- **علاقة اللغة بالرياضيات :**

تمثل اللغة للرياضيات تحدّيا قاسيا، فكيف لعلم صارم وقاطع أن يتعامل مع غموض اللغة والتباساتها ؟ لقد ظلت اللغة من دون الخضوع للمعالجة الرياضية، حيث عجزت رياضيات **إقليدس (ت:ق.م)** عن تناول إشكالية إبداعية اللغة، المتمثلة في قدرة الناطقين بها على توليد عدد لا نهائي من الجمل.

ترجع هذه الظاهرة إلى خاصية أساسية للتعبير اللغوي، التي يُعبّر عنها رياضيّا بمصطلح "**التداخل الحَلَقِي**" **recursion** ويقصد به أن الجملة الفعلية مثلا يمكن أن تتضمن شبه جملة اسمية، وشبه الجملة الإسمية هذه يمكن أن تتضمن جمل فعلية، وهكذا. نحو قولنا:

- **جاء الرجل الذي قابلْنا أخا له يعمل في المؤسسة التي...**

يلاحظ أن الجملة الفعلية المبدوعة بالفعل "**جاء**" تتضمن فاعلا في هيئة شبه الجملة الإسمية المبدوءة بالاسم "**الرجل**" التي تتضمن بدورها الجملة الفعلية المبدوءة بالفعل "**قابلنا**" التي تتضمن مفعولا في هيئة شبه الجملة الإسمية المبدوءة بالإسم "**أخا**" التي تتضمن بدورها جملة الوصف الفعلية المبدوءة بالفعل "**يعمل**".[1]

بمعنى أنّ من خصائص اللغة الإنسانية كثافة العلاقات وتشابكها ما بين الجمل والتراكيب التي تتداخل ما بين المستوى المبنوي والمستوى المعنوي.

لقد ظل هذا التكرار الحلقي حجر عثرة أمام إقدام أهل الرياضيات على تناول إشكالية اللغوي، فظلت دون حسم، إلى أن وضع **برتراند راسل (ت:1970م)** أسس

1

النظرية الصورية للغة **formal theory of language** التي كانت مدخلا أساسيا لتطوير اللغات الإصطناعية لبرمجة الكمبيوتر، ومهدت الطريق لكي يقيم **تشومسكي** نموذجه الرياضي للغات الإنسانية.

لقد أحال النحو التوليدي الذي أقامه **تشومسكي** النحوي اللغوي إلى سلسلة من المعادلات التي يمكن من خلالها توليد جميع التعابير اللغوية الممكنة على عكس ما عهدناه سابقا في النحو التقليدي، الذي يكتفي بإعطاء أمثلة من حالات الإطّراد والشذوذ، والتي مهما تعددت لا يمكن أن تغطي لا نهائية اللغة.[1]

- **آلية التكرار الرياضيّة للغة:**

يرى **تشومسكي** أن فرادة اللغة البشرية تتمثل في"**آلية التكرار**"، وربما يكون الفيلسوف الإيطالي **غاليليو غاليلي** (1564-1642) من أوائل العلماء في القرن السابع عشر الذين أدهشتهم هذه الآلية اللغوية، المتمثلة في قدرة الإنسان على توليج اللامحدود من الجمل من مجموعة محدودة من الأصوات. أما تشومسكي فقد أعاد رصد ظاهرة التكرار في بنية اللغة البشرية ضمن سياق رياضي معاصر؛ من الرياضي البريطاني **آلان تورينج** ويستخدمه في استنطاق قابلية البُنية الهرمية للغة على تكرار مستوياتها العميقة إلى ما لا نهاية.

- **لماذا تأثّر تشومسكي بـ:"آلان تورينج" ؟**[2]

1

[2] آلان تورينغ (1912- 1954) رياضي بريطاني أسس علوم الحاسوب الحديثة، وفكّ شفرة آلة إنجما (Enigma Machine) هو اسم يُطلق على أي آلة من عائلة الآلات الكهروميكانيكية الدوارة التي تستخدم لإنتاج الشيفرة السرية، وكلمة إنجما كلمة إنجليزية تعني "لغز"، ويقال أن "آلان تورينغ"، هو من قام بفك شيفرة آلة إنجما، وعنها كشف جميع القوات النازية في أوروبا وأنقذ أكثر من 14 مليون نسمة من الموت في الحرب العالمية الثانية وأطاح بهتلر.

إنّ الرياضي البريطاني "**آلان تورينج**"، لديه آلة تسمى "**آلة تورنج**"، برهن على خصائصها سنة **1936** وهي عبارةٌ عن نموذجٍ رياضيٍّ يُعبّرُ عن آلةٍ لها القدرة على القيام بعملياتٍ حَوسبيةٍ متعدّدةٍ كجمعِ الأرقامِ وتمييزِ النّصوص. وهي آلة تعمل بشكل منطقي رياضي بدون تدخّل بشري طبقًا لمدخلات وأوامر محددة مسبقًا. فهي قادرة على حل أي معضلة رياضية قابلة للحل، وهي تعتبر الأساس الرياضي لأيّ حاسوب، ويعتبر أي حاسوب يعمل ما هو إلا نجاح جديد لـ"**آلة تورنج**" بحيث تقوم بنفس مهامه في معالجة البيانات.[1]

أمّا لماذا تأثّر تشومسكي **بآلان تورينج** لأنّ **آلة تورينج** آلة عامة universal Turing machine، بمعنى أنّ ديناميكية جميع الحواسيب موحّدة، فإذا ربطناها بشبكة واحدة فإنها تعمل بنفس الآلية، عندها يمكن استخراج المعلومات في أي حاسوب في أي مكان في العالم. فكما أنّ جميع الحواسيب موحّدة بنظام الكتروني منظّم، لأنها موحدة في برمجتها وموحدة في شبكتها وبهذا تعمل على وتيرة واحدة وكأنها حاسوب واحد عبر العالم .

