الفصل الثاني

الخط العربي
Arabic Script

سنناقش في هذا الفصل عناصر الخط العربي المستخدمة في كتابة اللغة العربية المعاصرة، حيث سنبدأ بوصف لغوي لعناصر الخط العربي يتبعه نقاش عن ترميز الحرف العربي في الحاسب وكيفية إدخاله وعرضه. أيضاً سنتطرق للممارسات الشائعة في معالجة اللغات الطبيعية للتعامل مع خصائص الخط العربي. وأخيراً، سنستعرض بإيجاز أربع مهام حاسوبية متعلقة بالخط العربي وهي: النقل الكتابي/النقحرة (orthographic transliteration)، وتقليص الاحتمالات الهجائية (orthographic normalization)، والتعرف على خط اليد (Handwriting recognition) والتشكيل الآلي (Automatic diacritization). أما النقحرة المستخدمة في رومنة الخط العربي (Romanization) فسنناقشها في القسم ٢,٣,١.

٢,١ عناصر الخط العربي

يتكون الخط العربي من ألفبائية تكتب من اليمين إلى اليسار. وهناك نوعان من الرموز لكتابة الكلمات في الخط العربي، هما: الحروف والحركات. وبالإضافة إلى هذه الرموز، سنتناول في هذا القسم الأرقام وعلامات الترقيم والرموز الأخرى[١].

(١) الداخلة في ألفبائية الخط العربي (توضيح من المترجمة).

٢,١,١ الحروف

تكتب الحروف العربية بطريقة متصلة في كل من الطباعة والكتابة اليدوية. وعادة ما تتألف من جزئين : الرسم (شكل الحرف) والإعجام (علامة الحرف). ويمثل شكل الحرف عنصراً أساسياً للحرف، فهناك ما مجموعه ١٩ شكلاً للحروف العربية (انظر الشكل رقم ٢,١). كما يمكن تصنيف إعجام الحرف الذي يطلق عليه التشكيل الساكن (consonantal diacritics) إلى ثلاثة أنواع فرعية (انظر الشكل رقم ٢,٢) كالتالي :

الأولى : هي النقاط ويوجد منها خمسة أشكال : نقطة أو نقطتان أو ثلاث نقاط توضع فوق الحرف و نقطة أو نقطتان توضع تحت الحرف.

الثانية : الكاف القصيرة للدلالة على أشكال رسم الحرف كاف (انظر الشكل رقم ٢,٤).

الثالثة : علامة الهمزة. فالهمزة يمكن أن تظهر في أعلى حرف معين أو أسفله. ويقصد بالهمزة هنا شكلان، شكلها المجرد (ء) وشكلها المركب مع حرف آخر مثل (أ) (ؤ) (ئ). وتعتبر المدة (~) شكلاً من أشكال الهمزة[١].

ء ا ب ح د ر س ص ط ع ف ق ل م ن ه و ى

الشكل رقم (٢,١). شكل الأحرف الأساسية بدون نقاط.

(١) الوصلة وهي من أشكال الإعجام غير المألوفة ولها علاقة بالهمزة، وتظهر فقط مع حرف الألف بالشكل التالي : ألف– وصلة أو همزة الوصل : آ (Ä) . هذا الحرف لا يستخدم كثيرًا لذا نجد أن بعض المحارف لم تمثله، سنناقش هذا الموضوع في قسم ٣,٢,١.

الشكل رقم (٢,٢). النقاط المستخدمة في الحروف وأماكن رسم الهمزة وشكل الكاف القصيرة.

وبتركيب رسوم الحروف وإعجامها يمكن الحصول على ٣٦ حرفاً للألفبائية العربية المستخدمة في كتابة العربية المعاصرة (انظر الجدول رقم ٢.١ في نهاية هذا الفصل). تكتب بعض الحروف بشكلها المجرد بدون نقاط، حيث تستخدم النقاط في الغالب للتمييز بين أصوات الحروف الساكنة، (الشكل رقم ٢.٣) يوضح ذلك. في الفصل القادم سنناقش موضوع ربط الصوت بالحرف.

ب ت ث	س ش	ا أ إ آ ى ئ ؤ ء
/b/ /t/ /θ/	/s/ /š/	/'/

الشكل رقم (٢,٣). تنقيط الحروف في المجموعة الأولى والثانية من اليمين كوَّن أحرفاً بأصوات مميزة، بينما الأحرف في المجموعة الثالثة فتمثل الهمزة بمختلف سياقاتها المرسومة والمنطوقة.

تنبيه مصطلحي: لا يجدر أن تخلط النقاط المستخدمة في العربية مع النقاط المستخدمة في العبرية (Niqqud) ؛ لأن استخدامها في العبرية اختياري وليس ضرورياً كما في العربية (انظر القسم ٢,١,٢) فهي تقارن بالتشكيل في العربية. وعليه، فإن الباحثين في مجال التعرف الضوئي على الحروف (OCR) يشيرون في الغالب إلى التشكيل ليقصدوا به تنقيط الحروف، وهذا ما لا نستخدمه في هذا الكتاب وأيضاً ما لا يستخدمه معظم الباحثين في مجال معالجة اللغة العربية آلياً. فعلامة الهمزة إذا جاءت في بداية الكلمة تعامل كتشكيل، أما إذا جاءت في وسط الكلمة أو في آخرها فتعتبر حرفاً [5] (انظر قسم ٢,١,٢).

`	w	r	d	A	l	k	h	T	S	s	q	f	m	γ	j	y	n	b	
ء	و	ر	د	ا	ل	ك	ه	ط	ص	س	ق	ف	م	غ	ج	ي	ن	ب	معزولة
					لـ	كـ	هـ	طـ	صـ	سـ	قـ	فـ	مـ	غـ	جـ	يـ	نـ	بـ	في البداية
					ـل	ـك	ـه	ـط	ـص	ـس	ـق	ـف	ـم	ـغ	ـج	ـي	ـن	ـب	في الوسط
	ـو	ـر	ـد	ـا	ـل	ـك	ـه	ـط	ـص	ـس	ـق	ـف	ـم	ـغ	ـج	ـي	ـن	ـب	في النهاية

الشكل رقم (٢,٤). أمثلة على أحرف بأشكالها المختلفة.

أشكال الحرف

تعتمد أشكال الحروف العربية على موقعها في الكلمة (في أولها، أو أوسطها، أو آخرها، أو كحرف مستقل). ويستخدم شكل الحرف بلا تفريق بين الطباعة والكتابة على حد سواء. يطلق أيضاً على أشكال الحرف المختلفة مصطلح البديل الخطي أو ألوقراف (allograph)[1]، كما يطلق على شكل الحروف الكتابية مصطلح قرافيم

(١) ألوقرافي (Allography) هو تمثيل لأشكال الحرف المتنوعة التي يتخذها بحسب موضعه في الكلمة أو طريقة كتابته (توضيح من المترجمة).

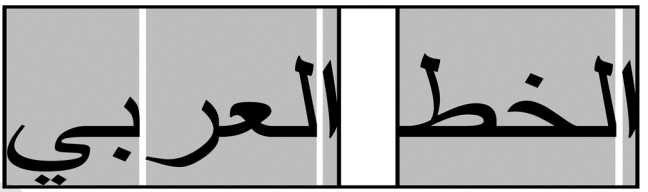
(graphemes)[١]، قياساً على الصوتيات أو الفونيمات (phonemes) وبديلها الصوتي الألفون (allophone) (انظر قسم ٣,١,١). وبالمثل، يعتمد اختيار شكل الحرف على سياقه في الكلمة وهذا ما يطلق عليه التكتيك الكتابي (graphotactics)، قياساً على التكتيك الصوتي (phonotactics).

يختلف المصطلح المستخدم في تصميم الخط والترميز: فالأحرف تسمى (characters) وشكلها يسمى (glyphs) (انظر قسم ٢,٢). وعادة ما تتشابه أشكال الحروف في بداية الكلمة ووسطها، كما تتشابه أشكالها إذا كانت في نهاية الكلمة وإذا كانت مستقلة. كما أن معظم أشكال الحروف تكتب متصلة بشكل كامل (أي من الجهتين)، إلا أن هناك بعضاً من أشكال الحروف المنفصلة من الآخر (post-disconnective)، فهي متصلة بما يسبقها من حروف وليس بما يليها. فجميع أشكال الحروف التي تلي الحروف المنفصلة من الآخر تظهر في بداية الكلمة أو تكون مستقلة. ويعتبر حرف الهمزة (ء) من الحروف المنفصلة بالكامل (من الجهتين) (انظر الشكل رقم ٢,٤).

تظهر عادة مساحة فارغة بعد الأحرف المنفصلة مكونة جزراً صغيرة من الأحرف المتصلة، وهذا ما يسمى بأجزاء الكلمة. في المثال الوارد في الشكل رقم (٢,٥)، نلاحظ وجود كلمتين وخمسة أجزاء للكلمة. هذه الفراغات تجعل من الصعب على أنظمة التعرف الضوئي (OCR) تمييز حدود الكلمة بشكل صحيح. كما أن الفراغات قد تؤدي إلى أخطاء في التهجئة لا يمكن تمييزها بصرياً، وذلك بسبب أن الكلمات تنقسم إلى أجزاء الكلمة أو عدد من الكلمات يتم ربطها من دون وجود فراغ بينها. وإلى حدٍّ ما، نجد أن مشكلة تحديد أجزاء الكلمة الصحيحة موجودة أيضاً في

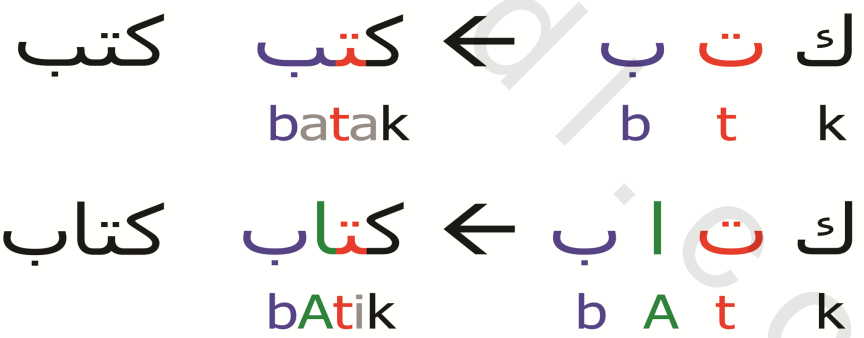
(١) القرافيم (Grapheme) هي أصغر وحدة كتابية في أي لغة، مثلا كلمة "الله" تتكون من ستة قرافيمات هي: ا ل ل ه ّ والألف الخنجرية (توضيح من المترجمة).

