البرمجة اللغوية وعلاقتها بالمعالجة الآلية :التنظيم والخصائص

مرّت الإنسانيّة بثلاث ثورات كبرى غيّرت مجرى حياتها وأعادت تشكّل الفكر البشريّ هي أوّلا : ثورة اكتشاف الكتابة في بلاد الرّافدين ومصر تلك التّي حدّدت بداية التّاريخ. وهي ثورة فكريّة لا يمكن إدراك تأثيرها في بنية الفكر البشريّ وتطوّره إلاّ بمقارنة الشّعوب التّي تكتب، بالشّعوب التّي لم تعرف الكتابة.
أمّا الثّورة الثّانية فكانت ثورة تصنيع الكتابة واكتشاف يوهان غوتمبرغ (**Johannes Gensfleisch**) الآلة الطّابعة التّي دعمت نهضة أوروبا. ثمّ ابتكار الطّابعة الشّخصيّة (أو ما يعرف بالآلة الرّاقنة) .وهي نقلة نوعيّة في نشر النّصوص المكتوبة تحوّل بموجبها كلّ فرد إلى ناشر مفترض.
وتجسّد الآلة الرّاقنة المرحلة الثّانية في المعالجة الآليّة للنّصوص العربيّة بعد دخول الطّباعة الوطن العربيّ بوساطة العثمانيّين. أمّا الثّورة الكبرى الثّالثة فهي المعالجة الإعلاميّة للكتابة.

المعالجة الآليّة وتحرير النّصوص:

عند بداية انتشار الحاسوب الشّخصيّ، كان مصطلح «المعالجة الآليّة للنّصوص» يشير إلى البرامج الحاسوبيّة الأولى لرقن النّصوص (word processor) وتحديدها وضبط بنط الخطّ ومظهره والفقرات والهوامش وما إلى ذلك . وتلك البرامج هي التّي أصبحت تسمّى اليوم برامج «تحرير النّصوص».
كانت برامج تحرير النّصوص كثيرة في الأوّل .ولم يثبت منها إلاّ القليل بعد مرور بضع سنوات، مثل وورد لماكروسوفت (Microsoft Word) .وكانت تلك البرامج تجسّد عندها ثورة تكنولوجيّة في الكتابة إذ كانت ،بعد الآلة الراقنة، أحدث قفزة نوعيّة تحقّقت في مجال الرّقن الذي كان يثقل كاهل المستعمل في حال وقوع أخطاء تجبره على إعادة رقن الصّفحة بكاملها وعلى إعادة النصّ بأكمله في حال إضافة شيء جديد إليه ، على حين تحوّل النّصّ المرقون من كتلة جامدة إلى جسم سائل يمكن تمطيطه وتغيير مواقع فقراته أو حذف صفحات كاملة فيه وإضافتها دون حاجة إلى إعادة طباعة ما

رُقن كلّه من جديد. و يتيح تحرير النّصوص أيضا إمكان تخزين ما يكتب مرّة واحدة لكي يعاد استعماله إلى ما لا نهاية له بفضل تقنية النّسخ واللّصق وسهولة البحث فيه عن الموادّ.
ورغم الإيجابيّات التّي لا تحصى لبرامج المعالجة الآليّة والتّي جعلت منها ضرورة لا يمكن تجاوزها، فإنّ بعض السّلبيّات غير المنتظرة برزت للعيان. فقد أصبحت البشريّة ولأوّل مرّة في التّاريخ في حاجة إلى **آلة للقراءة !** فالنصّ المكتوب على الحامل الألكترونيّ ، سواء كان القرص المضغوط أو قرص الحاسوب نفسه غير متاح مباشرة ، كما يفترض أن تكون حوامل الكتابة في الماضي، حين كان المكتوب في متناول القارئ مباشرة. فمن غير الآلة، ليس القرص المضغوط سوى صفيحة بكماء صمّاء.