وكما يمكن عبر برمجيات استخراج المعلومات الموجودة في أيّ حاسوب في أيّ مكان في العالم. كذلك يمكن استخراج القواعد الكلية للغات البشرية[2] وهي قيود ضمنية فطرية مصممة في جميع أدمغة البشر. ولقد اصطلح عليها تشومسكي بمصطلح "**النحو**

1

[2] يرى علي حرب أنّ الكليات اللغوية هي أشبه ما تكون بنظرية المثل عند أفلاطون. ويعني بها عالم ما قبل العالم الحسي، يكون فيه الإنسان على علم بجميع العلوم والخفايا، وحينما يولد يكون قد نسي كل هذه العلوم، وما عليه إلا أن يتذكرها في العالم الحسي.

الكلي" وهو التجهيز البيولوجي الفطري للطفل، بمعنى أنّ الطفل يولد مجهّز بيولوجيا بما يجعله قادرا على التمييز بين الأصوات اللغوية والأصوات الطبيعية والصناعيّة.[1]

ولأنّ معالجة الحاسوب للبيانات تشبه معالجة العقل للمعلومات، **تأثّر تشومسكي بآلان تورينج** لأنه حدس أنّ هناك تشابه ديناميكي بين آلة تورينج العالمية التي تعتبر القاعدة الرياضية التي تعمل بموجبها جميع الحواسيب، وديناميكية اللغة مع العقل الموحدة بين جميع اللغات البشرية، وبهذا أعطى تشومسكي للغة مفهوما رقميا عندما حصر فرادة اللغة البشرية في "آلية التكرار"، أي توليد اللامحدود من المحدود وبالتالي فيه إمكانية استثمار أفكار تورينج الرياضية في استنطاق البنية الهرمية اللغة.

- **حتميّة التقاء اللغة بالرياضيات:**

إنّ لقاء اللغة مع الرياضيات كان أمرا محتوما، وذلك لخمسة عوامل؛ وهي:[2]

1-تعقّد اللغة: وهو تعقّد نابع من داخلها، ومن شدّة العلاقات التي تُربط بخارجها. والتعقيد وهو يسعى إلى التبسيط، عادة ما يلوذ بالتجريد ليوقعه في شباك الرياضيات التي من مهامها تبسيط المعقّد.

2-لا نهائية التعبيرات اللغوية: فمن أبجدية محدودة العدد (30 حرفا في معظم اللغات)، يمكن تكوين مئات الآلاف من الكلمات، والجمل. ما يضاعف من هذه اللانهائية ما يعرف بخاصيّة "**التداخل الحلقي**" التي تتصف بها اللغة، وليس هناك ما هو أفضل من الرياضيات لاحتواء خاصية لا نهائية اللغة، وذلك بفضل قدرتها التوليدية **generative** الهائلة. بحيث يمكن من معادلة رياضية واحدة للخط المستقيم **(أ) س + (ب) ص + (ج) = صفر**. توليد العدد اللانهائي لجميع

1

2

الحالات الممكنة للخط المستقيم، وذلك بمجرد تغيير قيم عوامل المعادلة: أ، ب، ج.[1]

3-كثافة العلاقات[2]: بحيث تتمتع اللغة بشبكة كثيفة من العلاقات الصوتية والصرفية والتركيبية والدلالية والمنطقية والرياضيات علم قادر تمثيل العلاقات وإبراز عواملها الحاكمة. ويحتاج ذلك إلى درجة عالية من الصورية (**التجريدية**) على مستوى الحدود اللغوية، وعلى مستوى العلاقات التي تربط بين هذه الحدود.

4-صورية الحدود اللغوية: حيث تصنف الحدود اللغوية إلى أسماء وأفعال وحروف وصفات، وهو تصنيف يمكن إدراكه بالحدس، ولكي تنمذج اللغة رياضيا يجب أن تتخلّص من حدسيتها بما يسمح بصياغة حدودها بصورة صورية موحدة يمكن تطبيقها على جميع الحدود والمقولات اللغوية المركبة منها. وهو ما أتبعه بالفعل **تشومسكي** في نظريته عن المعمولية والرابط التي تستخدم حاليا في تمثيل نحو أي لغة تمثيلا رياضيا.

5-صورية العلاقات النحوية: إنّ العلاقات النحوية في مستوياتها السطحية حدسية بامتياز، كعلاقة الفعل بفاعله مثلا، بحيث يتطلب التنظير اللغوي صياغة أكثر صورية فمن علاقة العامل والمعمولية، كحرف يعمل في مجروره، (س يعمل في ص)، حتى تصبح التعبيرات اللغوية سلسلة من الرموز ترتبط فيما بينها بعلاقات مجردة.

---

[1] وقد كان **برتراند راسل** هو أول من نجح في تذليل ظاهرة التداخل الحلقيRecursion (وهو مصطلح رياضي، أي الدالة تستدعي نفسها بنفسها أي العودة إلى نفس المكان، أي تكرار نفس الشيئ) للتناول الرياضي في نظريته الصورية للغة وقد مهدت هذه النظرية إلى ما يعرف باللسانيات الرياضية التي مهدت بدورها لظهور اللسانيات الحاسوبية.

[2] نحو جاء محمدٌ، **فـ: محمد: فاعل**، لخمسة قرائن، وهي: لأنه **اسم**، ولأنه **مرفوع**، ولأنه **تقدمه فعل**، ولأنه **مبني للمعلوم**، وأخيرا قرينة **الإسناد**؛ أي هو من قام بالفعل.

