قياس درجة انتماء فئتي الأرقام الهندعربية إلى منظومة الكتابة العربية(*)

إعـــداد

د. محمد يسرى النحاس(**) أ.د. محمد يونس الحملاوى(***)

١- مقدمة:

يهدف هذا البحث إلى عمل دراسة مقارنة بين درجة انتماء كل من فئتى الأرقام الهندعربية المستخدمتين في منظومة الأعداد العشرية إلى منظومة الكتابة العربية. هاتان الفئتان تسميان خطأ فئة الأرقام الهندية وفئة الأرقام العربية[١] . بينما الأصح تسميتهما جغرافياً بأرقام المشرق وأرقام المغرب[٦]. وقد كان حافزنا إلى اجراء هذا البحث هو الأجابة عن السؤال أى من فئتى الأرقام أفضل فى الاستخدام فى منظومة الكتابة العربية؟

إن المقارنة بين فئتى الأرقام الهندعربية يمكن أن تتم بناء على مجموعة من المقاييس مثل قابلية الأرقام للتعرف الآلى أو قيمتها الجمالية أو نفعية استخدامها فى التقنيات المختلفة بالإضافة إلى قيمتها التراثية والحضارية. والخاصية التى نبحثها هنا هى درجة انتماء أشكال الأرقام فى كل فئة إلى منظومة الكتابة العربية، وبديهيا فإن انتماء أى من فئتى الأرقام إلى منظومة الكتابة العربية هو إنتماء تاريخى أولا وقبل كل شئ. ولكن التحقق من نتائج هذا المبحث التاريخى يمكن أن يتم أيضا بالمبحث الهندسى الذى يتيح لنا

(*) هذا البحث أجرى بالتعاقد مع الجمعية المصرية لتعريب العلوم .

(*) أستاذ مساعد، قسم هندسة النظم والحاسبات، كلية الهندسة، جامعة الأزهر .

(**) أستاذ هندسة الحاسبات، قسم هندسة النظم والحاسبات، كلية الهندسة، جامعة الأزهر .

استخدام القياس الكمى فى إجراء المقارنات الموضوعية. هذا الحكم الموضوعى على درجة الانتماء فى السؤال المطروح قد يخضع لعدة معايير منها:

١-درجة تشابه نمط أشكال الأرقام فى كل فئة مع نمط أشكال الحروف العربية.

٢-درجة توافق نمط أشكال الأرقام فى كل فئة مع نمط أشكال الحروف العربية.

٣-درجة توافق نمط كتابة الرقم مع نمط كتابة الحرف العربى.

...-

وإذ تقوم الجمعية المصرية لتعريب العلوم بعمل الدراسات والبحوث فى هذا المجال لوضع الأسس العلمية لتوصيف منظومة الكتابة العربية آخذة فى الاعتبار كل المعايير فإننا نستكمل فى هذا البحث الدراسة السابقة فى مقارنة فئتى الأرقام من حيث معدل التعرف على عناصر كل منهما[٦]. وأما المعيار الذى نقترحه للمقارنة بين مجموعتى الأرقام الهندعربية فى انتمائهما إلى منظومة الكتابة العربية يعتمد قياس درجة التشابه بين أنماط الأشكال فى كل مجموعة منهما مع أشكال حروف الكتابة العربية. ذلك أنه من الحقائق المعروفة فى علم التعرف على الأنماط أن انتماء كائن ما إلى طبقة من الطبقات يتناسب طردياً مع تشابهه وكائنات هذه الطبقة ويتناسب عكسيا مع تشابهه مع كائنات طبقته الخاصة.

سنستخدم في هذا البحث المدخل الإحصائى لعلم التعرف على الأنماط كوسيلة للمقارنة بين منظومتى الكتابة: منظومة كتابة المشرق العربية، ومنظومة كتابة المغرب العربية. هذا المدخل الإحصائى يعتمد على الاختيار الأفضل لثلاث أركان للمنهاج الموضوعى فى المقارنة وهى: التمثيل،

المعيار، والإحصائيات. ولإيضاح هذا المنهاج نبدأ بتعريف نمط الأشكال والكائنات المنتمية له وكيفية تمثيله من خلال متجه السمات، ونحدد المعيار المستخدم في عملية المقارنة بحيث تكون نتائج عملية المقارنة موضوعية غير متحيزة ومعتمدة عملياً. ثم نعرض الإحصائيات المختلفة الممكن استخدامها فى المقارنة المطروحة. بعد أن نحدد أركان منهاج المقارنة نعرض كيفية تطبيق هذا المنهاج لإجراء المقارنة الكمية المطلوبة ونعرض نتائج التطبيق ممثلة فى الإحصائيات المتنوعة ونوضح كيفية تفسير هذه النتائج.

