Arapçada Cümlenin Hesaplanması: Markov Modeli'nden Sayısal Dilbilgisi Modellerine Doğru

Soufiane Ouaki
ouaki.soufiane19@gmail.com
https://orcid.org/0000-0003-1758-8261
Dr. Mohammed Merzouk
medmerz@hotmail.fr
https://orcid.org/0000-0003-1758-8261

ÖZ

Bu araştırmamızda, Arapça cümlenin hesaplanması ve değerlendirilmesiyle ilgili olarak bilgisayarlı dilbilim veya hesaplamalı dilbilim alanındaki en önemli hesaplama yaklaşımlarını ve kullandıkları yöntemleri tanıtmayı amaçlamaktayız. Konu ile ilgili bilgisayarı bu görevi yerine getirmek için kullananların hesaplama ile ilgili kavramlar ve yaklaşım-modeller tanıtılacaktır. Ayrıca araştırmamızda cümleyi sayısal ve bilgisayar kullanarak analiz etmeye yardımcı olan dilbilimsel konular tartışılacaktır. Bu araştırmanın kavramlarını ve çıkarımlarını oluştururken benimsenen yöntem, Arapça cümleyi işlemek için kullanılan ve halen kullanılmakta olan betimleyici-analitik yöntemdir. Bu araştırmanın sonuçlara gelince; Bir yanda bilgisayarla hesaplama modellerinin ve karmaşıklığı, diğer yandan onların doğal dile uygun olmayışı söz konusu istatistiksel modellerin dil ile ilgili hesaplama için kısmen de olsa terk edilmesine yol açmıştır. Buna rağmen Sayısal Dilbilgisi modeli, Konuşma bölümleri modeline ek olarak Arapça cümlenin bilgisayarlı hesaplanmasında kullanılan en önemli istatistiksel modellerden biridir. Aslında, Konuşma bölümleri modeli, Sayısal Dilbilgisi modelinde gözlemlenen eksiklikleri gidermek için ortaya çıkmıştır.

Anahtar Kelimeler: *Hesaplamalı Dilbilim, Otomatik İşleme, Sayısal Sentaks, Derlem Dilbilimi, Konuşma Bölümleri*

Computing The Arabic Sentence: From Markov's Model To Numerical Grammar Models

ABSTRACT

This research aims to discover the most important computational approaches that are concerned with computing the Arabic sentence. Thus, it presents the

DOI: 10.17932/IAU.ARAP.2019.020/ayad_v03i1002

computational concepts and models that the computer users use to complete this task, in addition to the type of linguistics corpus that aids in dealing with the sentence. The method adopted in constructing the conceptualizations and inferences of this research was descriptive-analytical, as it attempted to explore the most important approaches and computer models that were used, and are still being used, for processing the Arabic sentence. As for some reached results, we may mention; The fact that both complexity rigor of computer models and natural language and that leads up to the abandonment, even partially, of linguistic models in order to calculate statistical models. The Numerical Grammar model is one of the most important statistical models used in processing the Arabic sentences, in addition to the Sections of Speech model. In fact, the Parts of Speech model came to cover the shortcomings observed in the Numerical Grammar model.

Keywords: *Computational Linguistics, Automatic Processing, Finite States, Numerical Grammar, Linguistics Corpus, Parts Of Speech*

حوسبة الجملة العربية: من نموذج ماركوف إلى نماذج النحو العددي

د. سفيان وعكي

د. محمد مرزوق

ملخص

يهدف هذا البحث إلى رصد أهم المقاربات الحاسوبية التي عُنيت بحوسبة الجملة العربية باعتبارها لغة طبيعية يمكن معالجتها آليا. فهو، بذلك، يعرض المفاهيم والنماذج الحاسوبية التي يستخدمها الحاسوبي للقيام بهذه المهمة، إضافة إلى نوع المدونات التي تُساعده على التعامل مع الجملة العربية. لقد كان المنهج المتبع في بناء تصورات هذا البحث واستدلالاته وصفيا- تحليليا، حيث حاول استكشاف أهم المقاربات والنماذج الحاسوبية التي استُعملت، وما زالت تُستعمل، لحوسبة الجملة العربية. أما ما يتعلق ببعض النتائج التي توصل إليها نذكر أن طبيعة التعقيد والصرامة في النماذج الحاسوبية من جهة، واللغة الطبيعية من جهة أخرى كانتا سببا في التخلي، ولو جزئيا، عن النماذج اللغوية لحساب نماذج إحصائية، ويعتبر نموذج النحو العددي من أهم النماذج الإحصائية المستعملة في حوسبة الجملة العربية، بالإضافة إلى نموذج أقسام الكلام، والواقع أن نموذج أقسام الكلام جاء ليغطّي القصور الذي لُوحظ في نموذج النحو العددي.

كلمات مفتاحية: لسانيات حاسوبية- معالجة آلية- نموذج حاسوبي- سلاسل متناهية- نحو عددي- مدونة لغوية- أقسام الكلام

مقدمة

هدف اللسانيات الحاسوبية هو إعادة إنتاج التحويل الطبيعي للمعلومات عن طريق نمذجة إنتاج المتكلم للغة الطبيعية من جهة، وتأويل إنتاج المستمع لها من جهة أخرى على نحو مناسب في الحاسوب. ويرقى هذا إلى بناء آلات معرفية مستقلة (روبوتات) يمكنها التواصل بحرية بلغة طبيعية. إن المهمة المركزية للسانيات الحاسوبية الموجهة نحو المستقبل هي تطوير الآلات المعرفية التي يمكن للبشر التحدث معها بحرية بلغتهم الطبيعية. فعلى المدى الطويل، ستضمن هذه المهمة تطوير نظرية وظيفية للغة، وطريقة موضوعية للتحقق من كفاءتها، ومجموعة واسعة من التطبيقات التي تساعد على تنفيذها حاسوبيا.

وفي سياق سعيها لتحقيق الأهداف السابقة تواجه الحاسوبي مجموعة من الإشكالات الحاسوبية التي تهتم بها المعالجة الآلية للغة، إذ يطرح منذ الوهلة الأولى فرضيات وتصورات متعلقة بحوسبة مكونات اللغة المختلفة؛ التركيب والصرف والأصوات والدلالة، فالنظر إلى اللغة على أساس أنها مكونات إجرائية جزئية من السمات الأساسية هو ما يميز نظرة الحاسوبي للغة، ويعني ذلك أنّ حوسبة كل مكون من المكونات السابقة لا يتم إلا بوضع برامج حاسوبية، وأنظمة تحليلية تحاول تحقيق الكفاية اللازمة بهذا الصدد.

