (النحو المعجمي lexical grammar)، وذلك بالمقارنة بين مستويات الوسوم الثلاثة في أكثر عشر تكرارات على تتابع كلمتين وثلاث كلمات وأربع كلمات n-grams، كاشفة عن تناسب نصوص مدونة النظام وانسجامها من خلال تطبيقها قانون زيف.

٤-٢ نتائج التقييم التقني:

ثمة العديد من خوارزميات التصنيف التي تستعمل في بناء أنظمة التوسيم النحوي الآلية. وهي متفقة معا في عملها لكنها مختلفة في نظامها الرياضي والتقني عند حل مشكلات التصنيف. ومن هذه الخوارزميات خوارزمية **الحقول العشوائية المشروطة** Conditional Random Fields (CRF) التي ستستعمل في هذا المقترح. وهي خوارزمية من خوارزميات تعلم الآلة تحت الإشراف تستعمل لبناء نماذج احتمالية تمييزية تتنبأ بوسوم كلمات متتالية

مع أخذ السياق بعين الاعتبار[١]، وقد استعملت أيضا في موسم ستانفورد[٢].

ويمر بناء الموسم النحوي الآلي باستعمال هذه الخوارزمية بعدد من المراحل هي كالتالي:

١- إعداد البيانات: وفيها قسمت المدونة التي وسمتها نحويا إلى قسمين: الأول يمثل ٨٠٪ من المدونة الأصلية الموسمة (١١٠٢١ كلمة فعلية)، وسيكون (مدونة التدريب)، وراعيت أن تكون ٨٠٪ من كل وعاء على حدة، والثاني يمثل الـ ٢٠٪ المتبقية من المدونة الأصلية (٢٧٩٢ كلمة فعلية) وسيكون (مدونة الاختبار). بعد ذلك قما بنسخ بيانات مدونة التدريب ومدونة الاختبار الممثلة على عمودين متجاورين بصيغة أكسل ونقلها بنفس التمثيل إلى ملفين منفصلين بصيغة نصية txt أحدهما مدونة التدريب والآخر مدونة الاختبار مع إضافة سطر جديد بعد الجمل بطريقة يدوية لحاجتنا لذلك في المرحلة التالية.

٢- تجهيز بيانات التدريب والاختبار: ونحتاج هنا إلى بيانات التدريب وفحصها، وتحديد الخصائص المستعملة في النموذج وتمثيلها للتقطيع وللتوسيم.

أ ـ بيانات التدريب للتقطيع: استعنت بلغة البرمجة (**البايثون** Python) وبمكتبة lazyme لتحويل البيانات التي أُعدّت إلى صورة تقبلها المكتبات الخاصة بالتصنيف في التقطيع (انظر الشكل ٤-٢). ثم استوردت مكتبة nltk الخاصة بمعالجة اللغات الطبيعية للاستفادة منها في تقطيع النصوص. وأما دالة features فاستعملتها لاستخلاص بعض

---

(١) Lafferty John D., McCallum A., Pereira, Fernando C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML ‹01 Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco: USA, 2001, p. 282

(٢) The Stanford Natural Language Processing Group. Arabic Natural Language Processing, 8-12-2018:
http://nlp.stanford.edu/software/corenlp.shtml

الخصائص المتعلقة بالتقطيع. وهذه الخصائص تسمى **خصائص الحالة**[١] State Features، ويشترط أن تفهرس من بيانات التدريب وتزود بها الخوارزمية لتعمل وفقها. لقد استخلصت خصائص التقطيع من خلال كل حرف داخل السلاسل الحرفية في الجمل، بعد تعريف الحرف وإعطائه موقعا في السلسلة الجُملية، ثم إسناد الرقم ١ لما بعد الحرف المستحق للفصل والرقم ٠ إذا لم يستحق الفصل (انظر الشكل ٤-٤). وكانت خصائص التقطيع كالتالي:

- الحرف نفسه والحرفان السابقان والحرفان اللاحقان.
- هل الحرف رقم أم لا.

ب ـ بيانات التدريب للتوسيم: وقد استعين فيها بنفس المكتبات في مرحلة التقطيع، وحوّلت فيها البيانات إلى نفس صورة تحويلها في التقطيع بحيث تقبلها المكتبات الخاصة بالتصنيف في التوسيم (انظر الشكل ٤-٣). واستعملت أيضا نفس الدالة features لاستخلاص بعض الخصائص المتعلقة بالتوسيم. وكانت الخصائص من خلال كل كلمة داخل جملة (انظر الشكل ٤-٥). وهي هنا كالتالي:

- الكلمة نفسها والكلمة السابقة واللاحقة لها.
- الكلمة نفسها وحروفها الثلاثة الأولى والأخيرة.
- ما إذا كانت الكلمة رقما أم لا.
- ما إذا كانت الكلمة في أول جملة أو نهايتها.

ويمكن أن يُحسن من هذا النموذج بزيادة خصائص أخرى مميِّزة للتقطيع والتوسيم حيث تعتمد جودة النموذج عليها. وقد نُقلت بعد ذلك بيانات التدريب والاختبار معا

---

(١) A. NLP Guide: Identifying Part of Speech Tags using Conditional ,Ramachandran :9-12-2018 ,Random Fields. Analytics Vidhya
https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31

لمجموعتي بيانات dataset. ففُصل في المجموعة الأولى الخاصة بالتقطيع ما أسند إلى الحروف عن الحروف لتصبح الأرقام المسندة في قائمة مستقلة عن الحروف، أما في المجموعة الثانية، ففصلت الوسوم عن الكلمات لتصبح الوسوم في قائمة مستقلة عن الكلمات. وبالتالي أصبحت القائمتان في ملفين منفصلين بصورة متسلسلة؛ إذ إن CRF تتعلم من **السلاسل المتوالية** sequences (انظر الشكل ٤-٦).

شكل (٤-٢) كود تجهيز مدونة النظام للتقطيع

```
from lazyme import per_section
import nltk
training_sentences = [[tuple(token.split('\t')) for token in
sent] for sent in
per_section(open('C:\\Test\\Seg\\TRAIN1.txt'))]
test_sentences = [[tuple(token.split('\t')) for token in sent]
for sent in per_section(open('C:\\Test\\Seg\\TEST1.txt'))]
```

شكل (٤-٣) كود تجهيز مدونة النظام للتوسيم

```
from lazyme import per_section
import nltk
training_sentences = [[tuple(token.split('\t')) for token in
sent] for sent in
per_section(open('C:\\Test\\Training\\L1\\Train_L1.txt'))]
test_sentences = [[tuple(token.split('\t')) for token in sent]
for sent in
per_section(open('C:\\Test\\Testing\\L1\\Test_L1.txt'))]
print(len(training_sentences))
print(len(test_sentences))
```

شكل (٤-٤) كود خصائص الحالة للتقطيع

```
def features(sentence, index):
       """sentence: [w1, w2, ...], index: the index of the
word"""
    return{
      'word': sentence[index],
      'prev_word1': '<s>' if index == 0 else sentence[index -
      1],
      'prev_word2': '<s>' if index == 0 else '<s>' if index ==
      1 else sentence[index - 2],
      'next_word1': '/<s>' if index == len(sentence) - 1 else
      sentence[index + 1],
      'next_word2': '/<s>' if index == len(sentence) - 1 else
      '/<s>' if index == len(sentence) - 2 else sentence[index
      + 2],
                    'is_numeric': sentence[index].isdigit(),
```

(٤-٥) شكل كود خصائص الحالة للتوسيم

```
def features(sentence, index):
    """ sentence: [w1, w2, ...], index: the index of the word
"""
    return {
        'word': sentence[index],
        'is_first': index == 0,
        'is_last': index == len(sentence) - 1,

        'prefix-1': sentence[index][0],
        'prefix-2': sentence[index][:2],
        'prefix-3': sentence[index][:3],
        'prefix-4': sentence[index][:4],
        'suffix-1': sentence[index][-1],
        'suffix-2': sentence[index][-2:],
        'suffix-3': sentence[index][-3:],
         prev_word1': '' if index == 0 else sentence[index -
1],
        'next_word1': '' if index == len(sentence) - 1 else
sentence[index + 1],
        'is_numeric': sentence[index].isdigit(),
```

شكل (٤-٦) كود نقل بيانات التدريب والاختبار لمجموعة بيانات dataset

```
from nltk.tag.util import untag
def transform_to_dataset(tagged_sentences):
    X, y[] ،[] =
     for tagged in tagged_sentences:
        X.append([features(untag(tagged), index) for index in
range(len(tagged))])
        y.append([tag for _, tag in tagged])
    return X, y
X_train, y_train = transform_to_dataset(training_sentences)
X_test,   y_test   =   transform_to_dataset(test_sentences)
print(len(X_train))
print(len(X test))
```

٣- تدريب نموذج CRF وقياس الأداء: وقد استعملت مكتبة sklearn_crfsuite لتدريب خوارزمية CRF باستخدام الإعدادات الافتراضية، وأُعدّ النموذج لتوليد كل **انتقالات** transitions الوسوم الممكنة، حتى تلك التي لم ترد في مدونة التدريب وأقصاها ١٠٠,٠٠٠ انتقال (انظر الشكل ٤-٧). ثم تم قياس دقة النموذج باستخدام مكتبة sklearn_crfsuite كما يوضح شكل (٤-٨).

شكل (٤-٧) كود تدريب نموذج CRF للتقطيع والتوسيم

```
from sklearn_crfsuite import metrics

y_pred = crf.predict(X_test)
print("Accuracy    =    ",metrics.flat_accuracy_score(y_test,
y_pred))
print ("Recall =", metrics.flat_recall_score(y_test, y_pred,
average='weighted', labels=labels))
print ("Precision =", metrics.flat_precision_score(y_test,
y_pred, average='weighted', labels=labels))
print ("F1 =", metrics.flat_f1_score(y_test, y_pred,
average='weighted', labels=labels))
```

شكل (٤-٨) كود اختبار نموذج CRF للتقطيع والتوسيم

```
from sklearn_crfsuite import CRF
import sklearn_crfsuite
from sklearn_crfsuite import scorers
from sklearn_crfsuite import metrics

crf = sklearn_crfsuite.CRF(algorithm='lbfgs',
    c1=0.01,
    c2=0.01,
    max_iterations=100000,
    all_possible_transitions=True)
crf.fit(X_train, y_train)
```

وجاءت نتائج الصحة بعد قياس الأداء للنظام بمقارنة مدونة الاختبار المقطعة يدويا بمدونة الاختبار التي قطعت بعد اختبار نموذج التقطيع المقترح عليها آليا بدقة عالية جدا في كل مقاييس الأداء الأربعة حيث بلغت في كلٍّ ٠,٩٩٣ (انظر الجدول ٤-١٢).

جدول (٤-١٢) مقاييس الأداء للمقطع

المقياس	درجته
الصحة	0.9925
الدقة	0.9928
الاسترجاع	0.9925
مقياس ف	0.9926

وتبين سلسلة الجداول (٤-١٣) على التوالي نتائج قياس الأداء للنظام بمقارنة مدونة الاختبار الموسمة يدويا بمدونة الاختبار التي وسمت بعد اختبار النموذج المقترح عليها آليا في المستويات الثلاثة لمجموعات الوسوم المقترحة. ويلاحظ أن أدق الوسوم وسوم المستوى الأول (الوسوم الرئيسة) حيث بلغت دقتها ٩٢ , ٠، وبلغت دقة وسوم المستوى الثاني (الوسوم الرئيسة والفرعية معا) ٨٢ , ٠، فيما جاءت الوسوم الموسعة (الوسوم الرئيسة والفرعية مع الخصائص التصريفية) بدقة ٧٢ , ٠. وجاءت نتائج الصحة بفوارق متشابهة قدرها ١٠٪، وهذا مؤشر على أن نسب الأخطاء لم تتشتت بقيم عشوائية لتشابه الفوارق بينها. فقد بلغت نسبة الخطأ محسوبة من الصحة في مجموعة الوسوم الرئيسة ٠,٠٨٤ وهي نسبة ضئيلة لا تكاد تذكر، وفي الوسوم الفرعية زادت نسبة الخطأ قليلا إلى ١٧٩ , ٠؛ لأن أغلب الوسوم قليلة التكرارات. أما في الوسوم الموسعة بالخصائص التصريفية فقد زادت نسبة الخطأ إلى ٢٧٧ , ٠؛ بسبب صغر حجم مدونة الاختبار الذي جعل أغلب الوسوم الموسعة بلا أمثلة.

جدول (٤-١٣-أ) مقاييس الأداء لموسم الأقسام الرئيسة

المقياس	درجته
الصحة	0.9158
الدقة	0.915
الاسترجاع	0.916
مقياس ف	0.913

جدول (٤-١٣-ب) مقاييس الأداء لموسم الأقسام الفرعية

المقياس	درجته
الصحة	0.8204
الدقة	0.816
الاسترجاع	0.820
مقياس ف	0.822

جدول (٤-١٣-ج) مقاييس الأداء لموسم الأقسام الموسعة

المقياس	درجته
الصحة	0.7210
الدقة	0.720
الاسترجاع	0.735
مقياس ف	0.713

وتوضح سلسلة الجداول (٤-١٤) أن كثرة الأمثلة في مدونة التدريب لم تؤثر في دقة الأسماء والصفات والأفعال (كلمات المحتوى)، حيث يظهر أن وسوم الكلمات الوظيفية في مجموعة الوسوم الرئيسة أعلى دقة منها، وذلك لأن الكلمات الوظيفية قوائم مغلقة (محدودة ومحصورة) وأكثر تكرارا من كلمات المحتوى. وفي قيم الدقة لمجموعة الوسوم الفرعية في الجدول (٤-١٤-ب) نلاحظ أن الأسماء المبهمة NI حين انفصلت عن باقي الأسماء ـ وهي كلمات وظيفية ذات قوائم مغلقة ـ انفصلت بدقة أعلى من باقي الأقسام الفرعية للأسماء التي ظلت تشابه دقتها دقة الأسماء في مجموعة الوسوم الرئيسة. وفي قيم مجموعة الوسم الموسعة في الجدول (٤-١٤-ج) ما زالت الكلمات الوظيفية تحافظ على دقة عالية رغم توسع الوسوم وتراجع دقة بعض وسوم كلمات المحتوى.

جدول (٤-١٤-أ) قيم الدقة والاسترجاع ومقياس ف لمجموعة الوسوم الرئيسة

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
N	1001	0.863	0.934	0.897
RP	887	0.991	0.982	0.986
V	263	0.889	0.856	0.872
A	261	0.783	0.609	0.685
PUNC	217	1.000	1.000	1.000

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
P	88	0.988	0.955	0.971
DIGIT	21	1.000	1.000	1.000
D	9	0.900	1.000	0.947
SYMB	7	1.000	0.286	0.444
FOREIGN	2	1.000	1.000	1.000
I	1	0.000	0.000	0.000
ABBREV	0	0.000	0.000	0.000

جدول (٤-١٤-ب) قيم الدقة والاسترجاع ومقياس ف لمجموعة الوسوم الفرعية

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
RP	887	0.983	0.988	0.985
NA	443	0.646	0.842	0.731
NC	357	0.667	0.594	0.628
PUNC	217	1	1	1
VS	150	0.803	0.787	0.795
AS	128	0.655	0.609	0.632
NI	125	0.811	0.72	0.763
VP	109	0.861	0.798	0.829
AA	55	0.711	0.491	0.581
AO	52	0.518	0.558	0.537
NV	39	0.773	0.436	0.557
PR	38	0.947	0.947	0.947
PD	33	0.97	0.97	0.97

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
NL	29	0.625	0.345	0.444
AC	25	0.571	0.32	0.41
DIGIT	21	1	1	1
PP	17	1	0.941	0.97
SYMB	7	1	0.429	0.6
NM	7	0	0	0
DT	6	1	1	1
VC	4	0	0	0
DL	3	1	1	1
FOREIGN	2	1	1	1
NT	1	0.5	1	0.667
IV	1	0	0	0
AE	1	0	0	0
ABBREV	0	0	0	0

جدول (٤-١٤-ج) قيم الدقة والاسترجاع ومقياس ف لمجموعة الوسوم الموسعة

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
AA_DLF	2	0	0	0	AC_ILF	0	0	0	0	NL_ILF_R3RSRF	2	0	0	0	VS_3SM_R3RSRM	0	0	0	0
NA_DLF	41	0.49	0.54	0.45	NC_ILF	14	0.29	0.14	0.19	NL_ISF	0	0	0	0	VS_A1LM	4	0.67	0.50	0.57
NA_DLF_AT	0	0	0	0	NC_ILF_R2RLRM	1	0	0	0	NL_ISF_R3RSRF	0	0	0	0	VS_A1LM_R3RSRM	1	1.0	1.0	1.0
NA_DLF_R1RLRM	0	0	0	0	NC_ILF_R3RLRF	0	0	0	0	NL_ISM	12	0.54	0.58	0.56	VS_A1SM	13	1.0	0.08	0.14
AA_DLM	5	0.80	0.80	0.80	NC_ILF_R3RLRM	3	0.25	0.33	0.29	NL_ISM_R1RSRM	1	0	0	0	VS_A2SF	0	0	0	0

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
NA_ DLM	0	0	0	0	NC_ILF_ R3RSRF	3	0	0	0	NL_ISM_ R3RSRF	0	0	0	0	VS_ A2SM	3	0.25	0.33	0.29
AA_ DLM_AT	0	0	0	0	NC_ILF_ R3RSRM	0	0	0	0	NL_ISM_ R3RSRM	0	0	0	0	VS_A3LF	3	0.25	0.33	0.29
NA_ DLM_AT	2	0	0	0	NC_ILF_ R3RURM	0	0	0	0	NL_IUM	0	0	0	0	VS_ A3LF_ R1RLRM	0	0	0	0
AA_DSF	5	0.50	0.40	0.67	AC_ILM	0	0	0	0	NM_DLF	0	0	0	0	VS_ A3LF_ R2RSRM	0	0	0	0
NA_DSF	40	0.49	0.68	0.38	NC_ILM	2	0	0	0	NM_DSF	0	0	0	0	VS_ A3LF_ R3RSRM	0	0	0	0
AA_ DSF_AT	0	0	0	0	NC_ ILM_AT	1	0	0	0	NM_ DSM	2	0	0	0	VS_ A3LM	5	1.0	0.60	0.75
NA_ DSF_AT	14	0.67	0.79	0.58	NC_ ILM_ R1RSRM	0	0	0	0	NM_ DSM_AT	0	0	0	0	VS_ A3LM_ R1RSRM	0	0	0	0
NA_ DSF_ R3RLRM	0	0	0	0	NC_ ILM_ R3RLRM	0	0	0	0	NM_ILF	2	0	0	0	VS_ A3LM_ R3RLRF	0	0	0	0
NA_ DSF_ R3RSRM	0	0	0	0	NC_ ILM_ R3RSRM	0	0	0	0	NM_ISF_ AT	0	0	0	0	VS_ A3LM_ R3RSRF	1	0	0	0
AA_DSM	14	0.52	0.43	0.67	AC_ISF	2	1.0	0.50	0.67	NM_ISM	2	0	0	0	VS_ A3LM_ R3RSRM	1	0	0	0
NA_DSM	88	0.64	0.81	0.53	NC_ISF	24	0.39	0.29	0.33	FOR-EIGN	2	0	0	0	VS_A3SF	15	0.50	0.67	0.57
AA_ DSM_AT	0	0	0	0	NC_ ISF_12	0	0	0	0	AO_DLF	2	0.67	1.0	0.80	VS_ A3SF_ R2RSRM	0	0	0	0
NA_ DSM_AT	9	0.44	0.44	0.44	NC_ ISF_22	0	0	0	0	AO_DLM	3	0	0	0	VS_ A3SF_ R3RSRF	0	0	0	0
NA_ DSM_ R3RLRM	0	0	0	0	NC_ISF_ AT	2	0	0	0	AO_DSF	8	0.6	0.38	0.46	VS_ A3SF_ R3RSRM	0	0	0	0
NA_ DSM_ R3RSRM	1	0	0	0	NC_ISF_ R1RSRM	0	0	0	0	AO_DSF_ R3RSRM	0	0	0	0	VS_ A3SM	74	0.57	0.70	0.63

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
NA_DUF	0	0	0	0	NC_ISF_R3RSRF	1	0	0	0	AO_DSM	10	0.33	0.60	0.43	VS_A3SM_R1RLRM	5	1.0	0.80	0.89
AA_DUM	0	0	0	0	NC_ISF_R3RSRM	0	0	0	0	AO_DUM	0	0	0	0	VS_A3SM_R1RSRM	0	0	0	0
NA_DUM	0	0	0	0	AC_ISM	10	0.43	0.30	0.35	AO_ILF	4	0.67	0.50	0.57	VS_A3SM_R3RLRF	1	0	0	0
AA_ILF	0	0	0	0	NC_ISM	102	0.54	0.68	0.60	AO_ILF_R3RSRM	0	0	0	0	VS_A3SM_R3RSRF	2	0	0	0
NA_ILF	21	0.35	0.43	0.29	NC_ISM_12	11	0.43	0.55	0.48	AO_ILM	0	0	0	0	VS_A3SM_R3RSRM	4	0.40	1.0	0.57
NA_ILF_AT_R1RLRM	0	0	0	0	NC_ISM_13	0	0	0	0	AO_ISF	9	0	0	0	VS_A3UM	0	0	0	0
NA_ILF_AT_R3RSRF	0	0	0	0	NC_ISM_22	11	0.43	0.55	0.48	AO_ISF_R3RSRM	0	0	0	0	AS_DLF	12	0	0	0
NA_ILF_R1RLRM	1	1.0	1.0	1.0	NC_ISM_23	0	0	0	0	AO_ISM	14	0.56	0.36	0.44	AS_DLF_12	0	0	0	0
NA_ILF_R1RSRM	0	0	0	0	NC_ISM_33	0	0	0	0	AO_ISM_R2RSRM	0	0	0	0	AS_DLF_22	0	0	0	0
NA_ILF_R2RLRM	1	1.0	1.0	1.0	NC_ISM_AT	1	0	0	0	AO_ISM_R3RSRM	0	0	0	0	AS_DLM	25	0.52	0.68	0.59
NA_ILF_R3RLRM	0	0	0	0	NC_ISM_R1RLRM	0	0	0	0	AO_IUF	1	1	1.0	1.0	AS_DLM_AT	0	0	0	0
NA_ILF_R3RSRF	4	0	0	0	NC_ISM_R1RSRM	0	0	0	0	RP	819	0.97	0.99	0.98	AS_DLM_R3RSRF	0	0	0	0
AA_ILF_R3RSRM	0	0	0	0	NC_ISM_R2RSRM	1	0	0	0	PP_1LM	2	1	1.0	1.0	AS_DSF	11	0.57	0.73	0.64
NA_ILF_R3RSRM	0	0	0	0	NC_ISM_R3RLRF	0	0	0	0	PP_1SM	3	1	1.0	1.0	AS_DSF_13	0	0	0	0
AA_ILM	2	0	0	0	NC_ISM_R3RLRM	0	0	0	0	PP_2SM	0	0	0	0	AS_DSF_23	0	0	0	0

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
NA_ILM	2	0	0	0	AC_ ISM_ R3RSRF	1	0	0	0	PP_3LF	1	0	0	0	AS_ DSF_33	0	0	0	0
NA_ ILM_AT	0	0	0	0	NC_ ISM_ R3RSRF	0	0	0	0	PP_3LM	0	0	0	0	AS_DSF_ AT	3	0	0	0
AA_ ILM_ R1RSRM	0	0	0	0	NC_ ISM_ R3RSRM	1	0	0	0	PP_3SF	0	0	0	0	AS_DSM	29	0.54	0.45	0.49
NA_ ILM_ R3RSRF	0	0	0	0	NC_IUF	3	0	0	0	PP_3SM	11	1	0.82	0.90	AS_ DSM_12	1	0	0	0
AA_ ILM_ R3RSRM	0	0	0	0	NC_IUF_ R3RSRF	0	0	0	0	PP_3UF	0	0	0	0	AS_ DSM_22	1	0	0	0
AA_ISF	4	0.29	0.25	0.33	NC_IUF_ R3RSRM	1	1.0	1.0	1.0	PP_3UM	0	0	0	0	AS_ DSM_AT	0	0	0	0
NA_ISF	67	0.69	0.81	0.60	NC_IUM	0	0	0	0	VP_ A1LM	4	0.13	0.25	0.17	AS_DUF	1	0	0	0
NA_ISF_ AT	6	0.80	0.67	1.0	NC_ IUM_ R2RSRF	0	0	0	0	VP_ A1LM_ R3RSRF	0	0	0	0	AS_DUM	1	0	0	0
NA_ISF_ R1RSRM	0	0	0	0	PD	1	1.0	1.0	1.0	VP_ A1LM_ R3RSRM	1	0	0	0.67	AS_ILF	2	1.0	0.50	0.67
NA_ISF_ R3RLRM	2	0	0	0	PD_LF	2	0	0	0	VP_ A1SM	12	1	0.08	0.15	AS_ILM	3	0	0	0
NA_ISF_ R3RSRF	3	0.50	0.67	0.40	PD_SF	10	0.91	1.0	0.95	VP_ A1SM_ R2RSRF	0	0	0	0	AS_ILM_ R1RSRM	0	0	0	0
NA_ISF_ R3RSRM	2	0.36	1.0	0.22	PD_SM	20	0.95	1.0	0.98	VP_ A1SM_ R3RSRM	1	0	0	0	AS_ILM_ R3RSRM	1	0	0	0
NA_ISF_ R3RURM	0	0	0	0	AE_DLF	0	0	0	0	VP_ A2LM	0	0	0	0	AS_ILM_ R3RURM	0	0	0	0
AA_ISM	21	0.64	0.67	0.61	AE_DLM	0	0	0	0	VP_ A2LM_ R3RSRF	0	0	0	0	AS_ISF	11	0.54	0.64	0.58
NA_ISM	99	0.68	0.80	0.59	AE_DSM	0	0	0	0	VP_ A2SM	1	0	0	0	AS_ISF_ R3RSRF	1	0	0	0
NA_ ISM_AT	3	0	0	0	AE_ILF	0	0	0	0	VP_ A2SM_ R2RSRM	0	0	0	0	AS_ISF_ R3RSRM	0	0	0	0