الصينية، حيث تتكون الكلمات الصينية من حرف أو أكثر من حرف، وهذه الأحرف قد تكون هي نفسها كلمات [6].



الشكل رقم (٢,٥). تأتي الكلمات العربية في الغالب متصلة، لكنها قد تحتوي على فراغات صغيرة ناتجة من الحروف المنفصلة.

يوضح الشكل رقم (٢,٦) كيفية بناء كلمة عن طريق ربط حروفها. وتذكر أن اللغة العربية تكتب من اليمين لليسار، وذلك عند مطابقة النقل الكتابي مع الحروف كما هو مبين في الشكل السابق. فالحرف ألف يعتبر من الحروف المنفصلة، لذلك نجده قسَّم الكلمة إلى جزئين.



الشكل رقم (٢,٦). طريقة كتابة كلمة "كتب" و"كتاب" بشكل حروف منفصلة وحروف متصلة.

ومن حيث المبدأ، يرتبط شكل الحرف بمكان الحرف في الكلمة، فبعض الحروف مثل التاء المربوطة (ة) والألف المقصورة (ى)[١]، بشكلها المنفصل تختلف عن شكلها إذا اتصلت بأحرف أخرى. بالإضافة إلى أن بعض أشكال الحروف، مثل الكاف التي تأتي في أول الكلمة أو في وسطها، تفقد علامتها (وهي الهمزة) وتظهر إذا كتب الحرف في آخر الكلمة أو بشكل مستقل (انظر الشكل رقم ٢,٤ والشكل رقم ٢,٦).

سؤال شائع: كم عدد حروف اللغة العربية؟

هناك خلاف واسع في تحديد عدد الحروف العربية، ويرجع السبب في ذلك إلى تصنيف الحروف وحركاتها. ففي الغالب يمكن القول إن الألفباء العربية تتكون من ٢٨ حرفاً (أحياناً باستبدال حرف «ا» بـ «أ») أو ٢٩ حرفاً (٢٨ حرفاً أساسية + الهمزة على السطر وهي إحدى أشكال الهمزة). وبهذه الحسبة، تعتبر الهمزة كأحد علامات التشكيل، بينما بحسبة أخرى، تضاف التراكيب الأربع للحرف لام ألف (لا) للستة والثلاثين حرفاً للأبجدية العربية (٢٨ حرفا أساسية + علامة الهمزة + التاء المربوطة + الألف المقصورة) ليصبح المجموع ٤٠ حرفاً. وفي هذا الكتاب سنتبع النمط المألوف للترميز المعياري الحاسوبي الذي لا يحسب أشكال الهمزة كنوع من التشكيل، أو حتى اللام ألف (لا) كحرف مستقل، بينما تمثل التاء المربوطة والألف المكسورة كأحرف لها ترميز مستقل. سنناقش موضوع الترميز في قسم (٢,٢).

(١) هناك العديد من الطرق الممكنة لرومنة الأسماء العربية، بما في ذلك أسماء الحروف العربية. انظر قسم ٣,٣,١. في هذا الكتاب، سنحاول أن نكون متسقين داخلياً، ولكن على القراء أن يدركوا أنهم سيواجهون طرقاً إملائية مختلفة، على سبيل المثال، الأسماء المعيارية لليونيكود المعروضة في الجدول رقم ٢,١.

تركيب/تشبيك الحروف

علاوة على تنوع أشكال الحروف في اللغة العربية، فإنها تمتلك قدراً كبيراً ومشتركاً من التراكيب/التشابيك، وأشكالاً مختلفة لتركيب/لتشبيك حرفين أو حتى ثلاثة. وتتضمن التراكيب/التشابيك في العادة الوضعية العامودية للحرف (الشكل رقم ٢,٧) وتتنوع بحسب نوع الخط المستخدم الشكل رقم (٢,١٣). جميع التراكيب/التشابيك اختيارية وتعتمد على نوع الخط ماعدا تشبيك (لام الف) فهو إلزامي: ل + ا تمثل بالشكل (لا) (ـلا إن كانت متوسطة) وليس لـا. ويمثل بالشكل (لا). وهذا التراكب والتشابك المتصل من اليمين فقط، له ثلاثة أشكال وهي: لام- ألف- همزة فوق السطر وتكتب (لأ)، لام- ألف -همزة تحت السطر وتكتب (لإ) ولام -ألف - مدة وتكتب (لآ). تشكل التراكيب/التشابيك تحدياً كبيراً في ترميز اللغة العربي، وهذا ما سنناقشه في قسم ٢,٢.

الأنواع المختلفة للأحرف العربية

يمكن تصنيف الأحرف الستة والثلاثين المكونة للغة العربية المعاصرة إلى التالي:

١. **٢٨ حرفاً أساسياً**: تمثل الحروف الساكنة الصحيحة (الصوامت)، وترسم حسب موقعها في الكلمة، ماعدا الهمزة. ويعتبر الإعجام في جميع هذه الأحرف علامةً تميز حرفاً عن آخر.

٢. **الهمزة**: هناك ستة أشكال للهمزة، الهمزة على السطر (ء)، وهذا هو شكل الحرف المعروف. والأشكال الأخرى تستخدم إعجام الهمزة وحرف المد مع رسم الأحرف الأخرى (ئ، إ، ؤ، أ، آ، ء). حينما تأتي الألف المهموزة (إ، أ، آ) بعد لام، فيصبح تراكب/تشابك اللام ألف إلزامياً. وقد جرت

العادة، ألا تظهر أشكال الهمزة المختلفة في الألفباء العربية، بل تمثل هذه الأشكال بحرف واحد وهي همزة القطع (انظر قسم ٣,١,٢). ويحكم رسم الهمزة بأشكالها المختلفة قواعد إملائية معقدة تعكس موقعها في الكلمة وعلاقتها بالأحرف المصاحبة [7، 8].

سؤال شائع: ما هو الترتيب الألفبائي لحروف اللغة العربية؟

هناك طريقتان شائعتان لترتيب حروف اللغة العربية. (١) الطريقة الألفبائية المعتمدة على جمع الأشكال المتشابهة للحروف بعضها مع بعض.

أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
y w h n m l k q f γ ç Ď T D S š s z r ď d x H j θ t b Â

(٢) الطريقة الأبجدية، والمبنية بشكل غير قوي على أساس ترتيب الأبجدية الفينيقية (الكنعانية) التي لا زالت مستخدمة مع العبرية، مع إضافة ستة حروف إلى آخرها. ويستخدم هذا الترتيب الأبجدي في ترتيب القوائم الصغيرة التي لديها عادة أقل من عشرة عناصر. أما القوائم الطويلة فتستخدم الأعداد (انظر قسم ٢,١,٣).

عادة ما تسرد القواميس العربية التقليدية كلماتها في مجموعات بناء على جذرها ومرتبة ألفبائياً، مع وجود بعض القواميس التي لا تقوم بعمل ذلك. وكما هو الحال في الجيماتريا (Gematria)[1]، يأتي ترتيب الأبجدية العربية مع أرقام مصاحبة لها كما هو موضح في التالي:

(١) وهو نظام **يحدد قيمة رقمية لكل حرف في اللغة العبرية** ويقابله نظام حساب الجُمل في اللغة العربية (توضيح من المترجمة).

غ	ظ	ض	ذ	خ	ث	ت	ش	ر	ق	ص	ف	ع	س	ن	م	ل	ك	ي	ط	ح	ز	و	د	ج	ب	أ
γ	Ď	D	ǒ	x	θ	t	š	r	q	S	f	ç	s	n	m	l	k	y	T	H	z	w	h	d	j	b	Â
1000	900	800	700	600	500	400	300	200	100	90	80	70	60	50	40	30	20	10	9	8	7	6	5	4	3	2	1

كلتا الطريقتين في الترتيب لا تغطي إلا الـ٢٨ حرفاً الأساسية. بينما يعتمد ترتيب الترميز المعاصر للعربية (انظر قسم ٢,٢) على الشكل مع الإضافات لتستوعب الحروف الإضافية في العربية المعاصرة (انظر الجدول رقم ٢,١ والجدول رقم ٢,٢).

أما الامتدادات للحروف غير العربية (Non-MSA) فعادة ما توضع خارج نطاق ترتيب الحروف العربية المعاصرة. لذا، عند ترتيب الحروف العربية المعاصرة بناء على شكلها يمكن الاعتماد على قيمة ترميز الحرف، بينما يتطلب ترتيب امتدادات الحروف غير العربية والترتيب حسب التسلسل الأبجدي إجراءات خاصة.

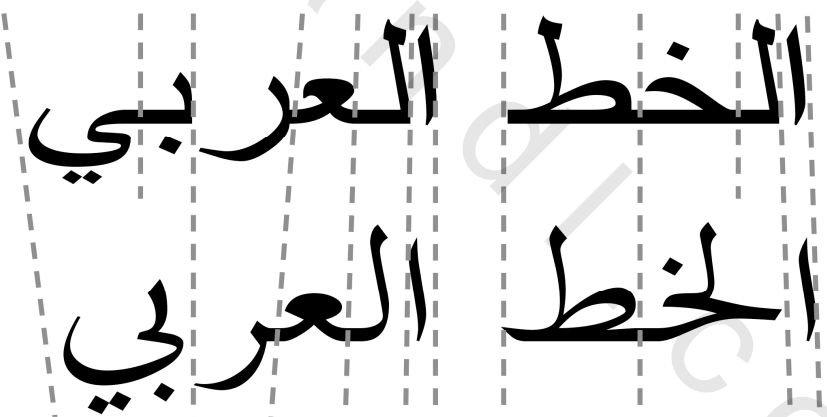
٣. **التاء المربوطة (ة ħ)**: من العلامات الصرفية التي تستخدم في التأنيث، وهو حرف هجين يدمج شكل حرف الهاء (هـ) مع التاء (ت). ويظهر الحرف فقط في آخر الكلمة، أما عندما تظهر التاء المربوطة في وسط المورفيم (أو الكلمة)، فإنها تكتب تاء (ت). على سبيل المثال، كلمة مكتبة+هم (mktbħ+hm) تكتب مكتبتهم (mktbthm). وعلى الرغم من أن شكل التاء المربوطة متصلة بالكامل (من الجهتين)، إلا أن الحرف غير متصل في نهايته (post-disconnective).