توسّع مفهوم المعالجة الآليّة:

بعد تعدّد الاكتشافات والبرامج وتنوّعها في ميدان الحاسوب الشّخصيّ، أصبحت فضفاضة تضمّ ، كما هو الحال في الوقت الرّاهن ، مجموعة كبيرة من التّطبيقات. فقد توسّع المفهوم ليشمل عدّة عمليّات تتراوح بين تحرير النّصوص ورقمنتها وإدخالها وتحليلها وكذلك إنتاجها. وتنقسم البرامج إلى ثلاثة أقسام رئيسة هي : **برامج إدخال المعطيات** (تحرير النّصوص والنّشر المكتبيّ والمسح الضّوئيّ والقراءة الضوئيّة وتعرّف الأصوات وما إلى ذلك) **وبرامج تحليل المعطيات وتخزينها أو تحويلها** (الأرشيف والمحلّلات والمدوّنات، والتشكيل، والمدقّق الإملائيّ الخ) و**برامج إخراج المعطيات** (إنتاج الأصوات، القراءة الآليّة، والكتابة المستعانة).

1- برامج إدخال المعطيات:

- **برامج النّشر المكتبيّ:** تطوّرت برامج تحرير النّصوص إلى أن انقسمت قسمين: قسم مبسّط يضاهي الآلة الرّاقنة ويضطلع بالعمليّات المبتذلة لإدخال النّصوص التّي لا تتجاوز بضع صفحات ولا يهمّها شكلها الخارجيّ وعرضها وقسم يعالج إخراج النّصوص وتنضيدها بدقّة متناهية لإعدادها للطّبع المحترف .وهو يهمّ برامج النّشر المكتبيّ (Desktop Publishing Software) وتستعمله خاصّة الجرائد

اليوميّة والمجلاّت والكتب العلميّة التّي تسعى للمحافظة على دقّة إخراج منشوراتها.

- **المسح الضّوئيّ والقارئ الآليّ:** تجري عمليّة التعرّف الضّوئيّ على نصّ ([OCR] Optical Character Recognition) محوْسب في صيغة الوثيقة المتبادلة (امتداد .pdf) أو بعد المسح الضّوئيّ (scanning) لإدخال ذلك النّصّ. وبعد أن يقرأ البرنامج النّص، يحلّل أشكال الصّور ويتعرّف إلى الحروف تلقائيّا أو يتعلّمها إذا كان المستعمل يرغب في جودة أفضل. ثمّ يحوّل البرنامج النّتيجة إلى ملفّ ألكترونيّ يحدّد شكلَه المستعمل ويكون في الغالب نصّا يمكن لبرامج تحرير النّصوص إعادة معالجته.

ومن مشكلات القراءة الضّوئيّة للنّصوص عامّة وللنّصوص العربيّة خاصّة نسبة الأخطاء حتّى لو كانت منخفضة في الظّاهر. وهذا يهمّ بالخصوص القارئ الآليّ العربيّ الذي يسهم بنسبة 96 ٪ من نسبة النّجاح بعد التعلّم .فيبدو للمرء أنّها نسبة محترمة. لكنّ عدد الأخطاء مرتفع جدّا إذا علمنا أنّ السّطر يحوي أكثر من مائة وحدة .وهو ما يعني حصول أكثر من عشرة أخطاء في كلّ سطر. ومن مشكلات القراءة الضّوئيّة للنّصوص أيضا تلك المتعلّقة بالكتابة العربيّة وأشكال حروفها وتداخلها، إذ تكتب الحروف متّصلة ، خلافا للحروف اللاّتينيّة ، فضلاً عن استعمال العربيّة المشبّكات الحرفيّة كحرف اللاّم- ألف «لا» واللاّم- ميم «ﻟﻤ» وغيرها وكذلك الحروف المركّبة في شكل عموديّ ، كما تظهر ذلك كلمة «ﻟﻤ» التي لا تلتزم حروفها بالسّطر القاعديّ.

وتتوفّر للمستعمل العربيّ بعض البرامج المهمّة في هذا المجال كبرنامج ريديريس (Readiris) الذي تنتجة شركةI.R.I.S. الأمريكيّة وبرنامج تسّيراكت (Tesseract) الذي تنتجه هولت بكارد (Hewlett Packard) .أمّا البرامج التّي أنتجت في الدّول العربيّة فنذكر منها برنامج القارئ الآليّ الذي تنتجه شركة صخر والبرامج التّي أنجزتها مخابر البحوث في بعض الأقطار كمصر وتونس والسّعوديّة .لكنّها لم تنتشر لعدم قدرتها على المنافسة.

وثمّة دراسات وأبحاث تختَبر حاليا لمعالجة الكتابة اليدويّة وتعرّفها آليا.