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
AA_ ISM_ R1RLRM	0	0	0	0	AE_ILM	0	0	0	0	VP_ A2SM_ R3RSRM	0	0	0	0	AS_ISM	26	0.50	0.54	0.52
NA_ ISM_ R1RLRM	0	0	0	0	AE_ISF	0	0	0	0	VP_A3LF	7	0	0.29	0.29	AS_ISM_ AT	0	0	0	0
NA_ ISM_ R1RSRM	0	0	0	0	AE_ISM	0	0	0	0	VP_ A3LF_ R3RSRF	0	0	0	0	AS_ISM_ R3RSRF	0	0	0	0
NA_ ISM_ R2RLRM	0	0	0	0	DIGIT	21	1.0	1.0	1.0	VP_ A3LF_ R3RSRM	0	0	0	0	AS_ISM_ R3RSRM	0	0	0	0
AA_ ISM_ R2RSRM	0	0	0	0	NI_DLF	3	0	0	0	VP_ A3LM	7	1	0.43	0.60	AS_IUM	0	0	0	0
NA_ ISM_ R2RSRM	0	0	0	0	NI_DSF	4	0.25	0.25	0.25	VP_ A3LM_ R1RLRM	0	0	0	0	VS_P1SM	0	0	0	0
NA_ ISM_ R3RFRM	0	0	0	0	NI_DSF_ AT	1	0	0	0	VP_ A3LM_ R2RSRM	0	0	0	0	VS_P3LF	2	0	0	0
NA_ ISM_ R3RLRF	0	0	0	0	NI_DSM	6	0.50	0.50	0.50	VP_ A3LM_ R3RLRF	0	0	0	0	VS_P3LM	0	0	0	0
NA_ ISM_ R3RLRM	8	0.33	0.25	0.50	NI_ DSM_AT	0	0	0	0	VP_ A3LM_ R3RSRF	0	0	0	0	VS_P3SF	1	0	0	0
NA_ ISM_ R3RSRF	6	0.62	0.67	0.57	NI_DUM	0	0	0	0	VP_ A3LM_ R3RSRM	0	0	0	0	VS_P3SM	8	0.33	0.13	0.18
NA_ ISM_ R3RSRM	16	0.58	0.69	0.50	NI_ILF	1	1.0	1.0	1.0	VP_A3SF	12	0.4	0.50	0.44	DT	6	1.0	0.83	0.91
NA_IUF	0	0	0	0	NI_ILF_ R1RLRM	0	0	0	0	VP_ A3SF_ R1RLRM	0	0	0	0	NT_DLF	0	0	0	0
NA_IUF_ R3RSRF	0	0	0	0	NI_ILF_ R3RLRM	0	0	0	0	VP_ A3SF_ R3RLRM	0	0	0	0	NT_DSF_ AT	0	0	0	0
NA_IUM	1	1.0	1.0	1.0	NI_ILF_ R3RSRM	0	0	0	0	VP_ A3SF_ R3RSRF	0	0	0	0	NT_DSM	1	0	0	0
ABBREV	0	0	0	0	NI_ILM	0	0	0	0	VP_ A3SF_ R3RSRM	4	0	0	0	NT_ DSM_AT	0	0	0	0

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
VC_2LM	0	0	0	0	NI_ISF	9	1.0	0.44	0.62	VP_ A3SM	37	0.65	0.92	0.76	NT_ISM	0	0	0	0
VC_2SF	0	0	0	0	NI_ISF_ R3RLRM	0	0	0	0	VP_ A3SM_ R1RLRM	0	0	0	0	NT_ISM_ R3RSRM	0	0	0	0
VC_2SM	2	0	0	0	NI_ISF_ R3RSRF	0	0	0	0	VP_ A3SM_ R1RSRM	0	0	0	0	PUNC	217	1.0	1.0	1.0
VC_2SM_ R1RLRM	0	0	0	0	NI_ISF_ R3RSRM	1	0	0	0	VP_ A3SM_ R3RLRM	1	0	0	0	IV	0	0	0	0
VC_2SM_ R1RSRM	1	0	0	0	NI_ISF_ R3RURF	0	0	0	0	VP_ A3SM_ R3RSRF	1	0	0	0	NV_DLF	4	1.0	0.50	0.67
VC_2SM_ R3RLRM	0	0	0	0	NI_ISM	80	0.82	0.84	0.83	VP_ A3SM_ R3RSRM	4	1	0.75	0.86	NV_DLF_ AT	1	0	0	0
AC_DLF	0	0	0	0	NI_ ISM_13	0	0	0	0	VP_ A3SM_ R3RURM	0	0	0	0	NV_DLM	6	0.50	0.33	0.40
NC_DLF	21	0.42	0.48	0.37	NI_ ISM_23	0	0	0	0	VP_ A3UM	0	0	0	0	NV_ DLM_AT	1	0	0	0
NC_ DLF_12	0	0	0	0	NI_ ISM_33	0	0	0	0	VP_P3LF	0	0	0	0	NV_ DLM_ R3RSRF	0	0	0	0
NC_ DLF_22	0	0	0	0	NI_ISM_ AT	1	0	0	0	VP_ P3LF_ R3RSRM	0	0	0	0	NV_DSF	5	1.0	0.40	0.57
AC_ DLF_AT	0	0	0	0	NI_ISM_ R1RLM	0	0	0	0	VP_ P3LM_ R2RSRM	0	0	0	0	NV_DSF_ AT	3	1.0	0.67	0.80
NC_ DLF_AT	0	0	0	0	NI_ISM_ R1RLRM	0	0	0	0	VP_P3SF	2	0	0	0	NV_DSM	7	0.75	0.43	0.55
NC_ DLF_ R3RLRM	0	0	0	0	NI_ISM_ R1RSRM	3	0	0	0	VP_P3SM	7	0.25	0.14	0.18	NV_ DSM_AT	1	0	0	0
AC_DLM	2	0	0	0	NI_ISM_ R2RSRF	0	0	0	0	VP_ P3SM_ R3RSRM	0	0	0	0	NV_DUM	0	0	0	0
NC_DLM	5	0	0	0	NI_ISM_ R3RLRM	1	1.0	1.0	1.0	RP_ R1RLRM	2	0.5	0.50	0.50	NV_ILF	1	0.33	1.0	0.50
NC_ DLM_AT	1	0	0	0	NI_ISM_ R3RSRF	5	0.80	0.80	0.80	RP_ R1RSRM	8	1	0.25	0.40	NV_ILF_ R3RSRM	0	0	0	0
AC_DSF	0	0	0	0	NI_ISM_ R3RSRM	4	0.67	0.50	0.57	RP_ R2RLRM	2	0	0	0	NV_ILM	0	0	0	0

الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف	الوسم	عدد الأمثلة	الدقة	الاسترجاع	مقياس ف
NC_DSF	37	0.27	0.22	0.36	NI_ISM_R3RURM	2	1.0	0.50	0.67	RP_R2RSRM	1	0	0	0	NV_ILM_AT	0	0	0	0
AC_DSF_AT	0	0	0	0	NI_IUF	0	0	0	0	RP_R3RLRM	2	0	0	0	NV_ILM_R2RSRM	0	0	0	0
NC_DSF_AT	11	0.56	0.46	0.71	NI_IUM	0	0	0	0	RP_R3RSRF	11	0.58	1.0	0.73	NV_ISF	0	0	0	0
AC_DSM	8	0.50	0.38	0.75	DL	3	1.0	1.0	1.0	RP_R3RSRM	41	0.91	0.95	0.93	NV_ISF_AT	0	0	0	0
NC_DSM	71	0.31	0.28	0.35	NL_DLF	2	0	0	0	RP_R3RURM	0	0	0	0	NV_ISF_R3RSRM	0	0	0	0
NC_DSM_AT	18	0.52	0.44	0.62	NL_DSF	2	0	0	0	PR	1	0	0	0	NV_ISM	7	1.0	0.29	0.44
AC_DSM_R3RLRF	0	0	0	0	NL_DSF_AT	1	0	0	0	PR_LF	2	0.33	0.50	0.40	NV_ISM_AT	0	0	0	0
NC_DUF	5	0.33	0.20	1.0	NL_DSM	5	1.0	0.20	0.33	PR_LM	0	0	0	0	NV_ISM_R1RSRM	0	0	0	0
NC_DUM	1	0	0	0	NL_DSM_AT	0	0	0	0	PR_SF	5	0.57	0.80	0.67	SYMB	7	1.0	0.29	0.44
AC_ILF	2	0	0	0	NL_ILF	4	1.0	0.50	0.67	PR_SM	30	0.96	0.90	0.93					

وهكذا، دربت خوارزمية CRF على ٨٠٪ من مدونة النظام الموسومة يدويا، ثم اختبرتها على الـ٢٠٪ المتبقية من المدونة، وقست بالمقاييس الأربعة المعروفة: الصحة، والدقة، والاسترجاع، ومقياس ـ ف أداء نظام التوسيم بمجموعات الوسوم الثلاثة. وجاءت النتائج بفوارق متشابهة قدرها ١٠٪ إذ لم تتشتت بالقيم العشوائية.

٤-٣ تحسين الفجوة بين التوسيم الآلي والتوسيم اليدوي:

يُشير جونسون إلى أنه من الضروري أن يكون هناك تسارع في التحليل التقني، وتسارع في التحليل اللغوي؛ من أجل تجاوز القواعد **المقيَّدة** constraint سواء أكانت اللغوية أم التقنية، والدخول في التطبيقات التجريبية البحتة في التوسيم(١).

(١) = How Linguistic Issues in Language Technology LiLT, 2011, p. 4

وفي مثل هذه المحاولات يرغب اللغوي دائما في أن تكون بيانات التوسيم جيدة وأن يكون الاسترجاع لجميع المعلومات النحوية المهمة، فيما يرجو الحاسوبي أن تفيد البيانات في عمليات فك الغموض وزيادة دقة التوسيم. وحتى تتضح لنا بشكل كامل قدرة مدونة النظام والنظام الآلي الذي بني على أساسها على التقطيع والتوسيم النحوي الآلي، فقد جربت عليه عينة من نصوص متنوعة غير منحازة من الأدب والقرآن والصحف تتكون من٥٠٠ كلمة فعلية تقريبا لم يرها النظام من قبل، فظهرت نتائج التقطيع والتوسيم بمجموعات الوسوم بمستوياتها الثلاثة كما في سلسلة الجداول الآتية (٤-١٥). وقد راجعت التقطيع يدويا لوجود أخطاء فيه عائدة لقلة البيانات التي درب عليها نظام التقطيع رغم دقته العالية في نصوص مدونة الاختبار. وأدخلت النصوص في نظام التوسيم النحوي الآلي بعد مراجعة تقطيعها يدويا وظهرت نتائج توسيمها كما توضحه الجداول وبالأخطاء المحددة بخط أسفلها.

ويلاحظ في الجدول (٤-١٥-أ) أن أغلب الأخطاء في مجموعات الوسوم هي في التوسيم بالأسماء وذلك لأن كثرة الأمثلة غذت الخوارزمية باحتمالات انتقالية جعلتها تسم حتى ما ليس باسم على أنه اسم. فالخطأ مثلا في كلمة (الحبيبة) بسبب الاحتمالية الأكبر لورود الاسم بعد الصفة والاحتمالية الأكبر لمجيء الاسم قبل الضمائر من ورود الصفة بعد الصفة (انظر جدول ٤-١٦-أ).

= والقواعد المقيدة هي شرط يوضع على القاعدة العامة التي تنطبق على الكلمات أو التراكيب لتحديد السياق الذي يمكن أن تنطبق فيه القاعدة وذلك لمنع إنتاج الجمل أو التراكيب غير المقبولة أو غير الصحيحة. فالقاعدة تتناول أحكام الأبواب الكبيرة، كالاسم والصفة والفعل، أما القيد فيتناول أحكام الأنواع الفرعية، كالاسم المؤنث والاسم المذكر والصفة المؤنثة والصفة المذكرة والفعل والفاعل المؤنث والمذكر والمفرد والجمع والمثنى، وغير ذلك. ومن أمثلة القيود النحوية: شرط وصف الاسم المؤنث بصفة مؤنثة في العربية.

انظر: قاموس مصطلحات المعلوماتية واللغويات الحسابية، ٢٠٠٣، ص. ٧١

جدول (٤-١٥-أ) تجريب النظام على عينة من نصوص متنوعة غير منحازة (نص ١)

نص (صحفي) مقطع آليا (١)
الرياض عاصمتنا الحبيبة هي عاصمة القرار العربي ب امتياز، ف منها تصدر الخطوط العريضة ال رسم السياسات التي تكفل تحقيق المصالح العربية على مختلف ملفاتها، ف ي هذا الأسبوع تشهد الرياض ـ ك ما هي عادتها ـ أحداثاً ب الغة الأهمية ب انعقاد قمة مجلس التعاون والاحتفال ب الذكرى الرابعة ال تول ي خادم الحرمين الشريفين المل ك س المان ب ن عبدالعزيز ـ حفظه الله ـ مقاليد الحكم الذي شهدت البلاد ف ي عهده ن قلات نوعية غير مسبوقة انعكست إيجاباً على ك ال مخرجات التنمية الاقتصادية ال الاجتماعية من خلال رؤية ٢٠٣٠.
النص مقطعا بعد المراجعة اليدوية **عدد الكلمات الفعلية: ٨٦**
الرياض عاصمتنا الحبيبة هي عاصمة القرار العربي ب امتياز، ف منها تصدر الخطوط العريضة ل رسم السياسات التي تكفل تحقيق المصالح العربية على مختلف ملفاتها، في هذا الأسبوع تشهد الرياض ـ ك ما هي عادتها ـ أحداثاً بالغة الأهمية ب انعقاد قمة مجلس التعاون والاحتفال ب الذكرى الرابعة ل تولي خادم الحرمين الشريفين الملك سلمان بن عبدالعزيز ـ حفظه الله ـ مقاليد الحكم الذي شهدت البلاد في عهده نقلات نوعية غير مسبوقة انعكست إيجاباً على كل مخرجات التنمية الاقتصادية والاجتماعية من خلال رؤية ٢٠٣٠.
توسيمه بالوسوم الرئيسة **عدد الأخطاء: ٨/ نسبة الخطأ: ٠,٠٦٩**
[('الرياض', 'N'), ('عاصمتنا', 'A'), ('الحبيبة', 'N'), ('هي', 'P'), ('عاصمة', 'A'), ('القرار', 'N'), ('العربي', 'N'), ('ب', 'RP'), ('امتياز', 'PUNC'), ('،', 'N'), ('ف', 'RP'), ('منها', 'RP'), ('تصدر', 'N'), ('الخطوط', 'N'), ('العريضة', 'N'), ('ل', 'RP'), ('رسم', 'N'), ('السياسات', 'N'), ('التي', 'P'), ('تكفل', 'V'), ('تحقيق', 'N'), ('المصالح', 'N'), ('العربية', 'N'), ('على', 'RP'), ('مختلف', 'A'), ('ملفاتها', 'N'), ('في', 'RP'), ('هذا', 'P'), ('الأسبوع', 'N'), ('تشهد', 'N'), ('الرياض', 'PUNC'), ('-', 'N'), ('ك', 'RP'), ('ما', 'RP'), ('هي', 'P'), ('عادتها', 'N'), ('-', 'PUNC'), ('أحداثاً', 'N'), ('بالغة', 'N'), ('الأهمية', 'N'), ('ب', 'RP'), ('انعقاد', 'N'), ('قمة', 'N'), ('مجلس', 'N'), ('التعاون', 'N'), ('و', 'RP'), ('الاحتفال', 'N'), ('ب', 'RP'), ('الذكرى', 'N'), ('الرابعة', 'A'), ('ل', 'RP'), ('تولي', 'V'), ('خادم', 'A'), ('الحرمين', 'N'), ('الشريفين', 'A'), ('الملك', 'A'), ('سلمان', 'N'), ('بن', 'N'), ('عبدالعزيز', 'PUNC'), ('-', 'N'), ('حفظه', 'V'), ('الله', 'PUNC'), ('-', 'N'), ('مقاليد', 'A'), ('الحكم', 'N'), ('الذي', 'P'), ('شهدت', 'V'),

('البلاد', 'N'), ('في', 'RP'), ('عهده', 'N'), ('نقلات', 'N'), ('نوعية', 'N'), ('غير', 'N'), ('مسبوقة', 'A'), ('انعكست', 'V'), ('إيجاباً', 'N'), ('على', 'RP'), ('كل', 'N'), ('مخرجات', 'A'), ('التنمية', 'N'), ('الاقتصادية', 'N'), ('و', 'RP'), ('الاجتماعية', 'N'), ('من', 'RP'), ('خلال', 'N'), ('رؤية', 'PUNC', '.'), ('2030', 'DIGIT'), ('N')]
تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('N', 'الحبيبة')	('A', 'الحبيبة')
('N', 'تصدر')	('V', 'تصدر')
('N', 'العريضة')	('A', 'العريضة')
('N', 'تشهد')	('V', 'تشهد')
('N', 'بالغة')	('A', 'بالغة')
('V', 'تولي')	('N', 'تولي')

توسيمه بالوسوم الفرعية **عدد الأخطاء: ١٤ / نسبة الخطأ: ٠,١٢٠**

[('الرياض', 'NA'), ('عاصمتنا', 'AS'), ('الحبيبة', 'NA'), ('هي', 'PP'), ('عاصمة', 'AS'), ('القرار', 'NA'), ('العربي', 'NV'), ('ب', 'RP'), ('امتياز', 'PUNC', '،'), ('NA'), ('ف', 'RP'), ('منها', 'RP'), ('تصدر', 'NA'), ('الخطوط', 'NA'), ('العريضة', 'NV'), ('ل', 'RP'), ('رسم', 'NA'), ('السياسات', 'NA'), ('التي', 'PR'), ('تكفل', 'VS'), ('تحقيق', 'NA'), ('المصالح', 'NC'), ('العربية', 'NV'), ('على', 'RP'), ('مختلف', 'AO'), ('ملفاتها', 'NA'), ('في', 'RP'), ('هذا', 'PD'), ('الأسبوع', 'NI'), ('تشهد', 'NA'), ('الرياض', 'PUNC', '–'), ('NA'), ('ك', 'RP'), ('ما', 'RP'), ('هي', 'PP'), ('عادتها', 'PUNC', '–'), ('PUNC'), ('–', 'NA'), ('أحداثاً', 'NC'), ('بالغة', 'NC'), ('الأهمية', 'NA'), ('ب', 'RP'), ('انعقاد', 'NA'), ('قمة', 'NA'), ('مجلس', 'NL'), ('التعاون', 'NA'), ('و', 'RP'), ('الاحتفال', 'NA'), ('ب', 'RP'), ('الذكرى', 'NA'), ('الرابعة', 'AS'), ('ل', 'RP'), ('تولي', 'VP'), ('خادم', 'AS'), ('الحرمين', 'NC'), ('الشريفين', 'AA'), ('الملك', 'AA'), ('سلمان', 'NC'), ('بن', 'NC'), ('عبدالعزيز', 'PUNC', '–'), ('NC'), ('حفظه', 'VS'), ('الله', 'PUNC', '–'), ('NC'), ('مقاليد', 'AO'), ('الحكم', 'NA'), ('الذي', 'PR'), ('شهدت', 'VS'),

('البلاد', 'NC'), ('في', 'RP'), ('عهده', 'NA'), ('نقلات', 'NA'), ('نوعية', 'NA'), ('غير', 'NI'), ('مسبوقة', 'AS'), ('انعكست', 'VS'), ('إيجاباً', 'NA'), ('على', 'RP'), ('كل', 'NI'), ('مخرجات', 'AS'), ('التنمية', 'NA'), ('الاقتصادية', 'NA'), ('و', 'RP'), ('الاجتماعية', 'NA'), ('من', 'RP'), ('خلال', 'NI'), ('رؤية', 'PUNC', '.', 'DIGIT', '2030', 'NA')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('NA', 'الرياض')	('NC', 'الرياض')
('NA', 'الحبيبة')	('AA', 'الحبيبة')
('NA', 'تصدر')	('VP', 'تصدر')
('NV', 'العريضة')	('AA', 'العريضة')
('VS', 'تكفل')	('VP', 'تكفل')
('NC', 'المصالح')	('NA', 'المصالح')
('NA', '،ملفاتها')	('NC', '،ملفاتها')
('NA', 'تشهد')	('VP', 'تشهد')
('NA', 'الرياض')	('NC', 'الرياض')
('NC', 'أحداثاً')	('NA', 'أحداثاً')
('NC', 'بالغة')	('AS', 'بالغة')
('VP', 'تولي')	('NA', 'تولي')
('AO', 'مقاليد')	('NM', 'مقاليد')
('AS', 'مسبوقة')	('AO', 'مسبوقة')
('AS', 'مخرجات')	('AO', 'مخرجات')

توسيمه بالوسوم الموسعة بالخصائص التصريفية **عدد الأخطاء: ٢٤ / نسبة الخطأ: ٠,٢١٥**

[('الرياض', 'NA_DSM'), ('عاصمتنا', 'RP_R3RSRF'), ('الحبيبة', 'NA_DSF'), ('هي', 'PP_3SF'), ('عاصمة', 'AS_ISF'), ('القرار', 'NA_DSM'), ('العربي', 'NV_DSM_AT'), ('ب', 'RP'), ('امتياز', 'PUNC', '،', 'NA_ISM'), ('ف', 'RP'), ('منها', 'RP_R3RSRF'), ('تصدر', 'VP_A3SF'), ('الخطوط', 'NC_DSF'), ('العريضة', 'AA_DSF'), ('ل', 'RP'), ('رسم', 'NA_ISM'), ('السياسات', 'NA_DLF'), ('التي', 'PR_SF'), ('تكفل', 'VS_A3SM'), ('تحقيق', 'NA_ISM'), ('المصالح', 'NC_DSM'), ('العربية', 'NV_DSF_AT'), ('على', 'RP'), ('مختلف', 'AO_ISM'), ('ملفاتها', 'AO_ISM'), ('في', 'RP'), ('هذا', 'PD_SM'), ('الأسبوع', 'NC_DSM'), ('تشهد', 'NA_ISM'), ('الرياض', 'PUNC', '-', 'NA_DSM'), ('ك', 'RP'), ('ما', 'RP'), ('هي', 'PP_3LF'), ('عادتها', 'PUNC', '-', 'NA_ILF_R3RSRF'), ('أحداثاً', 'NC_ISM_22'), ('بالغة', 'NA_ISF'), ('الأهمية', 'NA_DSF'), ('ب', 'RP'), ('انعقاد', 'NA_ISM'), ('قمة', 'NA_ISF'), ('مجلس', 'NL_ISM'), ('التعاون', 'NA_DSM'), ('و', 'RP'), ('الاحتفال', 'NA_DSM'), ('ب', 'RP'), ('الذكرى', 'NA_DSF'), ('الرابعة', 'NA_DSF'), ('ل', 'RP'), ('تولي', 'VP_A3SF'), ('خادم', 'NA_ISM'), ('الحرمين', 'NC_DUM'), ('الشريفين', 'AA_DUM'), ('الملك', 'AS_DSM'), ('سلمان', 'NC_ISM'), ('بن', 'NC_ISM'), ('عبدالعزيز', 'PUNC', '-', 'NC_ISM'), ('حفظه', 'VS_A3SM'), ('الله', 'PUNC', '-', 'NC_ISM'), ('مقاليد', 'NC_ISM'), ('الحكم', 'NC_DSM'), ('الذي', 'PR_SM'), ('شهدت', 'VS_A3SF'), ('البلاد', 'NC_DLF'), ('في', 'RP'), ('عهده', 'NA_ISM_R3RSRM'), ('نقلات', 'VS_A3SF'), ('نوعية', 'NA_ISF'), ('غير', 'NI_ISM'), ('مسبوقة', 'AS_ISF'), ('انعكست', 'VS_A3SF'), ('إيجاباً', 'NC_ISF'), ('على', 'RP'), ('كل', 'NI_ISM'), ('مخرجات', 'NA_ILF'), ('التنمية', 'NA_DSF'), ('الاقتصادية', 'NA_DSF_AT'), ('و', 'RP'), ('الاجتماعية', 'NA_DLF_AT'), ('من', 'RP'), ('خلال', 'NI_ISM'), ('رؤية', 'NA_ISF'), ('2030', 'DIGIT'), ('.', 'PUNC')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('الرياض', 'NA_DSM')	('الرياض', 'NC_DSF')
('عاصمتنا', 'RP_R3RSRF')	('عاصمتنا', 'AS_ISF_R1RLRM')
('الحبيبة', 'NA_DSF')	('الحبيبة', 'AA_DSF')

('الخطوط', 'NC_DSF')	('الخطوط', 'NA_DLF')
('التي', 'PR_SF')	('التي', 'PR_LF')
('تكفل', 'VS_A3SM')	('تكفل', 'VP_A3LF')
('المصالح', 'NC_DSM')	('المصالح', 'NA_DLF')
('ملفاتها', 'AO_ISM')	('ملفاتها', 'NC_ILF_R3RLRF')
('تشهد', 'NA_ISM')	('تشهد', 'VP_A3SF')
('الرياض', 'NA_DSM')	('الرياض', 'NC_DSF')
('هي', 'PP_3LF')	('هي', 'PP_3SF')
('أحداثاً', 'NC_ISM_22')	('أحداثاً', 'NA_ILF')
('بالغة', 'NA_ISF')	('بالغة', 'AS_ISF')
('الرابعة', 'NA_DSF')	('الرابعة', 'AS_DSF')
('تولي', 'VP_A3SF')	('تولي', 'NA_ISM')
('خادم', 'NA_ISM')	('خادم', 'AS_ISM')
('الملك', 'AS_DSM')	('الملك', 'AA_DSM')
('مقاليد', 'NC_ISM')	('مقاليد', 'NM_ILF')
('شهدت', 'VS_A3SF')	('شهدت', 'VS_A3LF')
('نقلات', 'VS_A3SF')	('نقلات', 'NA_ILF')
('مسبوقة', 'AS_ISF')	('مسبوقة', 'AO_ILF')
('انعكست', 'VS_A3SF')	('انعكست', 'VS_A3LF')
('إيجاباً', 'NC_ISF')	('إيجاباً', 'NA_ISM')
('مخرجات', 'NA_ILF')	('مخرجات', 'AO_ILF')

ويظهر في الجدول (٤-١٥-ب) ما يؤكد أن الأخطاء التي يقع فيها النظام يأتي غالبها من الاحتمالات الانتقالية، حيث وسم كلمة (هجوم) خطأ في مجموعة الوسوم

الأساسية بالوسم (P ضمير)، ثم وسمها وسما صحيحا في مجموعة الوسوم الفرعية بالوسم (NA اسم معنى)، وفي مجموعة الوسوم الموسعة بالوسم (NA_ISM اسم ـ معنى_غير ـ معرف ـ بأل ـ مفرد ـ مذكر). وذلك لأن احتمالية أن يلي الضمير فعلا كما في سياق الكلمة أكثر من احتمالية مجيء أي قسم كلامي بعده وإن كان قد أخطأ في الكلمة التي تلت كلمة هجوم وكان الخطأ فيها للسبب نفسه (انظر سلسلة الجداول ٤-١٦).