٢- منهاج المقارنة:

إن المقارنة بين منظومتين للمفاضلة بينهما يجب أن تتم من خلال منهاج موضوعى. والمنهاج الذى نعتمده فى هذه الدراسة يتحدد بثلاث أركان:

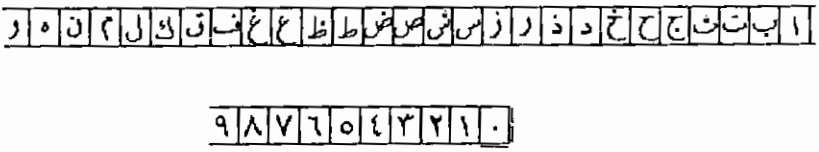
أولاً: تمثيل كائنات المنظومة، ثانياً: المعايير الثابتة للحكم على كائنات المنظومة، ثالثاً: الخطوات الواجب اتباعها لإجراء عملية المقارنة المطلوبة. وقد راعينا فى اختيارنا لكل ركن من هذه الأركان أن لايؤثر بصورة متحيزة على نتائج المقارنة.

أولاً: التمثيل:

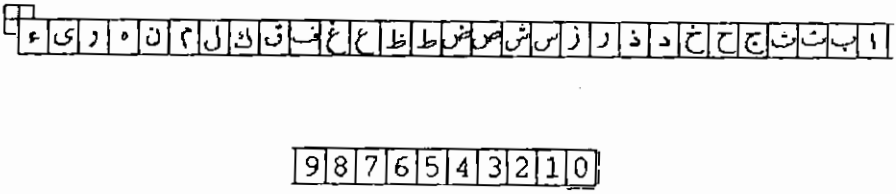
تمثيل منظومة الكتابة هنا يعنى وضع النموذج الرياضياتى لأشكال الرموز المستخدمة فى الكتابة. هذا النموذج يجب أن يسمح بالتحليل الكمى وأن يكون شاملا لجميع السمات الخاصة بالمنظومة؛ ومنظومتى الكتابة العربية يمكن تمثيلهما كما فى الشكل ١ والشكل ٢. حيث أن أشكال الأرقام والحروف هى التى تهمنا فى هذا البحث فإننا سنمثل منظومة الأرقام

والحروف بمجموعة الأشكال النمطية المستخدمة ؛ ر؛ والمكونة من تسع وثلاثون شكلاً كالآتى :

ر = { ش١، ش٢، ...، ش١، ...، ش ٣٩ } (١,١)



شكل ١ . أشكال منظومة أرقام المشرق والحروف العربية .



شكل ٢ . أشكال منظومة أرقام المغرب والحروف العربية .

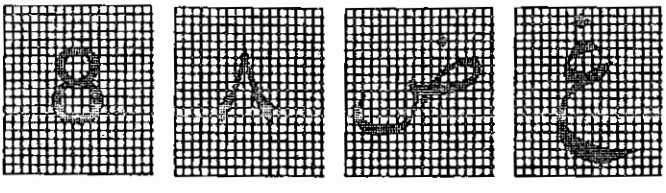
وحيث أن الإختلافات المطروحة واقعياً فى البنط أو الفنط المستخدم ليست ضرورية لهذه المقارنة لأنها بالرغم من تنوعها وكثرتها إلا فإنها تحتوى دائما على السمات الأساسية لنمط الكتابة العربية، وما غير ذلك فهو

تشويه لنمطها، فقد اختيرت هذه الأشكال ممثلة لأرقام وحروف منظومة الكتابة العربية حيث أنها تمثل إلى حد كبير نمط الحروف والأرقام العربية المستخدمة فى الكتابة والطباعة. والفنط المستخدم هو الخط العربى التقليدى وكل رمز فيه مرسوم بخط نحيف، أما حجم الرمز فهو أكبر من الحد الأدنى الضرورى لتمييز الرمز مما يسمح بدراسة الفروق الدقيقة فى بنية الرموز.