٤. **الألف المقصورة (ى ý)**: الألف المقصورة (ى)[١]، هي علامة صرفية خاصة يميز مجالاً من المعلومات الصرفية المتعددة بدءاً من الأسماء المؤنثة وحتى جذور الأفعال. وهو حرف هجين يدمج شكل حرف ألف (ا) مع الياء (ي). وتظهر الألف المقصورة في نهاية الكلمة كحرف ياء من دون نقاط.

(١) يسميها البعض الألف اللينة المتطرفة (توضيح من المترجمة).

عندما تظهر الألف المقصورة في وسط المورفيم يكتب كألف (ا) أو ياء (ي). على سبيل المثال، كلمة مستشفى+هم (mstšfý+hm) تكتب مستشفاهم (mstšfAhm)، غير أن إلى+هم (Âlý+hm) تكتب إليهم (Â lyhm). وعلى الرغم من أن شكل الألف المقصورة متصل بالكامل (من الجهتين)، إلا أن الحرف غير متصل في نهايته (post-disconnective).

هناك بعض الحروف الإضافية التي لا تعتبر رسمياً جزءاً من ألفباء اللغة العربية المعاصرة، مثل (ڤ و گ وپ وچ) هذه الحروف استعيرت من لغات أخرى، والهدف الرئيسي منها هو تمثيل الأصوات غير الموجودة في اللغة العربية المعاصرة أو لهجاتها، اطلع على قسم ٢,١,٥.



الشكل رقم (٢,٧). مثال على تمثيلين مختلفين للأحرف المركبة ناتجة عن استخدام خط من نوع معين. وكما هو ظاهر، فالحرف الثاني والثالث والحرفين الأخيرين في المثال السفلي تركبت بشكل عمودي، عكس المثال العلوي.

٢,١,٢ الضبط بالشكل (التشكيل)

المجموعة الثانية من الرموز الموجودة في الخط العربي هي التشكيل. فبينما تكتب الأحرف دائماً إلا أن التشكيل اختياري: وعند الكتابة يمكن تشكيل النص بالكامل أو تشكيله جزئياً أو عدم تشكيله، وفي قسم ٢,٣,٤ سوف نستعرض طرق التشكيل الآلي.

يأتي النص العربي في الغالب غير مشكل إلا في الكتب الدينية وكتب الأطفال الدراسية وأحيانا الشعر. وتشكل بعض الكلمات في النص العربي المعاصر عند الكتابة وذلك لكشف غموض نطق الكلمة. وفي بنك بنسلفانيا للتحاليل النحوية (Penn Arabic Treebank) (الفصل الثالث) [9]، نجد أن ١,٦٪ من جميع الكلمات شُكل حرف واحد منها على الأقل من قبل كاتبها، ومن هذه النسبة نجد أن ٩٩,٣٪ من هذه الحركات تعتبر صحيحة وذلك بسبب وضعها على الحرف الصحيح.

هناك ثلاثة أنواع من التشكيل وهي: الحركات والتنوين والشدة كما في الشكل رقم ٢,٨.

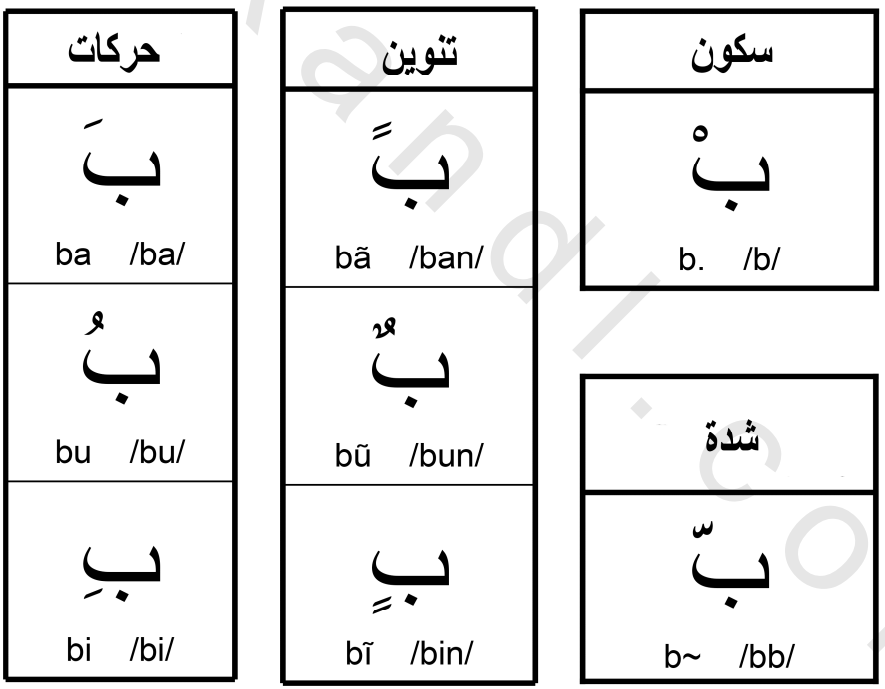
تمثل الحركات طريقة نطق الحرف من حيث المد الصوتي القصير (الفتحةَ /a/ والضمة ُ/u/ والكسرةِ /i/) ويُضاف إليها السكون للدلالة على غياب الحركة ككل.

أما التنوين فيظهر في آخر الكلمات الاسمية (الأسماء والصفات والظروف) وتدل على كون الاسم نكرة (انظر قسم ٤,٢,٢). ينطق التنوين كحرف مد متبوع بصوت النون غير المكتوبة، على سبيل المثال "بٌ" (*bũ*) تنطق "بن" (/bun/). ويكتب التنوين عن طريق مضاعفة حركة نهاية الاسم المطلوب وتأتي على ثلاثة أشكال: تنوين ضم وتنوين فتح وتنوين كسر. وعليه فإن كتابة التنوين بمضاعفة تشكيل الحركة هو مجرد عارض هجائي وليس له أهمية لغوية[١].

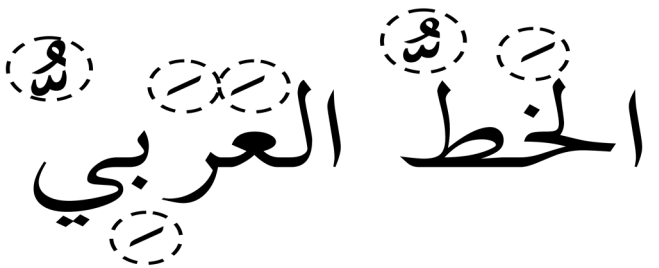
(١) وذلك حسب قول المؤلف (تعليق من المترجمة).

الشدة هي تكرار الحرف مرتين الحرف الأول ساكن والحرف الثاني متحرك مثال ذلك (بّ ~(/bb) b/).وغالبا ما تأتي الشدة مع التنوين أو حرف لين، مثال ذلك بٌّ b~u (/bbu) أو بٌّ bbun) b~ũ). على سبيل المثال، كلمة عَبَّرَ (*ʕab~ara*) تنطق (/ʕabbara/). وفي الفصل الثالث سوف نتطرق بشيء من التفصيل لكيفية نطق العربية. كما أن الشكل رقم ٢,٩ يوضح بعض الكلمات المشكلة بالكامل.

ومن علامات التشكيل غير المألوفة، الألف الخنجرية أو الألف الصغيرة التي تفيد إثبات نطق الألف الممدودة (/ā/) مع الحلول محلها، وتظهر في التهجئة القديمة لكلمات قليلة مثل كلمة الله (*All~áh*) وهذا (*háðA*).



الشكل رقم (٢,٨). أنواع التشكيل في العربية.



الشكل رقم (٢,٩). كلمة مشكلة بالكامل.

ويستخدم الإملاء القرآني مجموعة من علامات التشكيل للمساعدة في قراءة القرآن. وفي هذا الكتاب لن نتطرق للطريقة الإملائية المستخدمة في النص القرآني لخصوصيته التي تختلف عن اللغة العربية المعاصرة [10].

٢,١,٣ الأرقام

تكتب الأرقام العربية بالنظام العشري. وهناك مجموعتان من الأرقام المستخدمة في كتابة الأعداد في العالم العربي. الأرقام العربية التي يشيع استخدامها في أوروبا والأمريكتين ومعظم دول العالم، وفي البلدان العربية الغربية (المغرب والجزائر وتونس). أما دول الشرق الأوسط (على سبيل المثال، مصر، سوريا، العراق ، المملكة العربية السعودية) فهي تستخدم الأرقام العربية- الهندية.

سؤال شائع: هل الهمزة إحدى علامات التشكيل؟

كما ذكرنا سابقاً، تعد الهمزة رمزاً شبيهاً بحركات التشكيل يظهر مع عدد محدود من أشكال الحرف. والاتفاق العام في ترميز الهمزة هو في اعتباره إعجاماً (وجزءاً من الحرف)، عوضاً عن تمثيله كعلامة تشكيل. وفي الواقع، أدى تجاهل البعض

كتابة الهمزة خاصة مع الألف في بداية الجذع [٥] إلى جعل كتابتها اختيارية وتعامل معاملة التشكيل.