جدول (٤-١٥-ب) تجريب النظام على عينة من نصوص متنوعة غير منحازة (نص ٢)

نص (صحفي) مقطع آليا (٢)
أحبطت القوات الأفغانية اليو م، محاول ة مسلحين ل تنفيذ هجو م صاروخي على العاصمة كابول. وذكر بيان صادر عن شرطة كابو ال ن قلته وكالة أنباء « خاما برس » الأفغانية، أن قوات الشرطة اكتشفت صاروخًا من طراز ب ي إ م - ١ ف ي محيط منطقة الشرطة التاسعة ف ي المدينة صباح اليو م. وأضاف البيان أن الصاروخ تم اكتشافه ف ي منطقة دشت ب ادول ا، وتم إبطال مفعول ه من قبل فريق شرطة التخلص من المتفجرات، وال م تعلق الجماعات المسلحة المناهضة ال الحكومة ومن بينها حركة طالبان ب شأن الحادث حتى الآن.
النص مقطعا بعد المراجعة اليدوية **عدد الكلمات الفعلية: ٩٤**
أحبطت القوات الأفغانية اليوم، محاولة مسلحين ل تنفيذ هجوم صاروخي على العاصمة كابول. وذكر بيان صادر عن شرطة كابول نقلته وكالة أنباء « خاما برس » الأفغانية، أن قوات الشرطة اكتشفت صاروخًا من طراز بي إم - ١ في محيط منطقة الشرطة التاسعة في المدينة صباح اليوم. وأضاف البيان أن الصاروخ تم اكتشافه في منطقة دشت بادولا، وتم إبطال مفعوله من قبل فريق شرطة التخلص من المتفجرات، ولم تعلق الجماعات المسلحة المناهضة ل الحكومة ومن بينها حركة طالبان ب شأن الحادث حتى الآن.
توسيمه بالوسوم الرئيسة **عدد الأخطاء: ١٥/ نسبة الخطأ: ١٣٢ , ٠**
[('أحبطت', 'V'), ('القوات', 'N'), ('الأفغانية', 'N'), ('اليوم', 'PUNC'), ('،', ','), ('N'), ('محاولة', 'N'), ('مسلحين', 'A'), ('ل', 'RP'), ('تنفيذ', 'V'), ('هجوم', 'P'), ('صاروخي', 'V'), ('على', 'RP'), ('العاصمة', 'A'), ('كابول', 'PUNC'), ('.', 'A'), ('و', 'RP'), ('ذكر', 'V'), ('بيان', 'N'),

('صادر', 'A'), ('عن', 'RP'), ('شرطة', 'N'), ('كابول', 'A'), ('نقلته', 'N'), ('وكالة', 'N'), ('أنباء', 'PUNC', ''''), ('N'), ('خاما', 'A'), ('برس', 'PUNC', ''''), ('N'), ('الأفغانية', 'N'), PUNC', '،'), ('أن', 'RP'), ('قوات', 'N'), ('الشرطة', 'N'), ('اكتشفت', 'V'), ('صاروخًا', 'A'), ('من', 'RP'), ('طراز', 'N'), ('بي', 'N'), ('إم', 'PUNC', '-'), ('N'), ('١', 'DIGIT'), ('في', 'RP'), ('محيط', 'N'), ('منطقة', 'N'), ('الشرطة', 'N'), ('التاسعة', 'N'), ('في', 'RP'), ('المدينة', 'N'), ('صباح', 'N'), ('اليوم', 'PUNC', '.'), ('N'), ('و', 'RP'), ('أضاف', 'V'), ('البيان', 'N'), ('أن', 'RP'), ('الصاروخ', 'N'), ('تم', 'V'), ('اكتشافه', 'N'), ('في', 'RP'), ('منطقة', 'N'), ('دشت', 'V'), ('بادولا', 'PUNC', '،'), ('N'), ('و', 'RP'), ('تم', 'V'), ('إبطال', 'N'), ('مفعوله', 'N'), ('من', 'RP'), ('قبل', 'N'), ('فريق', 'N'), ('شرطة', 'N'), ('التخلص', 'N'), ('من', 'RP'), ('المتفجرات', 'PUNC', '،'), ('A'), ('و', 'RP'), ('لم', 'RP'), ('تعلق', 'V'), ('الجماعات', 'N'), ('المسلحة', 'N'), ('المناهضة', 'A'), ('ل', 'RP'), ('الحكومة', 'N'), ('و', 'RP'), ('من', 'RP'), ('بينها', 'N'), ('حركة', 'N'), ('طالبان', 'N'), ('ب', 'RP'), ('شأن', 'N'), ('الحادث', 'A'), ('حتى', 'RP'), ('الآن', 'PUNC', '.'), ('N')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح	الوسم الخاطئ	الوسم الصحيح
('P', 'هجوم')	('A', 'هجوم')	('N', 'إم')	('ABBREV', 'إم')
('V', 'تنفيذ')	('N', 'تنفيذ')	('N', 'بي')	('ABBREV', 'بي')
('V', 'صاروخي')	('N', 'صاروخي')	('N', 'محيط')	('A', 'محيط')
('A', 'كابول')	('N', 'كابول')	('N', 'التاسعة')	('A', 'التاسعة')
('A', 'كابول')	('N', 'كابول')	('V', 'دشت')	('N'', 'دشت')
('N', 'نقلته')	('V', 'نقلته')	('N', 'مفعوله')	('A', 'مفعوله')
('A', 'خاما')	('N', 'خاما')	('N', 'المسلحة')	('A', 'المسلحة')

توسيمه بالوسوم الفرعية **عدد الأخطاء: ١٩ / نسبة الخطأ: ٠,٢٠٠**

[('أحبطت', 'VS'), ('القوات', 'NA'), ('الأفغانية', 'NC'), ('اليوم', 'PUNC', '،'), ('NI'),
('محاولة', 'NA'), ('مسلحين', 'AS'), ('ل', 'RP'), ('تنفيذ', 'VP'), ('هجوم', 'NA'),
('صاروخي', 'AS'), ('على', 'RP'), ('العاصمة', 'AS'), ('كابول', 'PUNC', '.'), ('NC'),
('و', 'RP'), ('ذكر', 'NA'), ('بيان', 'NA'), ('صادر', 'AS'), ('عن', 'RP'), ('شرطة', 'NA'),
('كابول', 'RP'), ('نقلته', 'NA'), ('وكالة', 'NA'), ('أنباء', 'PUNC', '"'), ('NA'), ('خاما',
'AA'), ('برس', 'PUNC', '"'), ('NC'), ('الأفغانية', 'PUNC', '،'), ('NC'), ('أن', 'RP'),
('قوات', 'NA'), ('الشرطة', 'NA'), ('اكتشفت', 'VS'), ('صاروخًا', 'AS'), ('من', 'RP'),
('طراز', 'NA'), ('بي', 'NC'), ('إم', 'DIGIT', '1'), ('PUNC', '-'), ('NC'), ('في', 'RP'),
('محيط', 'NL'), ('منطقة', 'NA'), ('الشرطة', 'NA'), ('التاسعة', 'NA'), ('في', 'RP'), ('المدينة',
'NC'), ('صباح', 'NA'), ('اليوم', 'PUNC', '.'), ('NI'), ('و', 'RP'), ('أضاف', 'VS'), ('البيان',
'NA'), ('أن', 'RP'), ('الصاروخ', 'AS'), ('تم', 'VS'), ('اكتشافه', 'NA'), ('في', 'RP'),
('منطقة', 'NA'), ('دشت', 'VS'), ('بادولا', 'PUNC', '،'), ('NC'), ('و', 'RP'), ('تم', 'VS'),
('إبطال', 'NA'), ('مفعوله', 'AO'), ('من', 'RP'), ('قبل', 'NI'), ('فريق', 'NV'), ('شرطة',
'NV'), ('التخلص', 'NA'), ('من', 'RP'), ('المتفجرات', 'PUNC', '،'), ('AS'), ('و', 'RP'),
('لم', 'RP'), ('تعلق', 'VP'), ('الجماعات', 'NV'), ('المسلحة', 'NA'), ('المناهضة', 'AO'),
('ل', 'RP'), ('الحكومة', 'NA'), ('و', 'RP'), ('من', 'RP'), ('بينها', 'NI'), ('حركة', 'NA'),
('طالبان', 'NI'), ('ب', 'RP'), ('شأن', 'NA'), ('الحادث', 'AS'), ('حتى', 'RP'), ('الآن', 'NI')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح	الوسم الخاطئ	الوسم الصحيح
('AS', 'مسلحين')	('AO', 'مسلحين')	('NA', 'الشرطة')	('NV', 'الشرطة')
('AS', 'صاروخي')	('NC', 'صاروخي')	('NA', 'التاسعة')	('AS', 'التاسعة')
('NA', 'نقلته')	('VS', 'نقلته')	('NA', 'صباح')	('NI', 'صباح')
('AA', 'خاما'),	('NC', 'خاما'),	('AS', 'الصاروخ')	('NC', 'الصاروخ')
('NA', 'الشرطة')	('NA', 'الشرطة')	('NA', 'منطقة')	('NC', 'منطقة')
('AS', 'صاروخًا')	('NC', 'صاروخًا')	('VS', 'دشت')	('NC', 'دشت')

('بي', 'NC')	('بي', 'ABBREV')	('المسلحة', 'NA')	('المسلحة', 'AO')
('إم', 'NC')	('إم', 'ABBREV')	('المناهضة', 'AO')	('المناهضة', 'AS')
('محيط', 'NL')	('محيط', 'AS')	('طالبان', 'NI')	('طالبان', 'NC')
('منطقة', 'NA')	('منطقة', 'NC')	('الشرطة', 'NA')	('الشرطة', 'NV')

توسيمه بالوسوم الموسعة بالخصائص التصريفية **عدد الأخطاء: ٢٨ / نسبة الخطأ: ٠,٢٦٣**

[('أحبطت', 'VS_A3SF'), ('القوات', 'NA_DLF'), ('الأفغانية', 'NC_DSF_AT'), ('اليوم', 'PUNC'), ('،', 'NI_DSM'), ('محاولة', 'NA_ISF'), ('مسلحين', 'AS_ILM'), ('ل', 'RP'), ('تنفيذ', 'VP_A3SF'), ('هجوم', 'NA_ISM'), ('صاروخي', 'AS_ISM'), ('على', 'RP'), ('العاصمة', 'AS_DSF'), ('كابول', 'PUNC'), ('.', 'RP'), ('و', 'RP'), ('ذكر', 'NA_ISM'), ('بيان', 'NA_ISM'), ('صادر', 'AS_ISM'), ('عن', 'RP'), ('شرطة', 'NA_ISF'), ('كابول', 'RP'), ('نقلته', 'NA_ISF_R3RSRM'), ('وكالة', 'NA_ISF'), ('أنباء', 'NA_ILF'), ('"', 'PUNC'), ('خاما', 'RP'), ('برس', 'NC_ISM'), ('"', 'PUNC'), ('الأفغانية', 'NA_DSF'), ('أن', 'RP'), ('قوات', 'NA_ILF'), ('الشرطة', 'NA_DSF'), ('اكتشفت', 'VS_A3LF'), ('صاروخًا', 'AS_ISM'), ('من', 'RP'), ('طراز', 'NA_ISM'), ('بي', 'NC_ISM'), ('إم', 'DIGIT'), ('1', 'PUNC'), ('-', 'NC_ ISM'), ('في', 'RP'), ('محيط', 'NA_ISM'), ('منطقة', 'NA_ISF'), ('الشرطة', 'NA_DSF'), ('التاسعة', 'NA_DSF'), ('في', 'RP'), ('المدينة', 'NC_DSM'), ('صباح', 'NA_ISM'), ('اليوم', 'PUNC'), ('.', 'NI_DSM'), ('و', 'RP'), ('أضاف', 'VS_A3SM'), ('البيان', 'NA_DSM'), ('أن', 'RP'), ('الصاروخ', 'NA_DSM'), ('تم', 'VS_A3SM'), ('اكتشافه', 'NA_ISM_R3RSRM'), ('في', 'RP'), ('منطقة', 'NC_ISF'), ('دشت', 'NC_ISM'), ('بادولا', 'PUNC'), ('،', 'NC_ISM'), ('و', 'RP'), ('تم', 'VS_A3SM'), ('إبطال', 'NA_ISM'), ('مفعوله', 'NA_ISM_R3RSRM'), ('من', 'RP'), ('قبل', 'NI_ISM'), ('فريق', 'NA_ISM'), ('شرطة', 'NA_ISF'), ('التخلص', 'NA_DSM'), ('من', 'RP'), ('المتفجرات', 'PUNC'), ('،', 'NA_DLF'), ('و', 'RP'), ('لم', 'RP'), ('تعلق', 'NA_ISM'), ('الجماعات', 'NV_DLF'), ('المسلحة', 'NA_DSF'), ('المناهضة', 'AO_DSF'), ('ل', 'RP'), ('الحكومة', 'NA_ISF'), ('و', 'RP'), ('من', 'RP'), ('بينها', 'NI_ISM_R3RSRF'), ('حركة', 'NA_ISF'), ('طالبان', 'NC_ISM'), ('ب', 'RP'), ('شأن', 'NA_ISM'), ('الحادث', 'NA_DSM'), ('حتى', 'RP'), ('الآن', 'NI_DSM')]

تصحيح أخطاء التوسيم:	
الوسم الخاطئ	الوسم الصحيح
('VS_A3SF', 'أحبطت')	('VS_A3LF', 'أحبطت')
('AS_ILM', 'مسلحين')	('AO_ILM', 'مسلحين')
('VP_A3SF', 'تنفيذ')	('NA_ISM', 'تنفيذ')
('AS_ISM', 'صاروخي')	('NC_ISM_AT', 'صاروخي')
('RP', 'كابول')	('NC_ISF', 'كابول')
('NA_ISM', 'ذكر')	('VS_A3SM', 'ذكر')
('NA_ISF', 'شرطة')	('NV_ILF', 'شرطة')
('RP', 'كابول')	('NC_ISF', 'كابول')
('NA_ISF_R3RSRM', 'نقلته')	('VS_A3LF_R3RSRM', 'نقلته')
('NA_DSF', 'الأفغانية')	('NC_DSF_AT', 'الأفغانية')
('NA_DSF', 'الشرطة')	('NV_DSF', 'الشرطة')
('NA_ISM', 'محيط')	('AS_ISM', 'محيط')
('NA_ISF', 'منطقة')	('NC_ISF', 'منطقة')
('NA_DSF', 'الشرطة')	('NV_DSF', 'الشرطة')
('NA_DSF', 'التاسعة')	('AS_DSF', 'التاسعة')
('NA_ISM', 'صباح')	('NI_ISM', 'صباح')
('NA_DSM', 'الصاروخ')	('NC_DSM', 'الصاروخ')
('NA_ISM_R3RSRM', 'مفعوله')	('AO_ISM_R3RSRM', 'مفعوله')
('NA_ISM', 'فريق')	('NV_ISM', 'فريق')
('NA_ISF', 'شرطة')	('NV_ISF', 'شرطة')
('NA_DLF', 'المتفجرات')	('AS_DLF', 'المتفجرات')

('تعلق', 'NA_ISM')	('تعلق', 'VP_A3LF')
('المسلحة', 'NA_DSF')	('المسلحة', 'AO_DSF')
('المناهضة', 'AO_DSF')	('المناهضة', 'AS_DSF')
('الحادث', 'NA_DSM')	('الحادث', 'AS_DSM')
('صاروخًا', 'AS_ISM')	('صاروخًا', 'NC_ISM')
('بي', 'NC_ISM')	('بي', 'ABBREV')
('إم', 'NC_ISM')	('إم', 'ABBREV')

وليس غريبا ألا يميز النظام الأعلام من أسماء الذوات التي انتقلت من الوصفية للعلمية، كالاسم المؤنث (دانيا) الظاهر في الجدول (٤-١٥-ج)، حيث يسمه باعتباره صفة فاعل في كل مجموعات الوسوم، وقد يفيد النظام هنا تزويده بقوائم الأعلام المذكرة والمؤنثة للاستفادة منها في تحديد وسم الكلمة الصحيح. ويظهر كذلك أن النظام لا يميز صفات الفاعل التي تأتي في صيغة جمع تكسير، على نحو: سكان، فيسمها على أنها اسم. أما صفة الفاعل (قاطني) فينظر إلى نهايتها المختومة بالياء الشبيهة بحالة الفعل الماضي مع ياء المتكلم، إلى أن يتغير الحال في مجموعة الوسوم الموسعة فيسمها مركبة مع الاسم الذي يليها بسبب الانتقالات الاحتمالية لتركيب الكلمات السابقة واللاحقة لها.

جدول (٤-١٥-ج) تجريب النظام على عينة من نصوص متنوعة غير منحازة (نص ٣)

نص (أدبي) مقطع آليا (٣)
تبدأ الرواية ف ي أغسطس ١٩٨٩ ف ي ب يروت حيث تدور طواحين الحرب الأهلية اللبنانية على دانيا الطالبة ف ي المرحلة الثانوية وأسرتها. ال أواء الحرب والقذائف الطائشة تدفع بسكان البناية إلى التجمع ف ي ملجأ واحد، وهنا تعرض لنا الروائية كيف تذوب الفروقات بين السكان، ليظهر الخيط الإنساني الرفيع الذي يشبكهم. يتقاسمو ن الاحتياجات، ويتشاطرون الخوف والقلق. ف ي خضم ذلك، تتوثق علاقة دانيا ب جارهم رياض الصحافي صاحب المبادئ الذي يلحظ موهبة دانيا ف ي الكتابة ويتبناها ثقافيا. يتفرّق معظم قاطني الملجأ، إما ب الموت، أو ب الرحيل إلى أماكن آمنة. يحصل والد دانيا على عرض من الشركة التي كان يعمل فيها ل ينضم ل فرعها ف ي الكويت.

النص مقطعا بعد المراجعة اليدوية	عدد الكلمات الفعلية:١٢٢

تبدأ الرواية في أغسطس ١٩٨٩ في بيروت حيث تدور طواحين الحرب الأهلية اللبنانية على دانيا الطالبة في المرحلة الثانوية وأسرتها. لأواء الحرب والقذائف الطائشة تدفع ب سكان البناية إلى التجمع في ملجأ واحد، وهنا تعرض لنا الروائية كيف تذوب الفروقات بين السكان، ل يظهر الخيط الإنساني الرفيع الذي يشبكهم. يتقاسمون الاحتياجات، ويتشاطرون الخوف والقلق. في خضم ذلك، تتوثق علاقة دانيا ب جارهم رياض الصحافي صاحب المبادئ الذي يلحظ موهبة دانيا في الكتابة ويتبناها ثقافيا. يتفرّق معظم قاطني الملجأ، إما ب الموت، أو ب الرحيل إلى أماكن آمنة. يحصل والد دانيا على عرض من الشركة التي كان يعمل فيها ل ينضم ل فرعها في الكويت.

توسيمه بالوسوم الرئيسة:	عدد الأخطاء:٩ / نسبة الخطأ: ١١ , ٠

['تبدأ', 'V'), ('الرواية', 'N'), ('في', 'RP'), ('أغسطس', 'N'), ('١٩٨٩', 'DIGIT'), ('في', 'RP'), ('بيروت', 'N'), ('حيث', 'D'), ('تدور', 'V'), ('طواحين', 'N'), ('الحرب', 'N'), ('الأهلية', 'N'), ('اللبنانية', 'N'), ('على', 'RP'), ('دانيا', 'N'), ('الطالبة', 'N'), ('في', 'RP'), ('المرحلة', 'N'), ('الثانوية', 'A'), ('و', 'RP'), ('أسرتها', 'PUNC', '.'), ('N'), ('لأواء', 'N'), ('الحرب', 'N'), ('و', 'RP'), ('القذائف', 'N'), ('الطائشة', 'N'), ('تدفع', 'V'), ('ب', 'RP'), ('سكان', 'N'), ('البناية', 'N'), ('إلى', 'RP'), ('التجمع', 'N'), ('في', 'RP'), ('ملجأ', 'N'), ('واحد', 'PUNC', '،'), ('N'), ('و', 'RP'), ('هنا', 'P'), ('تعرض', 'V'), ('لنا', 'RP'), ('الروائية', 'N'), ('كيف', 'RP'), ('تذوب', 'V'), ('الفروقات', 'N'), ('بين', 'N'), ('السكان', 'PUNC', '،'), ('N'), ('ل', 'RP'), ('يظهر', 'V'), ('الخيط', 'N'), ('الإنساني', 'N'), ('الرفيع', 'N'), ('الذي', 'P'), ('يشبكهم', '.'), ('V'), ('يتقاسمون', 'V'), ('الاحتياجات', 'PUNC', '،'), ('N'), ('و', 'RP'), ('يتشاطرون', 'V'), ('الخوف', 'N'), ('و', 'RP'), ('القلق', 'PUNC', '.'), ('N'), ('في', 'RP'), ('خضم', 'N'), ('ذلك', 'PUNC', '،'), ('P'), ('تتوثق', 'V'), ('علاقة', 'N'), ('دانيا', 'A'), ('ب', 'RP'), ('جارهم', 'N'), ('رياض', 'N'), ('الصحافي', 'N'), ('صاحب', 'A'), ('المبادئ', 'A'), ('الذي', 'P'), ('يلحظ', 'V'), ('موهبة', 'N'), ('دانيا', 'A'), ('في', 'RP'), ('الكتابة', 'N'), ('و', 'RP'), ('يتبناها', 'V'), ('ثقافيا', 'PUNC', '.'), ('N'), ('يتفرّق', 'V'), ('معظم', 'N'), ('قاطني', 'V'), ('الملجأ', 'PUNC', '،'), ('N'), ('إما', 'RP'), ('ب', 'RP'), ('الموت', 'PUNC', '،'), ('N'), ('أو', 'RP'), ('ب', 'RP'), ('الرحيل', 'N'), ('إلى', 'RP'), ('أماكن', 'N'), ('آمنة', '.'), ('N'), ('يحصل', 'V'), ('والد', 'N'), ('دانيا', 'A'), ('على', 'RP'), ('عرض', 'N'), ('من', 'RP'), ('الشركة', 'N'), ('التي', 'P'), ('كان', 'RP'), ('يعمل', 'V'), ('فيها', 'RP'), ('ل', 'RP'), ('ينضم', 'V'), ('ل', 'RP'), ('فرعها', 'N'), ('في', 'RP'), ('الكويت', 'PUNC', '.'), ('N')]

تصحيح أخطاء التوسيم:	
الوسم الخاطئ	**الوسم الصحيح**
('N', 'الطالبة')	('A', 'الطالبة')
('N', 'الطائشة'),	('A', 'الطائشة'),
('N', 'سكان'),	('A', 'سكان'),
('N', 'السكان')	('A', 'السكان')
('N', 'الرفيع')	('A', 'الرفيع')
('A', 'دانيا')	('N', 'دانيا')
('A', 'المبادئ')	('N', 'المبادئ')
('V', 'قاطني')	('A', 'قاطني')
('N', 'آمنة')	('A', 'آمنة')
('A', 'دانيا')	('N', 'دانيا')
توسيمه بالوسوم الفرعية	**عدد الأخطاء: ٢٥/ نسبة الخطأ: ٢٨٠ , ٠**