هذا في ما يتعلّق بالبرامج التّي نعالج المكتوب لإدخاله الذاكرة الحاسوبيّة. وهناك برامج تعالج المنطوق وتحوّله إلى مكتوب.
- برامج التّعرّف إلى الأصوات (انظرفصل : "المعالجة الآليّة للأصوات")

2- برامج تحليل المعطيات وتخزينها أو تحويلها:

أ- **برامج تحليل المعطيات** :

- **المدقّق الإملائيّ**: يحلّل هذا البرنامج النّصوص كلمة بعد أخرى إذا كان يعتمد على محلّل نحويّ أو يقارنها بقاعدة بيانات محفوظة في ذاكرته لكلّ الكلمات العربيّة الممكنة .وإذا كان يشير إلى الكلمات التّي لا «يعرفها» باعتبارها خاطئة أو محلّ شكّ : فإمّا أن يقترح البديل أو يترك الخيار للمستعمل في استبدالها أو المحافظة عليها وإدخالها في قاموس المستعمل.

نظريّا، يمكن أن يحلّل المدقّق الإملائيّ النّصوص المشكولة أو غير المشكولة، إلاّ أنّ مدقّقات النّصوص المشكولة لم تر النّور لأنّها تفترض محلّلات نحويّة جيّدة لا تعرفها الصّناعة اللغّويّة إلى اليوم. بل إنّ طريقة العرب في الكتابة (عدم رسم الهمزة وإهمال الشدّة، مثلا) تجبر برامج التّدقيق الإملائيّ على إتاحة الاختيار بين إمكان التّدقيق الإملائيّ الصّارم والتّدقيق الإملائيّ المتساهل.تساهل.

ولا توجد حسب علمنا مدقّقات إملائيّة عربيّة مستقلّة عدا برنامج «هنسبيل» (Hunspell) المجَريّ الذي يمكن استعماله على حواسيب أبل وويندوز ولينوكس، إذ عادة ما تدمج في برامج تحرير النّصوص (على غرار وورد (Word) لمايكروسوفت أو وينتكست (Wintext) في نظام أبل) أو في برامج النّشر المكتبيّ كانديزاين (InDesign)، مثلا.

-**المحلّلات النّحويّة**: تنقسم المحلّلات النّحويّة إلى محلّلات صرفيّة ومحلّلات إعرابيّة أو تركيبيّة.

تضطلع برامج **التّحليل الصّرفيّ** بتعرّف مكوّنات الكلمات مثل جذورها وجذوعها وتفكيكها إلى سوابق ولواحق وأحشاء. وتستفيد هذه البرامج

كثيرا من الشّكل في عمليّة التّحليل لأنّ ذلك يساعد على تحديد أيّ قسم من من أقسام الكلام تنتمي إليه الوحدة المحلّلة.
وتوجد في الوطن العربيّ بعض النّماذج من هذه البرامج أفضلها برنامج «الخليل للتّحليل الصّرفيّ» الذي تنتجه مدينة الملك عبد العزيز للعلوم بالسّعوديّة (kacst) وكذلك برنامج «Stemmer» (لشيرين خوجة) .وهو يسمح بتعرّف جذور الكلمات ومن ثمّة تحديد الزوائد. ونذكر كذلك «AraMorph» الذي طوّره مجمّع البيانات اللّسانيّة (Linguistic Data Consortium) ويُعرف باسم «Buckwalter Arabic Morphological Analyzer». وهو يقوم بما يقوم به البرنامجان اللّذان سبق ذكرهما.
أمّا برامج التّحليل الإعرابيّ فوظيفتها التّعرّف الآليّ إلى أقسام الكلام بالنّسبة إلى كلّ وحدة معجميّة في النّصّ. ولا يتعلّق الأمر بأقسام الكلام الثّلاثيّة ، كما يعرفها العرب (اسم وفعل وحرف) بل يتجاوزها التّحليل إلى تحديد العدد (جمع أو مثنّى أو مفرد) والجنس (مؤنّث أو مذكّر أو ملتبس) وكذلك زمن الفعل وصيغته.
وقد اشتهر برنامج «ستنفورد لوسم أجزاء الكلام» (Stanford parser). ومن المحلّلات الإعرابيّة المعروفة في الوطن العربيّ برنامج «الخليل للتعرّف إلى أقسام الكلام» الذي أنجزته جامعة محمّد الأوّل بالمغرب. وثمّة برامج كثيرة معروضة للاستعمال أو هي بصدد الإعداد مثل برنامج «أميرة» (Amira) الذي ينتجه باحثان (منى ذياب وياسين بن عجيبة) من جامعة كولومبيا، كما نذكر «كلمسوفت» (kalmasoft) الذي يبشّر بكل خير.
-**المحلّلات المصطلحيّة** (Term extraction) **:** تعالج هذه البرامج النّصوص للتّعرّف إلى المصطلحات انطلاقا من خاصيّاتها التّركيبيّة في الجملة أو من مؤشّرات أخرى كورودها بين ظفرين أو بالبنط الغامق أو بحرف بارز. وعادة ما يتطلّب العمل الآليّ تدخّل المستعمل لتحديد الألفاظ المرشّحة للمعالجة أو لمراجعة النّتائج للتثبّت منها. ولا نعرف أيّ برنامج أنجز لهذا الغرض في الوطن العربيّ.
-**المفهرس الآلي** : (Concordancer) تحلّل برامج الفهرسة الآليّة النّصوص والمدوّنات الضّخمة للبحث عن كلمات أو جذوع أو جذور يحتاجها المستخدم وتُقدّم في شكل سياقات دنيا (معدّل خمس كلمات قبلها