وفي الأنظمة الحاسوبية، يمكن اعتبار أو عدم اعتبار إعادة كتابة الهمزة كجزء من مشاكل التشكيل الآلي [11، 12، 13]. ويستبدل حرف الألف- همزة بحرف الألف المجردة في عمليات المعالجة الآلية. أما الهمزة المصاحبة للأحرف الأخرى غير الألف، فإنها في الغالب لا تسقط أو تتجاهل، انظر القسم ٢,٣,٢.

بعض الدول غير العربية مثل إيران وباكستان تستخدم بدائل لمجموعة الأرقام العربية-الهندية تختلف في شكل الأرقام ٤ و ٥ و ٦ فقط. الشكل رقم ٢,١٠ يوضح الاختلافات بين مجموعة الأرقام الثلاثة.

الأرقام العربية تونس، المغرب، إلخ.	0	1	2	3	4	5	6	7	8	9
العربية-الهندية الشرق الأوسط	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
العربية-الهندية الشرقية إيران، باكستان، إلخ.	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

الشكل رقم (٢,١٠). مجموعة من الأرقام المستخدمة في الخط العربي.

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

الشكل رقم (٢,١١). تكتب الأرقام العربية من اليسار لليمين، بينما النص يكتب من اليمين لليسار.

على الرغم من أن اتجاه الكتابة العربية هو من اليمين لليسار، فإن نظام كتابة الأرقام في العربية هو من اليسار لليمين كما هو الحال في اللغات الأوروبية. وعند كتابة الأرقام من لوحة المفاتيح فإن الأرقام تكتب من اليسار لليمين (انظر شكل ٢,١١). أما عند الكتابة بخط اليد، فإن الأرقام الثنائية (أي التي تحتوي على رقمين) تكتب من اليمين لليسار، أما الأعداد الكبيرة فتبدأ من اليسار متجه لليمين. وطريقة الكتابة هذه هي انعكاس لكيفية نطق الأعداد العربية، ففي الأعداد الصغيرة (حتى ١٠٠) تنطق الخانة الصغرى أولا وأيضا تكتب. أما في الأعداد الكبيرة، فتنطق الخانات الكبيرة أولاً. على سبيل المثال ينطق العدد ٢,٣٤٥ كالتالي: ألفان وثلاثمائة وخمسة وأربعون[١].

إن مطابقة الرقم بنطقه مهم لبرامج تحويل النص إلى صوت وأيضا لنمذجة اللغة في برامج التعرف الآلي على الصوت Automatic Speech Recognition (ASR) [14].

٢,١,٤ علامات الترقيم والرموز الأخرى

يستخدم الخط العربي علامات ترقيم مشابهة لتلك المستخدمة في اللغات الأوروبية وهي (، ! ؟ : " .). إلا أن بعض هذه العلامات تبدو مختلفة قليلا عن نظيرتها الأوربية وذلك لتتواكب مع طبيعة الكتابة العربية من اليمين لليسار، مثل علامة الاستفهام (؟ ?) والفاصلة (، ,) والفاصلة المنقوطة (؛ ;). كما تستخدم الفاصلة العربية (،) في الأرقام لتحديد الجزء العشري، وفي بعض الحالات يستخدم الحرف (ر) لنفس الغرض مثال ذلك: ١,٥ أو ١ر٥.

(١) وللمعلومية فإن قراءة الأرقام في اللغة العربية الفصحى يكون من اليمين لليسار، ففي المثال السابق يقرأ العدد كالتالي: خمسة وأربعون وثلاثمائة وألفان.

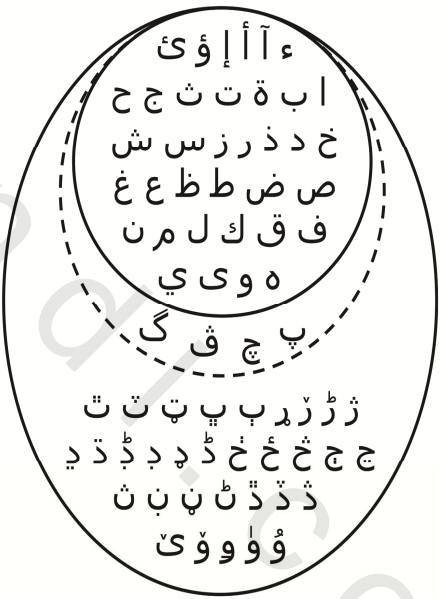
وهناك رمز خاص هو الكشيدة (أو التطويل)، حيث تضاف (ـ) بين حروف الكلمة الواحدة في الكتابة العربية بغرض تطويلها أو تمييزها أو لمساواة النص في الخط العربي. وبما أن الخط العربي لا يستخدم الحروف الكبيرة كما هو الحال في الكتابة الإنجليزية، تستخدم الكشيدة لنفس الغرض وهو للتأكيد والتمييز. فيما يلي مثال لكلمة بدون تطويل (قال) (*qAl*)، وعند إضافة كشيدة واحدة تصبح (قـال)، أما عند إضافة كشيدتين فإنها تصبح (قـال). ومن الواضح أن الكشيدة تعمل على تطويل الحروف المتصلة إذا ظهرت في وسط الكلمة. وأحيانا تستخدم الكشيدة لإجبار الحروف على الظهور بشكلها الأولي مثال ذلك (هـ) (*h_*) والمستخدم كاختصار للتاريخ الهجري. وعند معالجة اللغة العربية آلياً تحذف الكشيدة للحد من التناثر (sparsity) وذلك عند بناء نماذج اللغة (language models) (قسم ٢,٣,٢).

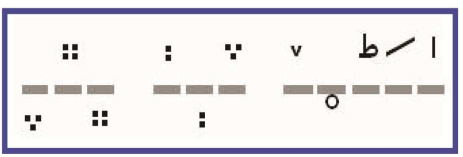
٢,١,٥ توسعات الخط العربي

يعتبر الخط العربي خطاً متنوعاً استخدم في كتابة العديد من اللغات التي تنتمي لعوائل لغوية مختلفة مثل: البلوشية، والدري، والهوسا، والقبائل، والكشميري، والقازاقية ، والكردية، والقرغيزستانية، والمالايوية، والموريسكية، والباشتو، والفارسية، والبنجابية، والسندية ، والسّيرايكيّة ، والتتارية، والتركية العثمانية، والأويغورية، والأوردية. وكما هو الحال مع الخط اللاتيني، فقد أدى استخدام الحروف العربية في اللغات الأخرى غير العربية إلى إضافة علامات إعجام متنوعة وإعادة تحديد القيمة الصوتية لبعض الحروف [15]. وبما أن التركيز في هذا الكتاب هو على اللغة العربية المعاصرة في مجال اللسانيات الحاسوبية والمعالجة الآلية للغة، فإن الهدف من هذا القسم هو إعطاء القارئ المبادئ الأساسية للتعرف على اللغة العربية،

على غرار ما يقوم به العديد من قراء الخط اللاتيني للتعرف على النص عندما يواجهون نصاً لاتيني غير مألوف، بمعنى آخر، أن يكون القارئ قادراً على تخمين أن بعض النصوص هي إنجليزية، أو فرنسية، أو ألمانية ، أو "إحدى اللغات الإسكندينافية".

تستخدم جميع اللغات التي وسعت الخط العربي مجموعة الحروف من اللغة العربية المعاصرة وتضيف عليها. يوضح الشكل رقم (٢,١٢) قائمة بالحروف وعلامات الإعجام المضافة إلى اللغة العربية وأمثلة عليها. بعض الحروف الممتدة تستخدم كتلاً من





الشكل رقم (٢,١٢). مجموعة الحروف المضافة على اللغة العربية المعاصرة. الدائرة الداخلية تحتوي على الحروف العربية الأساسية، تليها الدائرة الوسطى ذات الإطار المنقط التي تحتوي على أحرف موسعة لا تظهر باستمرار في اللغة العربية المعاصرة، وأخيرا الدائرة الخارجية التي تحتوي على أحرف لا تنتمي للغة العربية المعاصرة.

النقـاط إمـا فـوق الحـرف أو تحتـه ، ولا تظهـر هـذه النقـاط عـادة في الحـروف العربيـة المعاصرة. كما أن بعض الحروف الإضافية تستخدم حتى أربع نقاط أو تغير اتجاه النقـاط (مثل وضع نقطتين عمـوديتين). ومـن بعـض الحـروف الإضـافية العجيبـة : الهفـت (v) والحلقة (o) والطاء الصغيرة (ط).

سؤال شائع: ما الفرق الجوهري بين الخط العربي واللاتيني من ناحية معالجـة اللغـة آلياً؟

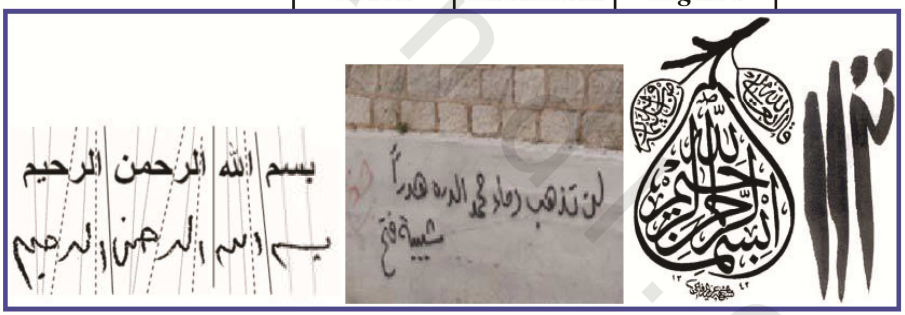
تكمـن بعـض الفـروق بـين الخطـين في التـالي : اتجـاه الكتابـة ، ورسـم الحـرف ، ووجـود التشـبيك الإلزامـي ، ويمكـن إزالـة هـذه الفـروق بفعاليـة في التطبيقـات الحاسـوبية (انظـر القسـم ٢,٢) وبالتـالي تعتـبر هـذه الفـروق لا علاقـة لهـا بمعالجـة اللغـة آليـاً. أمـا الفرقـان الجوهريان بين الخطين فتتضح في اختيارية وضع التشكيل وأيضـا عـدم وجـود الأحـرف الكبيرة (Capitalization). فعدم وضع التشكيل يضع حمـلاً كـبيراً علـى القـارئ العـادي لفك اللبس ناهيك عـن صـعوبتها للآلـة ، مقارنـة بـالخط اللاتـيني. وفي الفصـل الخـامس سـنناقش فـك الغمـوض الصـرفي للعربيـة. كمـا أن عـدم وجـود أحـرف صـغيرة وكـبيرة للعربية كما هـو الحـال في الخـط اللاتـيني ، سـيجعل مـن بعـض التطبيقـات مثـل تقنيـات التعـرف علـى الأسمـاء (Named Entity Recognition) وتعـيين أقسـام الكـلام (Part of Speech Tagging) صعبة على الآلة مقارنة بالخط اللاتيني.