إذن؛ فعلاقة اللغة بالحاسوب كان بسبب الرياضيات، والرياضيات بمعناها الواسع هي علم التعامل مع المفاهيم المجردة، وفي نفس الوقت نجد أنّ النظام اللغوي يزخر بالمجردات، مثل علاقة الفاعلية والمفعولية مثلا، وصلة الرياضيات بالحاسوب وهو آلة التعامل مع الرموز ليست بحاجة إلى توضيح، ويمكن أن نختصرها بمقولة **آلان تورينج** مؤسس نظرية الأوتوماتيات، "**ما يمكن تمثيله رياضيا يمكن برمجته آليا وبشكل قاطع**".[1]

وأهم فروع الرياضيات التي تتعامل معها اللغة والحاسوب، وهي:

- **الجبر**: المجموعات، العلاقات، نظرية الدوال، جبر التمثيلات المنطقية.
- **الهندسة**: نظرية الأشكال (طوبوغرافيا)، الشجريات.

وإذا كان تطبيق القوانين الرياضية والفيزيائية لإدراك الأشياء الممكنة في المحيط الخارجي، فإن وصف اللغة رياضيا يجعلنا قادرين على تجسيدها لدى الحاسوب الذي يتخذ الأداء الإنساني نموذجا له يسعى إلى مناظرته.

ويقتضي تمثيل اللغة رياضيّا من خلال الوصف والإستقراء الإستناد إلى مدونة لغوية تعكس الواقع اللغوي، بحيث يمكن استنباط القوانين والأحكام التي ترسم الخصائص العامة للظاهرة اللغوية.[2]

- **عملية الرقمنة (digitization) المفهوم والآليات :**

1

2

لا يختلف الحاسوب الذي يوجه الصواريخ عن الحاسوب الذي يستخدمه الأطفال في ألعابهم، ولا فرق بين الحاسوب الذي يستخرج كشوفات الحسابات وبين الحاسوب الذي يظهر الأشكال ويعرب الجمل ويحلل النصوص، فالحاسوب آلة عمومية. وبفضل جبرية برامج الحاسوب وقطعيتها في أعلى مستويات التجريد البحت فإنه يمكن رقمنة الواقع بكل تضاريسه وغموضه وسحق جميع مشاكله رقميا، وكما نعرف أنه لا يوجد ما هو أكثر تجريدا من الأرقام وهي الحقيقة التي أسبغت على الكومبيوتر صفة الرقمية (**digital computer**) ولكونه رقميا يلزم تحويل كل ما يغذى له إلى أرقام.[1]

هناك أشياء بحكم طبيعتها هي أرقام مثل عدد صفحات الكتاب أو عدد السكان، وهناك خصائص يمكن أن نعبر عنها بقيم رقمية باستخدام طرق القياس المختلفة كالمسافة والزمن والطول والوزن والحجم. ما أن نتجاوز هذه الحالات البسيطة حتى يبدو الأمر أكثر صعوبة؛ فكيف نحيل النصوص والكلام المنطوق والموسيقى والأشكال والقواعد إلى أرقام. وهو ما سنحاول الإجابة عنه هنا في إيجاز.

ترتكز عملية الرقمنة بصورة أساسية على عدة أساليب تستخدم مفردة أو متضافرة، وهي:[2]

1- **التكويد أو التشفير codification** : يستخدم لتمثيل النصوص المكتوبة، حيث يعطي لكل حرف من حروف الألفباء كودا رقميا، لتحل سلاسل الأرقام محل سلاسل الحروف في الكلمات، ومن ثم الجمل وما عداها من نصوص.

1

2- **التبسيط simplification** : يستخدم التبسيط في عدة أمور منها تمثيل الألوان بحيث يعطي كل لون أصلي رقما معينا، وهو الشيء نفسه بالنسبة لدرجة اللون حيث يتم تصنيفها في تسلسل رقمي من الشدة حتى الخفوت.

3- **التوصيف بدلالة الملامح (السمات) features -based specification** : يستخدم في توصيف الرموز اللغوية، بحيث يتم تمثيل الأصوات اللغوية بدلالة عدد محدود من السمات الصوتية مثل الهمس والجهر والشدة والرخاوة وهكذا.

4- **الصياغة الرسمية (الصورية) formalism** : بحيث يتم صياغة أحكام اللغة الصورية في قواعد رياضية أو منطيقية، حيث يسهل بعد ذلك رقمنتها.

تعتبر الرقمنة أحد مصادر قوة تكنولوجيا المعلومات بحيث نجحت في تحويل أنماط المعلومات من نصوص وأصوات وصور ورموز إلى سلاسل رقمية قوامها "الصفر والواحد" وهو لغة الحاسوب التي يقوم عليها. بمعنى أنّ الحاسوب آلة رقمية بامتياز والرقمنة هي جوهر الوظيفة الأساسية التي تقوم بها وحدات الإدخال input devices التي تحول ما يغذى إلى االحاسوب مهما كان أصله إلى أرقام، في حين تقوم وحدات الإخراج output devices برد الأرقام إلى الصورة الطبيعية من نصوص وأشكال وأرقام وغير ذلك.[1]

- **أهميّة رقمنة اللغة في عصر تكنولوجيا المعلومات:**

تعتبر الرقمنة من أهم الأفكار الذهبية لتكنولوجيا المعلومات بحيث استطاعت تجريد أنماط معلومات العالم الخارجي في عالم رقمي الكتروني، أكثر تفاعليةً ومرونة وأعمق تواصلا وسهولة. أما أهمية رقمنة اللغة فهو فتح الكتروني بالغ الأهمية في عالم تكنولوجيا المعلومات، وذلك لأهمية موقعية اللغة في خريطة العلوم والمعارف والفنون.