وخمس بعدها). وتكمن أهميّة الفهرسة الآليّة في تيسير أعمال البحث العلميّ في مختلف إمكانات تراكيب الألفاظ وسياقاتها وقيس مدى شيوعها في الاستعمال حسب نسبة ورودها. ويمكن ملاحظة تاريخ ظهورها واختفائها أو بروزها وانكفائها. لذلك تعدّ هذه البرامج ضروريّة في البحث العلميّ المتعلّق بالمعجميّة والمصطلحيّة وكذلك في اللّسانيات التّاريخّة والمعجم التّاريخيّ على وجه الخصوص.
وتوجد عدّة برامج للفهرسة الآليّة نذكر منها «الخليل للفهرسة الآليّة» الذي أنتجته مدينة الملك عبد العزيز للعلوم والتّقنية بالتعاون مع منظمة الألكسو، وكذلك «أرابيكربوس» (ArabiCorpus) الذي طوّره أستاذ العربيّة ديلوورث بركنسون (Dilworth Parkinson) من جامعة بريغام (Brigham University). ومن خاصّياته أنّ مدوّنته تحوي نصوصا قديمة ونصوصا حديثة مستخرجة من الكتب والجرائد والمجلاّت. ولا ننسى كذلك المفهرس الآليّ «aConCorde» الذي أنتجه أندي روبرتس (Andy Roberts) من جامعة ليدز الإنجليزيّة. ونذكر كذلك «Concordance» المجانيّ الذي طوّرته مجموعة وات (R.J.C. Watt) البريطانيّة ، كما نشير إلى المفهرس الآليّ «Concapp» الذي وضعته مجموعة غريفز (Chris Greaves) الكنديّة.
ب- **برامج تخزين المعطيات** : تتفرّع برامج تخزين المعطيات حسب وظيفتها وطبيعة المعطيات المخزّنة. وتكون في شكل مدوّنات **و**قواعد بيانات وبنوك أرشيف ومكتبات رقميّة.
-**المدوّنات وقواعد البيانات:** ظهرت أوّل مدوّنة لغويّة في ستّينات القرن العشرين في الولايات المتّحدة .وقد وضعتها جامعة براون .فسمّيت باسمها «مدوّنة براون» (Brown Corpus). ولئن كان حجمها يعتبر اليوم هزيلا (يتجاوز معدّل المدونات الكبرى الخمسمائة مليون كلمة) فإنها كانت تعتبر مدوّنة ضخمة في ذلك الوقت باعتبارها تحوي أكثر من مليون كلمة.
تنتمي مدوّنات اللّغة إلى مجال البحث في ما يسمّى بلسانيات المدوّنة ، لأنّها توفّر بنك المعطيات الذي تُستخرج منه الشّواهد ، كما هي مستعملة في الواق وتجرى عليها الإحصائيات وتستنتج منها التوّجهات العامّة لبعض المظاهر اللّغويّة. لذلك لا تبنى المدوّنات دون هدف محدد