['تبدأ', 'VP'), ('الرواية', 'NA'), ('في', 'RP'), ('أغسطس', 'NC'), ('١٩٨٩', 'DIGIT'), ('في', 'RP'), ('بيروت', 'NC'), ('حيث', 'DL'), ('تدور', 'VP'), ('طواحين', 'NI'), ('الحرب', 'NA'), ('الأهلية', 'NA'), ('اللبنانية', 'NA'), ('على', 'RP'), ('دانيا', 'NC'), ('الطالبة', 'NC'), ('في', 'RP'), ('المرحلة', 'NA'), ('الثانوية', 'AS'), ('و', 'RP'), ('أسرتها', 'PUNC', '.'), ('لأواء', 'NA'), ('الحرب', 'NA'), ('و', 'RP'), ('القذائف', 'NA'), ('الطائشة', 'NA'), ('تدفع', 'VP'), ('ب', 'RP'), ('سكان', 'NC'), ('البناية', 'NA'), ('إلى', 'RP'), ('التجمع', 'NA'), ('في', 'RP'), ('ملجأ', 'NA'), ('واحد', 'PUNC', '،'), ('NI'), ('و', 'RP'), ('هنا', 'PD'), ('تعرض', 'NA'), ('لنا', 'RP'), ('الروائية', 'NA'), ('كيف', 'RP'), ('تذوب', 'VP'), ('الفروقات', 'NA'), ('بين', 'NI'), ('السكان', 'PUNC', '،'), ('NC'), ('ل', 'RP'), ('يظهر', 'VP'), ('الخيط', 'NC'), ('الإنساني', 'NC'), ('الرفيع', 'NA'), ('الذي', 'PR'), ('يشبكهم', 'PUNC', '.'), ('VP'), ('يتقاسمون', 'VP'), ('الاحتياجات', 'PUNC', '،'), ('NA'), ('و', 'RP'), ('يتشاطرون', 'VP'), ('الخوف', 'NA'),

('و', 'RP'), ('القلق', 'PUNC', '.'), ('NA'), ('في', 'RP'), ('خضم', 'NA'), ('ذلك', 'PD'), ('،', 'PUNC'), ('تتوثق', 'VP'), ('علاقة', 'NA'), ('دانيا', 'AS'), ('ب', 'RP'), ('جارهم', 'NC'), ('رياض', 'NC'), ('الصحافي', 'NC'), ('صاحب', 'AS'), ('المبادئ', 'AO'), ('الذي', 'PR'), ('يلحظ', 'VP'), ('موهبة', 'AS'), ('دانيا', 'NC'), ('في', 'RP'), ('الكتابة', 'NA'), ('و', 'RP'), ('يتبناها', 'VP'), ('ثقافيا', 'PUNC', '.'), ('NC'), ('يتفرّق', 'VP'), ('معظم', 'NC'), ('قاطني', 'VS'), ('الملجأ', 'PUNC', '،'), ('AO'), ('إما', 'RP'), ('ب', 'RP'), ('الموت', 'PUNC', '،'), ('NC'), ('أو', 'RP'), ('ب', 'RP'), ('الرحيل', 'NC'), ('إلى', 'RP'), ('أماكن', 'NL'), ('آمنة', '.'), ('NA'), ('يحصل', 'VP'), ('والد', 'NC'), ('دانيا', 'NC'), ('على', 'RP'), ('عرض', 'NA'), ('من', 'RP'), ('الشركة', 'NA'), ('التي', 'PR'), ('كان', 'RP'), ('يعمل', 'VP'), ('فيها', 'RP'), ('ل', 'RP'), ('ينضم', 'VP'), ('ل', 'RP'), ('فرعها', 'NA'), ('في', 'RP'), ('الكويت', 'PUNC', '.'), ('NC')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح	الوسم الخاطئ	الوسم الصحيح
('NI', 'طواحين')	('NM', 'طواحين')	('AS', 'دانيا')	('NC', 'دانيا')
('NA', 'اللبنانية')	('NC', 'اللبنانية')	('NC', 'الصحافي')	('NA', 'الصحافي')
('NC', 'الطالبة')	('AS', 'الطالبة')	('AO', 'المبادئ')	('NA', 'المبادئ')
('NA', 'أسرتها')	('NV', 'أسرتها')	('AS', 'موهبة')	('NA', 'موهبة')
('NA', 'القذائف')	('NC', 'القذائف')	('NC', 'ثقافيا')	('NA', 'ثقافيا')
('NA', 'الطائشة')	('AS', 'الطائشة')	('NC', 'معظم')	('AO', 'معظم')
('NC', 'سكان')	('AS', 'سكان')	('VS', 'قاطني')	('AS', 'قاطني')
('NA', 'البناية')	('NA', 'البناية')	('AO', 'الملجأ')	('NL', 'الملجأ')
('NA', 'ملجأ')	('NL', 'ملجأ')	('NC', 'الموت')	('NA', 'الموت')
('NA', 'تعرض')	('VP', 'تعرض')	('NC', 'الرحيل')	('NA', 'الرحيل')
('NC', 'السكان')	('AS', 'السكان')	('NA', 'آمنة')	('AS', 'آمنة')
('NA', 'الأهلية')	('NV', 'الأهلية')	('NA', 'الرفيع')	('AA', 'الرفيع')

[('تبدأ', 'VP_A3SM'), ('الرواية', 'NA_DSF'), ('في', 'RP'), ('أغسطس', ',', '1989', 'NC_ISM'), ('DIGIT', 'في', 'RP'), ('بيروت', 'VS_A3SF'), ('حيث', 'DL'), ('تدور', 'VP_A3SF'), ('طواحين', 'NI_ISM'), ('الحرب', 'NA_DSF'), ('الأهلية', 'NA_DSF_AT'), ('اللبنانية', 'NA_DSF_AT'), ('على', 'RP'), ('دانيا', 'AS_ISM'), ('الطالبة', 'NA_DSF'), ('في', 'RP'), ('المرحلة', 'NA_DSF'), ('الثانوية', 'NC_DSF_AT'), ('و', 'RP'), ('أسرتها', 'NA_ILF_'), ('.', 'PUNC', 'R3RSRF'), ('لأواء', 'NA_ISM'), ('الحرب', 'NA_DSM'), ('و', 'RP'), ('القذائف', 'NC_DLF'), ('الطائشة', 'NA_DLF'), ('تدفع', 'VP_A3LF'), ('ب', 'RP'), ('سكان', 'NA_ISM'), ('البناية', 'NA_DSF'), ('إلى', 'RP'), ('التجمع', 'NA_DSM'), ('في', 'RP'), ('ملجأ', 'NI_ISM'), ('واحد', 'PUNC', '،', 'NI_ISM'), ('و', 'RP'), ('هنا', 'PD'), ('تعرض', 'NA_ISM'), ('لنا', 'RP_R1RLRM'), ('الروائية', 'NA_DSF_AT'), ('كيف', 'RP'), ('تذوب', 'NA_ISM'), ('الفروقات', 'NA_DLF'), ('بين', 'NI_ISM'), ('السكّان', ',', 'NC_DSM'), ('PUNC', '،', ')'), ('ل', 'RP'), ('يظهر', 'VP_A3SM'), ('الخيط', 'NC_DSM'), ('الإنساني', 'NC_DSM_AT'), ('الرفيع', 'NA_DSM'), ('الذي', 'PR_SM'), ('يشبكهم', ',', 'VP_A3SM'), ('PUNC', '.', ')'), ('يتقاسمون', 'VP_A3LM'), ('الاحتياجات', 'PUNC', '،', 'NA_DLF'), ('و', 'RP'), ('يتشاطرون', 'VP_A3LM'), ('الخوف', 'NA_DSM'), ('و', 'RP'), ('القلق', 'PUNC', '.', 'NA_DSM'), ('في', 'RP'), ('خضم', 'NI_ISM'), ('ذلك', ',', '،', 'PD_SM'), ('PUNC'), ('تتوثق', 'VP_A3SF'), ('علاقة', 'NC_DSF'), ('دانيا', 'AS_ISM'), ('ب', 'RP'), ('جارهم', 'NC_ISF'), ('رياض', 'NC_ISM'), ('الصحافي', 'NC_DSM_AT'), ('صاحب', 'AS_ISM'), ('المبادئ', 'AS_DSM'), ('الذي', 'PR_SM'), ('يلحظ', 'VP_A3SM'), ('موهبة', 'NA_ISF'), ('دانيا', 'NC_ISF'), ('في', 'RP'), ('الكتابة', 'NA_DSF'), ('و', 'RP'), ('يتبناها', 'VP_A3SM'), ('ثقافيا', 'PUNC', '.', 'NC_ISF'), ('يتفرّق', 'VP_A3SM'), ('معظم', 'NC_ISM'), ('قاطني', 'NC_ISM_12'), ('الملجأ', 'PUNC', '،', 'NC_ISM_22'), ('إما', 'RP'), ('ب', 'RP'), ('الموت', 'PUNC', '،', 'NC_DSM'), ('أو', 'RP'), ('ب', 'RP'), ('الرحيل', 'NC_DSM'), ('إلى', 'RP'), ('أماكن', 'NL_ILF'), ('آمنة', 'PUNC', '.', 'NC_ISF'), ('يحصل', 'VP_A3SM'), ('والد', 'NC_ISM'), ('دانيا', 'AS_ISM'), ('على', 'RP'), ('عرض', 'NA_ISM'), ('من', 'RP'), ('الشركة', 'NA_DSF'), ('التي', 'PR_SF'), ('كان', 'RP'), ('يعمل', 'VP_A3SM'), ('فيها', 'RP_R3RSRF'), ('ل', 'RP'), ('ينضم', 'VP_A3SM'), ('ل', 'RP'), ('فرعها', 'NA_ISM_R3RSRF'), ('في', 'RP'), ('الكويت', 'PUNC', '.', 'NC_DSF')]

تصحيح أخطاء التوسيم:	
الوسم الخاطئ	الوسم الصحيح
('VP_A3SF', 'تدور')	('VP_A3LF', 'تدور')
('NI_ISM', 'طواحين')	('NM_ILF', 'طواحين')
('NA_DSF_AT', 'اللبنانية')	('NC_DSF_AT', 'اللبنانية')
('AS_ISM', 'دانيا')	('NC_ISF', 'دانيا')
('NA_DSF', 'الطالبة')	('AS_DSF', 'الطالبة')
('NC_DSF_AT', 'الثانوية')	('AS_DSF_AT', 'الثانوية')
('NA_ISM', 'لأواء')	('NA_ISF', 'لأواء')
('NA_DSM', 'الحرب')	('NA_DSF', 'الحرب')
('NA_ILF_R3RSRF', 'أسرتها')	('NV_ISF_R3RSRF', 'أسرتها')
('NA_DLF', 'الطائشة')	('AS_DSF', 'الطائشة')
('NA_ISM', 'سكان')	('AS_ILM', 'سكان')
('NI_ISM', 'ملجأ')	('NL_ISM', 'ملجأ')
('NA_ISM', 'تعرض')	('VP_A3SF', 'تعرض')
('NA_ISM', 'تذوب')	('VP_A3LF', 'تذوب')
('NC_DSM', 'السكّان')	('AS_DLM', 'السكّان')
('NA_DSM', 'الرفيع')	('AA_DSM', 'الرفيع')
('NC_DSF', 'علاقة')	('NA_DSF', 'علاقة')
('AS_ISM', 'دانيا')	('NC_ISF', 'دانيا')
('AS_DSM', 'المبادئ')	('NA_DLF', 'المبادئ')
('NC_ISF', 'جارهم')	('NC_ISM_R3RLRM', 'جارهم')

('الصحافي', 'NC_DSM_AT')	('الصحافي', 'NA_DSM_AT')
('يتبناها', 'VP_A3SM')	('يتبناها', 'VP_A3SM_R3RLRF')
('ثقافيا', 'NC_ISF')	('ثقافيا', 'NA_ISM_AT')
('معظم', 'NC_ISM')	('معظم', 'NI_ISM')
('قاطني', 'NC_ISM_12')	('قاطني', 'AS_ILM')
('الملجأ', 'NC_ISM_22')	('الملجأ', 'NL_DSM')
('الموت', 'NC_DSM')	('الموت', 'NA_DSM')
('الرحيل', 'NC_DSM')	('الرحيل', 'NA_DSM')
('آمنة', 'NC_ISF')	('آمنة', 'AS_ISF')
('دانيا', 'AS_ISM')	('دانيا', 'NC_ISF')
('الأهلية', 'NV_DSF_AT')	('الأهلية', 'NV_DSF_AT')

ويلاحظ في الجدول (٤-١٥-د) أن ثمة أخطاء تسبب بها الغموض الكتابي للكلمات الواردة في النص، فكلمة (سِنة) تشابه الاسم المبهم (سَنة) الموسوم في مدونة التدريب بـ (NI)، وقد وسم النظام المقترح كلمة (سِنة) في هذا النص على أنها اسم مبهم رغم أنها هنا اسم معنى. ولأنه يندر أن يتوالى فعلان في العربية، فقد وسم الموسم (وسع) في كل مجموعات الوسوم باعتبارها اسما لأنها مسبوقة بـفعل آخر (شاء).

جدول (٤-١٥-د) تجريب النظام على عينة من نصوص متنوعة غير منحازة (نص ٤)

نص (قرآني) مقطع آليا (٤)
الله ل ا إله إلا هو الحي القيو م ل ا تأخذه سنة وال ا نو م ل ه ما ف ي السماوات وما ف ي الأرض من ذا الذي يشفع عنده إلا ب إذنه يعل م ما بين أيديهم وما خلفهم وال ا يحيطون بشيء من علمه إلا ب ما شاء وس ع كرسيه السماوات والأرض وال ا يئوده حفظهما وهو العلي العظيم
النص مقطعا بعد المراجعة اليدوية **عدد الكلمات الفعلية: ٦٠**

الله لا إله إلا هو الحي القيوم لا تأخذه سنة ولا نوم له ما في السماوات وما في الأرض من ذا الذي يشفع عنده إلا ب إذنه يعلم ما بين أيديهم وما خلفهم ولا يحيطون ب شيء من علمه إلا ب ما شاء وسع كرسيه السماوات والأرض ولا يئوده حفظهما وهو العلي العظيم
توسيمه بالوسوم الرئيسة **عدد الأخطاء:٢/ نسبة الخطأ: ٠,٠١٢**
['الله', 'N'), ('لا', 'RP'), ('إله', 'N'), ('إلا', 'RP'), ('هو', 'P'), ('الحي', 'N'), ('القيوم', 'N'), ('لا', 'RP'), ('تأخذه', 'V'), ('سنة', 'N'), ('و', 'RP'), ('لا', 'RP'), ('نوم', 'N'), ('له', 'RP'), ('ما', 'P'), ('في', 'RP'), ('السماوات', 'N'), ('و', 'RP'), ('ما', 'P'), ('في', 'RP'), ('الأرض', 'N'), ('من', 'RP'), ('ذا', 'P'), ('الذي', 'P'), ('يشفع', 'V'), ('عنده', 'N'), ('إلا', 'RP'), ('ب', 'RP'), ('إذنه', 'N'), ('يعلم', 'V'), ('ما', 'P'), ('بين', 'N'), ('أيديهم', 'N'), ('و', 'RP'), ('ما', 'P'), ('خلفهم', 'V'), ('و', 'RP'), ('لا', 'RP'), ('يحيطون', 'V'), ('ب', 'RP'), ('شيء', 'N'), ('من', 'RP'), ('علمه', 'N'), ('إلا', 'RP'), ('ب', 'RP'), ('ما', 'P'), ('شاء', 'V'), ('وسع', 'N'), ('كرسيه', 'N'), ('السماوات', 'N'), ('و', 'RP'), ('الأرض', 'N'), ('و', 'RP'), ('لا', 'RP'), ('يئوده', 'V'), ('حفظهما', 'N'), ('و', 'RP'), ('هو', 'P'), ('العلي', 'N'), ('العظيم', 'N')]
تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('V', 'خلفهم')	('N', 'خلفهم')
('N', 'وسع')	('V', 'وسع')

توسيمه بالوسوم الفرعية **عدد الأخطاء: ٩/ نسبة الخطأ: ٠,٠٥٤**
['الله', 'NC'), ('لا', 'RP'), ('إله', 'NC'), ('إلا', 'RP'), ('هو', 'PP'), ('الحي', 'NA'), ('القيوم', 'NC'), ('لا', 'RP'), ('تأخذه', 'VP'), ('سنة', 'NI'), ('و', 'RP'), ('لا', 'RP'), ('نوم', 'VP'), ('له', 'RP'), ('ما', 'PR'), ('في', 'RP'), ('السماوات', 'NA'), ('و', 'RP'), ('ما', 'PR'), ('في', 'RP'), ('الأرض', 'NC'), ('من', 'RP'), ('ذا', 'PD'), ('الذي', 'PR'), ('يشفع', 'VP'), ('عنده', 'NI'), ('إلا', 'RP'), ('ب', 'RP'), ('إذنه', 'NA'), ('يعلم', 'VP'), ('ما', 'PR'), ('بين', 'NI'), ('أيديهم', 'NC'), ('و', 'RP'), ('ما', 'PR'), ('خلفهم', 'VS'), ('و', 'RP'), ('لا', 'RP'), ('يحيطون', 'VP'), ('ب', 'RP'), ('شيء', 'NA'), ('من', 'RP'), ('علمه', 'NA'), ('إلا', 'RP'), ('ب', 'RP'), ('ما', 'PR'), ('شاء', 'VS'), ('وسع', 'NA'), ('كرسيه', 'RP'), ('السماوات', 'NA'), ('و', 'RP'), ('الأرض', 'NC'), ('و', 'RP'), ('لا', 'RP'), ('يئوده', 'VP'), ('حفظهما', 'NA'), ('و', 'RP'), ('هو', 'PP'), ('العلي', 'NA'), ('العظيم', 'NC')]

تصحيح أخطاء التوسيم:	
الوسم الخاطئ	الوسم الصحيح
('الحي', 'NA')	('الحي', 'NC')
('سنة', 'NI')	('سنة', 'NA')
('نوم', 'VP')	('نوم', 'NA')
('السماوات', 'NA')	('السماوات', 'NC')
('خلفهم', 'VS')	('خلفهم', 'NI')
('وسع', 'NA')	('وسع', 'VS')
('كرسيه', 'RP')	('كرسيه', 'NC')
('السماوات', 'NA')	('السماوات', 'NC')
('العلي', 'NA')	('العلي', 'NC')
توسيمه بالوسوم الموسعة بالخصائص التصريفية	**عدد الأخطاء: ١١ / نسبة الخطأ: ٠,٠٥٤**
['الله', 'NC_ISM'), ('لا', 'RP'), ('إله', 'NC_ISM'), ('إلا', 'RP'), ('هو', 'PP_3SM'), ('الحي', 'NA_DSM'), ('القيوم', 'NC_DSM'), ('لا', 'RP'), ('تأخذه', 'VP_A2SM'), ('سنة', 'NI_ISF'), ('و', 'RP'), ('لا', 'RP'), ('نوم', 'VP_A1LM'), ('له', 'RP_R3RSRM'), ('ما', 'PR_SM'), ('في', 'RP'), ('السماوات', 'NC_DLF'), ('و', 'RP'), ('ما', 'PR_SM'), ('في', 'RP'), ('الأرض', 'NC_DSM'), ('من', 'RP'), ('ذا', 'PD_SM'), ('الذي', 'PR_SM'), ('يشفع', 'VP_A3SM'), ('عنده', 'NA_ISM_R3RSRM'), ('إلا', 'RP'), ('ب', 'RP'), ('إذنه', 'NA_ISM_R3RSRM'), ('يعلم', 'VP_A3SM'), ('ما', 'PR_SM'), ('بين', 'NI_ISM'), ('أيديهم', 'RP_R3RLRM'), ('و', 'RP'), ('ما', 'PR_SM'), ('خلفهم', 'VS_A3SM'), ('و', 'RP'), ('لا', 'RP'), ('يحيطون', 'VP_A3LM'), ('ب', 'RP'), ('شيء', 'NA_ISM'), ('من', 'RP'), ('علمه', 'VS_A3SM_R3RSRM'), ('إلا', 'RP'), ('ب', 'RP'), ('ما', 'PR_SM'), ('شاء', 'VS_A3SM'), ('وسع', 'NA_ISM'), ('كرسيه', 'NA_ISM_R3RSRM'), ('السماوات', 'NC_DLF'), ('و', 'RP'), ('الأرض', 'NC_DSM'), ('و', 'RP'), ('لا', 'RP'), ('يئوده', 'VP_A3SM_R3RSRM'), ('حفظهما', 'NA_ISM'), ('و', 'RP'), ('هو', 'PP_3SM'), ('العلي', 'NC_DSM_AT'), ('العظيم', 'NC_DLF')]	

تصحيح أخطاء التوسيم:	
الوسم الخاطئ	الوسم الصحيح
('الحي', 'NA_DSM')	('الحي', 'NC_DSM')
('تأخذه', 'VP_A2SM')	('تأخذه', 'VP_A3SM')
('سنة', 'NI_ISF')	('سنة', 'NA_ISF')
('نوم', 'VP_A1LM')	('نوم', 'NA_ISM')
('عنده', 'NA_ISM_R3RSRM')	('عنده', 'NI_ISM_R3RSRM')
('أيديهم', 'RP_R3RLRM')	('أيديهم', 'NC_ILF_R3RLRM')
('خلفهم', 'VS_A3SM')	('خلفهم', 'NI_ISM_R3RLRM')
('وسع', 'NA_ISM')	('وسع', 'VS_A3SM')
('كرسيه', 'NA_ISM_R3RSRM')	('كرسيه', 'NC_ISM_R3RSRM')
('حفظهما', 'NA_ISM')	('حفظهما', 'NA_ISM_R3RURM')
('العلي', 'NC_DSM_AT')	('العلي', 'NC_DSM')

وفي الجدول (٤-١٥-هـ) يلاحظ على الأخطاء في مجموعة الوسوم الموسعة أنها تقع في القسم الرئيس غالبا وليست في الخصائص وحسب. ويظهر أن بعض الأخطاء التي ترد في الوسوم الرئيسة في نص معين تظهر على مستوى الوسوم الفرعية أو الوسوم الموسعة صحيحة، ويحدث كذلك العكس. فكلمة (المعدنية) ترد خطأ في نص (٥) بمستوى الوسوم الرئيسة بوصفها صفة، ثم في مستوى الوسوم الفرعية بوصفها صفة مفعول، وفي مستوى الوسوم الموسعة ترد بوسم صحيح (NC_DSF_AT اسم ذات_معرف بأل/ مفرد/ مؤنث_منسوب). وفي ذات النص يسم الموسَّم كلمة (فروعها) في المستوى الأول على أنها اسم وفي الثاني على أنها اسم معنى ثم يخطئ في توسيمها في المستوى الموسع ويسمها باعتبار أنها أداة متصلة بضمير،

وليست اسم معنى متصلا بضمير، وهو يقرر ذلك حسب ما زود به من الاحتمالات الانتقالية كما يظهر في الجدول (٤-١٦-ج). ويحدث ألا يتعرف الموسم على الكلمة في كل المستويات وهذا في حالات قليلة جدا، كما في: ضئيلة ـ الغابرة.