٢,١,٦ رسم الخط العربي

للخط العربي عدد كبير ومتزايـد مـن الخطـوط والأنمـاط المختلفـة ، انظـر الشـكل رقم (٢,١٣) لأمثلة على اسـتخدام الخـط العربـي. وتـدعم معظـم أنظمـة التشـغيل مثـل

الويندوز والماك واللينكس العربية وخطوطها المختلفة. ولتحرير العربية في نظام لاتك ($L^{A}TEX$)، ننصح باستخدام حزمة لاتك العربية ($ArabTEX$) [16].

Traditional Arabic	عربي	محمد	الجبر
Simplified Arabic	عربي	محمد	الجبر
Tahoma Arabic	عربي	محمد	الجبر
Andalus	عربي	محمد	الجبر
	çarabiy~ *Arabic*	muHam~ad *Muhammad*	Aljabr *Algebra*



الشكل رقم (٢,١٣). أمثلة على خطوط عربية مختلفة تستخدم في الحاسب والخط اليدوي والرسم الجداري.

٢,٢ ترميز اللغة العربية وإدخالها وعرضها

الترميز، الذي يعرف أيضاً باسم مجموعة الأحرف (character set)، أو المحارف (charset)، أو خريطة الأحرف (character map) أو صفحة الشفرة (code page)، هو

التمثيل المنهجي لرموز خطٍّ ما لغرض التخزين الثابت والوصول إليه (إدخال البيانات وعرضها) بواسطة الآلة. ويجب أن تكون الخيارات التمثيلية للترميز متوافقة مع أدوات إدخال البيانات والعرض. ويجلب الخط العربي معه تحديات معينة لمسألة تصميم الترميز وكيفية تفاعله مع وحدة التخزين والوصول للبيانات. ويرجع السبب في المقام الأول نتيجة أن الخط العربي يختلف عن الخطوط الأوروبية التي تستخدم افتراضياً منذ نشأة الحاسوب. ويمكن تلخيص التحديات الأساسية في التالي: اتجاه الكتابة العربية من اليمين إلى اليسار، والأشكال المختلفة للحرف العربي حسب السياق، والحروف المتشابكة، واستخدام علامات التشكيل ومعالجة الاتجاه الثنائي في الأرقام والحروف اللاتينية في السياقات العربية.

من ناحية، يمكن للترميز أن يمثل تشبيك كل حرف مع حركاته على شكل رمز معقد منفصل، ومن ثَم يصبح عدد الرموز المختلفة في نظام الترميز كبيراً جداً. ومن جهة أخرى، يمكن ترميز كل حرف على شكل رمز منفصل عن تشكيله. من الترميزات الشائعة المستخدمة للعربية هي: اليونيكود (Unicode وترميز CP-1256وترميز ISO-8859)، والتي تُرمّز (encode) العربية على شكل قرافيم مكون من حروف وحركات بترتيب منطقي (من الأول للآخر). وفي الأصل، تعتبر طريقة عرض اللغة العربية في اتجاه مختلف على الشاشة عن الخط اللاتيني ليس له علاقة بالترميز، كما هو الحال في مسائل التشكيل وتغير شكل الحرف الذي يعتمد على السياق. أدى هذا الخيار في تصميم الترميز إلى فاعلية تخزين العربية في جهاز الحاسب، على الرغم من أنه يضع عبء الإدخال الصحيح والعرض على نظام التشغيل أو البرنامج الذي يعمل على معالجة اللغة العربية.

٢,٢,١ دعم الإدخال والإخراج في اللغة العربية

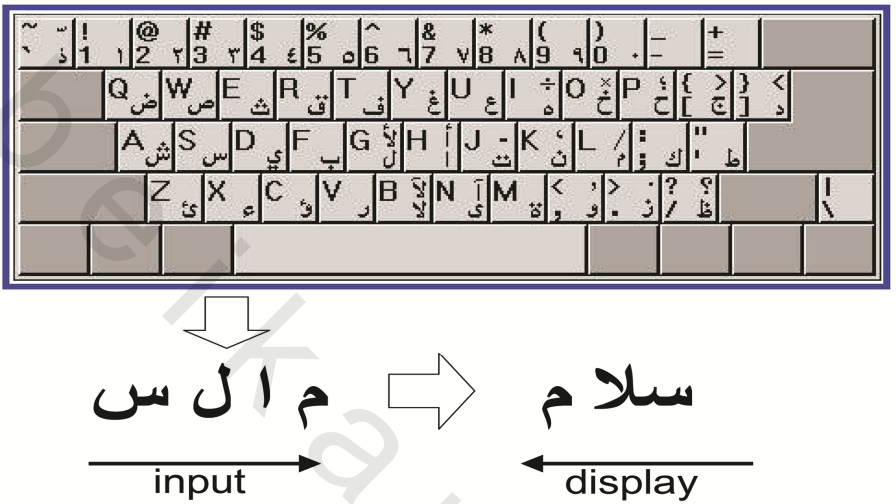
اُعتبر دعم العربية لمدة طويلة وخارج حدود الخطوط المتاحة غير متوفر في العديد من أنظمة التشغيل وبرامج التحرير. أما حالياً، فإن معظم أنظمة التشغيل تدعم العربية، ولكنها يجب أن تفعل في بعض الحالات. لذا قد يكون الدعم الجزئي للعربية مضللاً لغير القارئين بها، فوجود خط عربي في النظام قد يساهم في التعرف على الحروف، لكن قد يظهر الخط من اليسار لليمين أو تصبح أشكال الحروف غير صحيحة، انظر الشكل رقم (٢,١٤).

Logical Order	→ÔÇÑßÊ ÝáÓØíà (Palestine) Ýí ÇæáãÈíÇÏ (Olympics) 2000 æ 2004. .4002 æ 0002)scipmylO(ÏÇíÈãáæÇ íÝ)enitselaP(äíØÓáÝ ÊßÑÇÔ ←
Visual Order	← ش ا ر ك ت ف ل س ط ي ن (Palestine) في ا و ل م ب ي ا د (Olympics) 2000 و 2004 . ← ش ا ر ك ت ف ل س ط ي ن (Palestine) في ا و ل م ب ي ا د (Olympics) 2000 و 2004.

الشكل رقم (٢,١٤). مثال يوضح شكل ظهور النص العربي في ذاكرة الحاسب وعلى الشاشة. في الذاكرة، ترتب الحروف منطقيا (Logical Order) (من الأول للآخر). أما في الشاشة (Visual Order)، فتستخدم خوارزمية بسيطة لضبط اتجاه النص وشكله. إلا أن المعالجة الخاصة لوجود اتجاهات مختلفة في النص (مثل وجود أرقام أو حروف لاتينية) سيجعل مهمة عرض النص العربي معقدة.

في التسعينيات، استخدمت حلول عديدة وغير نافعة الآن لتجاوز النقص في الدعم العالمي لإدخال وعرض العربية. حيث تمثلت فكرة الحلول في ترميز العربية بشكل ألوقرافي ليس فقط عن طريق إسناد تشفير الرموز المختلفة لأشكال مختلفة من الحروف العادية والمتشابكة، بل القيام أيضاً بترميز العربية داخلياً بترتيب بصري. ويقصد بالترتيب البصري هنا ترميز العربية عكسياً بحيث يظهر من اليمين إلى اليسار في الأنظمة التي تستخدم طريقة العرض من اليسار لليمين [17]. وقد عانت هذه الطرق في الترميز الكثير من المشاكل والقصور. وتجدر الإشارة إلى أن الترميز البصري للأرقام لا

زال موجوداً إلى يومنا هذا في بعض النصوص، مما يتسبب بمشكلة في مجال المعالجة الآلية للغة[١].



الشكل رقم (٢,١٥). مخطط لوحة المفاتيح المعيارية للأجهزة الشخصية، في المقابل تختلف أجهزة الماكنتوش قليلا في مخططها عن المعتاد. لوحة المفاتيح تحتوي على الحروف، بالإضافة إلى الحرف المركب لام-ألف. عندما يقوم الشخص بالكتابة على لوحة المفاتيح بالترتيب المنطقي للأحرف، سيظهر النص من اليمين لليسار بشكله السليم والمركب.

قبل ظهور الحاسبات، كانت الآلات الكاتبة وآلات الطباعة تستخدم شرائط طباعة مختلفة لتمثيل الأشكال المختلفة من الحروف العادية والمتشابكة. حيث تتطلب الكتابة على الآلة الكاتبة تحديد الشكل الصحيح للحرف لكتابته بشكل صحيح، ويعتبر اتباع هذه الطريقة في اللغة العربية أكثر تعقيداً من اللاتينية التي تعتمد على

(١) على سبيل المثال حاول البحث بواسطة المحرك قوقل عن "سنة ١٩٩٩" وقارنه بنتيجة البحث عن "سنة ٩٩٩١".

الأحرف الكبيرة والصغيرة فقط. لذا يشكل الترميز في أجهزة الحاسب الحديثة تبسيطا ملحوظا لعملية الكتابة العربية بواسطة لوحة المفاتيح. فالعربية تكتب حرفاً بحرف حسب ترتيب ظهورها في الكلمة، انظر الشكل رقم (٢,١٥).

فيما يلي سنناقش بعضا من الترميزات الأكثر استخداما للغة العربية.