يبرّر إنشاءها ، كما تتفرّع المدوّنات وتصنّف حسب محتوياتها والمجالات المدوّنة فيها وعدد اللّغات المستعملة أو شكل بنائها أو طريقة اقتنائها أو حتّى إمكان إغنائها من عدمه وتاريخ أوّل نصّ مُحَوْسَب فيها.
وتُخزّن المدوّنة في شكل نصوص محوسبة تُدخل بشتّى الأشكال (نص امتداد doc أو pdf أو html أو rtf) يديرها محرّك بحث بوساطة بروتوكول يربط بينها جميعا.
وقد أدرك العرب وغيرهم في وقت مبكّر نسبيّا أهميّة المدوّنات في اللّغة العربيّة . فأنشؤوا عددا منها يراه بعضهم دون المرجوّ.
وتأتي في مقدّمتها مدوّنة مدينة الملك عبد العزيز للعلوم والتّقنية (Kacst) التّي تحوي أكثر من سبعمائة مليون كلمة. وتخزن المجلاّت والجرائد والمخطوطات والدّوريّات العلميّة. وهناك أيضا المدوّنة العربيّة المفتوحة المصدر
OSAC : Open Source Arabic Corpora التّي وضعها المعتزّ سعد من كليّة تقنيات الإعلام بجامعة غزّة الإسلاميّة. وتجمع هذه المدوّنة نصوصا نشرتها وسائل إعلام بريطانيّة (BBC) وأمريكيّة (lang="AR-SA"><) وأمريكيّة (CNN) إلى جانب صحف ومجلاّت عربيّة. ولا بدّ من ذكر مدوّنة نملار (Nemlar) التّي تحوي أكثر من نصف مليار كلمة من ميادين مختلفة وفيها جزء كبير من النّصوص المشّكولة. ويمكن أن نشير كذلك إلى موقع الفهرسة الآليّة «أرابيكربوس» (ArabiCorpus) الذي يحوي أيضا مدوّنة مهمّة.
- **بنوك الأرشيف**: تدخل بنوك الأرشيف المكوّنة من وثائق نصيّة (مخطوطات ومطبوعات مختلفة) وغير نصّية (صور أو تسجيلات) ضمن المعالجة الآليّة للنّصوص إذا كانت تلك النّصوص مُرَقْمنة بحيث يسهل البحث فيها واستدعاؤها عند الحاجة. وتختصّ المكتبات الرّقميّة بتوفير خدمات للباحثين في ميدان اللّغة على وجه الخصوص.
- **المكتبات الرّقميّة** : كان أوّل موقع عربيّ يوفّر هذه الخدمة ويضع على ذمّة القرّاء مئات الكتب المُحَوْسبة من التّراث ومن خزانة الأدب هو موقع «الورّاق» الإماراتي: «www.alwaraq.net». وهو يوفّر البحث كذلك في أمّات القواميس العربية كـ**لسان العرب** و**القاموس المحيط** و**تاج العروس** و**أساس البلاغة** و**مقاييس اللّغة** وغيرها .ثمّ

التحق به في توفير خدمات الكتب الألكترونية كلّ من «ويكي مصدر» : (http://ar.wikisource.org) ومكتبة الإسكندريّة: (www.bib-alex.com).

ج- برامج تحويل المعطيات :

- **برامج التشكيل**: تتدارك برامج تشكيل النّصوص عادة العرب في الكتابة دون شكل الكلمات. ويفترض أن تيسّر برامج التّشكيل الآليّة هذه العمليّة المُكْلفة في الوقت والجهد. لكنّ برامج التّشكيل المتوفّرة حاليّا تنقصها الدقّة لأنّها تشترط وجود محلّلات نحويّة جيّدة تعتمدها للتعرّف إلى وظائف كلّ كلمة في سياق النّص .وهو ما لم يتوفّر بعد. ومع ذلك تظل برامج التّشكيل مهمّة لاعتماد النّصوص المشكولة بكثرة في تدريس الصّغار وكذلك في برامج القراءة الآلية للنّصوص ، إذ إنّ النّصّ غير المشكول يثير في حدّ ذاته إشكالا حقيقيّا.
وتوجد في الوطن العربيّ ، على حدّ علمنا، ثلاثة برامج تسمح بالتّشكيل هي على التّوالي: برنامج «الخليل للتّشكيل الآليّ» - وقد أنتجته جامعة محمّد الأوّل بالمغرب وهو مجانيّ - وبرنامج «المشكّل الآليّ» لشركة صخر - وهو الأفضل بالنّظر إلى نسبة الأخطاء - وبرنامج «مشكال» للتّشكيل الآليّ .وهو برنامج مفتوح المصدر قابل للتّطوير لكنّ إمكاناته محدودة.