جدول (٤-١٥-هـ) تجريب النظام على عينة من نصوص متنوعة غير منحازة (نص ٥)

نص (ثقافي) مقطع آليا (٥)
تستخدم مواد التراث الشعبي والحياة الشعبية ف ي إعادة بناء الفترات التاريخية الغابرة ال الأمم والشعوب والتي ل ا يوجد ل ها إلا شواهد ضئيلة متفرقة وتستخدم أيضا ال إبراز الهوية الوطنية والقومية والكشف عن ملامحها. التراث والمأثورات التراثية ب شكلها ومضمونها أصيلة ومتجذرة إلا أن ف روعها تتطور وتتوسع مع مرور الزمن وبنسب مختلفة وذلك ب فعل التراكم الثقافي والحضاري وتبادل التأثر والتأثير مع الثقافات والحضارات الأخرى وعناصر التغيير والحراك ف ي الظروف الذاتية والاجتماعية ل كل مجتمع. ويتنوع التراث ب اختلاف ما تحمل ه الجذور إلى الشجرة. فقد تحمل إليها قوتها المتمثل ف ي الأملاح المعدنية وهو ب مثابة ما دون من التراث ف إن فقد ف س تصير الأمة ك شجرة حبست عنها الأملاح المعدنية ف س تذبل حتما شيئا ف شيئا ث م تضمحل.
النص مقطعا بعد المراجعة اليدوية **عدد الكلمات الفعلية:١٤٥**
تستخدم مواد التراث الشعبي والحياة الشعبية في إعادة بناء الفترات التاريخية الغابرة ل الأمم والشعوب والتي لا يوجد لها إلا شواهد ضئيلة متفرقة وتستخدم أيضا ل إبراز الهوية الوطنية والقومية والكشف عن ملامحها التراث والمأثورات التراثية ب شكلها ومضمونها أصيلة ومتجذرة إلا أن فروعها تتطور وتتوسع مع مرور الزمن وب نسب مختلفة وذلك ب فعل التراكم الثقافي والحضاري وتبادل التأثر والتأثير مع الثقافات والحضارات الأخرى وعناصر التغيير والحراك في الظروف الذاتية والاجتماعية ل كل مجتمع ويتنوع التراث ب اختلاف ما تحمله الجذور إلى الشجرة. ف قد تحمل إليها قوتها المتمثل في الأملاح المعدنية وهو ب مثابة ما دون من التراث ف إن فقد ف س تصير الأمة ك شجرة حبست عنها الأملاح المعدنية ف س تذبل حتما شيئا ف شيئا ثم تضمحل.
توسيمه بالوسوم الرئيسة **عدد الأخطاء: ٨ / نسبة الخطأ: ١١٦ , ٠**

[('تستخدم', 'V'), ('مواد', 'N'), ('التراث', 'N'), ('الشعبي', 'N'), ('و', 'RP'), ('الحياة', 'N'), ('الشعبية', 'N'), ('في', 'RP'), ('إعادة', 'N'), ('بناء', 'N'), ('الفترات', 'N'), ('التاريخية', 'N'), ('الغابرة', 'N'), ('ل', 'RP'), ('الأمم', 'N'), ('و', 'RP'), ('الشعوب', 'N'), ('و', 'RP'), ('التي', 'P'), ('لا', 'RP'), ('يوجد', 'V'), ('لها', 'RP'), ('إلا', 'RP'), ('شواهد', 'N'), ('ضئيلة', 'N'), ('متفرقة', 'A'), ('و', 'RP'), ('تستخدم', 'V'), ('أيضا', 'N'), ('ل', 'RP'), ('إبراز', 'N'), ('الهوية', 'N'), ('الوطنية', 'N'), ('و', 'RP'), ('القومية', 'N'), ('و', 'RP'), ('الكشف', 'N'), ('عن', 'RP'), ('ملامحها', 'N'), ('التراث', 'N'), ('و', 'RP'), ('المأثورات', 'N'), ('التراثية', 'N'), ('ب', 'RP'), ('شكلها', 'N'), ('و', 'RP'), ('مضمونها', 'N'), ('أصيلة', 'A'), ('و', 'RP'), ('متجذرة', 'A'), ('إلا', 'RP'), ('أن', 'RP'), ('فروعها', 'N'), ('تتطور', 'N'), ('و', 'RP'), ('تتوسع', 'V'), ('مع', 'N'), ('مرور', 'N'), ('الزمن', 'N'), ('و', 'RP'), ('ب', 'RP'), ('نسب', 'N'), ('مختلفة', 'A'), ('و', 'RP'), ('ذلك', 'P'), ('ب', 'RP'), ('فعل', 'N'), ('التراكم', 'N'), ('الثقافي', 'N'), ('و', 'RP'), ('الحضاري', 'N'), ('و', 'RP'), ('تبادل', 'N'), ('التأثر', 'N'), ('و', 'RP'), ('التأثير', 'N'), ('مع', 'N'), ('الثقافات', 'N'), ('و', 'RP'), ('الحضارات', 'N'), ('الأخرى', 'A'), ('و', 'RP'), ('عناصر', 'N'), ('التغيير', 'N'), ('و', 'RP'), ('الحراك', 'N'), ('في', 'RP'), ('الظروف', 'N'), ('الذاتية', 'N'), ('و', 'RP'), ('الاجتماعية', 'N'), ('ل', 'RP'), ('كل', 'N'), ('مجتمع', 'A'), ('و', 'RP'), ('يتنوع', 'V'), ('التراث', 'N'), ('ب', 'RP'), ('اختلاف', 'N'), ('ما', 'P'), ('تحمله', 'V'), ('الجذور', 'N'), ('إلى', 'RP'), ('الشجرة', 'PUNC'), ('.', 'N'), ('ف', 'RP'), ('قد', 'RP'), ('تحمل', 'V'), ('إليها', 'RP'), ('قوتها', 'N'), ('المتمثل', 'A'), ('في', 'RP'), ('الأملاح', 'N'), ('المعدنية', 'A'), ('و', 'RP'), ('هو', 'P'), ('ب', 'RP'), ('مثابة', 'N'), ('ما', 'RP'), ('دون', 'N'), ('من', 'RP'), ('التراث', 'N'), ('ف', 'RP'), ('إن', 'RP'), ('فقد', 'V'), ('ف', 'RP'), ('س', 'RP'), ('تصير', 'V'), ('الأمة', 'N'), ('ك', 'RP'), ('شجرة', 'N'), ('حبست', 'V'), ('عنها', 'RP'), ('الأملاح', 'N'), ('المعدنية', 'A'), ('ف', 'RP'), ('س', 'RP'), ('تذبل', 'V'), ('حتما', 'N'), ('شيئا', 'N'), ('ف', 'RP'), ('شيئا', 'N'), ('ثم', 'RP'), ('تضمحل', 'PUNC'), ('.', 'V')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('الغابرة', 'N'),	('الغابرة', 'A'),
('شواهد', 'N'),	('شواهد', 'A'),
('ضئيلة', 'N'),	('ضئيلة', 'A'),

('المأثورات', 'N'),	('المأثورات', 'A'),
('مضمونها', 'N'),	('مضمونها', 'A'),
('تتطور', 'N'),	('تتطور', 'V'),
('المعدنية', 'A'),	('المعدنية', 'N'),
('دون', 'N'),	('دون', 'V'),

توسيمه بالوسوم الفرعية **عدد الأخطاء: ٢٠ / نسبة الخطأ: ٠,٣٠٥**

['تستخدم', 'VP'), ('مواد', 'NA'), ('التراث', 'NA'), ('الشعبي', 'NV'), ('و', 'RP'), ('الحياة', 'NA'), ('الشعبية', 'NV'), ('في', 'RP'), ('إعادة', 'NA'), ('بناء', 'NA'), ('الفترات', 'NA'), ('التاريخية', 'NA'), ('الغابرة', 'NA'), ('ل', 'RP'), ('الأمم', 'NC'), ('و', 'RP'), ('الشعوب', 'NV'), ('و', 'RP'), ('التي', 'PR'), ('لا', 'RP'), ('يوجد', 'VP'), ('لها', 'RP'), ('إلا', 'RP'), ('شواهد', 'NC'), ('ضئيلة', 'NA'), ('متفرقة', 'AS'), ('و', 'RP'), ('تستخدم', 'VP'), ('أيضا', 'NA'), ('ل', 'RP'), ('إبراز', 'NC'), ('الهوية', 'NC'), ('الوطنية', 'NA'), ('و', 'RP'), ('القومية', 'NA'), ('و', 'RP'), ('الكشف', 'NC'), ('عن', 'RP'), ('ملامحها', 'NA'), ('التراث', 'NA'), ('و', 'RP'), ('المأثورات', 'AO'), ('التراثية', 'NA'), ('ب', 'RP'), ('شكلها', 'NA'), ('و', 'RP'), ('مضمونها', 'RP'), ('أصيلة', 'NA'), ('و', 'RP'), ('متجذرة', 'AS'), ('إلا', 'RP'), ('أن', 'RP'), ('فروعها', 'NA'), ('تتطور', 'NA'), ('و', 'RP'), ('تتوسع', 'VP'), ('مع', 'NI'), ('مرور', 'AO'), ('الزمن', 'NI'), ('و', 'RP'), ('ب', 'RP'), ('نسب', 'NA'), ('مختلفة', 'AS'), ('و', 'RP'), ('ذلك', 'PD'), ('ب', 'RP'), ('فعل', 'NA'), ('التراكم', 'NA'), ('الثقافي', 'NA'), ('و', 'RP'), ('الحضاري', 'NA'), ('و', 'RP'), ('تبادل', 'NA'), ('التأثر', 'NA'), ('و', 'RP'), ('التأثير', 'NA'), ('مع', 'NI'), ('الثقافات', 'NA'), ('و', 'RP'), ('الحضارات', 'NA'), ('الأخرى', 'AC'), ('و', 'RP'), ('عناصر', 'NC'), ('التغيير', 'NA'), ('و', 'RP'), ('الحراك', 'NA'), ('في', 'RP'), ('الظروف', 'NA'), ('الذاتية', 'NA'), ('و', 'RP'), ('الاجتماعية', 'NA'), ('ل', 'RP'), ('كل', 'NI'), ('مجتمع', 'AS'), ('و', 'RP'), ('يتنوع', 'VP'), ('التراث', 'NA'), ('ب', 'RP'), ('اختلاف', 'NA'), ('ما', 'PR'), ('تحمله', 'VP'), ('الجذور', 'NA'), ('إلى', 'RP'), ('الشجرة', 'PUNC'), ('.', 'NA'), ('ف', 'RP'), ('قد', 'RP'), ('تحمل', 'VS'), ('إليها', 'RP'), ('قوتها', 'NA'), ('المتمثل', 'AS'), ('في', 'RP'), ('الأملاح', 'NA'), ('المعدنية', 'AO'), ('و', 'RP'), ('هو', 'PP'), ('ب', 'RP'), ('مثابة', 'NL'), ('ما', 'RP'), ('دون', 'NI'), ('من', 'RP'), ('التراث', 'NA'), ('ف', 'RP'), ('إن', 'RP'), ('فقد', 'VS'), ('ف', 'RP'), ('س', 'RP'), ('تصير', 'VP'),

('الأمة', 'NA'), ('ك', 'RP'), ('شجرة', 'NC'), ('حبست', 'VS'), ('عنها', 'RP'), ('الأملاح', 'NA'), ('المعدنية', 'AO'), ('ف', 'RP'), ('س', 'RP'), ('تذبل', 'VP'), ('حتما', 'NA'), ('شيئا', 'NA'), ('ف', 'RP'), ('شيئا', 'VS'), ('ثم', 'RP'), ('تضمحل', 'PUNC', '.'), ('VS')]

تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح	الوسم الخاطئ	الوسم الصحيح
('NA', 'الغابرة')	('AS', 'الغابرة')	('NA', 'الذاتية')	('NI', 'الذاتية')
('NC', 'الأمم')	('NV', 'الأمم')	('AS', 'مجتمع')	('AO', 'مجتمع')
('NC', 'شواهد')	('AS', 'شواهد')	('NA', 'الشجرة')	('NC', 'الشجرة'),
'NA', 'ضئيلة')	('AA', 'ضئيلة')	('NA', 'الأملاح')	('NV', 'الأملاح'),
('NC', 'إبراز')	('NA', 'إبراز')	('AO', 'المعدنية')	('NC', 'المعدنية')
('NC', 'الهوية')	('NA', 'الهوية')	('NI', 'دون')	('VS', 'دون')
('NC', 'الكشف')	('NA', 'الكشف')	('NA', 'الأمة')	('NV', 'الأمة')
('RP', 'مضمونها')	('AO', 'مضمونها')	('NA', 'الأملاح')	('NV', 'الأملاح')
('NA', 'أصيلة')	('AA', 'أصيلة')	('VS', 'شيئا')	('NA', 'شيئا'),
('NA', 'تتطور')	('VP', 'تتطور')	('VS', 'تضمحل')	('VP', 'تضمحل')
('AO', 'مرور')	('NA', 'مرور')		

توسيمه بالوسوم الموسعة بالخصائص التصريفية **عدد الأخطاء: ٣٤/ نسبة الخطأ: ٠,٥١**

['تستخدم', 'VS_A3SM'), ('مواد', 'NA_ISM'), ('التراث', 'NA_DSM'), ('الشعبي', 'NC_DSM_AT'), ('و', 'RP'), ('الحياة', 'NA_DSF'), ('الشعبية', 'NC_DSF_AT'), ('في', 'RP'), ('إعادة', 'NA_ISF'), ('بناء', 'NA_ISM'), ('الفترات', 'NA_DLF'), ('التاريخية', 'NA_DSF_AT'), ('الغابرة', 'NA_DSF'), ('ل', 'RP'), ('الأمم', 'NC_DLF'), ('و', 'RP'), ('الشعوب', 'NA_DSM'), ('و', 'RP'), ('التي', 'PR_LF'), ('لا', 'RP'), ('يوجد', 'VP_P3SM'), ('لها', 'RP_R3RSRF'), ('إلا', 'RP'), ('شواهد', 'NC_ISM'), ('ضئيلة', 'NA_ISF'), ('متفرقة', 'AS_ISF'), ('و', 'RP'), ('تستخدم', 'VS_A3SM'), ('أيضا', 'NA_ISM'), ('ل', 'RP'), ('إبراز', 'NA_ISM'),

('الهوية', 'NA_DSF'), ('الوطنية', 'NA_DSF_AT'), ('و', 'RP'), ('القومية', 'NA_DSF'), ('و', 'RP'), ('الكشف', 'NC_DSM'), ('عن', 'RP'), ('ملامحها', 'NA_ISM_R3RSRF'), ('التراث', 'NA_DSM'), ('و', 'RP'), ('المأثورات', 'NA_DLF'), ('التراثية', 'NA_DSF_AT'), ('ب', 'RP'), ('شكلها', 'NA_ISM_R3RSRF'), ('و', 'RP'), ('مضمونها', 'RP_R3RSRF'), ('أصيلة', 'NA_ISF'), ('و', 'RP'), ('متجذرة', 'AS_ISF'), ('إلا', 'RP'), ('أن', 'RP'), ('فروعها', 'RP_R3RSRF'), ('تتطور', 'VP_A3SF'), ('و', 'RP'), ('تتوسع', 'VP_A3SF'), ('مع', 'NI_ISM'), ('مرور', 'NI_ISM'), ('الزمن', 'NI_DSM'), ('و', 'RP'), ('ب', 'RP'), ('نسب', 'NA_ISM'), ('مختلفة', 'NA_ISF'), ('و', 'RP'), ('ذلك', 'PD_SM'), ('ب', 'RP'), ('فعل', 'NA_ISM'), ('التراكم', 'NA_DSM'), ('الثقافي', 'NC_DSM_AT'), ('و', 'RP'), ('الحضاري', 'NC_DSM_AT'), ('و', 'RP'), ('تبادل', 'NA_ISM'), ('التأثر', 'NA_ DSM'), ('و', 'RP'), ('التأثير', 'NA_DSM'), ('مع', 'NI_ISM'), ('الثقافات', 'NA_DLF'), ('و', 'RP'), ('الحضارات', 'NA_DLF'), ('الأخرى', 'AC_DSF'), ('و', 'RP'), ('عناصر', 'NC_ILF'), ('التغيير', 'NA_DSM'), ('و', 'RP'), ('الحراك', 'NC_DSM'), ('في', 'RP'), ('الظروف', 'NA_DLF'), ('الذاتية', 'NC_DSF_AT'), ('و', 'RP'), ('الاجتماعية', 'NA_DSF_AT'), ('ل', 'RP'), ('كل', 'NI_ISM'), ('مجتمع', 'AS_ISM'), ('و', 'RP'), ('يتنوع', 'VP_A3SM'), ('التراث', 'NA_DSM'), ('ب', 'RP'), ('اختلاف', 'NA_ISM'), ('ما', 'PR_SM'), ('تحمله', 'VS_A3SM'), ('الجذور', 'NA_DSM'), ('إلى', 'RP'), ('الشجرة', 'PUNC'), ('.', 'NC_DSF'), ('ف', 'RP'), ('قد', 'RP'), ('تحمل', 'VP_A3LF'), ('إليها', 'RP_R3RSRF'), ('قوتها', 'NA_ISM_R3RSRF'), ('المتمثل', 'AS_DSM'), ('في', 'RP'), ('الأملاح', 'NC_DLF'), ('المعدنية', 'NC_DSF_AT'), ('و', 'RP'), ('هو', 'PP_3SM'), ('ب', 'RP'), ('مثابة', 'NL_ISF'), ('ما', 'RP'), ('دون', 'NI_ISM'), ('من', 'RP'), ('التراث', 'NA_DSM'), ('ف', 'RP'), ('إن', 'RP'), ('فقد', 'VS_P3SM'), ('ف', 'RP'), ('س', 'RP'), ('تصير', 'NA_ISM'), ('الأمة', 'NA_DSF'), ('ك', 'RP'), ('شجرة', 'NA_ISF'), ('حبست', 'VS_A3SF'), ('عنها', 'RP_R3RSRF'), ('الأملاح', 'NA_DSF'), ('المعدنية', 'NC_DSF_AT'), ('ف', 'RP'), ('س', 'RP'), ('تذبل', 'NA_ISM'), ('حتما', 'NA_ISM'), ('شيئا', 'NA_ISM'), ('ف', 'RP'), ('شيئا', 'NA_ISM'), ('ثم', 'RP'), ('تضمحل', 'PUNC'), ('.', 'VS_A3SM')]
تصحيح أخطاء التوسيم:

الوسم الخاطئ	الوسم الصحيح
('تستخدم', 'VS_A3SM')	('تستخدم', 'VP_P3LF')

('مواد', 'NA_ISM')	('مواد', 'NC_ILF')
('الشعبي', 'NC_DSM_AT')	('الشعبي', 'NV_DSM_AT')
('الشعبية', 'NC_DSF_AT')	('الشعبية', 'NV_DSF_AT')
('الغابرة', 'NA_DSF'),	('الغابرة', 'AS_DSF')
('الشعوب', 'NA_DSM')	('الشعوب', 'NV_DLF')
('شواهد', 'NC_ISM')	('شواهد', 'AS_ILF')
('ضئيلة', 'NA_ISF')	('ضئيلة', 'AA_ISF')
('تستخدم', 'VS_A3SM')	('تستخدم', ' VP_P3LF')
('الكشف', 'NC_DSM')	('الكشف', 'NA_DSM')
('ملامحها', 'NA_ISM_R3RSRF')	('ملامحها', 'NA_ILF_R3RSRF')
('المأثورات', 'NA_DLF')	('المأثورات', 'AO_DLF')
('مضمونها', 'RP_R3RSRF')	('مضمونها', 'AO_ISM_R3RLRF')
('أصيلة', 'NA_ISF')	('أصيلة', 'AA_ISF')
('فروعها', 'RP_R3RSRF')	('فروعها', 'NA_ILF_R3RLRF')
('تتطور', 'VP_A3SF')	('تتطور', 'VP_A3LF')
('تتوسع', 'VP_A3SF')	('تتوسع', 'VP_A3LF')
('مرور', 'NI_ISM')	('مرور', 'NA_ISM')
('مختلفة', 'NA_ISF')	('مختلفة', 'AS_ISF')
('الثقافي', 'NC_DSM_AT')	('الثقافي', 'NA_DSM_AT')
('الحضاري', 'NC_DSM_AT')	('الحضاري', 'NA_DSM_AT')
('الحراك', 'NC_DSM')	('الحراك', 'NA_DSM')
('الذاتية', 'NC_DSF_AT')	('الذاتية', 'NI_DSF_AT')

('مجتمع', 'AS_ISM')	('مجتمع', 'AO_ISM')
('الجذور', 'NA_DSM')	('الجذور', 'NC_DLF')
('تحمله', 'VS_A3SM')	('تحمله', 'VP_A3LF')
('دون', 'NI_ISM')	('دون', 'VS_P3SM')
('الأمة', 'NA_DSF')	('الأمة', 'NV_DSF')
('الأمم', 'NC_DLF')	('الأمم', 'NV_DLF')
('تصير', 'NA_ISM')	('تصير', 'RP')
('شجرة', 'NA_ISF')	('شجرة', 'NC_ISF')
('حبست', 'VS_A3SF')	('حبست', 'VS_P3LF')
('الأملاح', 'NA_DSF')	('الأملاح', 'NV_DLF')
('تذبل', 'NA_ISM')	('تذبل', 'VP_A3SF')

والملاحظ أن نسبة الخطأ في الكلمات الموسمة ظهرت على تنوع هذه النصوص وتفرقها بنفس سلوك الأخطاء في عينة الاختبار. حيث تقل نسب الخطأ في الوسوم الرئيسة ثم تبدأ في الزيادة بالتناسب تدريجيا من الوسوم الرئيسة وحتى الموسعة في كل نص. ومن المهم أن نشير إلى أن النسب متقاربة لأن عدد الوسوم الرئيسة ١٢ وعدد الوسوم الفرعية ٢٧ وعدد الوسوم الموسعة المستعملة فقط ٣٩٧، وازدياد النسب في الوسوم الموسعة لا يعني أن الفارق ازداد، لأن تلك النسب تعكس نسبتها مع تناسب العدد المعروف **بالتكرار المقيس** normalized frequency [١]، فنسبة الخطأ ٦ , ١١٪ في المستوى الأول من وسوم النص (٥) مثلا تتوزع على ٧ أقسام كلامية فقط، ولكن نسبة الخطأ ٥١٪ في المستوى الثالث من وسوم النص (٥) تتوزع على ٣٩٧ وسم.

(١) لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق، ٢٠١٦، ص. ٩٥

جدول (٤-١٦-أ) الانتقالات الاحتمالية لأقسام الكلام الرئيسة العشر الأقل والأكثر تكرارا في مدونة النظام

الأقل	الأكثر
('A', 'FOREIGN'), -0.917759	('N', 'N'), 2.807295
('D', 'P'), -0.986816	('V', 'N'), 2.534715
('P', 'A'), -1.095815	('P', 'V'), 2.433003
('I', 'RP'), -1.12504	('RP', 'N'), 2.070789
('PUNC', 'P'), -1.342092	('N', 'A'), 1.71685
('N', 'D'), -1.393871	('PUNC', 'N'), 1.472447
('SYMB', 'V'), -1.486854	('N', 'PUNC'), 1.352487
('P', 'D'), -1.634189	('A', 'N'), 1.328307
('A', 'ABBREV'), -1.67904	('V', 'RP'), 1.308271
('A', 'DIGIT'), -1.721349	('PUNC', 'V'), 1.198329

جدول (٤-١٦-ب) الانتقالات الاحتمالية لأقسام الكلام الفرعية العشر الأقل والأكثر تكرارا في مدونة النظام

الأقل	الأكثر
('DT', 'PR'), -1.888575	('FOREIGN', 'FOREIGN'), 4.35533
('AC', 'NC'), -1.908955	(‘PR’, ‘VS’), 3.494444
('VS', 'VP'), -1.940151	(‘PR’, ‘VP’), 2.580604
('PR', 'PR'), -2.006652	(‘IV’, ‘VC’), 2.455391
('PD', 'AS'), -2.054006	(‘NC’, ‘NC’), 2.400208
('AA', 'PR'), -2.106171	(‘SYMB’, ‘DIGIT’), 2.230793
('AS', 'DIGIT'), -2.11015	(‘NA’, ‘NA’), 2.176482
('VP', 'AO'), -2.264127	(‘NI’, ‘NI’), 2.149397
('PR', 'NA'), -2.866001	(‘DIGIT’, ‘ABBREV’), 2.105638
('PR', 'NC'), -2.918679	(‘PR’, ‘PP’), 2.014306

جدول (٤-١٦-ج) الانتقالات الاحتمالية لأقسام الكلام الموسعة العشر الأقل والأكثر تكرارا في مدونة النظام

الأقل	الأكثر
('AA_DSM', 'PR_SM'), -1.793966	('NC_ISM_12', 'NC_ISM_22'), 9.61
('RP', 'NA_ISF_AT'), -1.793995	('FOREIGN', 'FOREIGN'), 5.463013
('NC_DLF', 'PR_SF'), -1.795486	('NC_ISF_12', 'NC_ISF_22'), 5.02
('RP', 'AS_DSF_33'), -1.795839	('PR_SM', 'VS_A3SM_R3RSRM'), 3.8
('AS_DSM', 'NI_DSM'), -1.866243	('AS_DSF_13', 'AS_DSF_23'), 3.82
('NI_ISM','NC_ILF_R2RLRM'),2.03	('NC_DLF', 'AA_DLF'), 3.67493
('AS_ISM', 'RP_R3RSRF'), -2.21	('AS_DSF_23', 'AS_DSF_33'), 3.40
('VP_A1LM', 'RP_R3RSRF'), -2.25	('VS_P3SF', 'NV_ILF'), 3.36600
('RP_R3RSRM', 'NI_ISM_R3RSRM'), -2.73	('NC_ISM', 'NC_ISM'), 3.346888
('PD_SM', 'AS_DLM'), -3.23972	('PR_LM', 'NI_ISM_R3RSRM'), 3.25

أما ما يتعلق بنوع الأخطاء، فيلاحظ أن معظم الأخطاء هي في الأسماء والأفعال والصفات، وأن الأخطاء الكثيرة ليست في أكثر الوسوم تكرارا في مدونة التدريب التي وردت أمثلة كثيرة لها، حيث لم ترد الأخطاء مثلا في الأدوات إلا في ٨ مواضع فقط، وتندر جدا الأخطاء في الضمائر بأنواعها لأنها قوائم مغلقة، رغم الغموض الذي قد يحدث في (ما الموصولة) لتشابهها الكتابي مع (ما) الأداة لكن أمثلتها كثيرة في مدونة التدريب. وكثرة الأخطاء في الأسماء والصفات قد تكون بسبب تقسيماتها المتعددة في مجموعتي الوسوم الفرعية والموسعة.

ويمكن تحسين أنظمة التوسيم النحوي الآلية بتحسين دقتها، وقد يكون ذلك بأخذ نتائج توسيم كلمات مدونة الاختبار في المقترح، والنظر في الأخطاء التي أخطأ فيها النظام الآلي للتوسيم النحوي، ومحاولة تحسين الخوارزمية المستعملة في النظام بإضافة خصائص تميزية عالية للخوارزمية واستبعاد الخصائص التميزية ذات الدقة الأقل، ثم

إعادة اختبار الخوارزمية أكثر من مرة حتى تتحسن نتائجها. ولكن النظام سيظل محصورا بقاعدة بيانات محدودة.