1

يقول نبيل على :"أينما يكن مسلكك في دنيا المعرفة، فابحث عن اللغة: قمة العلوم الإنسانية ورفيقة العلوم الطبيعية، وركيزة الفلسفة عبر القرون، ورابطة عقد الفنون، ومحور تكنولوجيا المعلومات، وهندسة معرفتها ولغات برمجتها".[1]

فإذا كانت الثقافة محورية في منظومة المجتمع، فإنّ اللغة محورية في منظومة الثقافة، فقد ثبت أن الثقافة أصبحت محور عملية التنمية في مجتمع المعلومات، وأكّدت اللغة بفضل عالم المعلوماتية كونها، محور منظومة الثقافة بلامنازع، ونتيجة لذلك، فقد أصبحت معالجة اللغة آلياً بواسطة الحاسوب هي محور تكنولوجيا المعلومات.[2]

إذن؛ رقمنة اللغة العربية **خيار استراتيجي** وذلك لمزاحمة وحش العولمة اللغوية الإنجليزية في الشبكة العنكبوتية كقوة احتكارية لتكنولوجيا المعلومات وهي قوة ناعمة غير صدامية لا تدان ، فكلما كان لنا موقعا استراتجيا في الشبكة العنكوبتية عبر المواقع والصفحات والمنصات (نصا وصورة وفيديو) كلما زادت مكانة الحرف العربي ليكون مصدرا لتلقي المعرفة.

المدوّنات اللغوية المرقمنة (النشأة، المصطلح، والمفهوم، الآليات):

1- نشأة المدوّنات اللغوية Linguistics corpora :

إن المدوّنات اللغوية بمفهومها الحديث لم تبدأ إلا في الستينيّات من القرن الماضي، حيث جُمعت أول مدوّنة لغوية محوسبة سنة **1961**وهي مدونة **براون** نسبة إلى جامعة **براون** الأميركية التي كانت تحوي مليون كلمة من الإنجليزية الأميركية المعاصرة في ذلك الوقت. ثم تبعتها مدوّنات عدة بعد ذلك بنفس الحجم والتصميم، مثل مدونة **LOB**

1

2

التي جمعت سنة **1980**، والتي كانت نتيجة تضافر جهود جامعة **لانكستر** في بريطانيا وجامعة **أوسلو في النرويج**.

ومع تزايد قدرات الحاسوب وإمكانية رقمنة النصوص توالت المدوّنات اللغوية الغربية وبالأخص الإنجليزية منها؛ فظهرت مدوّنات عدة مثل :[1]

1- **مدونة كوبيلد Cobuild** التي تحوي ما يقرب من ثلاثة مليارات كلمة استخدمت في بناء معاجم **كوبيلد** المتعددة، جامعة برمنغهام في بريطانيا سنة 1987.

2- **المدونة اللغوية الوطنية البريطانية** التي تحوي مئة مليون كلمة، في مطلع **ق 20** .

3- **مدونة أكسفورد** اللغوية التي تحوي 2 مليار كلمة وتستخدم خصيصا في بناء معاجم أكسفورد الشهيرة.

1- المدوّنة اللغوية Linguistic corpus (المصطلح والمفهوم):

كلمة **corpus** كلمة لاتينية تعني الجسد وجمعها corpora ومن مقابلاتها في العربية ؛ منها: المدونات اللغوية، الذخائر اللغوية، المتون اللغوية، الذخائر النصية، المكتنزات النصية...الخ. أم مفهوم **المدوّنة اللغوية** فهي رصيد ضخم من نصوص اللغة في صورة إلكترونية، تُجمع اعتمادا على معايير خارجية؛ لتمثل قدر المستطاع اللغة أو أحد صورها لتكون مصدرا للأبحاث اللغوية؛ فهذه أهم ثلاثة معايير لها؛

- فمعنى أنها في صورة إلكترونية، أي أنها بصيغة نصية بسيطة، وقابلة للمعالجة الآلية المباشرة من الحاسوب، وليست بصيغة صور أو بصيغة **Pdf**.

1

- أما معنى أنها تجمع اعتمادا على معايير خارجية أي أنها تمثل فترة زمنية معينة وألا تطغى نصوص ذات طبيعة معينة على باقي نصوص المدونة كأن تطغي مؤلفات كاتب معيّن مثلا، ما لم يكن المقصود هو دراسة فكر هذا الكاتب بالأساس.
- وكذلك من شروط المدوّنة قدرتها على تمثيل واقع اللغة في مجال الدراسة، أو الغرض الذي من أجله جمعت النصوص.[1]

إذن فالمدوّنات اللغوية تحوي نصوصا تعكس الاستعمال الحقيقي للغة في مجال معين في شكل مقروءٍ آليّ **machine readable** لغرض الدراسة والتحليل.

- **ملاحظة علمية:**

المدوّنة لغة :اسم مفعول مشتقٌّ من دوّن يدوّن تدوينًا بمعنى كتب .والفعل دون مشتقٌّ بدوره، من كلمة فارسية معربة هي ديوان التي استعملها العرب لتدل على الدفتر التي تكتب فيه أسماء العمال والجند وأهل العطية، وكذلك على المكان التي تُحتفظ فيه هذه الدفاتر، ودون الكتب والصحف :جمعها ورتّبها. ويقال إن الخليفة عمر بن الخطاب أول من دون الدواوين في الدولة الإسلامية، أي أنشأها ونظّمها.