برامج الاختزال والتّرجمة الآليّة :

-**الاختزال الآليّ** : باتت التّخمة المعلوماتيّة - إذ تتطلّب قراءة ما يكتبه الإنسان اليوم في ساعة أكثر من ستّ سنوات - تشكّل هاجس ضياع المعلومة وسط هذا الكمّ الهائل من المكتوب. لذلك ظهر طلب ملحّ على البرامج التّي تلخّص آليّا محتوى آلاف الوثائق على نحو ناجع ،سريع. ويعتمد التّلخيص الآليّ على برنامج يسبق تطبيقه عمليّةَ التلخيص. وهو برنامج استخراج المعلومات الموضوعيّة وتجريدها لتركيز العناية على المعطيات المهمّة. ويبدو أنّ النّصوص تحتوي على مؤشرات تدلّ على مواقع الاستطراد وأخرى تدلّ على الموضوعات المركزيّة.

ولم نسمع عن تطبيق مّا جُرّب على العربيّة في هذا المجال.
- **الترجمة الآليّة** : انظر «باب التّرجمة الآليّة».
- **أكروبات وتبادل الوثائق:** كان الهدف من وراء إنتاج برنامج أكروبات صيغة «ب.د.ف.» (pdf) لنقل النّصوص ومختلف الوثائق بين أنظمة التّشغيل هو تخليصُ المستعمل من قيود الحروف وتجاوز اختلاف أجهزة الحاسوب وأنظمتها التّي كانت غير موحّدة وغير منمّطة وكان بعضها يتطلّب إعادة تشغيل الحاسوب لتغيير نظام التّشغيل، باعتبار العربيّة تكتب من اليمين إلى اليسار واللاّتينيّة من اليسار إلى اليمين والصينيّة من فوق إلى أسفل، الخ. وكان كلّ نظام يستعمل شكلا من الحروف ومن البرمجة المناسبة. وبفضل صيغة الوثيقة المتبادلة أصبح بإمكان المستعمل العربيّ قراءة نصّ صينيّ أو روسيّ أو عبريّ دون حاجة إلى اقتناء الحروف اللاّزمة التّي تستعصي قراءة النّص من دونها وكذلك الشأن بالنّسبة إلى المستعمل الغربيّ مع العربيّة. وبهذا أصبحت صيغة الوثيقة المتبادلة حاملا مهمّا لترويج المعطيات النّصيّة دون الخوف من إتلاف تنضيدها وأيضا صيغة لا يمكن تجنّبها لنشر المقالات والكتب الالكترونيّة.

3- برامج إخراج المعطيات:

تُدخَل المعطيات في الحاسوب من ثلاث قنوات هي لوحة الكتابة أو الفأرة أو الشّاشة إذا كانت حسّاسة للّمس أو المسح الضّوئي أو الكاميرا بالنّسبة إلى المكتوب والميكروفون أو لاقط الصّوت بالنّسبة إلى المسموع. أمّا إخراج المعطيات فيكون من الشّاشة والآلة الطّابعة بالنّسبة إلى المكتوب ومخارج الصّوت (مكبّرات صوتيّة مدمجة أو ملحقة) بالنّسبة إلى المسموع.
وبهذا تنقسم برامج إخراج المعطيات إلى برامج إخراج المنطوق وبرامج إخراج المكتوب. وقد نظرنا بعدُ في برامج إخراج المكتوب، إذ تدخل في هذا الباب كلّ برامج النّشر المكتبيّ وتحرير النّصوص العامّة أو الخاصّة كالكتابة المستعانة وكتابة السّيناريو ومقارنة النّصوص وغيرها ...بقي لنا أن ننظر في برامج إخراج الأصوات. وسنتناوله ضمن « المعالجة الآليّة للأصوات».