ويمكن أيضا أن توسع قاعدة البيانات نفسها فتُوسَّم يدويا نصوص من أوعية أوسع وموضوعات أكثر تنوعا وعددا. وتسمح تقنيات **التمهيد** bootstrapping الذاتية بتطوير أداة التوسيم النحوي الآلية دون الحاجة إلى مدونة كبيرة الحجم موسمة يدويا. وفيها تعد سريعا أداة توسيم نحوية ذات أداء ضعيف جدا وتستعمل لتحليل مدونة. ومن ثم تستعمل بيانات المدونة لتحسين الأداة باعتبارها بيانات تدريب. ثم تستعمل الأداة المحسنة لتحليل بيانات المدونة مرة أخرى، وهذه المرة ستكون بأخطاء أقل. ويعاد العمل مرارا حتى تتحسن الأداة[١]. وقد تُعلم مجموعة من الموسمات المزودة بكمية قليلة من **بيانات التوسيم الأولية** seed data، ثم يستعمل كل موسِّم لتوسيم بعض البيانات غير الموسمة مسبقا. وبعد ذلك تضاف بعض البيانات الفرعية للبيانات الموسمة حديثا في مدونة التدريب ولكل موسم. ثم تعاد هذه العملية وتكرر هذه العملية حتى تتحسن الدقة[٢].

ويمكن أن يحسن نموذج CRF الاحتمالي نفسه بدمجه مع خوارزميات تعلم الآلة العميق لاستخراج الخصائص التي تزود بها الخوارزمية بطريقة آلية لا يدوية، حيث إن تقنيات تعلم الآلة التقليدية لا تفيد في العمل مع بيانات ذات أبعاد عالية حتى وإن كانت ضخمة، إذ يبدأ أداؤها يستقر عند درجة معينة بلا تحسن على عكس تقنيات تعلم الآلة العميق التي تزيد من تحسن الأداء كلما تضخمت البيانات، وتُضعفه كلما قلت البيانات.

وحيث قد وسمت ٥٠٠ كلمة أخرى وأبنت فيها الأخطاء وصححتها، أضفت هذه

---

(١) Baker, P., Hardie, A., McEnery, A. A Glossary of Corpus Linguistics. Edinburgh University Press, Edinburgh: UK., 2006, pp. 22-23

(٢) Clark S., James R., Osborne M. Bootstrapping POS taggers using Unlabelled Data. Proceeings of the Seventh CoNLL conference, Canada: USA, 2003

البيانات لمدونة التدريب للتقطيع وكذلك مدونة التدريب للتوسيم للكشف عن مدى التحسن في النتائج إذا ما زاد حجم المدونة، وقد تحسنت فعلا نتائج التقطيع تحسنا طفيفا مقارنة بنتائج التقطيع قبل إضافة الخمسمائة كلمة (انظر جدول ٤-١٧)، كما تحسنت النتائج تباعا في كل مجموعات الوسوم (الرئيسة والفرعية والموسعة) بنفس درجة التحسن فيما بينها (انظر سلسلة الجداول ٤-١٨). وحتى تظهر نتائج التحسن بشكل واضح نحتاج لبيانات أكثر من نصوص مقطعة ونصوص موسمة بكل مجموعات الوسوم النحوية في المقترح، وهو الأمر الذي سأعمل عليها مستقبلا في سبيل تحسين أداء النظام الآلي بمجموعات الوسوم النحوية المقترحة.

جدول (٤-١٧) مقاييس الأداء للمقطع بعد إضافة الخمسمائة كلمة

المقياس	درجته
الصحة	0.9925
الدقة	0.9928
الاسترجاع	0.9925
مقياس ف	0.9926

جدول (٤-١٨-أ) مقاييس الأداء لموسم الأقسام الرئيسة بعد إضافة الخمسمائة كلمة

المقياس	درجته
الصحة	0.9173
الدقة	0.916
الاسترجاع	0.917
مقياس ف	0.915

جدول (٤-١٨-ب) مقاييس الأداء لموسم الأقسام الفرعية بعد إضافة الخمسمائة كلمة

المقياس	درجته
الصحة	0.8229
الدقة	0.528
الاسترجاع	0.522
مقياس ف	0.819

جدول (٤-١٨-ج) مقاييس الأداء لموسم الأقسام الموسعة بعد إضافة الخمسمائة كلمة

المقياس	درجته
الصحة	0.7269
الدقة	0.721
الاسترجاع	0.738
مقياس ف	0.715

الفصل الخامس
النتائج والتوصيات

يتناول الفصل الحالي ملخصا للنتائج والإسهامات التي قدمها البحث في مجال التوسيم النحوي للنصوص العربية وقد توصلت إليها بعد أن سعيت إلى تقديم نماذج حاسوبية لبناء نظام آلي للتوسيم النحوي في العربية منطلقة من نظرية لغوية عربية قدمها تمام حسان لأقسام الكلام، وتطبيق هذه النماذج على مدونة فصيحة متعددة الأوعية وممتدة الأزمنة أطرت بالإطار العام لمدونة الملك عبد العزيز للعلوم والتقنية. وبعد ذلك يقدم أهم التوصيات والدراسات المستقبلية المقترحة التي تستكمل جوانب هذا المقترح في ضوء هذه النتائج.

أولا: تلخيص النتائج:

إن ثمة اضطرابا وتداخلا في تقسيم الكلام عند النحاة أنفسهم قديمهم وحديثهم راجع للمعايير التي استندوا إليها في التقسيم. ولقد تتبعت أنظمة التوسيم النحوي الآلي بشكل عام، ثم وقفت على ما هو موجود من هذه الأنظمة في المدونات العربية الموسومة حاليا متخذة من الوصف التحليلي منهجا في ذلك. وقد بدا أن مجموعات الوسوم النحوية المطبقة على المدونات العربية غير صالحة لمعالجة النصوص حاسوبيا لافتقارها للدقة، ولاستبعادها بعض الأقسام التي تؤدي إلى خلط الأقسام بعضها ببعض. فمعظم مجموعات الوسوم المطبقة على المدونات العربية هي محاولات لاستيعاب العربية في إطار الإنجليزية؛ إذ اعتمدت على مجموعة وسوم باكولتر غير الملائمة للعربية، حيث إن

بعض الأقسام الكلامية الموجودة فيها لا وجود له في العربية، مثل: الأسماء المكممة. فضلا عن أن بعض المجموعات التوسيمية لا توفر ملفا توثيقيا documentation يشرح فيه كل ما يتعلق بالوسوم، وبعضها ضخم جدا، وإن كان صالحا نظريا فهو غير قابل للتطبيق عمليا.

واعتمادا على ما وجدته من قصور في أنظمة التوسيم النحوي الآلي الخاصة بالعربية فقد جمعت مدونة شاملة ومتوازنة تتخذ من الإطار النموذجي للمدونة العربية (المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية) KACSTAC إطارا لها. والسعي للحصول على مدونة ضخمة موسمة يدويا مهمة صعبة تستغرق وقتا طويلا يتجاوز الفترة المحددة لإنجاز هذا الهدف، وهو السبب الذي جعل من مدونة النظام ١٠ آلاف كلمة. لقد قسمت مدونة النظام عشرة أوعية، وقطعت نصوصها الخام وفق قائمة من متغيرات التقطيع التي تعتمد عليها مجموعات الوسوم المقترحة المنطلقة من نظرية تمام حسان لتقسيم الكلام، ثم وسمت توسيما يدويا. واستُعملت هذه المدونة المقطعة والموسومة يدويا بالاعتماد على منهج تعلم الآلة **تحت الإشراف** supervised لبناء نظامين آليين أولهمها للتقطيع والثاني للتوسيم النحوي.

لقد قدمت ١٣,٦٠٦ كلمة فعلية موسومة في سياقاتها، ويمكن الاستفادة منها في مشاريع بحثية متعددة ومتنوعة. وفي مراحل توسيم المدونة بدا أن نظرية تمام حسان في أقسام الكلام تحتاج إلى مزيد من التفصيل خاصة فيما يتعلق بقسمي الأدوات والجهة (الزمن النحوي) Aspect، ولذا لم توسم الأدوات باعتبارها أصلية ومحولة ولم توسم الأفعال باعتبار الزمن النحوي. كما توصلت إلى ثلاث مجموعات توسيمية مقترحة زوّد بها نظام آلي معتمد على خوارزمية من خوارزميات تعلم الآلة بإشراف وهي خوارزمية CRF. وقد بدا لي أن نماذج التوسيم المقترحة من نظريته تحتاج إلى مدونة ضخمة للدخول في تفاصيلها.

وبالنظر إلى معالجة مدونة النظام بمتغيرات التقطيع المقترحة في هذا الكتاب من

حيث مقاييس الأداء، فإن درجة **الصحة** accuracy ظهرت عالية جدا حيث بلغت ٩٩٣ , ٠ . وبالنظر إلى معالجة مدونة النظام بالوسوم المقترحة في هذا الكتاب من حيث مقاييس الأداء لوسوم الأقسام الرئيسة والفرعية والموسّعة، فإن نسبة **الصحة** accuracy هي نتيجة لقسمة عدد الوسوم الصحيحة على مجموع الوسوم كلها، ونسبة الخطأ هي نتيجة قسمة عدد الوسوم الخاطئة على مجموع الوسوم كلها. وبذلك فإن معدل نسبة الصحة لمقاييس وسوم الأقسام الرئيسة قد كان ٩١٥٨ , ٠، و٨٢٠٤ , ٠ لمقاييس وسوم الأقسام الفرعية، و٧٢١٠ , ٠ لمقاييس وسوم الأقسام الموسعة، ونسبة عدد الوسوم كلها هو (١)، وعليه تكون نسب الخطأ بطرح ١ من نسب الصحة، على النحو الآتي:

- معدل الخطأ في مقاييس الأداء لوسوم الأقسام الرئيسة: ٠٨٤٢ , ٠
- معدل الخطأ في مقاييس الأداء لوسوم الأقسام الفرعية ١٧٩٦ , ٠
- معدل الخطأ في مقاييس الأداء لوسوم الأقسام الموسعة ٢٧٩٠ , ٠

ولمعرفة دقة تجريب النظام التوسيمي النحوي الآلي المصمم، أجريت اختبارا مقارنا لمعدلات الصحة والخطأ عن طريق تطبيق جميع الوسوم الرئيسة والفرعية والموسعة المطبقة على عينة من نصوص غير منحازة تفوق الخمسمائة كلمة من الصحف والأدب والقرآن الكريم. وقد ظهرت نسب الخطأ في تلك العينة على الوسوم الرئيسة والفرعية والموسعة على النحو الآتي:

- نسبة الخطأ في الوسوم الرئيسة:

٠٦٩ , ٠+١٣٢ , ٠+١١٠ , ٠+٠١٢ , ٠+١١٦ , ٠= ٤٣٩ , ٠/ ٥ = ٠٨٧٨ , ٠

٤٣٩ , ٠/ ٥ = (٠٨٧٨ , ٠ معدل الخطأ)

- نسبة الخطأ في الوسوم الفرعية:

١٢٠ , ٠+٢٠٠ , ٠+٢٦٣ , ٠+٢٨٠ , ٠+٠٥٤ , ٠+٣٠٥ , ٠= ٩٤٢ , ٠

٠,٩٤٢/ ٥ = (١٨٨٤, ٠ معدل الخطأ)

- نسبة الخطأ في الوسوم الموسعة:

١,٤٠٨=٠,٥١٠+٠,٠٥٤+٠,٣٦٦+٠,٢٦٣+٠,٢٥١

١,٤٠٨/ ٥ = (٢٨١٦, ٠ معدل الخطأ)

وبمقارنة نسب معدلات الخطأ بين مقاييس الأداء لوسوم الأقسام الرئيسة والفرعية والموسعة التي ظهرت من مدونة النظام وبين النصوص غير المنحازة وفق نظام التوسيم النحوي الآلي المطبق، فإن الفروقات طفيفة جدا تكاد لا تُذكر، وهذا مؤشر على أن النظام الآلي يمكن تطبيقه على نصوص متنوعة بنفس الدقة التي أظهرها على نصوص مدونة الاختبار في النظام.

وفي مسعى مني لتحسين دقة المقطع والموسم الآلي، استخدمت نتيجة هذه الاختبارات لتدعيم مدونة التدريب، وتدريب النظام مرة أخرى. وقد أدى ذلك إلى تحسن في أداء المقطع بنسبة ٠,٠٠٠٣ وفي أداء الموسم بمجموعة الوسوم الرئيسة بنسبة ٠,٠٠٠٧، وفي مجموعة الوسوم الفرعية بنسبة ٠,٠٠٣٣ وفي مجموعة الوسوم الموسعة بنسبة ٠,٠٠٣٧. هذا التحسن الطفيف قد يزيد بزيادة حجم التدعيم لمدونة التدريب.

إن وجود الأخطاء في نظام التوسيم النحوي الآلي لا يعني أنه سيء، فحجم الاختلافات بين العقول البشرية النحوية كبير فكيف بالآلة، وعلى الرغم من ذلك، فإن دقة النظام التوسيمي النحوي الآلي لم تتأثر، حتى بعد تطبيق هذا النظام على عينة أخرى.

ثانيا: التوصيات والدراسات المستقبلية المقترحة:

١- تحسين دقة النظام الآلي المقترح عن طريق تطبيقه على عدة نصوص عربية متنوعة الأوعية والمجالات والأزمنة. ومن المهم أن يكون هذا السعي مرتبطا بالإجراء المنهجي الآلي الذي اقترح وصمم وطُبق في هذا الكتاب.

٢- تتبع الوسوم التي يظهر النظام خطأها في مراحل الاختبار المقترحة في التوصية الثانية، والعمل على تحسين مجموعة الوسوم الفرعية وفق الحالات النحوية على الرغم من صحتها في استعمالات لغوية أخرى، فعلى سبيل المثال: كلمة (الحي) في قوله تعالى: (الله لا إله إلا هو الحي القيوم) قد وسمها النظام الآلي على أنها اسم معنى مع أنها اسم ذات، وهو أمر طبيعي، لأن ورودها لوحدها قد يُكسبها سمة اسم المعنى لا سمة اسم الذات، أما ورودها في النص القرآني فقد أكسبها السمة الثانية.

٣- تقييم حالات الوسوم الخاطئة ذات النسب المنخفضة جدا بشكل تجريبي موسّع في دراسات مستقبلة، ويتطلب هذا النوع من الدراسات تقييمات ذات مراحل متعددة ومتتابعة لا تُسعِف بها دراسة واحدة. فعلى سبيل المثال: أظهر الموسم وسم الكلمة (شيئا) على أنها فعل ماض في: شيئا فشيئا، والصحيح هو أنها اسم معنى، والجملة الفعلية: وَسِع كرسيّه على أنها اسم + أداة، والصحيح هي أنها فعل ماضي + اسم ذات، والسبب في ذلك قد يعود إلى صغر حجم المدونة. غير أن الاحتمالات الانتقالية بين الوسوم الموسعة قد صححت الخطأ (شيئا) في المثال السابق، وجعلتها مصدرية، ولم تصحح الخطأ في (وَسِع) في كل مجموعات الوسوم الثلاثة بسبب قد يكون عائدا إلى عدم تداركه توالي الأفعال في العربية أحيانا؛ لأن ما قبل (وسع) هو الفعل (شَاء).

٤- التوسع في الدراسات المستقبلة بزيادة حجم مدونة النظام وإضافة خصائص تصريفية لمجموعات الوسوم الأساسية.

٥- التـوسع في الدراسات المستقبلية بتطوير نظام آلي للموسمات التركيبية syntactic taggers، والموسمات الدلالية semantic taggers، والموسمات الإحالية cataphor taggers.

المراجع العربية

الثبيتي، عبد المحسن، وآخرون. **طريقة تعتمد على المدونات اللغوية لتجهيز بيانات تدريب واختبار أنظمة الوسوم النحوية**. في المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية في دورته الثامنة، القاهرة: مصر، ٢٠١٢

الثبيتي، عبد المحسن. **تصميم المدونات اللغوية وبناؤها**. في: المدونات اللغوية العربية، بناؤها وطرق الإفادة منها، تحرير: صالح العصيمي. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة ١٤٨-١٧٨

جمال الدين، مصطفى. **رأي في تقسيم الكلمة**. مجلة تراثنا، العدد ٦، مؤسسة آل البيت، قم: إيران، ١٩٨٦

الحاج، يحيى؛ وآخرون. **إعداد وتجهيز نظام إحصائي للتعرف الآلي على المفردات القرآنية: الخصائص والسمات الصرف ـ نحوية وآلية مستحدثة لوسمها**. الدورة التاسعة للمؤتمر الدولي لعلوم وهندسة الحاسوب، الحمامات: تونس، ٢٠١٣

حبش، نزار. **مقدمة في المعالجة الطبيعية للغة العربية**، ترجمة: هند الخليفة، جامعة الملك سعود، الرياض: السعودية، ط١، ٢٠١٤

حسان، تمام. **اللغة العربية، معناها ومبناها**. الدار البيضاء: المغرب، دار الثقافة، ١٩٩٤

الزجاجي، أبو القاسم. **اللامات في اللغة**. تحقيق: مازن المبارك، دار الفكر، دمشق: سوريا، ط٢، ١٩٨٥

الزهيري، نبيل. **قاموس مصطلحات المعلوماتية واللغويات الحسابية**. ناشرون، بيروت: لبنان، ٢٠٠٣

الساقي، فاضل. **أقسام الكلام العربي من حيث الشكل والوظيفة**. القاهرة: مصر، مكتبة الخانجي، ١٩٧٧

طباع، نبيل؛ شيخو معمو، محمد. **دليل شعاع لمصطلحات الحاسب**. حلب: سوريا، ٢٠٠٤

عبد الواحد، عبد الحميد. **الكلمة في اللسانيات الحديثة**. كلية الآداب والعلوم الإنسانية وحدة بحث اللسانيات والنظم المعرفية المتصلة بها، صفاقس: تونس، ٢٠٠٧

ابن عقيل، عبد الله بن عبد الرحمن العقيلي الهمداني المصري. **شرح ابن عقيل على ألفية ابن مالك**. دار التراث، القاهرة: مصر، ١٩٨٠

ماكنري، توني؛ هاردي، أندرو. **لغويات المدونة الحاسوبية، المنهج والنظرية والتطبيق**. ترجمة: سلطان المجيول، جامعة الملك سعود، السعودية: الرياض، ٢٠١٦

المالكي، أبو محمد بدر الدين حسن بن قاسم بن عبد الله بن عليّ. **الجنى الداني في حروف المعاني**. تحقيق: د فخر الدين قباوة ـ الأستاذ محمد نديم فاضل، دار الكتب العلمية، بيروت: لبنان، ط١، ١٩٩٢

المجيول، سلطان. **البحث اللغوي الآلي في المدونة الحاسوبية واللغة العربية**. التواصل اللساني، م.١٩، ع. ١-٢، فاس: المغرب، ٢٠١٨، ص. ٥٥-٨٢

المجيول، سلطان. **مناهج التهيئة المعجمية في تعليم العربية لغير الناطقين بها**. المؤتمر الدولي الثاني اتجاهات حديثة في تعليم العربية لغة ثانية، معهد اللغويات العربية بجامعة الملك سعود، مجلد٢، الرياض: المملكة العربية السعودية، ٢٠١٦، ص. ٦٠١-٦٣٣

المجيول، سلطان. **التخطيط اللغوي والسياسة: مفاهيم شمولية**. حولية كلية الدراسات الإسلامية والعربية بالإسكندرية، م. ٢، ع. ٣١، الإسكندرية: مصر، ٢٠١٥

المملكة العربية السعودية. **نظام حماية حقوق المؤلف بالمملكة العربية السعودية**. مسترجع من: https://www.boe.gov.sa/ViewSystemDetails.aspx?lang=ar&SystemID=16&VersionID=24

المراجع الأجنبية:

Abdul-Mageed, M., Diab, M. , Kubler, S. **ASMA: A System for Automatic Segmentation and Morpho-Syntactic Disambiguation of Modern Standard Arabic**. Proceedings of Recent Advances in Natural Language Processing, Hissar: Bulgaria, 2013

Abuleil, S., Evens, M. **Discovering lexical information by tagging Arabic newspaper text**. In Proceedings of the Workshop on Computational

Approaches to Semitic Languages, Quebec: Canada, 1998

Acedánsk, S., Przepiórkowski, A. **Towards the Adequate Evaluation of Morphosyntactic Tagger**. Proceedings of the 23rd International Conference on Computational Linguistics, Beijing: China, 2010, p. 1-8

Ali, B., Jarray, F. **Genetic approach for Arabic part of speech tagging**. International Journal on Natural Language Computing (IJNLC), Vol. 2, No.3, India: USA, 2013

Aliwy A. **Arabic Morphosyntactic Raw Text Part of Speech Tagging System**. Thesis, University of Warsaw, Poland, 2013

Attia, M.A. **Arabic Tokenization System**. Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague: Czech Republic, 2007

Biber, D, Conrad, S., Repper R. **Corpus Linguistics: Investigating Language Structure and Use**. Cambridge University Press, USA, 2006

Baker, P., Hardie, A., McEnery, A. **A Glossary of Corpus Linguistics**. Edinburgh University Press, Edinburgh: UK., 2006

Albared, M., Omar, N., Ab Aziz, M. J. **Improving Arabic part-of-speech tagging through morphological analysis**. In Intelligent Information and Database Systems, Springer Berlin Heidelberg, 2011

Banko, M., R.C. Moore. **Part of speech tagging in context**. Proceeding of the 20th international conference on Computational Linguistics, Association for Computational Linguistics Morristown, Article No. 556, New Jersey: USA., 2004

Byon, A. S., **Teaching Korean Honorifics**. The Fifth National Conference on Korean Language Education (The Korean Language in America), Penn.: USA, 2000

Brezina, V. **Statistics In Corpus Linguistics (A practical Guide)**. Cambridge University Press, Cambridge: UK, 2018

Btoush MH., Alarabeyyat A., Olab I. **Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition**. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, UK, 2016

David M. W. Powers. **Applications and Explanations of Zipf's Law**. In D. M. W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL, 1998, pp.151-160.

