المبحث الثالث: تصميم المدونات اللغوية وبناؤها: عبدالمحسن بن عبيد الثبيتي

مدينة الملك عبدالعزيز للعلوم والتقنية

الرياض، المملكة العربية السعودية

aalthubaity@kacst.edu.sa

مستخلص

بغض النظر عن التعريفات المتعددة للمدونات اللغوية (المدونة فيما بعد) باختلاف توجهات الباحثين في هذا المجال، فإنها بيانات يحاول الباحثون بواسطتها دراسة اللغة من خلال استخدامها الفعلي كما فعل علماء اللغة من قبل، وما زالوا في دراستهم للغة ووضعهم لقواعدها ولكن بمنهج جديد يعتمد على شواهد كثيرة أصبحت متاحة هذا العصر بسبب تطور الحاسب وتوافر النصوص الإلكترونية بشكل كبير. ولأنه من المستحيل عمليا جمع كل ما قيل وكُتب للحصول على أحكام دقيقة وفاصلة، فإننا أمام خيار آخر يبدو أكثر واقعية ويستند إلى أساس علمي ألا وهو جمع عينة متوازنة وممثلة للغة أو إحدى صورها أو ظواهرها محل الدراسة. ومثلما هو حاصل في مجالات علمية مختلفة، مثل: الطب أو علم النفس؛ فإن النتائج المتحصل عليها اعتمادا على دراسة العينات وتحليلها يمكن الوثوق بها متى ما توفرت ثلاثة شروط رئيسة، هي: 1) ألا تكون العينة منحازة وأن تكون كافية للدراسة، 2) أن تدرس هذه العينة وتحلل باتباع منهج علمي، 3) الحصول على نفس النتائج متى ما استخدم المنهج نفسه على ذات العينة. يقدم هذا الفصل إطارا عمليا للإجراءات التي تمكّن الباحث من تحقيق الشرط الأول من خلال اتباع معايير واضحة لتصميم وجمع نصوص المدونة لتكون عينة متوازنة وممثلة لمجال الدراسة اللغوية مع تعريف مبسط لبعض الأدوات التي تساعد في دراسة المدونات بشكل ميسر متى ما توفر المنهج العلمي لدى الباحث (الشرط الثاني). ويمكننا تحقيق الشرط الثالث عندما تكون جميع المعلومات الخاصة بتصميم وبناء المدونة موثقة ومتاحة للباحثين الآخرين مع إتاحة نصوص المدونة مجاناً أو بمقابل مادي- بما لا يخرق قوانين وأنظمة الحقوق الفكرية أو الخصوصية الشخصية – ليتمكن الباحثون من التحقق من نتائج الدراسات الأخرى التي أجريت على المدونة.

مقدمة

قبل أن يفكر الباحث في تصميم وبناء مدونته اللغوية الخاصة، عليه أن يجيب عن السؤالين التاليين:

1- ما الغرض الذي من أجله أريد أن استخدم المدونة؟

2- هل توجد مدونة أو مدونات تجيب عن أسئلة بحثي؟

عادة ما يتم تحديد الغرض من استخدام المدونة بوضع سؤال أو عدة أسئلة تكون مدار البحث في المدونة، وتوجهُ هذه الأسئلة الباحث لنوعية المعلومات التي يريدها وكيفية الحصول عليها. وتتطلب الإجابة عن السؤال الثاني معرفةً بتصميم ومحتويات المدونات المتوفرة وأدوات البحث فيها واطلاعاً على الأبحاث التي تمت باستخدام هذه المدونات؛ ليطلع الباحث على جوانب القوة والقصور فيها، وليعرف مدى قدرتها على الإجابة عن أسئلة بحثه. وإن لم يجد الباحث مدونة تفي بالإجابة عن أسئلته فقد يضطر إلى أن يعدّل أسئلته أو محددات بحثه ليتمكن من استخدام المدونات المتوفرة، أو أن يغيرها تماما لتتناسب مع ما هو متوفر، أو قد يلجأ – إن كان لديه الوقت والإمكانات – إلى بناء مدونته الخاصة.