من هنا، تحاول هذه الورقة البحث في إشكال حوسبة الجملة في اللغة الطبيعية، وفي اللغة العربية من بين لغات أخرى. وعليه، تنقسم الورقة إلى ثلاثة أقسام أساسية، الأول: يهتم بعرض مجموعة من المفاهيم الأساسية التي ترتبط بحوسبة الجملة في اللغة الطبيعية، وإبراز بعض المشاكل التي تفرضها الطبيعة الوِجاهية للغة. والثاني: يسعى إلى رصد أهم المقاربات

الحاسوبية التي عُنيت بحوسبة الجملة. أما الثالث: فيسعى إلى الكشف عن نوع المدونات التي تُسعف الحاسوبي في التعامل مع الجملة في اللغة الطبيعية، واللغة العربية خصوصا.

1. تحديدات مفهومية

1.1. مفهوم اللغة

من المفيد بدايةً تحديد المادة الأساس التي تشتغل عليها حوسبة الجملة وهي اللغات الطبيعية، واللغة، بهذا الصدد، "مجموعة من الملفوظات، كل واحدة منها مبنية من مجموعة من الرموز والأصوات Alphabets، حيث يعتبر الصوت نفسه مجموعة... واللغات المقصودة ههنا هي اللغات العادية "Regular languages (Brain, Roark and Richard, Sproat, 2007, p2). إنّ هذا التحديد يشير إلى أنّ اللغة العادية لها بنيات مخصوصة يمكن التعامل معها حاسوبيا انطلاقا من مكوناتها، ويعني ذلك أن الانطلاق من الصوت كوحدة أولى تسمحُ ببناء وحدات أخرى هي الكلمة والجملة، وبعبارة أخرى؛ تعتبر البنيات الصوتية والصرفية والتركيبية وجاهية في اللغة.

2.1. التحليل التركيبي للجلمة

يلاحظ مصطفى غلفان أنّ تحليل الجملة "يتطلب الوقوف على بعض أساسيات التحليل النحوي، ومن ضمنها تحديد طبيعة الوحدات المكونة للجملة، وما يرتبط بها من مقولات ووظائف وأدوار دلالية" (غلفان وآخرون، 2010، ص 261). إنّ النموذج الأساسي في تحليل الجملة، في الأدبيات التوليدية خصوصا، موروث من الفكرة القائمة على التصور التوزيعي للوحدات اللغوية، بمعنى توزيع الكلمات في الجلمة، وسُمي هذا النموذج بالنموذج المركبي.

وقد "تم اقتراح قواعد لإعادة كتابة مستقلة عن السياق تدمج بعض المنطلقات المنهجية الأساسية المعروفة في التحليل للمكونات المباشرة على نحو ما نجدُ في [القاعدة التالية]" (غلفان وآخرون 2010، ص 262):

-ج ⟵ ف+ م س+ م س

من الملاحظ، إذا، أنّ الجملة تتكون من مجموعة من الكلمات، وهي، أي الجملة، كلٌّ يتَسَامَى عن مجموع أجزائه (الكلمات)، ويعني هذا أنّ الجملة بنية تشمل مجموعة من الكلمات، وهذه الكلمات هي أقسام الكلام في اللغة الطبيعية، كما سنرى في فقرة لاحقة، وأي تحليل للجملة هو محاولة لفهم طبيعة الوحدات المكونة لها من جهة، وما يربطها من علائق من جهة أخرى.

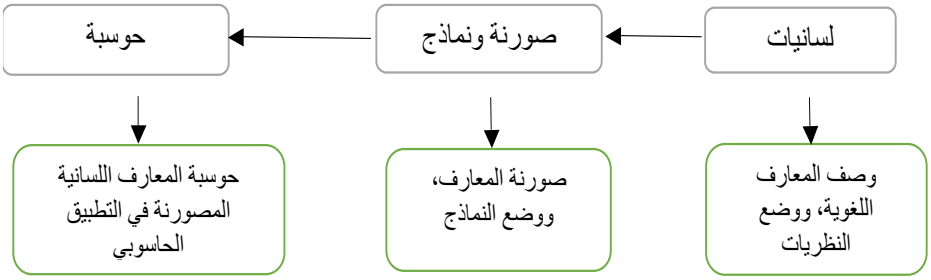
إنّ فهم العلائق التي تربط بين وحدات الجملة حاسوبيا يتم بمنهجين أو طريقتين، الأولى: هي التحليل التركيبي للجملة نفسها، ويعني ذلك ضرورة الالتزام بقواعد التأليف. أما الثانية: فهي اعتماد طرق إحصائية في مقاربة الجملة ومن ثم حوسبتها، ويمكن القول بأننا بهذه الخطوة ننتقل من النحو المركبي إلى النحو العددي، من هنا يكون السؤال مشروعا للحاسوبي، أي الطرق أسهل في الحوسبة؟ وأيهما يحقق الكفاءة اللازمة حاسوبيا؟

يذكر رالف غريشمان Ralph Grishman أن مجموعة من مطوري البرامج الحاسوبية يقترحون إمكانية القيام بالمعالجة الآلية للغة الطبيعية دون اللجوء إلى التحليل التركيبي... ويُرجع هذا الأمر إلى مجموعة من الملاحظات التي تتعلق أساسا بإشكالية حوسبة المكون الدلالي، وإمكانية تأثير المكون الدلالي على التحليل التركيبي...إلى آخره (Grishman 1987, p 12).

تبدو هذه الملاحظة خطوة منهجية مهمة لفهم طبيعة هندسة المكونات اللغوية في سيرورة المعالجة الآلية للغة، بمعنى أنّ كون الصوت والكلمة وحدات قاعدية تشكل الجملة، تسهِمُ في بناء برامج حاسوبية تساعد على تجاوز التعقيدات التركيبية وصرامتها القواعدية.

إنّ جعل النموذج المركبي مكون الدلالة معيارا أساسيا في تحديد المقولات النحوية يجعله صعبا في التطبيق الحاسوبي، ذلك أنّ المشكلات التي تطرحها الدلالة يصعب صورنتها أو صياغتها رياضيا، ومن المعروف أن أي انتقال من الوصف اللغوي إلى صياغة النماذج

الحاسوبية والبرامج الحاسوبية يقتضي المرور من الصياغة الرياضية للمكونات اللغوية، وهذا ما لا تُتيحه الدلالة بحال.



ورغم ذلك فإنّ غريشمان يرى أنّ التحليل التركيبي للجملة له نوعين أساسيين من الوظائف؛ الأول هو: تعين الفاعل والمفعول المتعلقان بكل فعل. والثاني هو: تنظيم البنية التركيبية للجملة (Grishman 1987, p 10). إنّ هاتين الوظيفتين تحيلان إلى نوع البرامج الحاسوبية الممكنة في حوسبة الجملة وهي المُعرِبَات Parsers التي تحدد اشتقاقات جملة معينة وإعرابها بالعودة إلى النحو الخاص للغة من جهة، والنماذج الحاسوبية الرياضية التي تسهل عملية الحوسبة من جهة أخرى، أي الطرق الرياضية والإحصائية الممكنة لوصف الجملة نظريا.