٢,٢,٢ الترميزات العربية

على مر السنين، طورت العديد من الترميزات المعيارية للغة العربية. سنناقش في هذا القسم الترميزات الثلاثة الأكثر شيوعاً والمدعومة دعماً جيداً في وحدات الإدخال والإخراج للمنصات المختلفة. يعرض الجدول رقم (٢,١) والجدول رقم (٢,٢) قيم الرموز المختلفة والمستخدمة في أشكال أحرف اللغة العربية المعاصرة جنبا إلى جنب مع ترميزاتها المختلفة. ولمزيد من النقاش حول معايير ترميز اللغة العربية، انظر [18، 19].

الترميز الفعلي	الترميز المعروض: CP-1256	ISO-8859	Unicode	Western
CP-1256	تدشين منطقة حرة في دبي للتجارة الالكترونية	ٍ□ ة حرة□تدشِل كلظ ترنلِة□دبٍ ففتجارة افاف	Y ɠ ɠ□ ψ□□Ŏɠāā□ʌ	ÊÏÔíä ãäØÞÉ ÍÑÉ Ýí ÏÈí ááÊÌÇÑÉ ÇáÇáßÊÑæäíÉ
ISO-8859	ة حرة â×و هوêتدش ننتجارة êدب êل ةêوèانانمتر	تدشين منطقة حرة في دبي للتجارة الالكترونية	Y□粢既 ɠ ɠ□ψ ŎɠGG ㈱親ɠ	ÊÏÔêæ åæ×âÉ ÍÑÉ áê ÏÈê ääÊÌÇÑÉ ÇäÇäãÊÑèæêÉ
Unicode	ï» †ظٹظ´ط¯ط؟ ©ط±ط ط©ط‚ط·ظ†ظ…ظ ظٹظپ ط¯ط¨ظٹ ©ط±ط§ط¬طھظ„ظ„ظ ظ„ظ§ط„ظ§ط±ƒ ظٹظ†ظˆط±طھ©	ع ع□ظ□ظ□ ؛ ؟ظ □ظ ع□ظ عع ع ع □ظ□ظ-ظ ع□ظ□ظ □ظ□ظ□ظ، ظ□ظ ع ع ع□ظ□ظ ع ع□ظ ع□ظ □ظ ع ع	تدشين منطقة حرة في دبي للتجارة الالكترونية	ï»¿Ø ªØ¯Ø´ÙŠÙ† Ù…Ù†Ø·Ù‚Ø© ØØ±Ø© Ù ÙŠ Ø¯Ø¨ÙŠ Ù„Ù„Ø ªØ¬Ø§Ø±Ø© Ø§Ù„Ø§Ù„ÙƒØ ªØ±Ù ˆÙ†ÙŠØ©

الشكل رقم (٢,١٦). مقارنة العرض الصحيح والخاطئ لمختلف الترميزات العربية.

الترميزات ذات ٨ بت (ISO-8859-6 و CP-1256)

يعتبر ترميز ISO-8859-6 وترميز CP-1256 من أوائل وأشهر الترميزات التي استخدمت في ترميز العربية. طور ترميز (ISO-8859-6) بواسطة منظمة المقاييس العالمية (ISO) وهو مطابق لترميز (ASMO-708) الذي طورته المنظمة العربية للمقاييس والمعايير (ASMO). أما ترميز CP-1256 وهو اختصار (Code Page 1256) أو ما يعرف بترميز ويندوز العربي، فقد طور بواسطة شركة مايكروسوفت وشاع استخدامه بسبب نظام التشغيل ويندوز. كلا الترميزين يستخدمان بايتاً واحداً (يعادل ثمانية بتات)[١] لتمثيل كل رمز على حدةٍ، أي بحد أقصى ٢٥٦ حرفاً. وكما هو الحال في بقية اللغات، يحتفظ ترميز العربية بأول سبعة بتات (ما يعادل ١٢٨ حرفاً) لترميز الآسكي

(١) البايت (Byte) هي وحدة رقمية تتألف من ثمانية بتات (Bit). والبت هي أصغر وحدة حاملة أو ناقلة للمعلومة على هيئة 0 أو 1 (توضيح من المترجمة).

(ASCII)[1] الإنجليزي، ويستخدم الـ ١٢٨ حرفا الأخرى لتمثيل لغات أخرى. بهذه الطريقة يمكن استخدام نفس الترميز لتمثيل خطين (أو أكثر من لغة) إذا دعت الحاجة لذلك. أما الجزء العربي في ترميز (ASMO-708) وترميز (ISO-8859-6) يعتمد على الترميز الأسبق (ASMO-449) المعتمد على سبعة بتات (فهو متوافق معه وليس مماثلاً له) [19].

تصف الحروف العربية في ترميز CP-1256 بالترتيب مع وجود فجوات فيما بين كل مجموعة من الأحرف وذلك للإبقاء على القيمة الرمزية لبعض الأحرف الأوربية وبالتحديد على الفرنسية، ومن ثَم تنتج صفحة محارف متعددة اللغات (إنجليزية، وفرنسية، وعربية). ليس بمقدور ترميزISO-8859-6 وترميز CP-1256 استيعاب كامل الحروف العربية الموسعة، ومع ذلك، ضمنت أحرفاً من الفارسية. تعمل هذه الترميزات على تحديد القرافيم فقط وتعتمد على خوارزميات منفصلة لعرض الرموز الصحيحة للخط.

كما يعتبر ترميز ISO-8859-6 وترميز CP-1256 غير متوافقين إلا في أول ٢٢ حرفاً. تعني هذه المعلومة البسيطة أن الكلمات المكونة من حروف مستلة من تقاطع حروف الترميزين سوف تبدو صحيحة في كلا الترميزين. على سبيل المثال، انظر كلمة "حرة" في الشكل رقم (٢.١٦). وللتحقق من ترميز قائمة مرتبة من الكلمات (كما في القاموس)، فإنه من الحكمة النظر إلى أبعد من الكلمات الأولى القليلة المرتبة وذلك لتجنب الوقوع في هذا اللبس.

---

(١) الآسكي اختصار لكلمة النظام الأمريكي الموحد لتبادل المعلومات (American Standard Code for Information Interchange) (توضيح من المترجمة).

اليونيكود

الترميز العالمي الموحد أو اليونيكود (unicode) هو المعيار الحالي المستخدم لترميز عدد كبير من اللغات والخطوط في وقت واحد. صمم هذا الترميز في الأصل لاستخدام بايتين اثنين من المعلومات أي لترميز ٦٥,٥٣٦ رمزاً مميزاً، وتوسع منذ ذلك الحين لتغطية أكثر من مليون رمز مميز. ويدعم اليونيكود ترميز الحروف العربية الموسعة. كما أنه يعطي أشكال الحروف العربية العادية والمركبة عناوين خاصة تحت ما يسمى بمخططات نماذج العرض أ و ب[١]. وبما أن ترميز اليونيكود يضم ترميز حروف أكثر من ترميز ISO-8859-6 وترميز CP-1256، فإن التحويل من هذين الترميزين إلى ترميز اليونيكود ممكن، ولكن التحويل العكسي قد يسبب ضياع بعض الحروف.

وعلى الرغم من أن ترميز اليونيكود يوفر حلاً هاماً لتمثيل الحروف العربية الموسعة، إلا أنه يقدم تحديات جديدة، وبالتحديد يقدم عدة طرق لتمثيل نفس مظهر الرمز. على سبيل المثال، تستنسخ كل أشكال الأرقام العربية- الهندية والعربية- الهندية الشرقية. بالمثل، بعض الحروف لها أشكال لا يمكن تمييزها بسهولة، مثال ذلك حرف "ك" العربي ورمزه (U+0643) وحرف "ک" الفارسي ورمزه (U+06A9)، كلاهما يملكان الشكل الأولي للحرف (كـ). وستظهر مشكلة اللبس عند القيام بكتابة العربية في لوحة مفاتيح فارسية. وأخيراً، سيسمح وجود مخططات نماذج العرض بترميز خاطئ للألوقراف[٢] لا يمكن كشفه بسهولة بمجرد النظر. وستجعل جميع هذه الحالات من الصعب مطابقة النصوص التي تبدو متطابقة على الشاشة على الرغم من اختلاف ترميزها.

---

(١) http://www.unicode.org/charts/PDF/U0600.pdf
http://www.unicode.org/charts/PDF/UFB50.pdf
http://www.unicode.org/charts/PDF/UFE70.pdf

(٢) انظر تعريف ألوقرافي (توضيح من المترجمة).

٢,٣ مهام المعالجة الآلية للغة

٢,٣,١ النقل الكتابي (النقحرة)

بالإضافة إلى الترميزات القياسية التي نوقشت آنفاً، هناك العديد من الباحثين في مجال المعالجة الآلية للغة العربية ممن يستخدمون النقل الكتابي أو ما يطلق عليها النقحرة (orthographic transliteration)، وعلى وجه التحديد الرومنة، في أبحاثهم في مجال المعالجة الآلية للغة العربية. سنتبع هنا تعريف مصطلحي الكتابة الصوتية (transcription)، والنقل الكتابي (transliteration) اللذين قدمهما بيزلي (Beesley) [20]: يرمز مصطلح الكتابة الصوتية إلى تمييز النظام الصوتي أو ما يسمى الفنولوجي[١] والصرف- الصوتي (morpho-phonology) للغة ما كتابياً، في المقابل، يرمز مصطلح النقل الكتابي إلى عملية نقل حروف من اللغة المصدر إلى ما يقابلها (حرف بحرف) في اللغة الهدف وذلك وفقاً لمعيار كتابتها مع إمكانية عكس نقلها مجدداً للغة المصدر بعد تحويلها. وهذا التعريف للنقل الكتابي يطلق عليه أحياناً نظام النقل الكتابي الصارم أو النقحرة.