ومادام أنّ المدونة هي مجموعة النصوص المكتوبة والموثقة من حيث **المصدر** و**التاريخ** و**النوع** كحد أدنى؛ يمكن اعتبار المصنفات القديمة مدوّنات لغوية تراثية معجمية ورقية تراثية كمعجم **لسان العرب** مدوّنة معجمية، والكتاب **لسيبويه** مدوّنة نحوية، والبحر المحيط لأبي حيّان الأندلسي مدونة تفسيرية...الخ.

فرغم الاختلاف بين المدونات اللغوية المرقمنة والمدونات اللغوية التراثية الورقية من حيث الكم والكيف وآلية المعالجة؛ بمعنى أنّ المدونات اللغوية المرقمنة كمّها

1

كبير، ومتنوعة النصوص، وتستعين بالتطبيقات الحاسوبية في البحث، إلا أنّ كلاهما يعدّ مصدرا للبحث اللغوي. وللعلم أنّ أنّ هناك كتاب موسوم، بـ: **المدوّنة**: وهو مجموعة أحكام فقهية للإمام مالك (ت:179هـ)، التي جمعها سحنون (ت:240هـ).

- **من آليّات التهيئة الحاسوبية للمدوّنة اللغوية:**

يلزم لتهيئة النصوص حاسوبيّا جعلها في صياغة رسمية، بحيث يكون لديها قابليّة التقنيّة للتعامل مع الأساليب البرمجية المختلفة للمعالجة الآلية للغة الطبيعية؛ وذلك تمهيدا لتوظيفها في التطبيقات[1] التي تناظر الأداء الإنساني. وتمر تهيئة النص اللغوي آليّا بعدة مراحل، وهي:[2]

1- **تحرير النص Text Editing**: يقصد به تحويل البيانات النصية المتناثرة على الشبكة العنكبوتية أو على صفحات الويب إلى بيانات نصية منتظمة في ملفات نصيّة **Text Document**، ليسهل التعامل معها بالتعديل أو الحذف. ويتم ذلك باستخدام المحررات النصية مثل **Notepad** أو **++Notepad**

2- **حذف المسافات الزائدة Remove Spaces**: تقتضي المعالجة الآلية للغة الطبيعية حذف المسافات الزائدة في أول السطر أو في نهايته أو بين الكلمات

[1] هناك فرق بين التطبيقات والبرامج، فالبرامج هي برامج النظام صممت من أجل جهاز الحاسوب والعمل على تشغيله، وتعمل أيضًا على التنسيق بين مكونات الحاسوب والنظام، فهي تعمل من تلقاء نفسها وبمجرد تشغيل جهاز الحاسوب وتبقى مستمرة طوال مدة تشغيل النظام، ومن أهم برامج الحاسوب؛ نظام التشغيل Windows فالبرامج لا تحتاج إلى التطبيقات لتشغيلها، أما التطبيقات فإنها لا تعمل دون برامج، ومن أمثلة التطبيقات معالجات النصوص كتطبيق مايكروسوفت وورد، ومثل تطبيق جداول البيانات Excel ومتصفحات الويب مثل فايرفوكس وغوغل كروم.

[2]

الناتجة عن عدم الانضباط في عملية إدخال النصوص؛ لأن المسافة الزائدة تعد حرفا زائدا، أو كلمة زائدة.

3- **توحيد علامات الترقيم Punctuation Normalization:** يلزم توحيد الرموز المتماثلة في الشكل مثل علامات الترقيم؛ حتى يتمكن الباحثون في معالجة اللغة العربية آليا -لا سيما في بناء النماذج الإحصائية للغة الطبيعية– من حد تناثر البيانات، حيث تتداخل علامات الترقيم اللاتينية (? , ;) مع علامات الترقيم العربية (؟ ، ؛) أثناء إدخال النصوص العربية إلى الحاسوب نظرا للتشابه الشكلي بين هذه العلامات.

4- **توحيد الأرقام Numbers Normalization:** إن المزج بين الأرقام الهندية والأرقام الإنجليزية التي أصلها عربي في النصوص العربية يمثل تحدّيا كبيرا في معالجة اللغة العربية آليا؛ لذا التزمت الدراسة بتوحيد الأرقام بصيغة الأعداد الهندية[1]، حيث إن الأعداد الهندية أكثر ملاءمة من حيث الشكل وطابعها اليميني في الكتابة. وقد تم توحيد الأرقام في نصوص مادة المدونة اللغوية من خلال عملية استبدال الأرقام الهندية بالأرقام الإنجليزية الواردة في النصوص.

5- **إزالة الكشيدة Tatweel removal:** وهي زائدة تضاف بين حروف الكلمة، بغرض مساواة النص في الخط العربي، وأحيانا تضاف في النص دون فائدة، إلا أن وجودها في النص يمثل تحديّا كبيرا في المعالجة الآلية اللغة العربية؛ لأنها تؤثر على شكل الكلمة أثناء المعالجة؛ وتظهر الكشيدة (التطويل بين الحروف) بهذا الشكل (مـلك)، وتعددها من حرف لآخر.

[1] الأرقام الهندية، وهي الرموز التالية: (٠ - ١ - ٢ - ٣ - ٤ - ٥ - ٦ - ٧ - ٨ – ٩).

6- **تشفير النصوص Text Encoding:** يقصد بها تحديد أكواد ثابتة لجميع الحروف، وعلامات التشكيل الأساسية، وعلامات الترقيم، والأرقام، والرموز المستخدمة في النص. وقد ظهرت مؤخّرا نظم تشفير موحدة تدعم العديد من ألفبائيات اللغات الطبيعية، وأهمها نظام التشفير العالمي الموحّد **Unicod** يحوي 65536 حرف.