ويمكن الاستدلال على هذا الطرح بما يُطلق عليه نظرية العامل في اللغة العربية، مثلا، حيث تتحكم المقولة الأولى في الجملة في إعراب المقولات التي تأتي بعدها، ومثال ذلك قولنا:

-علّم الأستاذُ الطالبَ

يشير الدرس النحوي إلى أنّ الفعل "علّم" يعمل في الفاعل "الأستاذ" ويُسندُ إليه الرفع، كما يعملُ في المفعول به "الطالبَ" ويُسندُ إليه النصب، غير أنّ هذه الملاحظة في تحديد العلاقة العامليّة قد تكون معقدة في بعض الجمل خصوصا الاسمية منها، إضافة إلى الجمل التي تتغير

فيها رتبة (ترتيب الكلمات في بنية الجملة) الجملة، بمعنى أنّه إذا كانت الرتبة المفترضة في اللغة العربية هي: فعل+ فاعل+ مفعول به، فإنّ أي تغيير فيها قد يجعل حوسبتها على أساس النظرية العاملية معقدة، فمن المعلوم أنّ التقديم والتأخير، مثلا، من خصائص الجملة العربية.

ورغم أنّ التحليل التركيبي يمكن اعتباره فعّالًا على نحو ما، وذلك بالتدقيق في المعلومات التي يمكن أن يقدمها للحاسوبي، وتحديد البرامج الحاسوبية التي يمكن استعمالها، إلا أنّ أدبيات المجال تُظهر أنّه يمكن الاستغناء عنه جزئيا لحساب مقاربات ونماذج حاسوبية أخرى، وهو ما سنتحدث عنه في الفقرة الموالية.

2. مقاربات ونماذج حوسبة الجملة

1.2 مقاربة السلاسل المتناهية finite states approaches

يرجع نموذج السلاسل المتناهية إلى الروسي أندري ماركوف Andrey Markov، و"يعد حالة خاصة من نموذج عام هو النموذج المركبي، وذلك من خلال تقييده بمجموعة من القيود الصارمة على القواعد التي يقترحها لإنتاج جمل اللغة، وينطلق نموذج ماركوف من النتائج المتوصل إليها في الرياضيات، ولاسيما في الجانب المتعلق بحوسبة البنيات اللغوية وإعادة إنتاجها آليا"(غلفان وآخرون، مرجع سابق، ص 53).

إنّ الصرامة التي يتطلّبُها نموذج ماركوف في تحديد الاحتمالات الممكنة لورود كلمة معينة في الجملة راجعٌ إلى طبيعة القواعد والقيود الصارمة التي تقرّ بأنّ طريقة اشتغال النموذج صوريا "عبارة عن آلة تمر عبر عدد من الحالات، بحيث ترسل الآلة في كل طور أو نقلة Transition رمزا واحدا لا أكثر، وتطابق كل نقلة من حالة إلى أخرى موقعا في الجملة" (غلفان وآخرون، مرجع سابق، 54).

من الملاحظ أنّ أهم سمة تميز نموذج ماركوف هي طبيعته العلاقية، بمعنى أنّ كل عنصر في السلسلة المقترحة يكون لها رمز واحد فقط، كما أنّ هذا العنصر تكون له علاقة

بعنصر واحدٍ فقط من العناصر التي تسبقهُ. وبعبارة أخرى، يقوم هذا النموذج على أساس العلائق الرياضية والعددية، وبكلمة أدق؛ إنّه يقوم على أساس الاحتمالات العددية.

ويبدو أن مقاربة السلاسل المتناهية من أكثر النماذج التي تعتمد التحليل التركيبي والإحصائي معًا، وإن كانت تولي الإحصاءَ أهمية كبرى فيما يتعلق بحوسبة اللغة، إنّ مقاربة السلاسل المتناهية تتضمن مجموعة من التقنيات والنماذج الصغرى، كنماذج N-gram، وأقسام الكلام parts of speech.

1.1.2. نموذج النحو العددي N-gram

يذكر ف.ج. سميث وآخرون F. J. Smith and others أنّه يمكن تعريف النحو العددي بأنّه نموذج يهتم بقياس n-1 وهو متغير يقيس حالة الكلمات السابقة أثناء حساب احتمالات كلمة معينة موجودة في جملة ما (Smith and others, 1994, Pp337-349)، ويعني هذا أنّ نماذج النحو العددي المتعددة مبنية على نتائج نموذج ماركوف العام، وتقرّ بأن كل عنصر في سلسلة من الوحدات يمكن افتراض ارتباطه بآخر العناصر التي وردت قبله مباشرة، وقد يكون عنصرا واحدا أو أكثر.

وبالعودة إلى مفهوم النحو العلاقي، يمكن تفسير أهمية العلاقات الكامنة بين العناصر الموجودة في كل سلسلة متناهية بغض النظر عن عدد الوحدات التي تتضمنها، غير أنّه من المفيد ههنا أيضًا ملاحظة أهمية حساب الاحتمالات الممكنة في ربط العناصر فيما بينها بغض النظر عن الصرامة التي تفرضها القاعدة اللغوية، ويعني هذا أن تقنيات النحو العددي تركز على جانب الاحتمالات العددية، أي الإحصائية، في النحو على حساب الجوانب اللغوية الأخرى، فهدف نمذجة اللغة هو معرفة احتمالات الجمل في اللغة الهدف وفي البرنامج الهدف، واحتمال تسلسل الكلمة كما هو مبين في المعادلة التالية (Jurafsky Daniel & Martin James H 2021, Pp 1-29):

$$P(w1, w2 \ldots wn\text{-}1wn) =$$

$$P(w1)\ xP(w2|w1)\ xP(w3|w1, w2)\ x \ldots xP(wn|w1, w2 \ldots wn\text{-}1)$$

يمكن إبداء ملاحظة مهمة بصدد حساب الاحتمالات الممكنة في كل سلسلة من جهة، وعدد العناصر التي يكون فيها للكلمة المحتملة الورود في السلسلة علاقة بالكلمات التي تسبقها، وهي الافتراض الذي يمكن أن نطلق عليه بالاستقلال المشروط Conditionally independency للكلمات، بمعنى أنّ الكلمة قد يكون لها في نماذج النحو العددي علاقة بكلمة أو أكثر قبلها، فإذا كانت معادلة حساب n-gram بقيمة n-1 تفترض حساب احتمال واحد فقط، فإنّ تعقيد المدونة نفسها يفرض تطوير نماذج أخرى، لاسيما أنّ النموذج العام يتيح التعامل مع قيم هائلة في المدونة، ويعد حساب احتمال واحد قاصرا في الإحاطة بالقيم التي يسمح بها النموذج.