وعلى الرغم من توافر عدد لا بأس به من المدونات العربية (انظر المبحث الأول من هذا الكتاب) إلا أن أغلبها لا يوضح المعايير التي أُتبعت في تصميمها، وجمع نصوصها، وما هو الفرق بين التصميم والمحتوى الفعلي للمدونة بشكل واضح، ولعلي أستثني من هذا ثلاثاً من المدونات العربية وهي مدونة اللغة العربية المعاصرة Corpus of Contemporary Arabic من جامعة ليدز (السليطي وأتول 2006، Al-Sulaiti and Atwell)، والمدونة العالمية للغة العربية International Corpus of Arabic من مكتبة الإسكندرية (الأنصاري وآخرون 2007، .Alansari et al)، والمدونة العربية لمدينة الملك عبدالعزيز للعلوم والتقنية (الثبيتي 2014–Al-Thubaity). وأفرّق هنا بين أن توضع معايير تستخدم مرشدًا لجمع نصوص المدونة واستخدامها لجمع النصوص وبين أن يتم توضيح محتويات المدونة وتوزيعها بعد جمع نصوصها دون الاسترشاد بأية معايير تضبط هذا الجمع.

يتطرق القسم الثاني من هذا المبحث بشكل مبسط ومختصر إلى أهم المعايير التي من المفترض وضعها وتوثيقها عند تصميم المدونات لتكون متوازنة وممثلة للغة محل الدراسة. ويتعرض القسم الثالث إلى خطوات بناء المدونات بحيث تحقق معايير التصميم. أما القسم الرابع فيتعرض لأهم الأدوات اللازمة لمعالجة المدونات. وفي القسم الخامس أضع مثالاً تطبيقيا لما ذكر في القسمين الثاني والثالث. وفي القسم السادس مثال تطبيقي على خطوات التصميم والبناء المذكورة في القسمين الثالث والرابع. أما خاتمة المبحث وخلاصته فهي في القسم السابع.

معايير التصميم

يتطرق هذا القسم إلى أهم المعايير التي يجب وضعها في الاعتبار وتحديدها بوضوح قبل البدء بجمع نصوص المدونة. تساعد هذه المعايير –إن طبقت بشكل دقيق عند جمع النصوص– على أن تكون المدونة قادرة على إجابة أسئلة البحث أو الغرض الذي بُنيت من أجله. تُعد هذه المعايير إطارًا عامًّا يحدد نوعية وكمية الجهد الذي يجب أن يبذل في الخطوات التي تلي التصميم. إنّ وضع هذه المعايير والاهتمام بها منذ البداية ينقل المدونات من كونها بيانات يهتم بها بعد إنجازها "مجموعة من البيانات اللغوية المكتوبة أو المنطوقة" كما عرفها كريستال (كريستال 1992، Crystal) إلى كونها عملاً منهجياً مخططاً ومدروساً له أهداف واضحة منذ البداية "مجموعة من نصوص اللغة في صورة إلكترونية تجمع اعتمادًا على معايير خارجية؛ لتمثل قدر المستطاع اللغة أو أحد صورها لتكون مصدرًا للأبحاث اللغوية" (سنكلير 2005، Sinclair).

والمقصود بالمعايير الخارجية هنا، المعايير التي تعتمد على الوظيفة التواصلية للنص في المجتمع الذي ظهر فيه، وهذه المعايير تشبه إلى حد كبير المعلومات الببليوغرافية للنص مثل الوعاء الذي صدر فيه النص وموضوعاته المختلفة والمنطقة الجغرافية التي صدر منها، وقد تشمل معلومات أخرى مثل جنس الكاتب وجنسيته ومستواه الاجتماعي. أما المعايير الداخلية للنص فإنها تتعلق بالبنية اللغوية الداخلية للنص ذاته مثل كونه يحتوي على تركيب نحوي أو صرفي بصيغة معينة. إن الاعتماد على المعايير الداخلية يعزز بشكل كبير ظهور هذه التراكيب وما يلازمها عادة وبالتالي لا يعكس صورة محايدة عن استخدامها الفعلي مقارنة بغيرها من الظواهر اللغوية.

إن أول ما يتحكم في معايير بناء المدونات التي سوف نتطرق إليها في الأقسام الفرعية التالية هو الغرض الذي تُبنى من أجله، هذا الغرض يجب أن يكون واضحاً ومحدداً بدقة منذ البداية. ويمكننا بشكل عام أن نقسم الأغراض التي تبنى لأجلها المدونات إلى قسمين رئيسين: أغراض خاصة/محددة وأغراض عامة/شاملة.