7- **التمثيل الكتابي Orthographic Transliteration (النقحرة: النقل الحرفي):** هو عملية نقل هجائي من لغة ما إلى هجاء لغة أخرى، وفقا لمعيار أنظمة كتابتها.أي محاولة للتوسط بين المنطوق والمكتوب، في معالجة اللغة العربية آليا؛ تجنبا للتحديات الناتجة عن ترميز اليونيكود **Unicode**. وتستخدم عدة أنظمة للتمثيل الكتابي في معالجة اللغة العربية آليا، أشهرها نظام باكولتر الكتابي **(للمطور البريطاني أندرو روبرتس)** الذي يتبع الترميز المعياري للحروف العربية، بحيث يقابلها أي الحروف والعلامات العربية رموز في أغلبها إنجليزية؛ لتكون أكثر موثوقية في اكتشاف أخطاء ترميز اليونيكود، نحو : أ مقابلها alef hamza above.

8- **إعادة تسمية الملفات Files Rename:** وتتجلى أهمية توحيد صياغة تسمية الملفات؛ لتهيئتها للعتاد البرمجي، بحيث تسهل قراءتها لدى أنظمة التشغيل المختلفة أثناء المعالجة الآلية، ويتم ذلك ببرامج، ليتم توحيد تسمية عدة ملفات في آن واحد.

- **نماذج من المدوّنات اللغوية العربية :**

من أشهر المدونات اللغوية العربية المحوسبة التي لها صفحة رئيسية على الويب نجد المدوّنات اللغوية، التالية :

1- **مدوّنة صخر:** كانت في بداية تأسيسها ضمن مجموعة العالمية للإلكترونيات سنة 1982. وكان من أهدافها تطوير اللغة العربية ودعمها، لتوائم العصر من تكنولوجيا المعلومات، وقد استثمرته **مدوّنة الرياض** السعودية التي لها موقعا في الويب.

2- **المدوّنة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية:**[1]

وهي من أكبر المدوّنات اللغوية العربية حيث يتجاوز حجم المدونة **المليار** كلمة، نشأت سنة **2012**، بحيث تغطي **481** موضوع في الدراسات اللغوية بمستوياتها المتنوعة، وتعتمد على خمسة ركائز أساسية في اختيار نصوص المدوّنة وهي:البعد الزمني، والبعد الجغرافي، والوعاء المعلوماتي، والمجال المعرفي، والتصنيف الموضوعي.

3- **المدوّنة اللغوية التاريخية للجامعة الأردنية :**

تهدف هذه المدونة إلى خدمة علماء اللغة ومتعلمي العربية بحيث يمكنهم استكشاف وفهم الاستعمال اللغوي وتطوره، والتحقق من التغير الدلالي عبر المراحل الزمنية المختلفة للأدب العربي من نثر وشعر وتاريخ وفلسفة ودين وعلوم ومعاجم، ويبلغ حجم هذه المدونة **45** مليون هيكل كلمة من مختلف العصور التاريخية للأدب العربي تمتد لأكثر من ستة عشر قرنا من الإستعمال اللغوي منذ العصر الجاهلي الأوّل إلى عصرنا هذا.

1

4- المدوّنة اللغوية العربية الدولية لمكتبة الإسكندرية:[1]

تتاح هذه المدونة اللغوية على الشبكة العنكبوتية تحت اسم المدونة اللغوية العربية العالمية، وهي تمثل أحد المشروعات الثقافية التابعة لمكتبة الإسكندرية الهادفة لبناء مدونة لغوية للعربية المعاصرة تحوي **100** مليون هيكل كلمة محللة صرفيا ونحويا ودلاليا، وممثلة القطاع إقليمي كبير من الدول الناطقة باللغة العربية المعاصرة، وعاكسة لأنماط استخدام اللغة العربية المعاصرة في العالم العربي.

5- مدونة عربي كوربص arabiCorpus :

تتيح هذه المدونة اللغوية إمكانية استرجاع الكلمات والعبارات وفقا لتكرار ترددها في عدد من الفئات. وتضم هذه الفئات خمسة أنواع أدبية رئيسة، هي: الصحف، الأدب الحديث، الأدب غير القصصي، العامية المصرية، الأدب قبل العصر الحديث .

6- المدونة العربية القرآنية :

وهي مدوّنة توضح اللغة العربية في ضوء قواعد النحو والصرف والأنطولوجيا الدلالية (**لتمثيل المعرفة**) لكل كلمة من كلمات القرآن الكريم.

7- مدوّنة المعجم التاريخي للغة العربية:

وهي مدوّنة معجمية لغوية محوسبة انطلقت في قطر في ديسمبر 2018، "**البوابة الإلكترونيّة لمعجم الدوحة التاريخيّ للغة العربيّة**"، تضمُّ كمّا وافيًا من النصوص التي تعكس واقع اللغة العربية في بيئاتها ومراكزها الثقافية والعلمية والحضارية التي شهدت نموّها وتطوُّر دلالات ألفاظها وتراكيبها.

فهي ديوان معجمي للعربية يُؤَرِّخ لألفاظها ومعانيها، ويُبَيِّن ما طرأ على تلك الألفاظ من تَحوُّل وتغيُّر، عبر تاريخها المديد، ما بين **480** قبل الهجرة إلى العام **1431** هجري،

1

وتستمدّ مادّتها من التراث العربي المكتوب عبرَ المراحل الزَّمنيَّة المُتعاقبة للُّغة العربيَّة، وتضمّ ما يزيد على 116 مليون كلمة مجمّعة في ثمانمائة وتسع وستين (869) وثيقة.[1]

- **مشروع الذخيرة اللغوية العربية :**

لقد عَرَضَ **عبد الرحمن الحاج صالح** : هذا المشروع في مؤتمر التعريب في عمان سنة **1986**، وهو عبارة عن بنك آلي من النصوص القديمة والحديثة ومن الجاهلية إلى وقتنا الحاضر، ويقوم تصوّر مشروع "**الذخيرة العربية**" على إدراج أجود الإنتاج العلمي العربي القديم والمعاصر، والإنتاج العلمي العالمي بعد ترجمته إلى العربية، في بنك آلي محوْسب يمكن لأي قارئ عربي أن ينهل منه في أي علم أو ميدان، بعد فتح موقع "الذخيرة" للقراء على الإنترنت.