ويمثل كل نمط من الأشكال بمتجه السمات ؛ ش ا ؛ المكون من عدد ؛ ن ؛ من السمات الأساسية ؛ س ل ؛ كالآتى :

ش ا = (س ١، س ٢، ...، س ل، ...، س ن) م (١،٢)

في صورتها البدائية تعتبر مجموعة نقاط الصورة الثنائية هى سمات ذلك الشكل كما هو موضح فى الشكل ٣، الذى يمثل أربع عينات لأشكال الأرقام والحروف. وفى هذه العينات فإن شبكة الصورة الثنائية الممثلة لكل شكل تتكون من ١٧×٢٠ من النقاط . وطبقاً للتعريف السابق (١،٢) فإن متجه سمات الشكل يتكون من ٣٤٠عنصراً .



شكل ٣. التمثيل الثنائى لعينات الحروف والأرقام فى منظومتى الكتابة العربية

هذه السمات لاتصلح للمقارنة كما هى حيث أنها تتغير بالنقل أوالدوران أو الإنكماش فى فراغ الصورة. لذلك فإن شكل الحرف يتم معالجته أولاً

باستخدام استحالة خطية؛ لقياس بعد نقاط الشكل عن مركز ثقله ومعايرتها بالنسبة لمحاور الشكل الأساسية وحجمه[٣]. ونتيجة لهذا فالشكل فى الصورة الثنائية الناتجة يتميز بأنه لايتغير مع أية تحويلات هندسية فى فراغ الصورة ويمكن تمثيل كل نمط بمتجه آخر للسمات، شَ، المكون من نفس العدد ن من السمات.

ثانياً : المعيار

اخترنا في هذا البحث معامل تانيموتو ؛ ت ؛ معيارا لقياس درجة التشابه بين أشكال الحروف والأرقام. ويعرف معامل تانيموتو ؛ ت ؛ بين كائنين ش١، ش٢ بالعلاقة الآتية [٢] :

$$ت = \frac{ش١ \cdot ش٢}{ش١ \cdot ش١ + ش٢ \cdot ش٢ - ش١ \cdot ش٢} \qquad (١,٣)$$

هذا المعامل يمثل النسبة بين عدد السمات المشتركة بين الكائنين إلى عدد السمات الخاصة بكليهما. وكما هو واضح من التعريف فهذا المعيار لايعتمد على عدد السمات الممثلة لكل كائن ولا يعتمد كذلك على قيم تلك السمات . ولهذا فهو شائع الاستخدام فى علوم تصنيف الكائنات.

ثالثاً : اجراء المقارنة

بعد أن عرفنا وضعيا نموذج تمثيل الرمز فى منظومة الكتابة العربية وكذلك المعيار المستخدم فى حساب درجات التشابه بين الرموز وبعضها نحدد الخطوات لحساب درجات انتماء كل فئة من فئتى الأرقام إلى منظومة الكتابة العربية والمقارنة بينهما. هذه الخطوات هى:

١- تولد أشكال منظومة المشرق العربية؛ رش ؛ ومنظومة المغرب العربية ؛ رغ ؛ بواسطة برنامج للرسم بالحاسوب .

٢-تعالج الأشكال الناشئة بحصرها داخل نافذة ثم تتحف هذه الأشكال وتحدد بنافذة مكونة من ١٧×٢٠ نقطة .

٣- يحسب لكل شكل من الأشكال ؛ ش١ ؛ متجه مركز الثقل ؛ م١ ؛ ومصفوفة الانتشار ؛ ك١ ؛ ودالة الاستحالة ؛ د١ .

٤-يعاد توقيع النقاط السوداء فى كل شكل من الأشكال ؛ ش١ ؛ باستخدام استحالة خطية لينشأ عن ذلك الشكل ؛ شَ١ الذى لايعتمد على التحولات فى فراغ الصورة.

٥-تحسب درجة التشابه بين جميع الثنائيات الممكنة للأشكال ؛ (شَ١، شَ٢) ؛ فى كل منظومة على حدة وتوضع فى صورة مصفوفة متماثلة .

٦-يجرى نوعان من الإحصائيات على كل مصفوفة على حدة :

- إحصائيات فردية تدرس علاقة كل شكل من أشكال المنظومة ببقية أشكال المنظومة .

- إحصائيات كلية تدرس توزيع درجات التشابه داخل المنظومة ككل .

٧-تحلل النتائج السابقة للمقارنة بين درجة انتماء كل فئة من فئات الأرقام إلى منظومة الكتابة العربية.