وبعبارة أخرى، يلاحظ براين وريتشارد Brain Roark and Richard "أنّ n في n-gram تشير إلى أن النموذج يجعل افتراض ماركوف لحصر مجال المتتاليات dependencies، وبهذا فالاحتمال المشروط لكل كلمة مستقل فقط عن n-1 في الكلمات السابقة...، وفي رتبة order نموذج ماركوف n+1" (Brain Roark and Richard, Ibid, P 141):

$$P(w_1 \ldots w_k) = \prod_{i=1}^{k} P(w_i \mid w_{i-n} \ldots w_{i-1})$$

وبالتالي فإنّ هذا الاحتمال يعطي إمكانية توسيع النموذج الأول الذي يقتصر على كلمة واحدة قبل الاحتمال أي n-1، إلى نماذج أخرى يضاف على أساسها عنصر آخر إلى العنصر الأول أي n+1، وعلى هذا الأساس يمكن صياغة هذه النماذج كالآتي (Jurafsky Daniel & Martin James H 2021, Pp 1-29):

1- **النموذج الأحادي** n=1: ويسمى Unigram، ويعني أنّ ورود كلمة في مكان ما في السياق ليس مرتبطا بما قبلها من كلمات:

$$P\ (wk|w1,w2...wk\text{-}1) = P\ (wk)$$

2- **النموذج الثنائي** n=2: ويسمىbigram، ويعني أن ورود كلمة ما في السياق يرتبط فقط بالكلمة التي قبلهُ:

$$P\ (wk|w1,w2...wk\text{-}1) = P(wk|wk\text{-}1)$$

3- **النموذج الثلاثي** n=3: ويسمى trigram، ويعني أن ورود كلمة ما في السياق يرتبط بكلمتين قبلها:

$$P\ (wk|w1,\ w2...wk\text{-}1) = P\ (wk|wk\text{-}2,wk\text{-}1)$$

4- **النموذج الرباعي**: n=4 ويسمى quadrigram، ويعني أن ورود كلمة ما في السياق يرتبط بثلاث كلمات قبلها:

$$P\ (wk|w1,\ w2...wk\text{-}1) = P\ (wk|wk\text{-}3,\ wk\text{-}2,\ wk\text{-}1)$$

إنّ نماذج النحو العددي تساعد في حوسبة الجملة من منطلقات إحصائية، ومن الحسنات التي يمكن ملاحظتها في هذه النماذج تسهيلها لإمكانية التعامل مع قيم هائلة من المعلومات اللغوية، أي المدونات، ويمكن التمثيل لهذه الخاصية التي تسمح بتوزيع w المتضمنة في المدونة على السلسة اللغوية بمعادلة يحيل فيها w على مجموع كلمات المدونة، فيما يحيل w1 على الكلمة القاعدية الأولى، فيما يحيل wk على أقصى عدد ممكن من الكلمات المفترض وجودها في السلسلة:

$$W= w1....wk$$

إضافة إلى ذلك، تجاوزها، أي نماذج النحو العددي، لفكرة نموذج ماركوف العام القائم على النحو المركبي واستبداله بفكرة النحو العلاقي الرياضي، وعموما فإنّ هذه النماذج يمكن ربطها بمفهومين أساسيين في علاقتها بحوسبة الجملة وهما: التوزيع distribution، والمقبولية Acceptance، بمعنى أنّ العمليات الإحصائية تقوم على احتمال الوحدات، ثم توزيعها في سياقها المناسب داخل السلسلة المفترضة، أما العملية الأخيرة فهي البحث في مقبولية التوزيع داخل السلسلة، وهنا يُطرح سؤال كفاءة النموذج وقدرتهُ.

يبدو أن التساؤل عن المشاكل التي تعترض نموذج النحو العددي سيكون مشروعا، وعموما فإنّ هنالك ثلاث مشاكل أساسية، من بين مشاكل أخرى، يمكن طرحها بهذا الصدد وهي، **أولا**: قصور نموذج ماركوف نفسه الذي قامت عليه نماذج النحو العددي. **ثانيا**: القصور الذي يكتنف المدونات. **ثالثا**: المشكل الذي يخلقه قصور المدونات نفسه، وهو ما يسمى بمشكل القيم الصفرية.

من بين الانتقادات التي وجهت لنموذج ماركوف المركبي اعتباره نموذجا غير اقتصادي لأنه لا يراعي طبيعة العناصر اللغوية من حيث النوع والعدد، إذ أنّ حساب اللوغاريتمات الضابط للآلة لا يمكنه، وهو ينتج الوحدات المكونة للجملة، أن يميّز بين المذكر والمؤنث، وبين المفرد والمثنى، أو أن يطابق بين الفعل وفاعله، فـ[قد] تحتاج الآلة مثلا إلى قاعدتين مختلفين لتوليد جملة واحدة في متغير موسوم فقط (غلفان وآخرون، مرجع سابق، ص55).

غير أنّ هذا الانتقاد يمكن تلافيه بتدريب البرامج المعتمدة على تحقيق كفاية التمييز التي يفتقر إليها النموذج، وحصول هذا الأمر ممكن عندما تكون المدونة التي يعتمدها الحاسوبي كافية، ومواردها تسمح بتغطية كل الاحتمالات الممكنة التي ترد فيها المشاكل السابقة، بمعنى أنّ تجاوز القاعدة اللغوية الصارمة، التي تُلزم بالتمييز بين المفرد والمثنى مثلا، يتم عن طريق تدريب الحاسوب على هذا النوع من الجمل الموجودة في اللغة.

إن افتراض وجود كل الجمل الممكنة في اللغة في مدونة ما يُعد أمرا عسيرا، لاسيما أن التوليد والابداع من سمات القدرة اللغوية، وعلى هذا الأساس فمشكل قصور تمييز الحاسوب بين مجموعة من القواعد اللغوية يمكن تفسيره حاسوبيا بعدم كفاية المدونة، وضعف تقدير الاحتمالات التي تردُ فيها الجمل، وقد أُطلق على هذا القصور مشكل القيم الصفرية.

في سعيها لتحقيق الكفاءة اللازمة في المدونة من جهة، وتلافي وجود قيم صفرية من جهة ثانية، يُقترح إقحام قيمة زائدة في السلسلة المقترحة، ويعني ذلك افتراض وجود الكلمة بإضافة الرقم 1 إلى السلسلة (Teemu, Hirsimäki and others, 2009, Pp 724-731):

$$P(w_i) = \frac{c_i}{N} \rightarrow P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V} \rightarrow c_i^* = (c_i + 1)\frac{N}{N + V}$$

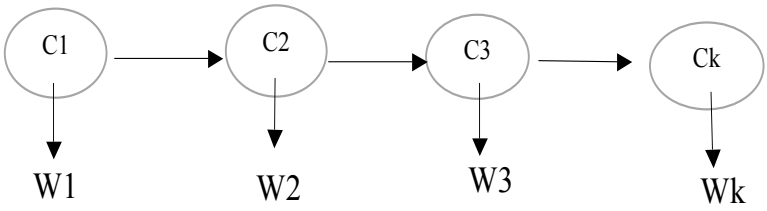
إضافة إلى ذلك، يذكر براين وريتشارد طريقتين إضافيتين لحل مشكل القيم الصفرية هما: طريقة التنازل التدريجي back-off في درجات النموذج، ويعني النزول درجة واحدة من نموذج إلى النموذج الذي تحته مباشرة، ثم طريقة المزج بين النماذج.