من أشهر مخططات النقل الكتابي للمعالجة الآلية للغة العربية لدى الغرب، هو نظام بكوالتر للنقل الكتابي (Buckwalter transliteration) [23] أو أحد أنواعه (انظر الجدول رقم ٢,١ والجدول رقم ٢,٢). يتبع نظام بكوالتر للنقل الكتابي الترميز المعياري المستخدم في ترميز الأحرف العربية في الحاسب، مثل اليونيكود. وقد استخدم نظام بكوالتر للنقل الكتابي في العديد من الأبحاث المنشورة في مجال المعالجة الآلية للغة العربية وفي مصادر طورت بواسطة مركز ائتلاف البيانات اللسانية (Linguistic Data Consortium). أما الميزة الرئيسية في نظام بكوالتر للنقل الكتابي فهو اتباعه لنظام النقل

(١) يقصد بالفونولوجي (Phonology) النظام الصوتي للغة (توضيح من المترجمة).

الكتابي الصارم (أي بطريقة واحد مقابل واحد) وأيضاً استخدامه لأحرف الآسكي عند الكتابة، مما يعني سهولة استنساخه من دون الحاجة لوجود خطوط مخصصة.

تنبيه مصطلحي: يستخدم العديد من الباحثين كلمة النقل الكتابي لتعني جميع أنواع النقل من كتابة لأخرى بصرف النظر عن نوع النقل. وقد يشمل ذلك الكتابة الصوتية (الصارمة والخاصة) والتي ينتج عنها عدة نقولات سليمة. من أشهر أنواع طرق النقل الكتابي نقل أسماء الأعلام (Proper Name)، حيث يستكشف هذه النوع من النقل طرق تمثيل أسماء الأعلام في اللغات المختلفة [21، 22]. سنناقش هذه الطريقة في قسم ٣,٣,١.

واحدة من أشهر الانتقادات على نظام بكوالتر للنقل الكتابي هو صعوبة قراءته. وفي هذا الكتاب نستخدم نظام (حبش – سودي– بكوالتر) الأكثر بديهية، وهو نوع من أنواع نظام بكوالتر للنقل الكتابي [4]. أما الانتقاد الآخر على نظام بكوالتر للنقل الكتابي فهو احتواؤه على بعض الحروف المحجوزة للغات برمجة حاسوبية مثل لغة السي (C) والبيرل (Perl) وأيضا في لغة الترميز الممتدة (XML) مثال ذلك الأقواس المتعرجة {} وعلامة الدولار ($). لمعالجة هذه المسألة، ظهرت العديد من أنواع نظام بكوالتر الآمنة، لكنها لم تخضع للمعايرة. وبما أن هناك أشكالاً متنوعة من هذا الترميز لاستخدامها في إعدادات مختلفة، فهي بحاجة لعناية خاصة حتى لا يتم الخلط فيما بينها. وأخيراً، فقد نقد نظام بكوالتر للنقل الكتابي لكونه أحادي اللغة، ذلك لأنه عند استخدام رموز الآسكي لكتابة العربية، فإنه لا يمكن استخدامه لتمثيل الإنجليزية. وقد

عالج بعض الباحثين هذه المشكلة عن طريق استخدام علامات خاصة لتحرير الأحرف غير العربية والمكتوبة بالترميز المعياري قبل تحويلها لنظام بكوالتر.

٢,٣,٢ تسوية/توحيد الاحتمالات الهجائية

يستخدم الباحثون في مجال المعالجة الآلية للغة العربية دوماً تسوية الاحتمالات الهجائية (Orthographic normalization) بهدف الحد من الضوضاء (noise) وتناثر البيانات (sparsity). وهذا صحيح بغض النظر عن المهمة سواء كانت: إعداد نص مواز للترجمة الآلية، أو مستندات لاسترجاع المعلومات أو نص لنمذجة اللغة، إضافة إلى ذلك، هناك طرق إضافية أكثر تعقيداً مثل تقطيع النص (tokenization) (قسم ٥,٣)، وتسوية الاحتمالات المعجمية للكلمات المكتوبة بطرق مختلفة، وتصحيح الأخطاء الإملائية ويمكن تطبيقها عادة بعد تسوية الحالات الهجائية.

هناك أنواع مختلفة من تسوية الاحتمالات الهجائية التي يمكن تطبيقها على انفراد أو مجتمعة على أي نص قيد الاهتمام. سوف نناقش في هذا الفصل الأمور المتعلقة بالخط العربي فقط. فمهام مثل فصل علامات الترقيم التي تطبق على الخط اللاتيني والعربي، تشكل نفس التحديات العامة.

- **تنظيف الترميز**: يشكل الترميز العربي، وبالتحديد في ترميز اليونيكود، العديد من التحديات الناتجة من إمكانية عرض نفس النص بطرق مختلفة وبأحرف مختلفة المصدر. أولاً، هناك طرق مختلفة لترميز الأحرف التي تبدو متشابهة، مثل الرموز للأرقام العربية- الهندية والعربية- الهندية الشرقية والأشكال المختلفة للحروف المتشابهة، مثال ذلك حرف الكاف العربي والفارسي (ك/ک). أيضا العديد من علامات الترقيم المتشابهة في الشكل تظهر في مخططات مختلفة تحت اليونيكود.

ويتطلب تنظيف الترميز تسوية الأشكال المختلفة للرموز إلى شكل واحد. ثانياً، يمكن ترميز نماذج عرض الخط العربي مباشرة، مما يؤدي إلى غموض في الحرف وشكله ومن ثَم لا يمكن الكشف عنها بسهولة على الشاشة. كما أن الحروف المركبة المعقدة أضافت إلى هذه المشكلة. وبواسطة تسوية المناسب منها يمكن تحويل الأحرف الألوقرافية إلى شكلها القرافيمي.

- **إزالة التطويل**: يزال رمز التطويل من النص بسهولة.
- **إزالة التشكيل**: بما أن التشكيل لا يظهر في النص بشكلٍ دائم، يعتبرها بعض الباحثين نوعاً من التشويش على النص لذا يقومون بإزالتها من النص.
- **تسوية الاحتمالات للحروف**: هناك أربعة حروف في العربية غالبا ما يتم كتابتها بشكل غير صحيح إملائيا، ويجد الباحثون أنه من الفائدة اختزال هذه الاختلافات في الكتابة. فيما يلي عرض للأربعة أحرف مرتبة حسب شيوع اختزالها من الأعلى شيوعاً في الاختزال إلى الأدنى (أول مجموعتين هما ما يقوم الباحثون باختزالها تلقائياً، أما المجموعتان الأخيرتان فهما الأقل شيوعاً في التطبيق).

١- أشكال الألف المهموزة (Â أ، Aˇ إ، A¯ آ) تختزل إلى حرف الألف المجردة (ا).

٢- الألف المقصورة (ý ى) تختزل إلى (y ي). في الغالب في مصر، وليس بالضرورة في بقية الدول العربية، تكتب الياء المتطرفة في الكلمة بدون نقاط (مثل

الألف المقصورة). إلا أنه في الآونة الأخيرة، يمكن مشاهدة العكس: حيث جميع الألفات المقصورة تكتب كياءات منقوطة[١].

٣- تختزل التاء المربوطة (h ة) إلى هاء (h ه).

٤- الأشكال الأخرى من الهمزة (ؤ w وَ y ئ) تختزل إلى الهمزة (ء).

سؤال شائع: بماذا تنصح: استخدام نظام بكوالتر للنقل الكتابي أم اليونيكود؟

على الرغم من وجود كل هذا النقد على نظام بكوالتر للنقل الكتابي، إلا أنه لا يزال يستخدم وذلك لسهولة قراءته وإمكانية تتبعه للباحثين غير القارئين للعربية. وعلى الرغم من عيوبه إلا أن نظام بكوالتر للنقل الكتابي يمكن أن يكون أكثر موثوقية في الكشف عن أخطاء الترميز التي قد تمر مرور الكرام في ترميز اليونيكود، مثل تمثيل الأحرف ألوقرافيا بدلاً من قرافيميا.

سؤال شائع: لماذا يستخدم نظام النقل الكتابي حرفاً بحرف؟

يسمح نظام النقل الكتابي الذي ينتهج طريقة واحد مقابل واحد إلى سهولة تناظر الخط العربي إلى مقابله اللاتيني والعكس صحيح. وقد يكون من الملائم استخدام بعضٍ من التناظرات المعروفة والمستخدمة للأحرف المتعددة إذا وضعت علامة على سلاسل الأحرف لتجنب أي غموض. على سبيل المثال، السلسة الحرفية (sh) تستخدم في الغالب لتمثيل حرف (š ش) وقد يساء تفسيرها إلى السلسلة (سه sh)، مثال ذلك، قارن كلمة "أشم" (*Âašum*) مع "أسهم" (*Âashum*).

(١) يمكن الاطلاع على إحدى صفحات جريدة الأهرام المصرية لتلاحظ تكرار استخدام هذا النمط من الكتابة، لذا لا يمكننا القول إنه خطأ إملائي وذلك لشيوعه في النص.

سؤال شائع: كيف يمكن كتابة نص عربي من دون وجود لوحة مفاتيح عربية؟
توجد العديد من الأدوات على الإنترنت تسمح للمستخدمين بكتابة بعض أشكال الرومنة الصارمة أو المتساهلة، على سبيل المثال، هناك موقع يملي (Yamli)، وتعريب (ta3reeb)[١] من قوقل ، ومرن (Maren)[٢] من مايكروسوفت. كما أن بعض أنظمة التشغيل توفر لوحة مفاتيح صوتية للغة العربية.

٢,٣,٣ التعرف على خط اليد

التعرف على خط اليد (Handwriting Recognition) هو تحويل النص المكتوب بخط اليد أو المطبوع إلى نص رقمي. يمكن تصنيف عملية التعرف على خط اليد إلى نوعين: مباشر (online) وغير مباشر(offline). في طريقة التعرف غير المباشر على خط اليد، يكون المدخل عبارة عن صورة رقمية لنص معين تم التقاطه بواسطة كاميرا رقمية أو مسحه بواسطة الماسح الضوئي. بينما تُعرف طريقة التعرف المباشر على خط اليد على أنها عملية التعرف التلقائي للنص المدخل كسلسة من النقاط ثنائية الأبعاد (عن طريق استخدام قلم رقمي أو قلم تأشير (stylus)). يتم أحياناً تمييز نظام التعرف الضوئي على الحروف ((Optical Character Recognition (OCR) عن نظام التعرف غير المباشر على خط اليد بالإشارة إلى النص المطبوع (في مقابل النص المكتوب يدوياً). ومع هذا نجد أن بعض الباحثين يستخدمون المصطلحين بالتبادل.