٣- التطبيق والنتائج

طبق هذا المنهاج لقياس درجة انتماء كل من فئتى الأرقام الهندعربية إلى منظومة الكتابة العربية، وذلك بقياس درجة التشابه بين ثنائيات الأشكال في داخل كل من منظومتى الكتابة العربية، وسجلت النتائج الخاصة بكل

منظومة في صورة مصفوفة تشابه تمثل درجة التشابه بين أنماط أشكال المنظومة، ثم استخدم المنهج الإحصائى فى تحليل هذه النتائج.

فى المنهج الإحصائى لعلم " التعرف على الأنماط " يتحقق الوصف الشامل للأنماط من خلال منحنى التوزيع التكرارى أو المنحنى التكراري التصاعدى. وفى هذه الدراسة وباعتبار أن درجة التشابه بين الأنماط هى المتغير المستقل فإن منحنى التوزيع التكرارى يمثل العلاقة بين درجة التشابه وعدد ثنائيات الأشكال التى لها نفس درجة التشابه . والمنحنى التكرارى التصاعدى يمثل العلاقة بين درجة التشابه وعدد ثنائيات الأشكال التى لاتزيد درجة التشابه بينها عن هذه الدرجة.

ولقياس درجة الانتماء وعمل المقارنة وفقا لهذا المنهج نحتاج هنا إلى حساب ورسم خمس منحنيات هى:

١-منحنى التوزيع التكرارى لفئة أرقام المشرق.

٢-منحنى التوزيع التكرارى لفئة أرقام المغرب.

٣-منحنى التوزيع التكرارى لفئة الحروف العربية.

٤-منحنى التوزيع التكرارى لمنظومة كتابة المشرق العربية.

٥-منحنى التوزيع التكرارى امنظومة كتابة المغرب العربية.

وبتحليل منحنى التوزيع التكرارى لكل منظومة على حدة يمكننا معرفة إذا ما كانت فئة الأرقام المستخدمة تمثل نمطا مستقلا عن نمط بقية عناصر المنظومة أم أنها تمثل مجرد فئة جزئية من عناصر المنظومة. ويستدل على ذلك بظهور وسطين حسابيين فى منحنى التوزيع التكرارى لدرجات التشابه، وكلما زاد الفرق بين هذين الوسطين كلما قلت درجة انتماء كل من النمطين إلى الآخر، أى نمط الحروف ونمط الأرقام. أما المقارنة بين منحنى التوزيع

التكرارى لكل منظومة ومنحنى التوزيع التكرارى لفئة الأرقام المستخدمة فيها يظهر مدى التوافق بين نمط الأرقام المستخدمة ونمط الكتابة العربية وذلك بقياس الزيادة فى التشتت فى درجة التشابه الناجم عن إضافة عناصر هذه الأرقام إلى عناصر الحروف.

٤- الخلاصة:

يقدم هذا البحث مقارنة كمية بين فئتى الأرقام الهندعربية من حيث درجة انتماء كل منهما إلى منظومة الكتابة العربية. هذه المقارنة الكمية صممت على أساس منهجى موضوعى . تحققت الموضوعية فى المنهاج من خلال حياد عناصره الأساسية؛ التمثيل، المعيار، والإحصاء . وكنتيجة ثانوية لهذا البحث تحددت السمات الرئيسية المشتركة لمنظومة الكتابة العربية، ومدى التوافق فى منظومة كتابة المشرق العربية بين حروف وأرقام المشرق العربى.

٥-المراجع:

[١] Encyclopedia Universalis, "Notation mathématique", vol. 11 , pp 908, France, 1977.

[٢] R.O. Duda & P. Hart, "Pattern Classification and Scene Analysis", A. Wiley Interscience Publication , New York , 1974.

[٣] K. Fukunaga , "Statistical Pattern Recognition", Academic Press, inc., San Diego, 1990.

[٤] M.Y. Mahmoud (El Nahas) et al : "A statistical approach for arabic character recognition", 12th international congress for statistics, computer science, social and demographic research, Cairo, 1987,

[٥] M.A. El Hamalaway & S.H. El Ramly , "A Novel Arabic Numerals Shapes for Optical Character Recognition", Thirteen International Congress for Statistics, Computer Science, Social and Demographic Research, 26-31 march 1988.

[٦] محمد يسرى النحاس ومحمد يونس الحملاوى؛ قياس درجة التشابه فى مجموعتى الأرقام الهندعربية"،المؤتمر الخامس عن الحاسب الآلى بين النظرية والتطبيق، الإسكندرية، ١٢-١٤ سبتمبر ١٩٩٥م؛ صفحة ٤٤-٥٠