2.1.2 نموذج أقسام الكلام اللغوي class-based language model

نموذج أقسام الكلام هو محاولة لتعويض المعطيات القليلة عن طريق القيام بتعميم خصائص كلمة انطلاقا من الكلمات المشابهة لها، وفي هذا النموذج المبسط يكون لكل كلمة Wi تعيين في خانة واحدة محددة، وترجع أهمية هذا النموذج إلى محاولته مسايرة الطبيعة الخلاقة للغة، وقابلية تطور موارد المدونات، فقد نجد أن حصر كلمات لغة مجال معيّن ووسْمها، كالصحافة والاعلام مثلا، خلال أسبوع واحد فقط، وحصر الاحتمالات التي ترد فيها جملة بشكل معين يحتاج عمليات كثيرة باستعمال نموذج n-gram، ومن المعلوم أن وسم المدونات يستغرق زمنا طويلا في سيرورة الحوسبة، وبذلك فاعتماد نموذج أقسام الكلام يتيح تصنيف

الكلمات وتعيينها تحت كلمة واحدة، بل في خانة واحدة فقط من أقسام الكلام المتاحة في اللغات الطبيعة.

وعلى هذا الأساس، يمكن ملاحظة فائدتين على الأقل باستعمال هذا النموذج، أما الفائدة الأولى فهي: تعيين الكلمات في خانة واحدة مرتبطة بقسم واحد من أقسام الكلام، وأما الفائدة الثانية فهي: تجاوز التكرار، ووسم الكلمات المتشابهة في الخانة نفسها، ويقترح براين وريتشارد الترسيمة التالية لتوضيح هذه العملية (Brain Roark and Richard, Ibid, P 141):



من الملاحظ، إذًا، أن كل كلمة تُصنف ضمن خانة واحدة فقط، ويتم ربطها بالسلسلة من منطلق التشابهات التي تربطها بكلمات مثلها، وبهدف أجرأة نموذج أقسام الكلام، التي تكون غالبا (اسم، فعل، حرف أو أداة) يتطلب البرنامجُ الحاسوبي القيام بمجموعة من العمليات منها: التوسيم (بأحد أقسام الكلام)، تعيين الوظيفة التركيبية (معرب، مبني، حالة الإعراب، حالة البناء...إلخ)، ثم تحديد الوسائل التركيبية (المعطيات اللغوية، خوارزمية التحليل، المدونة اللغوية).

يعتمد نموذج أقسام الكلام في حوسبة الجملة العربية على تحديد خوارزميات Algorithms التحليل التي تستوعب معطيات اللغة المتنافرة، بل والمتعارضة أحيانا، ثم تحديد نوع المُعرِبات parsers التي تحدد الوظائف التركيبية، بمعنى أن الحاسوبي بصدد مشكلتين في آن معًا هما:

المعطيات المتنافرة للغة، وضرورة تحديد المعربات، ويعني ذلك أنّ هذه الإشكالات ستؤثر على التحليل التركيبي، كما ستؤثر على وسائله أيضًا.

إضافة إلى ذلك، يُطرح مشكل كيفية التعامل مع المدونات الضخمة التي تُستقى منها النصوص، وبالتالي الجمل، ويعني هذا أنّ الحاسوبي بحاجة إلى تقنيات وأدوات تمكنّه من التحليل التركيبي للجمل المكونة لنصوص المدونات قبل التعامل مع الجمل بشكل مباشر.

2.2 الوسم والعنونة التركيبية

يمكن، بدايةً، طرح سؤال حول المادة اللغوية التي نريد وسمها وعنْونَتها على أساس أقسام الكلام، ويعني ذلك أنه يجب تحديد العلائق الممكنة بين الموارد اللغوية، والنصوص، والجمل، والوحدات المكونة لها، ذلك أنّ الانطلاق من المورد اللغوي والانتهاء إلى الوحدة اللغوية يقتضي المرور من عمليات عديدة، قبل الانتقال إلى أي عملية حوسبة الجملة.

يرى تشانغ نينغ وجوان تزي أنّ "الغرض الأساسي الذي يسعى إليه علماء اللغة من إنشاء الذخائر اللغوية هو استخدام تلك الذخائر في عمليات استقراء التراكيب اللغوية وتحليلها والتعرف على طرق استخدامها"(خوانغ، تشانغ نينغ وجوان تزي، لي، 2016، ص 65)، إنّ استقراء التراكيب اللغوية، أو الجمل، يقوم على وصف اللغة الطبيعة كما يتحدثها، أو يكتبها المتكلم الأصلي لها.

ويمكن للذخيرة أن تتفرع إلى قسمين اثنين هما: ذخيرة ثابتة وذخيرة متطورة؛ فالذخيرة الثابتة تكون عيّنة مُمثِّلة لفترة تاريخية محددة، ويمكن التعامل معها بواسطة نماذج n-gram أيضًا مهما كانت ضخامتها، أما الذخيرة اللغوية المتطورة فتسمح بحركية النصوص وإمكان إضافة نصوص أخرى إليها، وهي التي يمكن التعامل معها بواسطة نموذج أقسام الكلام، فالذخيرة اللغوية على هذا الأساس تمكن الحاسوبي من النصوص اللازمة لحوسبة تراكيب اللغة كما هي في الذخيرة نفسها.

غير أنّ الإشكال المطروح ههنا هو أن الحاسوب لا يمكنه التعامل مع كل نصوص الذخيرة اللغوية دفعة واحدة من جهة، لا سيما أن الذخيرة تكون ضخمة، كما أنّه لا يمكنه انتقاء الوحدات التي تشكل الجمل بشكل عشوائي من النصوص، حتى لو كانت هذه النصوص قصيرة، ومن هنا يكون التسويغ العقلاني لضرورة تحديد وسوم للنصوص اللغوية وعنونتها كي يتسنى للحاسوب الانتقاء على أساسها من النصوص.

من بين البرامج التي تُستعمل للوسم والعنونة التركيبية في اللغة العربية برنامج ستانفورد Stanford، ويمكن توضيح طريقة عمل هذا البرنامج كالتالي: (http://nlp.stanford.edu:8080/parser/)

تمّ اختيار نص قصير من مدونة من ألفي كلمة ونسخها في البرنامج، وقد توصل إلى النتائج التالية:

من الملاحظ أنّ الوسوم Tagges التي يُسندها البرنامج إلى كل مقولة من مقولة النص المنتقى من المدونة لا تخرج عن أي قسم من أقسام الكلام المعروفة، أي الاسم والتي يشير إليها ب N، أو الفعل والتي يشير إليها ب V، أو الأداة والتي يشير إليها IN.