إذن؛ فالهدف الرئيسي للذخيرة وكما يرى **الحاج صالح** أنّه يساعد العلماء والباحثين والأساتذة وحتى الأطفال من العثور على معلومات شتى من واقع استعمال العربية بكيفية آلية وفي وقت وجيز، علماً بأن المشروع كان في بدايته لغوياً لسانياً ثم توسع ليصبح ثقافياً علمياً شاملاً. ويرى كذلك **الحاج صالح** أنّ مشروع الذخيرة اللغوية هو "**نوع من التعريب للعلم والمعرفة**" وخاصة في ظل الترجمة الآلية الفورية اليومية، لكنه لم يتحقّق.

- **أنواع المدونات اللغوية:** يذكر **سنكلير** خمسة أنواع من المدونات، وهي:[2]

1

2

1- **المدوّنات اللغوية المرجعيّة Reference Corpora** : هي المدونات التي تُصمم بحيث تعطينا معلومات مفصلة بقدر الإمكان عن استخدامات اللغة، ويتحقق هذا من خلال احتوائها على عدد كبير من النصوص، بحيث تظهر صور التنوع اللغوي ذي العلاقة والمفردات المميزة له بشكل واضح.

2- **المدوّنات اللغوية الراصدة Monitor Corpora**: وهي مدونات متابعة، بحيث تُنقح بصفة مستمرة، لتعطي صورة عن التغيرات اللسانية التي قد تطرأ على استخدام اللغة.

3- **المدوّنات اللغوية المقارنة Comparable Corpora**: وهي مجموعات متشابهة من النصوص من لغات عدة أو من لغة واحدة مثل التعليقات السياسية حول قضية ما بين مختلف الصحافات، أو مقارنة بين لهجتين للغة واحدة.

4- **المدوّنات اللغوية المتوازية Parallel Corpora**: تشتمل على مجموعة من النصوص المتماثلة بلغتين مختلفين (مثلا أحد النصين ترجمة للنص من لغة أخرى).

5- **المدوّنات اللغوية المتخصصة Technical Corpora**: وهي مجموعة من النصوص ذات طابع محدد، مثل المقالات العلمية في مجال الفيزياء أو القانون.

- **الأسئلة التي يمكن أن تجيب عنها المدوّنات اللغوية**؛ وهي:[1]

1- ما أكثر الكلمات أو العبارات تردّدا ؟

[1]

2- ما أوجه الاختلاف بين النصوص المكتوبة والنصوص المنطوقة ؟

3- ما الأفعال، أو الأسماء، أو الحروف التي يستخدمها أهل اللغة أو أهل التخصص أكثر من غيرها ؟

4- ما حروف الجر أو الأفعال، أو الأسماء التي تسبق أو تلي كلمة بعينها ؟

5- كيف يستخدم أهل اللغة أو أهل التخصص كلمة أو مصطلحا معيّنا ؟

6- كم مرة تستخدم فيها التعبيرات الاصطلاحية بين أهل اللغة أو أهل تخصص ما ؟

- **الأسئلة التي لا يمكن أن تجيب عنها المدوّنات اللغوية**، وهي:[1]

1- ما هي البراهين أو الأدلة السلبية حول استعمال كلمة، أو مصطلح، أو عبارة معيّنة ؟

2- لماذا هذه الظاهرة اللغوية منتشرة في لغة الصحافة مثلا ؟

3- ما كافة الاستخدامات الممكنة لمصطلح أو كلمة أو عبارة في اللغة على إطلاقها ؟

فهذه أهم الأسئلة التي لا تستطيع المدوّنات اللغوية أن تجيب عليها، والإجابة عنها تتم عن طريق أهل اللغة أو أهل التخصص أنفسهم.

- **مزايا المدوّنات اللغوية:** إن المدونات اللغوية هي مقاربة منهجية **approach** الكترونية وبفضل البرمجيات التي صُممت لمعالجتها، ومن مزاياها مايلي:[2]

1- إنها عملية وتجريبية **empirical**، مبنية على نصوص حقيقية للاستعمال اللغوي، وليس على الحدس الشخصي، وتدرس نماذج واقعية للغة.

1

2

2- التنوع المبنى على أسس علميّة لنصوص المدوّنة لتمثل استخدامات اللغة المختلفة، وذلك بمراعاة التمثيل الأفقي (الجغرافي) والعمودي (التاريخي) والنوعي (الأسلوبي مثلا) للغة واستعمالاتها المختلفة.

3- تساعد الباحث اللغوي في الوصف والتحليل والإحصاء في شتى حقول المعرفة اللسانية في ضوء النصوص المتاحة في المدوّنة المحوسبة .