لا يزال مجال التعرف على خط اليد العربي المكتوب مجالاً نشطاً بحثياً، بنوعيه المباشر وغير المباشر، ويرجع السبب في ذلك للصعوبات المتأصلة في المهام المتعلقة بهذا المجال

---

(١) http://www.google.com/ta3reeb/

(٢) http://www.microsoft.com/middleeast/egypt/cmic/maren/

[24 ، 25، 26]. في المقابل نجد أن التعرف على خط اليد المطبوع الذي يتميز بانسجام أشكال حروفه بالإضافة لعوامل أخرى سمحت بسهولة التعرف عليه، يعتبر حالياً أقل اهتماماً لدى الباحثين [25]. يتميز الخط العربي بخصائص متعددة جعلت من مهمة التعرف على الخط المكتوب تحديا كبيرا [25، 27، 28]. تتمثل هذه التحديات في طبيعة الكتابة النسخية المتصلة في العربية مع وجود الأحرف المنفصلة، واستخدام التشكيل وعلامات عائمة بديلة للأحرف (التي تزاح في كثير من الأحيان أفقياً عند الكتابة) واستخدام الأحرف المركبة عمودياً والتطويل.

إن الكتابة المتصلة، واستخدام الأحرف المركبة جعلت مهمة تمييز الأحرف المكونة للنص شاقة على الأجهزة. وهي بالتأكيد ليست خاصية فريدة للغة العربية، فطرق مثل نموذج ماركوف المخفي (Hidden Markov Models)، الذي طور للخطوط النسخية في لغات أخرى أمكن تطبيقها بنجاح على العربية [25، 29، 30، 31]. بينما قد تسبب الأحرف المنفصلة في العربية بعض الصعوبة في تحديد حدود الكلمات، إلا أنه من الممكن أن تسهم بشكل معقول في تقليل الغموض بين الأشكال المماثلة. وتسبب الطبيعة العائمة للتشكيل وبعض علامات الأحرف مشاكل مختلفة لعملية التعرف على خط اليد بنوعيه المباشر وغير المباشر. في الطريقة غير المباشرة للتعرف على خط اليد، قد ينتج عن وجود الغبار أو الأوساخ على المستند الممسوح خطأ في التعرف على علامات التشكيل [27]. بدلا من ذلك، قد تكون هذه الرموز صغيرة جدا، أو متقاربة جدا، بحيث لا يمكن تمييزها بسهولة، مما تسبب في قيام النظام بإسقاطها بالكامل. أما في الطريقة المباشرة للتعرف على خط اليد، فإن التشكيل وعلامات الأحرف يجب أن تكون مقترنة بأشكال الحرف المناسبة للتعرف الصحيح عليها [24].

وقد أدى برنامج (MADCAT)[١] المدعوم من وكالة مشاريع الأبحاث المتقدمة لوزارة الدفاع الأمريكية (DARPA) الذي يستهدف الترجمة الآلية للوثائق العربية المكتوبة بخط اليد والممسوحة ضوئياً، إلى تكوين العديد من الموارد لتدريب وتقييم برامج التعرف على خط اليد العربي [32]. ولدى المعهد الوطني للمعايير والتكنولوجيا (NIST) نسخة مفتوحة لتقييم برنامج (MADCAT) تدعى (Open HaRT)[٢]. للمزيد من المصادر، انظر الملحق ج.

٢,٣,٤ التشكيل الآلي

التشكيل أو ما يسمى استعادة التشكيل، أو التحريك، هي عملية معالجة الحركات المفقودة مثل (التنوين، الضمة، الفتحة، الكسرة، علامة الشدة). وللتشكيل علاقة قوية بفك اللبس الصرف- نحوي (morphosyntactic) والتشذيب[٣] (انظر قسم ٥,١) وذلك لكون بعض علامات التشكيل تختلف بحسب الحالة النحوية (مثل التشكيل حسب السياق الإعرابي) وأخرى تختلف للدلالة على الاختلاف في المعنى.

إن تحديد اختيار التشكيل للحرف الأخير للكلمة (دون وجود ضمير متصل) صعب للغاية، وذلك لأنه بحاجة إلى معلومات نحوية: ففي الفعل المضارع، يعبر التشكيل عن

---

(١) اختصار لكلمة (Multilingual Automatic Document Classification Analysis and Translation) التصنيف التلقائي للتحليل والترجمة للمستندات متعددة اللغات.

(٢) اختصار لكلمة (Open Handwriting Recognition and Transcription) التعرف على الكتابة اليدوية والصوتية المفتوح.

(٣) يقصد بالتشذيب (lemmatization) عملية إزالة اللواصق من الكلمة وإعادتها إلى ساقها أو جذعها (توضيح من المترجمة).

الإعراب (mood) أما في الأسماء والصفات فيمثل الحالة النحوية (case). ومن ثَم، يوضع تشكيل بسيط لا يمثل التشكيل النهائي للكلمة.

وقد أنجزت الكثير من الأبحاث على التشكيل الآلي للغة العربية باستخدام مجموعة واسعة من التقنيات، انظر [11، 12، 33، 34، 35].

٢,٤ المزيد من القراءات

قد يبدو الخط العربي لغير المطلعين عليه مخيفاً، ومع ذلك، فإنه من السهل تعلمه مقارنة مثلاً بالصينية. ويوجد في السوق العديد من الكتب لتعليم القراءة والكتابة بالعربية. إن الألفة مع الخط العربي ستزيل بالتأكيد الغموض عن بعض جوانبه وهذا ما ينصح به الباحثون والمطورون في مجال اللغة العربية (لذا ينصح بالبحث عن كتب لتعليم العربية في موقع قوقل للكتب (http://books.google.com/books?q=arabic+script).

الجدول رقم (٢,١). الترميزات المختلفة للأحرف العربية.

Arabic	Unicode Letter Name	HSB	Buckwalter			CP-1256	ISO-8859-6	Unicode
			base	*xml*	*safe*			
ء	Hamza	’	’	’	C	C1	C1	0621
آ	Alef Madda Above	Ā	|	|	M	C2	C2	0622
أ	Alef Hamza Above	Â	>	O	O	C3	C3	0623
ؤ	Waw Hamza Above	ŵ	&	W	W	C4	C4	0624
إ	Alef Hamza Below	Ǎ	<	I	I	C5	C5	0625
ئ	Yeh Hamza Above	ŷ	}	}	Q	C6	C6	0626
ا	Alef	A	A	A	A	C7	C7	0627
ب	Beh	b	b	b	b	C8	C8	0628
ة	Teh Marbuta	ħ	p	p	p	C9	C9	0629
ت	Teh	t	t	t	t	CA	CA	062A
ث	Theh	θ	v	v	v	CB	CB	062B
ج	Jeem	j	j	j	j	CC	CC	062C
ح	Hah	H	H	H	H	CD	CD	062D
خ	Khah	x	x	x	x	CE	CE	062E
د	Dal	d	d	d	d	CF	CF	062F
ذ	Thal	ð	*	*	V	D0	D0	0630
ر	Reh	r	r	r	r	D1	D1	0631
ز	Zain	z	z	z	z	D2	D2	0632
س	Seen	s	s	s	s	D3	D3	0633
ش	Sheen	š	$	$	c	D4	D4	0634
ص	Sad	S	S	S	S	D5	D5	0635
ض	Dad	D	D	D	D	D6	D6	0636
ط	Tah	T	T	T	T	D8	D7	0637
ظ	Zah	Ď	Z	Z	Z	D9	D8	0638
ع	Ain	ς	E	E	E	DA	D9	0639
غ	Ghain	γ	g	g	g	DB	DA	063A
ف	Feh	f	f	f	f	DD	E1	0641
ق	Qaf	q	q	q	q	DE	E2	0642
ك	Kaf	k	k	k	k	DF	E3	0643
ل	Lam	l	l	l	l	E1	E4	0644
م	Meem	m	m	m	m	E3	E5	0645
ن	Noon	n	n	n	n	E4	E6	0646
ه	Heh	h	h	h	h	E5	E7	0647
و	Waw	w	w	w	w	E6	E8	0648
ى	Alef Maksura	ý	Y	Y	Y	EC	E9	0649
ي	Yeh	y	y	y	y	ED	EA	064A

الجدول رقم (٢,٢). الترميزات المختلفة للتشكيل وعلامات الترقيم والأحرف المستعارة في العربية.

Arabic	Unicode Letter Name	HSB	Buckwalter			CP-1256	ISO-8859-6	Unicode
			base	*xml*	*safe*			
ً	Fathatan	ā	F	F	F	F0	EB	064B
ٌ	Dammatan	ū	N	N	N	F1	EC	064C
ٍ	Kasratan	ī	K	K	K	F2	ED	064D
َ	Fatha	a	a	a	a	F3	EE	064E
ُ	Damma	u	u	u	u	F5	EF	064F
ِ	Kasra	i	i	i	i	F6	F0	0650
ّ	Shadda	~	~	~	~	F8	F1	0651
ْ	Sukun	.	o	o	o	FA	F2	0652
ٰ	Dagger Alef	á	`	`	e			0670
آ	Alef Wasla	Ä	{	{	L			0671
ـ	Tatweel	_	_	_	_	DC	E0	0640
،	Comma	,	,	,	,	A1	AC	060C
	Soft Hyphen	-	-	-	-	AD	AD	00AD
؛	Semicolon	;	;	;	;	BA	BB	061B
؟	Question Mark	?	?	?	?	BF	BF	061F
پ	Peh	p	P	P	P	81		067E
چ	Tcheh	c	J	J	J	8D		0686
ڤ	Veh	v	V	V	B			06A4
گ	Gaf	g	G	G	G	90		06AF