إضافة إلى ذلك، فإن البرنامج يحدد علامات الترقيم ويُسند إليها وسمًا خاصا بها، وإذا كانت أقسام الكلام السابقة هي الوحدات القاعدية التي يُعتمدُ عليها في الحوسبة، فإنّ المحلل التركيبي يقوم قبل ذلك بتحديد العبارات phrases الموجودة في النص، ويظهر ذلك من خلال وسوم برنامج ستانفورد الموضحة كالتالي:

1. مركب فعلي: verbal phrase (Vp)

2. مركب اسمي: Noun phrase (Np)

ومن ناحية أخرى يسمح البرنامج نفسه بتحديد العلائق الإعرابية بين الكلمات (أقسام الكلام) المعنونة في النص السابق، ويمكن توضيح المثال كالتالي:

وعموما فإنّ التحليل التركيبي له آلية تمر من ثلاث مراحل تعنى كل مرحلة منها بتحديد نوع من المعلومات، ويمكن اختزالها كالتالي:

- المرحلة الأولى: يتم فيها تحليل النص إلى مجموعة من الجمل واستخلاص المعلومات المرتبطة بها.
- المرحلة الثانية: يتم فيها تحليل الجمل الناتجة من المرحلة السابقة إلى مجموعة من المركبات/ العبارات.
- المرحلة الثالثة: يتم فيها تحليل المركبات إلى عناصرها الأولية مع تحديد أقسام الكلام المصاحبة لها، ويتم بعد ذلك دمج المعلومات المتعلقة بأقسام الكلام مع السمات الإعرابية المتعلقة بها.

3. المدونات اللغوية وحوسبة الجملة في اللغة العربية

تتم عملية حوسبة الجملة من منطلق نموذج أقسام الكلام بشكل نسقي، وعبر مراحل يؤدي بعضها إلى الآخر، فالعملية الأولى هي تحديد المدونة اللغوية التي يمكن أن تكون

ثابتة أو متطورة، ثم الانتقال إلى التحليل التركيبي القائم على أساس عنْصرَيْ التوسيم والعنونة التركيبة.

إنّ وسم كل العبارات التي ترِدُ في النصوص تحدد العلائق الإعرابية الكامنة في الوحدات التي تشكل العبارات، وهكذا تكون العلاقة جدلية، بمعنى أن التحليل ينطلق من الوحدة الكبرى هي نص المدونة، ثم العبارات، فالوحدات، ويتم التحليل بشكل عكسي خلال السعي لتحديد العلائق الإعرابية.

1.3 مفهوم وأنواع المدونات الحاسوبية

يشير المعتز بالله السعيد إلى أنه يمكن تعريف المدونة اللغوية بأنها "كتلة غير منتظمة من النصوص المكتوبة أو المنطوقة، يمكن التعامل معها آليا والتحكم في بياناتها ومدخلاتها بالإضافة أو الحذف أو التعديل من خلال محررات النصوص" (السعيد 2016، ص 11)، فالمدونة اللغوية بهذا المعنى تعتبر نصوصا أصيلة من اللغة الطبيعية التي يكتبها أو ينطقها متكلمو لغة ما، وهدف لسانيات المدونة الحاسوبية هو جمع تلك النصوص والتحكم في مدخلاتها ومخرجاتها آليا.

بعد تعريف المدونة اللغوية، يبدو أن التساؤل حول أساليب بنائها، وأقسامها مشروعا، إذ تعددت الأفكار والآراء بخصوص معايير بناء المدونة، وكذلك حول تسميّة أقسامها بين المختصين.

أولا: معايير بناء المدونة

يقوم اختيار نصوص المدونة اللغوية المعنية بالتحليل، كما رأينا في التعريف، على عدة معايير وأساليب يجب مراعاتها في ضوء الهدف المنشود من التحليل، وهي كما لخصها عبد الله الفيفي وآخرون كالتالي (الفيفي وآخرون 2017، ص 88):

1. انتماء النصوص لأي مستوى (الفصيح- المعاصر- العامي)

2. مصدر النصوص (الكتب- الصحف الورقية- الصحف الالكترونية).
3. طريقة اختيار النصوص (الحصر الشامل- العينات الإحصائية).
4. تنوع النصوص أو التركيز على نوع محدد من النصوص.

وبالإضافة إلى معايير اختيار المدونات، فإن الحاسوبي يستعين بثلاثة أساليب محدّدة لبناء المدونة اللغوية الحاسوبية وهي كما لاحظ السعيد: أسلوب الحصر الشامل، وأسلوب الاستبانة، وأسلوب العينات الإحصائية (السعيد 2017، ص 4).

إن المعايير والأساليب السابقة كما عددها الفيفي وآخرون (2017)، والسعيد (2017) تشير إلى ثلاث خطوات شاملة في جمع المدونات الحاسوبية وبنائها وهي، **أولا**: نوعية اللغة (فصيحة- عامية- قديمة- معاصرة). **ثانيا**: مصدر النصوص (ورقية، الكترونية). **ثالثا**: كيفية حصر النصوص وانتقائها (أسلوب الحصر الشامل، أسلوب الإحصاء).

ثانيا: أقسام المدونات اللغوية الحاسوبية

يذكر صالح العصيمي وآخرون أن المدونات اللغوية الحاسوبية يمكن تقسيمها بالنظر إلى؛ النوع، والغرض، والعدد، والتصميم، (العصيمي وآخرون 2015، ص 241)، ورغم تعدد المدونات اللغوية الحاسوبية واختلافها، ركزّ العصيمي على المدونات المتعلقة بالعدد، وذلك راجع في نظره إلى أن المدونات المتعلقة بالغرض والنوع تنحصر أبحاثهما فقط على مستويات لغوية محددة، وأجناس كتابية معيّنة، أما التصميم فمنتهى اهتمامه ينحصر في مجال تطوير تصميم المدونة ومعالجتها من حيث أنواع التوسيم والتحشية (صالح العصيمي وآخرون 2015، ص 241).

ويرى العصيمي وآخرون، كذلك، أن نوع المدونة الحاسوبية الذي يمكن اعتباره كافيا هو النوع المتعلق بالعدد، على الرغم من أهمية بقية الأنواع، وتُعزى أهمية مدونة العدد إلى كونها الأكثر شمولية لمستويات اللغة. (العصيمي وآخرون 2015، ص 241).