Darwish, K., Abdelali, A., Mubarak, H. **Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging**. In International Conference on Language Resources and Evaluation, Reykjavik: Iceland, 2014

Diab, M., Hacioglu, K., Jurafsky, D. **Automatic tagging of Arabic text: From raw text to base phrase chunks**. In Proceedings of HLT-NAACL, Association for Computational Linguistics, Boston: USA., 2004

Diab, M. **Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking**. In 2nd International Conference on Arabic Language Resources and Tools, Cairo: Egypt, 2009

Dukes K., Habash N. Morphological Annotation of Quranic Arabic. In The Language Resources and Evaluation Conference, Malta, 2010

Duh K., Kirchhoff K. **POS Tagging of Dialectal Arabic: A Minimally Supervised Approach**. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, University of Michigan, USA, 2005

Escartín C. P., Alonso H. M. **Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task**. Procesamiento de Lenguaje Natural, Vol. 54, Jaen: Spain, 2015, p. 29-36

Alfaifi, A., Atwell, E. **Computer-Aided Error Annotation A New Tool for Annotating Arabic Error**. The 8th Saudi Students Conference, London: UK, 2015

Freeman, A. **Brill's POS Tagger and a Morphology Parser for Arabic**. In: Proceedings of ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse: France, 2001

Garside, R. and Leech, G. **Running a grammar factory: the production of syntactically analysed corpora or "treebanks"**. In S. Johansson and A.-B. Stenström (eds.), English Computer Corpora: Selected Papers and Research Guide .Berlin & New York: Mouton de Gruyter, 1991

Garside, R., Leech, G., McEnery, T. **Corpus Annotation**, Routledge, USA, 2013

Habash N., Rambow O., Roth R. **MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization**. In Proc. of the International Conference on Arabic Language Resources and Tools, Cairo: Egypt, 2009

Habash N., Roth R. CATiB: The Columbia Arabic Treebank. Center for Computational Learning Systems, Columbia University, New York: USA, 2009

Habash, N., Owen R. **Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop**. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Michigan: USA, 2005

Hadni M., Ouatik SA., Lachkar A, Meknassi M. **Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text**. International Journal on Natural Language Computing (IJNLC), Vol. 2, No. 6, India, 2013

Elhadj, Y. **Statistical Part-of-Speech Tagger for Traditional Arabic Texts**. Journal of Computer Science, Vol.5, No. 11, Riyadh: KSA, 2009

El-Haj, M., Koulali, R. **KALIMAT a Multipurpose Arabic Corpus**. At the Second Workshop on Arabic Corpus Linguistics (WACL-2), UK, 2013

Elhadj, Y. O., Abdelali, A., Bouziane, R., Ammar, A. H**. Revisiting Arabic**

Part of Speech Tagsets. In Computer Systems and Applications (AICCSA), IEEE/ACS 11th International Conference on IEEE, Doha: Qatar, 2014

El-Haj, M., Rayson, P., Piao, S. and Wattam, S. **Creating and validating multilingual semantic representations for six languages: expert versus non-expert crowds**. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Valencia: Spain, 2017

Hajic, J., Smrz, O., Buckwalter, T., & Jin, H. **Feature-based tagger of approximations of functional Arabic morphology**. In Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT), Barcelona: Spain, 2005

Hammami S. , Belguith L., Ben Hamadou A.. **Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links**. The International Arab Journal of Information Technology, Vol. 6, No. 5, Jordan: Amman, 2009

Hasi, H., Nasun-Urt, N. **The Design and Application of Mongolian Homograph Words Information Dictionary**. IEEE - Institute of Electrical and Electronics Engineers, Inc., Tianjin: China, 2012

Hoey M. **What's in a word?**. English Teaching Professional, Issue 27, Hove: UK, 2003

Kanaan K., Al-Shalabi R., Sawalha M. **Full automatic Arabic text tagging system**. The proceedings of the International Conference on Information Technology and Natural Sciences, Jordan, 2003

Kang N., Yu Q. **Corpus-based Stylistic Analysis of Tourism English.** Journal of Language Teaching and Research, Vol. 2, No. 1, Finland, 2011

El-Kareh, S., Al-Ansary, S. **An Arabic Interactive Multi-feature POS Tagger**. In Proceedings of the, ACIDCA conference, Monastir: Tunisia, 2000

Fries, Charles Carpenter. **The Structure of English**. Harcourt Brace, New York: USA, 1952

Johnson, M. **How relevant is linguistics to computational linguistics?**. Linguistic Issues in Language Technology - LiLT, USA, Vol.6, No.7, 2011, p. 1-23

Kaszubski, P., Wojnowska A. **Corpus-informed exercises for learners of English: The TestBuilder program**. In E. Oleksy & B. Lewandowska-Tomaszczyk (Eds.), Research and scholarship in integration processes, Poland–USA–EU, 2003

Khoja, Sh. **APT: Arabic part-of-speech tagger**. In Proceeding of the Student Workshop at the 2nd Meeting of the NAACL, (NAACL'01), USA, 2001

Khoja, Sh., Garside R., Knowles G. A tagset for the morphosyntactic tagging of Arabic. In Proceedings of Corpus Linguistics Conference, UK, 2001

Kjellmer, G. **'The lesser man': observations on the role of women in modern English writings**. In J. Aarts and W. Meijs (eds) Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora, Rodopi; Amsterdam, 1986

Kukulska-Hulme, Agnes. **Language and Communication: Essential Concepts for User Interface and Documentation designed**. Oxford University Press, Oxford: UK, 1999

Kulick, S. Gabbard, R., Marcus, M. **Parsing the Arabic Treebank: Analysis and Improvements**. Proceedings of the Treebanks and Linguistic Theories Conference, Prague: Czech Republic, 2006

Kulick, S. **Exploiting separation of closed-class categories for Arabic tokenization and part-of-speech tagging**. ACM Transactions on Asian Language Information Processing (TALIP), Vol. 10, No. 1, 2011

Maamouri M., Cieri C. **Resources for Natural Language Processing at the Linguistic Data Consortium**. In Proceedings of the International Symposium on Processing of Arabic, Manouba: Tunisia, 2002

Maamouri M., Bies A., Buckwalter T., Mekki W. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo: Egypt, 2004

Mahafdah R.; Omar N.; Al-Omari O. **Arabic part of speech tagging using K-Nearest Neighbour and Naive Bayes classifiers combination**. Journal of Computer Science, New York: USA, Vol. 10, No. 10, 2014, p. 1865-1873

Manning, C. D. **Part-of-speech tagging from 97% to 100%: is it time for some linguistics?**. In Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 2011

McEnery, T, Xiao, R., Tono, Y. **Corpus-Based Language Studies**, Routledge, USA, 2006

McEnery, T. Wilson, A. **Corpus Linguistics (An Introduction)**. Edinburgh University press, Edinburgh: UK, 2011

Mohammad B., Abdulsalam A., Isa B. **Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition**. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, UK, 2016

Mohamed, E., Kübler, S. **Arabic Part of Speech Tagging**. Journal Natural Language Engineering, Vol. 18, No. 4, New York: USA, 2011

Nahar, Kh., Al-Muhtaseb, H., Al-Khatib, W., Elshafei, M., Alghamdi, M. **Arabic Phonemes Transcription using Data Driven Approach**. The International Arab Journal of Information Technology, Vol. 12, No. 3, Jordan: Amman, 2015

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Roth, R. **Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic**. In Proceedings of the

9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014

Plag, I., **Word-formation in English**. Cambridge University, Cambridge: UK, 2002, pp.13-14

Powers, David M W. **Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation.** Journal of Machine Learning Technologies, Australia, Vol. 2, No. 1, 2011, p. 37–63

Al Qady, M., Kandil, A. **Concept Relation Extraction from construction Documents Using Natural Language Processing**. Journal of Construction Engineering & Management, Vol. 136, No. 3, USA, 2010

Alqrainy, S., AlSerhan, H. M., Ayesh, A. **Pattern-based algorithm for Part-of-Speech tagging Arabic text**. Computer Engineering & Systems ICCES International Conference on IEEE, Cairo: Egypt, 2008

Alrabiah, M. **Building A Distributional Semantic Model for Traditional Arabic and Investigating its Novel Applications to The Holy Qur'an**. (Unpublished doctoral thesis), King Saud University, Riyadh: KSA, 2015

Roth, R., Rambow, O., Habash, N., Diab, M., Rudin, C. **Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking**. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, USA, 2008

Rousseau, R., **Goerge Kingsley Zipf: life, ideas, his law and informetrics**. Glottometrics, Germany, Vol. 3, 2002

Sampson, G. **The grammatical database and parsing scheme**. In Garside, R., Leech, G., and Sampson, G. (eds.), The Computational Analysis of English: A Corpus-Based Approach. London: Longman, 1987

Sawalha M. **Open-source Resources and Standards for Arabic Word**

Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora TAGGING. Leeds Uni., UK., 2011

Sawalha, M., Brierley C., Atwell E**. Tools for Arabic Natural Language Processing: a case study in qalqalah prosody**. In 9th International Conference on Language Resources and Evaluation, Reykjavik; Iceland, 2014

Sawalha, M. Brierley C., Atwell E. **Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning**. Proceedings of LRE-Rel'2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, Reykjavik; Iceland, 2014

Al Shamsi, F., and Guessoum A. **A hidden Markov model-based POS tagger for Arabic**. Des Journées internationales d'Analyse statistique des Données Textuelles, Besançon; France, 2006

Sherkawi L., Ghneim N., Al Dakkak O. **Arabic Speech Act Recognition Using Bootstrapped Rule Based System**. International Journal on Computer and Communications Networks, Computational Intelligence and Data Analytics, Vol. 1, No. 1, Rome: Italy, 2017

Al-Taani, A., Al-Rub, S. A. **A Rule-Based Approach for Tagging Non-Vocalized Arabic Words**. International Arab Journal of Information Technology (IAJIT), Vol. 6, No. 3, Amman: Jordan, 2009

Tlili-Guiassa, Y. **Hybrid method for tagging Arabic text**. Journal of Computer Science, Vol. 2, No. 3, NewYork: USA, 2006

Teuber, W. Cermakova, A. **Corpus Linguistics: A short introduction**, Continuum; UK, 2008

Van den Bosch, A., Marsi, E., Soudi, A. **Memory-based morphological analysis and part-of-speech tagging of Arabic**. In Arabic Computational Morphology, Springer Netherlands, 2007

Yonatan B., Nizar H., Kilgarriff A., Ordan N., Roth R., Suchomel V. **ar-**

TenTen: a new, vast corpus for Arabic. In Proceedings Of WACL'2 Workshop on Arabic Corpus Linguistics. Lancaster: UK, 2013

Zribi, C., Aroua, T., Ahmed, M. **A Multi-Agent System for POS-Tagging Vocalized Arabic Texts**. Int. Arab J. Inf. Technol. 4.4, 2007

المواقع الإلكترونية:

جامعة الملك سعود. **مدونة الذخيرة الفصحى**. ٢٣-٣-٢٠١٦:
http://ksucorpus.ksu.edu.sa/

القنيعير، فارس. **تعلم الآلة: مقدمة سريعة**. مقال في شبكة الإنترنت (موقع نمذجيات)، ١-١٠-٢٠١٧: https://www.nmthgiat.com/ تعلم ـ الآلة ـ مقدمة ـ سريعة/

مدينة الملك عبد العزيز للعلوم والتقنية. **المدونة العربية**. ٢-٩-٢٠١٦:
http://corpus.kacst.edu.sa/index.jsp

المملكة العربية السعودية. **نظام حماية حقوق المؤلف بالمملكة العربية السعودية**. ١٨-١٠-٢٠١٨:
https://www.boe.gov.sa/ViewSystemDetails.aspx?lang=ar&SystemID=16&VersionID=24

BNC. **How the BNC was created**. 10-9-2017:
http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=creation

BoE. **The Bank of English User Guide**. 20-12-2017:
http://titania.bham.ac.uk/docs/svenguide.html

Dukes K. **Quranic Arabic Corpus**. 9-6-2016:
http://corpus.quran.com/

Longman. **Longman Dictionary of Contemporary English**. 16-1-2019:
https://www.ldoceonline.com/dictionary/quantifier

Oxford University Press. **OED-Abbreviations**. 2-9-2017:
https://public.oed.com/how-to-use-the-oed/abbreviations/

Ramachandran, A. **NLP Guide: Identifying Part of Speech Tags using Conditional Random Fields.** Analytics Vidhya, 9-12-2018:
https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31

Sakhr. **Arabic corpora**. 10-9-2017:
http://www.sakhr.com/index.php/en/technology/arabic-resources

Sketch Engine. **English part-of-speech tagsets**. 2-1-2019:
https://www.sketchengine.eu/tagsets/english-part-of-speech-tagset/

Sketch Engine. **KSUCCA: King Saud University Corpus of Classical Arabic**. 2-8-2017:
https://www.sketchengine.co.uk/corpus-of-classical-arabic-ksucca/

Sketch Engine. **Arabic MADA system tagset**. 10-9-2017:
https://www.sketchengine.co.uk/arabic-mada-system-tagset/

Sketch Engine. **Arabic corpus (arWaC)**. 14-10-2017:
https://www.sketchengine.co.uk/arabic-web-corpus-wac/

Sketch Engine. **POS tag set for Modern Standard Arabic** .14-10-2017:
sketchengine.co.uk/pos-tag-set-for-modern-standard-arabic/

Sketch Engine. **Non open source**. 7-6-2018:
https://the.sketchengine.co.uk/corpus/wordlist?corpname=preloaded/artenten12_stanford;wlmaxitems=1000;wlattr=doc.urldomain;wlminfreq=1;include_nonwords=1;wlsort=f;wlnums=docf

The Stanford Natural Language Processing Group. **What POS tag set does the parser use?**. 23-3-2016**:**
http://nlp.stanford.edu/software/parser-arabic-faq.shtml#d

The Stanford Natural Language Processing Group. **Stanford CoreNLP**. 10-12-2017:
http://nlp.stanford.edu/software/corenlp.shtml

The Stanford Natural Language Processing Group. **Arabic Natural Language Processing**, 8-12-2018:
http://nlp.stanford.edu/software/corenlp.shtml

Steven Bird S., Klein E., Loper E. **Categorizing and Tagging Words**. 11-12-2018:
https://www.nltk.org/book/ch05.html

ملحق الكتاب

قائمة الوسوم الموسعة بالخصائص التصريفية المستعملة في المدونة مع تكراراتها النسبية المئوية

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١	RP	أداة	27.69
٢	PUNC	ترقيم	8.48
٣	NC_ISM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر	4.32
٤	NA_ISM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر	4.04
٥	NA_DSM	اسم_معنى_معرف^بأل ـ مفرد ـ مذكر	3.76
٦	NI_ISM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر	3.04
٧	VS_A3SM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر	2.32
٨	NC_DSM	اسم_ذات_معرف^بأل ـ مفرد ـ مذكر	2.14
٩	VP_A3SM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر	1.96
١٠	NA_ISF	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث	1.91
١١	NA_DSF	اسم_معنى_معرف^بأل ـ مفرد ـ مؤنث	1.71
١٢	RP_R3RSRM	أداة_رابط^إحالي^غائب ـ مفرد ـ مذكر	1.40
١٣	NA_DLF	اسم_معنى_معرف^بأل ـ جمع ـ مؤنث	1.29
١٤	PR_SM	ضمير_موصول_مفرد ـ مذكر	1.06

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٥	NA_ILF	صفة_فاعل_غير^معرف^بأل ـ مفرد ـ مذكر	0.98
١٦	AS_ISM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث	0.98
١٧	NC_DSF	اسم_ذات_معرف^بأل ـ مفرد ـ مؤنث	0.97
١٨	AS_DSM	صفة_فاعل_معرف^بأل ـ مفرد ـ مذكر	0.97
١٩	NC_ISF	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مؤنث	0.96
٢٠	PD_SM	ضمير_إشارة_مفرد ـ مذكر	0.85
٢١	DIGIT	رقم	0.84
٢٢	NC_DLF	اسم_ذات_معرف^بأل ـ جمع ـ مؤنث	0.78
٢٣	AA_ISM	أداة_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.65
٢٤	AS_DLM	صفة_فاعل_معرف^بأل ـ جمع ـ مذكر	0.65
٢٥	RP_R3RSRF	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر	0.65
٢٦	VP_A3SF	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مؤنث	0.65
٢٧	NC_DSM_AT	اسم_ذات_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.61
٢٨	NA_ISM_R3RSRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.59
٢٩	AA_DSM	صفة_مشبهة_معرف^بأل ـ مفرد ـ مذكر	0.57
٣٠	VS_A3SF	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مؤنث	0.56
٣١	NA_DSF_AT	اسم_معنى_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.54
٣٢	AS_DSF	صفة_فاعل_معرف^بأل ـ مفرد ـ مؤنث	0.53
٣٣	AS_ISF	صفة_فاعل_غير^معرف^بأل ـ مفرد ـ مؤنث	0.51
٣٤	AO_DSM	صفة_مفعول_معرف^بأل ـ مفرد ـ مذكر	0.47

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٥	NC_DSF_AT	اسم_ذات_معرف^بأل_مفرد_مؤنث_منسوب	0.46
٣٦	NC_ILF	ضمير_إشارة_مفرد_مؤنث	0.43
٣٧	PD_SF	اسم_ذات_غير^معرف^بأل_جمع_مؤنث	0.43
٣٨	AO_ISM	صفة_مفعول_غير^معرف^بأل_مفرد_مذكر	0.42
٣٩	NI_DSM	اسم_مبهم_معرف^بأل_مفرد_مذكر	0.38
٤٠	AC_ISM	صفة_تفضيل_غير^معرف^بأل_مفرد_مذكر	0.35
٤١	NL_ISM	اسم_مكان_غير^معرف^بأل_مفرد_مذكر	0.35
٤٢	VP_A3LF	فعل_مضارع_معلوم_غائب_جمع_مؤنث	0.34
٤٣	VP_A1LM	فعل_مضارع_معلوم_متكلم_جمع_مذكر	0.32
٤٤	PP_3SM	ضمير_شخصي_غائب_مفرد_مذكر	0.31
٤٥	NA_DSM_AT	اسم_معنى_معرف^بأل_مفرد_مذكر_منسوب	0.29
٤٦	VP_A3LM	فعل_مضارع_معلوم_غائب_جمع_مذكر	0.29
٤٧	NC_ISM_22	اسم_ذات_غير^معرف^بأل_مفرد_مذكر_٢٢	0.29
٤٨	NC_ISM_12	اسم_ذات_غير^معرف^بأل_مفرد_مذكر_١٢	0.29
٤٩	PR_SF	ضمير_موصول_مفرد_مؤنث	0.28
٥٠	VS_A3LM	فعل_ماض_معلوم_غائب_جمع_مذكر	0.28
٥١	NA_ISM_R3RSRF	اسم_معنى_غير^معرف^بأل_مفرد_مذكر_رابط^إحالي^غائب_مفرد_مؤنث	0.26
٥٢	NI_ISF	اسم_مبهم_غير^معرف^بأل_مفرد_مؤنث	0.26
٥٣	VP_P3SM	فعل_مضارع_مجهول_غائب_مفرد_مذكر	0.26
٥٤	VS_P3SM	فعل_ماض_مجهول_غائب_مفرد_مذكر	0.23

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٥٥	AC_DSM	صفة_تفضيل_معرف^بأل ـ مفرد ـ مذكر	0.23
٥٦	NI_ISM_ R3RSRM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.22
٥٧	AA_ISF	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مؤنث	0.22
٥٨	AS_DLF	صفة_فاعل_معرف^بأل ـ جمع ـ مؤنث	0.22
٥٩	AO_DSF	صفة_مفعول_معرف^بأل ـ مفرد ـ مؤنث	0.21
٦٠	NC_ISM_ R3RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.21
٦١	ABBREV	اختصار	0.20
٦٢	VS_A1SM	فعل_ماض_معلوم ـ متكلم ـ مفرد ـ مذكر	0.20
٦٣	VP_A1SM	فعل_مضارع_معلوم ـ متكلم ـ مفرد ـ مذكر	0.20
٦٤	NA_ISF_ R3RSRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.20
٦٥	NA_ISF_AT	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.19
٦٦	PR_LF	ضمير_موصول_جمع ـ مؤنث	0.19
٦٧	FOREIGN	أجنبي	0.19
٦٨	NI_ISM_ R3RSRF	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.19
٦٩	AA_DSF	صفة_مشبهة_معرف^بأل ـ مفرد ـ مؤنث	0.19
٧٠	AA_DLM	صفة_مشبهة_معرف^بأل ـ جمع ـ مذكر	0.18
٧١	NV_DSM	اسم_جنس_معرف^بأل ـ مفرد ـ مذكر	0.18
٧٢	NV_DLM	اسم_جنس_معرف^بأل ـ جمع ـ مذكر	0.18

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٧٣	NL_DSM	اسم_مكان_معرف^بأل ـ مفرد ـ مذكر	0.17
٧٤	NV_ISM	اسم_جنس_غير معرف^بأل ـ مفرد ـ مذكر	0.17
٧٥	NC_DLM	اسم_ذات_معرف^بأل ـ جمع ـ مذكر	0.16
٧٦	AO_ISF	صفة_مفعول_غير^معرف^بأل ـ مفرد ـ مؤنث	0.16
٧٧	DT	ظرف_زمان	0.15
٧٨	VS_A3SM_ R3RSRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.15
٧٩	NI_DSF	اسم_مبهم_معرف^بأل ـ مفرد ـ مؤنث	0.15
٨٠	NL_DLF	اسم_مكان_معرف^بأل ـ جمع ـ مؤنث	0.15
٨١	VS_A3LF	فعل_ماض_معلوم ـ غائب ـ جمع ـ مؤنث	0.15
٨٢	AO_DLF	صفة_مفعول_معرف^بأل ـ جمع ـ مؤنث	0.14
٨٣	DL	ظرف_مكان	0.14
٨٤	VP_A3SM_ R3RSRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.14
٨٥	VS_A1LM	فعل_ماض_معلوم ـ متكلم ـ جمع ـ مذكر	0.13
٨٦	NA_ISM_ R3RLRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.13
٨٧	NV_DLF	اسم_جنس_معرف^بأل ـ جمع ـ مؤنث	0.12
٨٨	NL_ILF	اسم_مكان_غير^معرف^بأل ـ جمع ـ مؤنث	0.12
٨٩	RP_R3RLRM	أداة_رابط^إحالي^غائب ـ جمع ـ مذكر	0.12
٩٠	VS_P3SF	فعل_ماض_مجهول ـ غائب ـ مفرد ـ مؤنث	0.12

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٩١	PP_3SF	ضمير_شخصي_غائب ـ مفرد ـ مؤنث	0.12
٩٢	AS_ILM	صفة_فاعل_غير^معرف^بأل ـ جمع ـ مذكر	0.12
٩٣	AO_ILF	صفة_مفعول_غير^معرف^بأل ـ جمع ـ مؤنث	0.12
٩٤	AO_DLM	صفة_مفعول_معرف^بأل ـ جمع ـ مذكر	0.11
٩٥	AE_ISM	صفة_مبالغة_غير^معرف^بأل ـ مفرد ـ مذكر	0.10
٩٦	PP_3LF	ضمير_شخصي_غائب ـ جمع ـ مؤنث	0.10
٩٧	SYMB	رمز	0.10
٩٨	NA_ILF_R3RSRF	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.10
٩٩	NC_ISM_AT	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_منسوب	0.10
١٠٠	RP_R1RLRM	أداة_رابط^إحالي^متكلم ـ جمع ـ مذكر	0.10
١٠١	RP_R1RSRM	أداة_رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.10
١٠٢	VP_A2SM	فعل_مضارع_معلوم ـ مخاطب ـ مفرد ـ مذكر	0.10
١٠٣	VC_2SM	فعل_أمر_مخاطب ـ مفرد ـ مذكر	0.10
١٠٤	AC_DSF	صفة_تفضيل_معرف^بأل ـ مفرد ـ مؤنث	0.10
١٠٥	NA_ILF_R3RSRM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.10
١٠٦	NV_DSF	اسم_جنس_غير^معرف^بأل ـ جمع ـ مؤنث	0.10
١٠٧	NV_ILF	اسم_جنس_معرف^بأل ـ مفرد ـ مؤنث	0.10
١٠٨	VS_A2SM	فعل_ماض_معلوم ـ مخاطب ـ مفرد ـ مذكر	0.10
١٠٩	PR_LM	ضمير_موصول_جمع ـ مذكر	0.09

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١١٠	NA_ISM_AT	اسم_معنى_غير^معرف^بأل_مفرد_مذكر_منسوب	0.09
١١١	NC_ILF_ R3RLRM	اسم_ذات_غير^معرف^بأل _ جمع _ مؤنث_ رابط^إحالي^غائب_جمع_مذكر	0.08
١١٢	VP_P3SF	فعل_مضارع_مجهول_غائب_مفرد_مؤنث	0.08
١١٣	NA_ISF_ R3RSRF	اسم_معنى_غير^معرف^بأل _ مفرد _ مؤنث_ رابط^إحالي^غائب_مفرد_مؤنث	0.08
١١٤	NC_ISM_ R3RSRF	اسم_ذات_غير^معرف^بأل _ مفرد _ مذكر_ رابط^إحالي^غائب_مفرد_مؤنث	0.08
١١٥	NV_DSF_AT	اسم_جنس_معرف^بأل_مفرد_مؤنث_منسوب	0.07
١١٦	NI_ILF	اسم_مبهم_غير^معرف^بأل_جمع_مؤنث	0.07
١١٧	PP_1SM	ضمير_شخصي_متكلم_مفرد_مذكر	0.07
١١٨	NC_ISF_AT	اسم_ذات_غير^معرف^بأل_مفرد_مؤنث_منسوب	0.07
١١٩	VP_P3LF	فعل_مضارع_مجهول_غائب_جمع_مؤنث	0.07
١٢٠	NL_DSF	اسم_مكان_معرف^بأل_مفرد_مؤنث	0.07
١٢١	NV_ISF	اسم_جنس_غير^معرف^بأل_مفرد_مؤنث	0.07
١٢٢	VP_A3SF_ R3RSRM	فعل_مضارع_معلوم _ غائب _ مفرد _ مؤنث_ رابط^إحالي^غائب_مفرد_مذكر	0.07
١٢٣	AS_DSF_AT	صفة_فاعل_معرف^بأل_مفرد_مؤنث_منسوب	0.06
١٢٤	AC_ISF	صفة_تفضيل_غير^معرف^بأل_مفرد_مؤنث	0.06
١٢٥	NA_ILF_ R3RLRM	اسم_معنى_غير^معرف^بأل _ جمع _ مؤنث_ رابط^إحالي^غائب_جمع_مذكر	0.06
١٢٦	PD	ضمير_إشارة	0.06

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٢٧	NC_DLF_AT	اسم_ذات_معرف^بأل ـ جمع ـ مؤنث_منسوب	0.06
١٢٨	NC_ILM	اسم_ذات_معرف^بأل ـ مثنى ـ مذكر	0.05
١٢٩	VS_A1LM_ R3RSRM	فعل_ماض_معلوم ـ متكلم ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.05
١٣٠	NA_ILM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مذكر	0.05
١٣١	VS_A3SM_ R1RLRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.05
١٣٢	VP_A3LM_ R3RSRM	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.05
١٣٣	AS_ILM_ R3RSRM	صفة_فاعل_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.05
١٣٤	VS_A3SM_ R3RSRF	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.05
١٣٥	AS_ILF	صفة_فاعل_غير^معرف^بأل ـ جمع ـ مؤنث	0.05
١٣٦	NI_DLF	اسم_مبهم_معرف^بأل ـ جمع ـ مؤنث	0.05
١٣٧	VS_A3LM_ R3RSRM	فعل_ماض_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.05
١٣٨	NC_DUF	**اسم_ذات_غير^معرف^بأل ـ مثنى ـ مذكر**	0.05
١٣٩	AA_DLF	صفة_مشبهة_معرف^بأل ـ جمع ـ مؤنث	0.04
١٤٠	NC_ISM_ R1RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.04
١٤١	PD_LF	ضمير_إشارة_جمع ـ مؤنث	0.04
١٤٢	NI_DSF_AT	اسم_مبهم_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.04