ومهما يكن؛ فإنّ المدونة اللغوية هي كتلة من النصوص المرقمنة التي تستخدم لدراسة جوانب اللغة، يمكن قراءتها والتعامل معها آليا والتحكم في بياناتها ومدخلاتها، بالحذف أو التعديل من خلال قواعد بيانات (**Databases**) التي صُممت للتعامل مع هذه النصوص. وتعد قاعدة البيانات الحاوية لنصوص المدونة اللغوية مخزنا كبيرا للغة، يرجع إليه وقت الحاجة، ويتحمل أي قدر من النصوص التي تضاف إلى المادة الأساسية مستقبلا.[1]

- **الذكاء الاصطناعي (المفهوم، النشأة، التطبيقات):**

يعتبر الذكاء الإصطناعي **Artificial intelligence** الجيل الخامس من برمجيات الحاسوب والذي رافقه ظهور كثير من لغات البرمجة الخاصة به، والهدف من برامج الذكاء هو مساعدة الخبير في أداء عمله بكفاءة متميزة، وهو عبارة عن برامج وأجهزة تتعاون لتؤدي عملية فهم معقدة يمكن أن تضاهي ذكاء البشر من فهم وسمع ورؤية وشم وكلام وتفكير، أي أنه برامج ذكية + أجهزة = ذكاء إصطناعي. أما العلوم الواجب دراستها

1

قبل الذكاء الاصطناعي هي الرياضيات المنطيقية+ لغات البرمجة + حقول معرفية أخرى.[1]

بدأ موضوع الذكاء الإصطناعي سنة 1947 مع الرياضي البريطاني **آلان تورينج** حين حدد بحوثه أن الذكاء الإصطناعي هو عمل برامج ذكية وليس بناء آلات ذكية. أما الميلاد الحقيقي لعلم الذكاء الإصطناعي فكان في مؤتمر دارتموث سنة **1956**، حول "**ذكاء الآلة**" وتم تبني مصطلح "الذكاء الاصطناعي" رسميا من طرف مجموعة علماء الكومبيوتر ومنذ ذلك الحين أصبح الذكاء الاصطناعي يتطوّر في شتى حقول المعرفة وببشر بمستقبل تكنولوجي مشرق للحضارة الإنسانية، وخير دليل تأسيس شركة علي بابا الصينية العملاقة سنة 1999 التي تبحث في شتى مجالات الذكاء الاصطناعي إلى يومنا هذا.[2]

2

- **التطبيقات اللغوية للذكاء الاصطناعي:**[1]

من أشهر التطبيقات المختلفة للذكاء الاصطناعي نجد تطبيقات **رؤية الكمبيوتر** وهي تتعامل التي مع المرئيَّات مثل الصور والفيديوهات، وتطبيقات **النظم الخبيرة** وهي أحد أقوى فروع الذكاء الاصطناعي وتستخدم غالبًا فى عالم المال والطب والتسويق، وتطبيقات **معالجة الصور والأشكال**، وتطبيقات **الألعاب**، وتطبيقات **التعليم**، وتطبيقات **تلخيص الأخبار**، وتطبيقات **الروبوتات** كتطبيق قياس درجة الحرارة...أما التطبيقات الذكاء الاصطناعي التي هي في خدمة اللغة، فمن أهمها مايلي:

1- **تمييز الكلام speech recognition** : هي برامج تستطيع تحويل الأصوات إلى كلمات text على الحاسوب، وهناك برامج تمكن المستخدم من توجيه أوامر وجمل للحاسوب (**السكرتير الآلي**)، بعض الأماكن ذات الوضع الأمني المميز تستخدم الصوت للتعرف على الموظفين أو العملاء في البنوك.

2- **صناعة الكلام speech synthesis** : هي برامج تستطيع تحويل الكلمات والجمل المكتوبة text إلى أصوات. وهناك برامج تمكن المستخدم من قراءة الجمل وترجمتها وهي تفيد جميع المستخدمين وخصوصا ذوي الإعاقة البصرية أو اليدوية.

3- **تمييز وقراءة الحروف character recognition** : هي برامج تستطيع قراءة حروف وكلمات مكتوبة باليد أو مطبوعة وتحويلها إلى حروف وكلمات وجمل على الحاسوب text بعد ذلك نستطيع استخدام هذا النص كما لو قد أدخلناه من لوحة المفاتيح.

1

4- **فهم اللغات الطبيعية natural language understanding** : برامج تكمن الحاسوب من فهم لغة طبيعية مكتوبة text مثل اللغة العربية أو أي لغة أخرى في مجال تطبيق معين. ونعني بالفهم هنا هو التعرف أولا على التركيب النحوي للجمل وموقع كل كلمة من الإعراب ثم فهم معنى الجملة والرد عليها سواء بإضافة معلومة جديدة إلى قاعدة المعرفة أو استخراج معلومة معينة مطلوبة من قاعدة المعرفة أو التحقق من صحة معلومة من عدمه. مثال ذلك نظام **AQAS** .

- **مفهوم نظام AQAS: arabic question answering system** أي نظام استعلام باللغة العربية الفصحى: وهو نظام مبني على المعرفة يقبل جملة مُدْخَلَة باللغة العربية. فإذا كانت الجملة المدخلة جملة خبرية تعلّم منها النظام وإن كانت سؤالا أنتج له النظام إجابة مناسبة.
- **مكونات نظام AQAS**: ويتكون من الأجزاء التالية:[1]
  - **معرب الجمل :** يقوم بمعالجة الجملة المدخلة بالتشكيل والإعراب.
  - **فاهم الجمل :** يقوم بفهم الجملة المدخلة وتصنيفها خبر أم سؤال.
  - **منتج الإجابة :** يقوم بالإجابة عن الأسئلة.
  - **قاعدة المعرفة :** وهي ذاكرة نظام **AQAS** وبياناته ليتمكن من تمثيل المعرفة.
  - **القاموس :** يقوم بالاحتفاظ بخصائص كل كلمة لتستخدم في مكانها المناسب في المعالجة.
  - **تمثيل المعنى الداخلي :** يقوم بالتنسيق الداخلي بين الجمل المدخلة.

[1]