وتنقسم أنواع المدونات الحاسوبية العربية المتعلقة بالعدد إلى ثلاثة أقسام: **أولا؛** يتضمن نصوص اللغة العربيّة المكتوبة والمنطوقة أو أحدهما، والمدخلة حاسوبيا مع توفر محرك بحث شبكي فيها من أجل البحث عن كلمة محددة في النصوص المُدخلة فيها حاسوبيا، وإمكانية إظهار عدد مرات تكرارها، وإتاحة إخراج كشافاتها السياقية، كله من أجل كشف السلوك البيئي السياقي اللغوي الطبيعي لتلك الكلمة وما يرج قبلها وبعدها من وحدات معجمية نظمية syntagmatic lexical units تعرف بمصطلح (المتصاحبات اللفظية collocations). **ثانيا؛** يتضمن مدونتين منفصلتين؛ الأولى: عربية، والثانية: لغة غير العربية، ويُعمَل على جمعها معًا وفق عدة معايير هي على النحو الآتي: تقابل الأوعية genres والمجالات Domains من حيث النوع والحجم، وتوافق أزمنة إنتاج النصوص الحيّة في كلتا اللغتين، وتقارب خيارات محرك البحث فيها. **ثالثا**: يُعرف بالمدونة المتوازية، وهذا النوع من المدونات مازال في بداية انطلاقته بين العربية والانجليزية.

وبخلاف التقسيم الذي اقترحه العصيمي وآخرون، يرى تشانغ نينغ وجوان تزي، لي أن المدونة اللغوية تعرف اختلافا واضحا في التقسيم والتسمية، ولا يوجد اصطلاح واحد لأنواعها وعددها (تشانغ نينغ وجوان تزي 2016-، ص 99-100)، وعلى هذا الأساس اقترح أربعة أقسام للذخيرة اللغوية وهي:

1. **الذخيرة المختلطة:**
   هي أبسط وسيلة لجمع المادة اللغوية داخل ذخيرة لغوية، حيث يسعى المصممون قدر الإمكان إلى جمع كل ما يمكن جمعه من أنواع النصوص المختلفة، ولا يكون هناك أية قواعد سابقة فيما يتعلق باختيار المادة اللغوية.
2. **الذخيرة المتجانسة:**
   وهي عكس النوع السابق، وتُعتبر ذخيرة تيبوتر tipoter التابعة للحكومة الأمريكية خير مثال على هذا النوع من الذخائر.
3. **الذخيرة المنظومية:**

الغرض من إنشاء هذا النوع من الذخائر اللغوية ضمان جمع الممادة اللغوية بما يُغطّي إطارا لغوية واسعا، بحيث تمثّل الذخيرة اللغوية أكبر مساحة من اللغة.

4. ذخيرة الاستخدام المتخصص:

يخزّن بتلك الذخيرة كل أنواع الذخائر المتخصصة، مثل ذخيرة العلوم الإنسانية بأمريكا الشمالية إلى آخره.

يبدو، إذا، أنه رغم اختلاف التسمية وتِعْداد أنواع المدونات اللغوية الحاسوبية كما رأينا عند العصيمي وآخرون (2015)، وتشانغ نينغ وجوان تزي (2016)، إلا أنهم يتفقون جميعًا على أن تلك الأقسام تغطي الجوانب الضرورية في كل مدونة لغوية من حيث نوع النصوص، وأغراضها، وسعتها، وتصميمها، فرغم اتجاههم إلى تغليب شمولية مدونة العدد، أو المدونة المنظومية، نظرًا للمجالات الواسعة التي تغطيها، والإمكانات الضخمة التي تُتيحها في معالجة البيانات اللغوية، إلا أنهم لا ينكرون أهمية الأنواع الأخرى وضرورتها بحال.

بعد الحديث عن مفهوم المدونة اللغوية، ومعايير بنائها، وأقسامها، أصبح لِزاما الحديث عن كيفية الإفادة منها في حوسبة الجملة العربية. ويمكن القول، أننا سنركز على جانب واحد فقط بهذا الصدد وهو الوسم والعنونة التركيبية.

2.3 الوسم والعنونة في المدونات اللغوية

يمكن حل الكثير من الإشكالات عن طريق المدونة اللغوية، ويبدو التساؤل عن نوع المدونة المقصود ههنا مشروعا، وكيف يمكن أن تساعد في حوسبة الجملة في اللغة العربية، من بين لغات أخرى.

يلاحظ السعيد أنّه "ثمة ثلاثة مناهج يعتمدُ عليها الباحثون في بناء أدوات التحليل التركيبي الموافقة لطبيعة اللغة العربية؛ حيث يقوم المنهج الأول على المعطيات اللغوية المستمدة من قواعد النحو العربي، سواء أكانت في صورة قوالب أم قواعد بيانات، ويقوم المنهج الثاني على خوارزمية التحليل التركيبي التي تمثل صورة رياضية لقواعد النحو العربي. أما المنهج الثالث؛ فيقوم على المعالجة الإحصائية لقواعد التركيب المستخلصة من المدونات اللغوية

العربية باعتبارها تمثيلا لواقع اللغة، حيث يتم تدريب نصوص المدونات اللغوية بهدف استخلاص الأنماط التركيبية، ثم تهيئة الآلة لاستقبال النتائج والتفاعل معها" (السعيد 2017، ص 311-344).

لقد رأينا سابقا أنّ النحو المركبي، أي القائم على القواعد اللغوية، لا يحقق الكفاية اللغوية اللازمة في حوسبة الجملة العربية، ذلك أنّ القواعد اللغوية صارمة وكثيرة من جهة، كما أنّها لا تمثل حقيقة اللغة من جهة ثانية، وإذا كان هذا القصور، أي الصرامة والكثرة، يطال النحو، فإنّ عمليات الصورنة تكون صعبة، بل إنّ صرامة الصورنة ذاتها تمنع من خاصية أساسية يجب أنّ تتوفر للحاسوبي في سيرورة الحوسبة وهي المرونة.

من الواضح إذن أن اعتماد منهج المدونات في التحليل التركيبي للغة العربية سيكون أكثر فائدة، وأكثر قابلية للتطبيق، غير أنّه يُطرح إشكال أساسي يرتبط بنوع المدونات المطلوب بناءها من جهة، وكيفية بنائها، وكيف يُستفاد منها في التحليل التركيبي، ومن ثم حوسبة الجملة في اللغة العربية.

يقترح السعيد المدونة الموسّمَة تركيبيا كطريقة تتوافق ومنهج تدريب نصوص المدونات اللغوية لاستخلاص الأنماط التركيبية، ذلك أنّ هذه المدونات تمثل حقيقة اللغة وواقعها، حيث أن وسم المدونات تركيبيا يعني إحصاء أنماط الجمل الممكنة في أي لغة طبيعة، واللغة العربية ههنا، واستخلاص خصائص كل جملها بهدف إسناد وسم لها.