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٤٣	VS_P3LF	فعل_ماض_مجهول ـ غائب ـ جمع ـ مؤنث	0.04
١٤٤	NC_ILF_ R3RSRM	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.04
١٤٥	PP_3LM	ضمير_شخصي_غائب ـ جمع ـ مذكر	0.04
١٤٦	NV_DSM_AT	اسم_جنس_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.04
١٤٧	NA_ILF_ R1RLRM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.04
١٤٨	NA_DLM_AT	اسم_معنى_معرف^بأل ـ جمع ـ مذكر_منسوب	0.04
١٤٩	NL_ISM_ R3RSRM	اسم_مكان_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.04
١٥٠	AA_ILM	صفة_مشبهة_غير^معرف^بأل ـ جمع ـ مذكر	0.04
١٥١	NA_IUM	اسم_معنى_غير^معرف^بأل ـ مثنى ـ مذكر	0.04
١٥٢	AA_DSF_AT	صفة_مشبهة_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.04
١٥٣	NC_DUM	اسم_ذات_غير^معرف^بأل ـ مثنى ـ مذكر	0.04
١٥٤	RP_R2RSRM	أداة_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.04
١٥٥	NI_ISM_ R3RLRM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.04
١٥٦	NA_ISM_ R1RSRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.04
١٥٧	NC_ILF_ R3RSRF	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.04
١٥٨	NA_DLF_AT	اسم_معنى_معرف^بأل ـ جمع ـ مؤنث_منسوب	0.04

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٥٩	NI_ISM_ R1RSRM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.04
١٦٠	NA_ISF_ R3RLRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.04
١٦١	NC_IUF	اسم_ذات_غير^معرف^بأل ـ مثنى ـ مؤنث	0.04
١٦٢	AS_ISM_ R3RSRM	صفة_فاعل_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.04
١٦٣	VP_A3SM_ R3RSRF	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.04
١٦٤	AE_DLF	صفة_مبالغة_معرف^بأل ـ جمع ـ مؤنث	0.04
١٦٥	NV_ISM_AT	اسم_جنس_غير^معرف^بأل ـ مفرد ـ مذكر_منسوب	0.04
١٦٦	RP_R2RLRM	أداة_رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.04
١٦٧	AC_DLF	صفة_تفضيل_معرف^بأل ـ جمع ـ مؤنث	0.04
١٦٨	NM_ISM	اسم_آلة_غير^معرف^بأل ـ مفرد ـ مذكر	0.03
١٦٩	AS_IUM	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.03
١٧٠	NA_DSM_ R3RSRM	اسم_معنى_معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.03
١٧١	VC_2SM_ R1RSRM	فعل_أمر_مخاطب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.03
١٧٢	NT_ISM	اسم_زمان_غير^معرف^بأل ـ مفرد ـ مذكر	0.03
١٧٣	VS_A3SM_ R1RSRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.03

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٧٤	NM_DLF	اسم_آلة_معرف^بأل ـ جمع ـ مؤنث	0.03
١٧٥	NA_IUF	اسم_معنى_غير^معرف^بأل ـ مثنى ـ مؤنث	0.03
١٧٦	NI_ISM_R3RURM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ مثنى ـ مذكر	0.03
١٧٧	NA_ISM_R3RLRF	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ جمع ـ مؤنث	0.03
١٧٨	IV	خالفة_إخالة	0.03
١٧٩	PP_1LM	ضمير_شخصي_مخاطب ـ جمع ـ مذكر	0.03
١٨٠	NV_DLM_AT	اسم_جنس_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.03
١٨١	NC_DLM_AT	اسم_ذات_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.03
١٨٢	AA_ISM_R1RLRM	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^متكلم ـ جمع ـ مذكر	0.03
١٨٣	VP_A3UM	فعل_مضارع_معلوم ـ غائب ـ مثنى ـ مذكر	0.03
١٨٤	NA_ISM_R1RLRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^متكلم ـ جمع ـ مذكر	0.02
١٨٥	NA_DUM	اسم_معنى_معرف^بأل ـ مثنى ـ مذكر	0.02
١٨٦	NL_ILF_R3RSRF	اسم_مكان_غير^معرف^بأل ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.02
١٨٧	VS_A3SF_R3RSRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.02
١٨٨	VS_A3SF_R3RSRF	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.02
١٨٩	NC_ILM_AT	اسم_ذات_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.02

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
١٩٠	NI_IUM	اسم_مبهم_غير^معرف^بأل_مثنى_مذكر	0.02
١٩١	VS_P3LM	فعل_ماض_مجهول_غائب_جمع_مذكر	0.02
١٩٢	NC_ISF_22	اسم_ذات_غير^معرف^بأل_مفرد_مؤنث_٢٢	0.02
١٩٣	NL_DSF_AT	اسم_مكان_معرف^بأل_مفرد_مؤنث_منسوب	0.02
١٩٤	VS_A1SM_R3RSRM	فعل_ماض_معلوم_متكلم_مفرد_مذكر_رابط^إحالي^غائب_مفرد_مذكر	0.02
١٩٥	NM_DSM	اسم_آلة_معرف^بأل_مفرد_مذكر	0.02
١٩٦	NC_ISF_12	اسم_ذات_غير^معرف^بأل_مفرد_مؤنث_١٢	0.02
١٩٧	NM_DSF	اسم_آلة_معرف^بأل_مفرد_مؤنث	0.02
١٩٨	NM_ILF	اسم_آلة_غير^معرف^بأل_جمع_مؤنث	0.02
١٩٩	AE_DSM	صفة_مبالغة_معرف^بأل_مفرد_مذكر	0.02
٢٠٠	RP_R3RURM	أداة_رابط^إحالي^غائب_مثنى_مذكر	0.02
٢٠١	NA_DLM	اسم_معنى_غير^معرف^بأل_جمع_مذكر	0.02
٢٠٢	VP_A3SM_R3RLRM	فعل_مضارع_معلوم_غائب_مفرد_مذكر_رابط^إحالي^غائب_جمع_مذكر	0.02
٢٠٣	NC_IUF_R3RSRM	اسم_ذات_غير^معرف^بأل_مثنى_مؤنث_رابط^إحالي^غائب_مفرد_مذكر	0.02
٢٠٤	VP_A1SM_R3RSRM	فعل_مضارع_معلوم_متكلم_مفرد_مذكر_رابط^إحالي^غائب_مفرد_مذكر	0.02
٢٠٥	NI_ISM_AT	اسم_مبهم_غير^معرف^بأل_مفرد_مذكر_منسوب	0.02
٢٠٦	PR	ضمير_موصول	0.02
٢٠٧	NA_ILF_R1RSRM	اسم_معنى_غير^معرف^بأل_جمع_مؤنث_رابط^إحالي^متكلم_مفرد_مذكر	0.02

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٠٨	NV_ILM	اسم_جنس_غير^معرف^بأل ـ جمع ـ مذكر	0.02
٢٠٩	AS_DSF_33	صفة_فاعل_معرف^بأل ـ مفرد ـ مؤنث_٣٣	0.02
٢١٠	AS_DSF_23	صفة_فاعل_معرف^بأل ـ مفرد ـ مؤنث_٢٣	0.02
٢١١	NV_DLF_AT	اسم_جنس_معرف^بأل ـ جمع ـ مؤنث_منسوب	0.02
٢١٢	VP_A3SM_R1RSRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.02
٢١٣	NT_DSM	اسم_زمان_معرف^بأل ـ مفرد ـ مذكر	0.02
٢١٤	AC_DLM	أفعل_تفضيل_غير^معرف^بأل ـ جمع ـ مذكر	0.02
٢١٥	AS_DSF_13	صفة_فاعل_معرف^بأل ـ مفرد ـ مؤنث_١٣	0.02
٢١٦	AC_ISM_R3RSRF	أفعل_تفضيل_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.02
٢١٧	VS_A3LM_R3RSRF	فعل_ماض_معلوم ـ غائب ـ جمع ـ مذكر_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.02
٢١٨	NC_ISM_R2RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.02
٢١٩	NA_DSF_R3RSRM	اسم_معنى_معرف^بأل ـ مفرد ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٢٠	VS_A3LF_R2RSRM	فعل_ماض_معلوم ـ غائب ـ جمع ـ مؤنث_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٢١	AO_ISM_R3RSRM	صفة_مفعول_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٢٢	NI_ISF_R3RSRM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٢٣	NI_IUM_ R3RSRM	اسم_مبهم_غير^معرف^بأل ـ مثنى ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٢٤	NC_ILF_ R3RLRF	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٢٢٥	VP_A3LF_ R3RSRM	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٢٦	VS_A3SM_ R3RLRF	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٢٢٧	NC_ILF_ R2RLRM	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٢٢٨	NI_ISM_ R1RLRM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٢٢٩	VP_A1LM_ R3RSRM	فعل_مضارع_معلوم ـ مخاطب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٣٠	VP_A3LM_ R1RLRM	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٢٣١	NC_ISM_33	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_٣٣	0.01
٢٣٢	AS_DUM	صفة_فاعل_معرف^بأل ـ مثنى ـ مذكر	0.01
٢٣٣	AS_DUF	صفة_فاعل_معرف^بأل ـ مثنى ـ مؤنث	0.01
٢٣٤	AO_IUF	صفة_مفعول_غير^معرف^بأل ـ مثنى ـ مؤنث	0.01
٢٣٥	NC_ISM_23	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_٢٣	0.01
٢٣٦	NC_ILM_ R3RLRM	اسم_ذات_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٣٧	NA_ILF_ R2RLRM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٢٣٨	AE_ILM	صفة_مبالغة_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٢٣٩	NC_ISM_13	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_١٣	0.01
٢٤٠	AE_ILF	صفة_مبالغة_غير^معرف^بأل ـ جمع ـ مؤنث	0.01
٢٤١	VP_A3SF_ R3RSRF	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٤٢	VP_A3SM_ R2RLRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٢٤٣	VS_A3UM	فعل_ماض_معلوم ـ غائب ـ مثنى ـ مذكر	0.01
٢٤٤	NL_ISM_ R3RSRF	اسم_مكان_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٤٥	VP_A3SM_ R1RLRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٢٤٦	NV_ISF_AT	اسم_جنس_غير^معرف^بأل_مفرد_مؤنث_منسوب	0.01
٢٤٧	NI_ILM	اسم_مبهم_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٢٤٨	VP_A1SM_ R2RSRM	فعل_مضارع_معلوم ـ متكلم ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٤٩	NV_DUM	اسم_جنس_معرف^بأل ـ مثنى ـ مذكر	0.01
٢٥٠	AC_ISM_ R3RLRM	أفعل_تفضيل_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٢٥١	AO_DUM	صفة_مفعول_معرف^بأل ـ مثنى ـ مذكر	0.01
٢٥٢	NC_ISM_ R3RLRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٥٣	NC_ISF_ R3RSRF	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٥٤	NL_ISM_ R1RSRM	اسم_مكان_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٢٥٥	NA_ILF_AT	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_منسوب	0.01
٢٥٦	AE_DLM	صفة_مبالغة_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٢٥٧	NC_IUM_ R3RSRM	اسم_ذات_غير^معرف^بأل ـ مثنى ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٥٨	AA_ILF	صفة_مشبهة_غير^معرف^بأل ـ جمع ـ مؤنث	0.01
٢٥٩	NA_DLF_ R3RSRF	اسم_معنى_معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٦٠	AS_ISF_ R3RSRF	صفة_فاعل_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٦١	NC_ISF_ R1RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٢٦٢	AA_ILF_ R3RSRM	صفة_مشبهة_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٦٣	VS_A3LF_ R1RLRM	فعل_ماض_معلوم ـ غائب ـ جمع ـ مؤنث_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٢٦٤	AS_DSM_22	صفة_فاعل_معرف^بأل ـ مفرد ـ مذكر_٢٢	0.01
٢٦٥	NV_ISM_ R3RSRM	اسم_جنس_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٦٦	AS_DSM_12	صفة_فاعل_معرف^بأل ـ مفرد ـ مذكر_١٢	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٦٧	VP_A3LF_ R3RLRM	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٢٦٨	AE_DSF_AT	صفة_مبالغة_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
٢٦٩	NA_DSM_ R3RLRM	اسم_معنى_معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٢٧٠	VS_A2SM_ R1RSRM	فعل_ماض_معلوم ـ مخاطب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٢٧١	VS_A3LM_ R1RSRM	فعل_ماض_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٢٧٢	AA_ISM_ R3RSRM	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٧٣	NL_IUM	اسم_مكان_غير^معرف^بأل ـ مثنى ـ مذكر	0.01
٢٧٤	PP_3UM	ضمير_شخصي_غائب ـ مثنى ـ مذكر	0.01
٢٧٥	PP_3UF	ضمير_شخصي_غائب ـ مثنى ـ مؤنث	0.01
٢٧٦	NV_ISM_ R1RSRM	اسم_جنس_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٢٧٧	NA_DUF	اسم_معنى_معرف^بأل ـ مثنى ـ مؤنث	0.01
٢٧٨	VC_2SM_ R1RLRM	فعل_أمر_مخاطب ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٢٧٩	NI_ISF_ R3RSRF	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٨٠	AC_DSM_ R3RLRF	أفعل_تفضيل_معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٨١	AA_ISM_ R2RSRM	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٨٢	VP_A2SM_ R3RSRM	فعل_مضارع_معلوم ـ مخاطب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٨٣	NC_ILF_AT	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_منسوب	0.01
٢٨٤	NT_ISM_ R3RSRM	اسم_زمان_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٨٥	AO_ISF_ R3RSRM	صفة_مفعول_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٢٨٦	AO_ISM_ R2RSRM	صفة_مفعول_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٨٧	NT_DLF	اسم_زمان_معرف^بأل ـ جمع ـ مؤنث	0.01
٢٨٨	NA_ISM_ R2RSRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٨٩	NL_ISF	اسم_مكان_غير^معرف^بأل ـ مفرد ـ مؤنث	0.01
٢٩٠	VP_A1LM_ R2RLRM	فعل_مضارع_معلوم ـ مخاطب ـ جمع ـ مذكر_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٢٩١	VS_A2LM	فعل_ماض_معلوم ـ مخاطب ـ جمع ـ مذكر	0.01
٢٩٢	AE_ISF	صفة_مبالغة_غير^معرف^بأل ـ مفرد ـ مؤنث	0.01
٢٩٣	VP_A3SM_ R3RURM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مثنى ـ مذكر	0.01
٢٩٤	VC_2SM_ R3RSRM	فعل_أمر_مخاطب_مفرد_مذكر_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٢٩٥	VP_A2SM_R2RSRM	فعل_مضارع_معلوم ـ مخاطب ـ مفرد ـ مذكر_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٢٩٦	VP_A3LF_R3RSRF	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٢٩٧	NV_ILM_AT	اسم_جنس_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.01
٢٩٨	VP_A3LM_R3RLRF	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مذكر_رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٢٩٩	VP_P3LF_R3RSRM	فعل_مضارع_مجهول ـ غائب ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٠٠	VP_P3LM_R2RSRM	فعل_مضارع_مجهول ـ غائب ـ جمع ـ مذكر_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٠١	VP_A3SF_R3RLRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مؤنث_رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٠٢	VS_A3SM_R3RLRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٠٣	NI_ISM_33	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_٣٣	0.01
٣٠٤	NI_IUF	اسم_مبهم_غير^معرف^بأل ـ مثنى ـ مؤنث	0.01
٣٠٥	AC_DSF_AT	أفعل_تفضيل_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
٣٠٦	NI_ISM_23	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_٢٣	0.01
٣٠٧	AS_DSM_AT	صفة_فاعل_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣٠٨	VS_A3SF_R2RSRM	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مؤنث_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٠٩	AO_DSF_ R3RSRM	صفة_مفعول_معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣١٠	AS_DLF_22	صفة_فاعل_معرف^بأل ـ جمع ـ مؤنث_٢٢	0.01
٣١١	VP_A1LM_ R3RSRF	فعل_مضارع_معلوم ـ مخاطب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣١٢	NI_ISM_13	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مذكر_١٣	0.01
٣١٣	NA_ISF_ R3RURM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مثنى ـ مذكر	0.01
٣١٤	NC_DLF_ R3RLRM	اسم_ذات_معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣١٥	NI_DSM_AT	اسم_مبهم_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣١٦	AS_DLF_12	صفة_فاعل_معرف^بأل ـ جمع ـ مؤنث_١٢	0.01
٣١٧	NM_ILM	اسم_آلة_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٣١٨	VP_A3SF_ R1RLRM	اسم_زمان_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
٣١٩	PP_2SM	فعل_مضارع_مجهول ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٢٠	NI_ISM_ R2RSRM	صفة_مفعول_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٣٢١	VP_P3SM_ R3RSRM	اسم_ذات_معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٢٢	NI_ISM_ R2RSRF	فعل_مضارع_مجهول ـ غائب ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٢٣	NL_DSM_AT	اسم_مبهم_غير^معرف^بأل - مفرد - مؤنث_ رابط^إحالي^غائب - مفرد - مؤنث	0.01
٣٢٤	VC_2SF	اسم_ذات_غير^معرف^بأل - جمع - مؤنث_ رابط^إحالي^غائب - مثنى - مذكر	0.01
٣٢٥	AA_DLM_AT	فعل_مضارع_معلوم - غائب - جمع - مؤنث_ رابط^إحالي^غائب - مفرد - مؤنث	0.01
٣٢٦	VP_A3LM_ R3RSRF	اسم_جنس_غير^معرف^بأل - مفرد - مذكر_ رابط^إحالي^متكلم - مفرد - مذكر	0.01
٣٢٧	NV_ILF_ R3RSRM	اسم_ذات_غير^معرف^بأل - مفرد - مؤنث_ رابط^إحالي^غائب - مفرد - مذكر	0.01
٣٢٨	IV_R2RSRM	فعل_مضارع_معلوم - غائب - جمع - مذكر_ رابط^إحالي^مخاطب - مفرد - مذكر	0.01
٣٢٩	NC_DLF_22	فعل_مضارع_معلوم - غائب - جمع - مؤنث_ رابط^إحالي^غائب - جمع - مذكر	0.01
٣٣٠	NC_IUF_ R3RSRF	اسم_زمان_معرف^بأل - مفرد - مذكر_منسوب	0.01
٣٣١	VP_A3LM_ R2RSRM	صفة_فاعل_غير^معرف^بأل - جمع - مذكر_ رابط^إحالي^غائب - مثنى - مذكر	0.01
٣٣٢	VS_P1SM	اسم_ذات_غير^معرف^بأل - مفرد - مؤنث_ رابط^إحالي^مخاطب - مفرد - مذكر	0.01
٣٣٣	NI_DUM	اسم_مبهم_غير^معرف^بأل - مفرد - مؤنث_ رابط^إحالي^غائب - مثنى - مؤنث	0.01
٣٣٤	NA_ILM_ R3RSRF	صفة_تفضيل_معرف^بأل - جمع - مؤنث_منسوب	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٣٥	NC_DLF_12	صفة_مفعول_غير^معرف^بأل ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣٣٦	NA_ISM_R3RURM	صفة_تفضيل_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
٣٣٧	AS_ISM_AT	صفة_تفضيل_معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٣٣٨	VP_A2LM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٣٩	AO_ILF_R3RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٣٤٠	AC_ILM	صفة_فاعل_غير^معرف^بأل ـ جمع ـ مذكر_رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٤١	AS_ILM_R1RSRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٤٢	AO_ILF_R3RSRF	اسم_مبهم_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣٤٣	AC_ILF	اسم_معنى_معرف^بأل ـ مثنى ـ مؤنث	0.01
٣٤٤	VP_A3SF_R1RSRM	اسم_آلة_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٣٤٥	NV_ILF_R1RSRM	اسم_زمان_معرف^بأل ـ جمع ـ مؤنث	0.01
٣٤٦	NI_ISF_R3RURF	فعل_أمر_مخاطب ـ جمع ـ مذكر	0.01
٣٤٧	NT_DSF_AT	صفة_مفعول_غير^معرف^بأل ـ جمع ـ مؤنث_رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٤٨	NA_ISF_ R3RLRF	اسم_مبهم_معرف^بأل ـ مثنى ـ مذكر	0.01
٣٤٩	NI_ILF_R3RL-RM	ضمير_شخصي_غائب ـ مثنى ـ مؤنث	0.01
٣٥٠	AA_DUM	فعل_ماض_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٣٥١	VP_A1SM_ R2RSRF	صفة_مشبهة_معرف^بأل ـ مثنى ـ مذكر	0.01
٣٥٢	NC_ILM_ R3RSRM	اسم_ذات_معرف^بأل ـ جمع ـ مؤنث_٢٢	0.01
٣٥٣	NL_ISF_ R3RSRF	فعل_مضارع_معلوم ـ متكلم ـ جمع ـ مذكر_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٣٥٤	NC_ISF_ R3RSRM	أداة_رابط^إحالي^غائب ـ جمع ـ مؤنث	0.01
٣٥٥	NA_DLF_ R1RLRM	اسم_مبهم_معرف^بأل ـ جمع ـ مؤنث_منسوب	0.01
٣٥٦	NC_ISM_ R3RLRF	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٥٧	NC_IUM	ضمير_شخصي_غائب ـ مثنى ـ مذكر	0.01
٣٥٨	VS_A3LM_ R3RLRF	اسم_ذات_معرف^بأل ـ جمع ـ مؤنث_١٢	0.01
٣٥٩	NA_IUF_ R3RSRF	صفة_فاعل_معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣٦٠	NV_ISF_ R3RSRM	اسم_جنس_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٦١	AS_ISM_ R3RSRF	فعل_مضارع_معلوم ـ غائب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣٦٢	NI_DLF_AT	صفة_مبالغة_معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
٣٦٣	RP_R3RLRF	فعل_ماض_معلوم ـ غائب ـ مفرد ـ مؤنث_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٦٤	VC_2LM	اسم_زمان_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٦٥	NT_DSM_AT	اسم_زمان_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣٦٦	NC_ISF_ R2RSRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٦٧	NA_ISF_ R1RLRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٣٦٨	AO_ILM	صفة_مفعول_غير^معرف^بأل ـ جمع ـ مذكر	0.01
٣٦٩	AC_DLF_AT	أفعل_تفضيل_معرف^بأل ـ جمع ـ مؤنث_منسوب	0.01
٣٧٠	NI_ILF_R1RL-RM	اسم_مبهم_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٣٧١	VP_A3SM_ R2RSRM	فعل_مضارع_معلوم ـ غائب ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٧٢	NV_ILM_ R2RSRM	اسم_جنس_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٧٣	AS_DLM_AT	صفة_فاعل_معرف^بأل ـ جمع ـ مذكر_منسوب	0.01
٣٧٤	AS_ISF_ R3RSRM	صفة_فاعل_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٧٥	NA_DSF_ R3RLRM	اسم_معنى_معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٧٦	NC_ILM_ R1RSRM	اسم_ذات_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٧٧	NM_DSM_AT	اسم_آلة_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣٧٨	NC_ISM_ R1RLRM	اسم_ذات_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٣٧٩	AA_ILM_ R3RSRM	صفة_مشبهة_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٨٠	NC_IUM_ R2RSRF	اسم_ذات_غير^معرف^بأل ـ مثنى ـ مذكر_ رابط^إحالي^مخاطب ـ مفرد ـ مذكر	0.01
٣٨١	NA_ILM_AT	اسم_معنى_غير^معرف^بأل ـ جمع ـ مذكر_منسوب	0.01
٣٨٢	NI_ILF_ R3RSRM	اسم_مبهم_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٨٣	NV_DLM_ R3RSRF	اسم_جنس_معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣٨٤	NA_ILF_AT_ R1RLRM	اسم_معنى_غير^معرف^بأل ـ جمع ـ مؤنث_ منسوب_رابط^إحالي^متكلم ـ جمع ـ مذكر	0.01
٣٨٥	NI_ISF_R3RL-RM	اسم_مبهم_غير^معرف^بأل ـ مفرد ـ مؤنث_ رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٨٦	VS_A2SF	فعل_ماض_معلوم ـ مخاطب ـ مفرد ـ مؤنث	0.01
٣٨٧	VP_A2LM_ R3RSRF	فعل_مضارع_معلوم ـ مخاطب ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01

م	الوسم الموسع	القسم الموسع	التكرار النسبي المئوي
٣٨٨	VS_A2SM_ R3RSRM	فعل_ماض ـ معلوم ـ مخاطب ـ مفرد ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مذكر	0.01
٣٨٩	AA_ISM_ R2RLRM	صفة_مشبهة_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٣٩٠	NC_ILF_ R3RURM	اسم_ذات_غير^معرف^بأل ـ جمع ـ مؤنث_ رابط^إحالي^غائب ـ مثنى ـ مذكر	0.01
٣٩١	NA_ISM_ R2RLRM	اسم_معنى_غير^معرف^بأل ـ مفرد ـ مذكر_ رابط^إحالي^مخاطب ـ جمع ـ مذكر	0.01
٣٩٢	AS_DLM_ R3RSRF	صفة_فاعل_معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مفرد ـ مؤنث	0.01
٣٩٣	NC_DLM_ R1RSRM	اسم_ذات_معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٩٤	AA_ILM_ R1RSRM	صفة_مشبهة_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^متكلم ـ مفرد ـ مذكر	0.01
٣٩٥	AS_ILM_ R3RURM	صفة_فاعل_غير^معرف^بأل ـ جمع ـ مذكر_ رابط^إحالي^غائب ـ مثنى ـ مذكر	0.01
٣٩٦	VC_2SM_ R3RLRM	فعل_أمر_مخاطب_مفرد_مذكر_رابط^إحالي^غائب ـ جمع ـ مذكر	0.01
٣٩٧	AA_DSM_AT	صفة_مشبهة_معرف^بأل ـ مفرد ـ مذكر_منسوب	0.01
٣٩٨	NM_ISF_AT	اسم_آلة_غير^معرف^بأل ـ مفرد ـ مؤنث_منسوب	0.01
المجموع		100	