إن التوسيم يعد تقنية لاستكشاف أنماط الجملة العربية، وتعيين مواطن الالتباس التركيبي فيها (السعيد 2017، ص 311-344)، وفي السياق ذاته يذكر السعيد أنّه "ثمة ثلاثة أساليب رئيسية لبناء المدونات اللغوية المحوسبة، وهي: أسلوب الحصر الشامل، وأسلوب الاستبانة، وأسلوب العينات الإحصائية" (السعيد 2017، ص 311-344).

يبدو أنّ المدونة الموسّمة تركيبيا ستكون لها أهمية كبرى في حوسبة الجملة العربية، فإذا كانت أنواع المدونات المختلفة تسمح بالاستفادة منها لأغراض متعددة، فإنّ النوع الُموسّم تركيبيا يمكن الحاسوبي من انتقاء المعلومات الضرورية لحوسبة الجملة، ذلك أنّ التوسيم قائم على تعيين أنماط الجمل في اللغة من جهة، وعلى إحصائها من جهة أخرى، إن التوسيم يبدأ من تحليل هذه الأنماط إلى وحداتها الصغرى، أي أقسام الكلام، وإسناد وسوم لها حسب كيفية ورودها في نمط الجملة.

لقد رأينا سابقا كيف يتم وسم وعنونة النصوص والجمل حسب برنامج ستانفورد، ويعني ذلك أنّ اللجوء إلى نماذج النحو العددي تسهل عملية وسم أي مدونة كيفما كان عدد الكلمات فيها، كما يتيح الوسم على أساس أقسام الكلام التحكم في المدونة، بمعنى أنّ المدونة إذا كانت ثابتة يسهل التعامل معها بنماذج النحو العددي، ووسمها تركيبيا على هذا الأساس، أمّا إذا كانت المدونة متطورة وديناميكية، فإن نموذج أقسام الكلام يسهل عملية التعامل مع هذا التطور والزيادة عن طريق تصنيف كل الكلمات الجديدة تحت قسم واحد، وواحدٍ فقط، من أقسام الكلام.

وبهذا، يمكن القول أنّ نماذج النحو العددي، ونموذج أقسام الكلام يُسعفان في حوسبة الجملة العربية من خلال إمكانية تطويع الحاسوبي لهما في إحصاء أنماط الجملة الممكنة في اللغة في النموذج الأول، ومن خلال تصنيف الوحدات المكونة للجمل في النموذج الثاني، إنّ تحديد أقسام الكلام يسهل عملية إسناد الوظائف الاعرابية الممكنة لكل جملة، وبالتالي تكون الجملة محوسبة بشكل إحصائي لكن مخرجاتها تكون مركبية، أي أنها تستجيب لقواعد اللغة.

خاتمـــــــــــة

لقد رأينا في ثنايا الفقرات السابقة ثلاث موضوعات أساسية؛ الأول هو: مفهوم اللغة والجدال الدائر حول مفهوم الجملة ومفهوم التحليل التركيبي. أما الثاني فهو: مقاربات ونماذج حوسبة الجملة في اللغة الطبيعية واللغة العربية، وقد ركزت الورقة على نموذجي النحو

العددي، ونموذج أقسام الكلام، وتأتي الفقرة الثالثة لتحدد أهمية المدونة في حوسبة الجملة، ونوع المدونات الممكنة من بين أنواع أخرى للمدونات لتسهيل هذه العملية.

إنّ تحديد تعريف اللغة ضروري لوضع مقدمة منطقية للمادة التي نتعامل معها، فتحديد موضوع العلم من أهم ما يجعلهُ دقيقا ومتماسكا منطقيا، أما الجزئية المتعلقة بالتحليل التركيبي فهو نقاش في الجملة ومكوناتها، وما تطرحها من إشكالات في بنيتها... إلى آخره.

ويبدو أن الفهم الدقيق للنماذج اللسانية التي تعاطت مع الجملة يبلور فكرة مدى صرامة القواعد اللغوية، والنماذج الرياضية التي تُصَورِنُ القواعد السابقة، ولما كان الحاسوبي يبحث أساسا عن أسهل وأكفأ الطرق الممكنة لنقل المعلومات الجُملية إلى الحاسوب، فقد كان التعقيد والصرامة السابقين سببا في التخلي، ولو جزئيا، عن النماذج اللغوية لحساب نماذج إحصائية، حيث إنّ نموذج النحو العددي من أهم النماذج المستعملة في حوسبة الجملة العربية، بالإضافة إلى نموذج أقسام الكلام، ولعل النتيجة التي يمكن ملاحظتها ههنا هو أن نموذج أقسام يكمل القصور الذي لُوحظ في نموذج النحو العددي.

المراجع

خوانغ، تشانغ نينغ وجوان تزي، لي. (2016) **علم الذخائر اللغوية**، ت: هشام موسى المالكي، لقاهرة: المركز القومي للترجمة.

سعيد، المعتز بالله. (2017) **"كيف نبني مدونة لغوية موسما تركيبيا للغة العربية بطريقة نصف آلية"**. في: مجلة الدراسات اللغوية، ع 3، مجلد 19، الرياض: مركز الملك فيصل للبحوث الإسلامية.

سعيد، المعتز بالله. (2016) **"توظيف المدونات اللغوية في تطوير مقررات اللغة العربية لمراحل التعليم العام"**، في: مجلة السياسة اللغوية والتخطيط، ع3، الرياض: مركز الملك عبد الله لخدمة اللغة العربية.

شراف، شناف. (2015) **"الترجمة الأدبية وسلطة الأنساق الثقافية"**، في: العربية والترجمة، ع 20، بيروت: المنظمة العربية للترجمة.

فيفي، عبد الله بن يحيى وآخرون. (2017) **مدخل إلى اللسانيات الحاسوبية**، الرياض: مركز الملك عبد الله لخدمة اللغة العربية.

عصيمي، خالد بن فهد وآخرون. (2015) **المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها**، الرياض: مركز الملك عبد الله لخدمة اللغة العربية.

غلفان، مصطفى وآخرون (2010). **اللسانيات التوليدية: من النموذج ما قبل المعيار إلى البرنامج الأدنوي: مفاهيم وأمثلة**، إربد: عالم الكتب الجديد.

Brain, Roark and Richard, Sproat. (2007) *Computational approaches to syntax and morphology*, New York: Oxford university press.

Grishman, Ralph. (1987)*Computational linguistics: an introduction*, Cambridge: Cambridge University press.

Jurafsky Daniel & Martin James H (2021*), N-gram Language Models*, Stanford: university of Stanford.

Teemu, Hirsimäki and others.(2009) *Importance of high- order N-gram models in morph- based speech recognition*, in: IEEE transaction on audio, and language processing, , V17, N4.

Smith, F. J. and others. (1994) "*A weighted average n-gram model of natural language*", in: *Computer speech and language*, V8.

http://nlp.stanford.edu:8080/parser/