يتميز القسم الأول بمحاولته الإجابة عن أسئلة محددة، وبالتالي فإن نتائج الدراسة لا يمكن تعميمها على اللغة ككل، مثل: أن يكون الغرض من بناء المدونة دراسة التراكيب النحوية في لغة الشعر الجاهلي أو دراسة الأفعال في أبحاث علوم الكيمياء أو دراسة أخطاء متعلمي اللغة العربية لغة ثانية، ويمكن أن يلحق بهذا القسم المدونات التي تبنى لأغراض معالجة اللغة أو نمذجتها فقط ولم يوضع في الحسبان عند تصميمها الدراسات اللغوية. ومن أمثلة هذه المدونات مدونة متعلمي العربية

من جامعة ليدز (الفيفي وآخرون 2014، Alfaifi et al.) (انظر المبحث الثاني من هذا الكتاب)، ومدونة مدينة الملك عبدالعزيز للعلوم والتقنية لتصنيف النصوص (خورشيد والثبيتي 2013، Khorsheed and Al-Thubaity)، ومدونة تصحيح الأخطاء الإملائية (الكنهل وآخرون 2012، Alkanhal et al.).

أما القسم الثاني من الأغراض التي تبنى لأجلها المدونات فيتميز بشكل عام بتنوع موضوعاته ومحاولته الوصول إلى مدونة تستطيع الإجابة عن أسئلة متنوعة، ويندرج تحت هذا النوع من المدونات المدوناتُ المرجعية. ومن أمثلة هذه المدونات المدونة العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، ومدونة آر تن تن arTenTen على موقع سكتش إنجن[7] Sketch Engine.

فيما يلي شرح لأهم معايير تصميم المدونات التي جمع شتاتها سنكلير (سنكلير 2005) وكانت نتاجاً لخبرته الطويلة في العمل على المدونات وإشرافه المباشر على المدونة الشهيرة بنك اللغة الإنجليزية The Bank of English مع بعض النقاش والرجوع إلى مراجع أخرى يتم الإشارة إليها في حينها حسبما تقضي الحاجة. وهذه المعايير هي:

1.3 لغة المدونة

لا يقتصر تحديد لغة المدونة على التحديد العام للغة مثل كونها اللغة العربية أو الإنجليزية، بل يتعدى ذلك إلى تحديد تفاصيل أكثر دقة. ومن أمثلة هذه التفاصيل أن تكون لغة المدونة هي اللغة الفصحى المعاصرة أو القديمة التراثية (إن جازت هذه التسمية) أو حتى اللهجات إن كان هذا هو مطلب الدراسة، كما أن التفاوت داخل هذه الأنواع الثلاثة يجب أن يُنظر إليه ويؤخذ بالحسبان أيضاً. فاللغة العربية المعاصرة –على سبيل المثال –على الرغم من أن لها ملامح عامة عند متحدثيها في البلدان العربية إلا أن هناك اختلافات واضحة على المستوى اللفظي –على الأقل– فيما بينهم. ومن أمثلة هذه الاختلافات اللفظية في العربية المعاصرة بين لغة الصحافة السعودية والمغربية كلمتا "بنوك" و "أبناك".

2 .3 طبيعة النصوص

تختلف الطبيعة التي تظهر فيها اللغة البشرية، فهي قد تكون منطوقة وهذا الأغلب وقد تكون مكتوبة في صور متعددة وقد تكون أيضًا لغة إشارة، ومهما يكن الأصل الذي ظهرت فيه اللغة فيجب تحويل هذا الأصل إلى صورة إلكترونية قابلة للمعالجة الآلية، وفي الأغلب فإن هذه الصورة تكون على صيغة TXT.

إن تحديد نسبة ما سوف تحتويه المدونة من نصوص منطوقة أو مكتوبة أو كليهما له أثر كبير على اختيار الأوعية والفترات الزمنية وكذلك في الجهد الذي سوف يستغرق في جمع محتويات المدونة، فالجزء المنطوق من المدونة يستغرق وقتاً أطول بكثير في الحصول عليه ومن ثم تحويله إلى نص مكتوب مقارنة بالجزء المكتوب أصلاً، وتزداد الصعوبة لو كانت اللغة هي اللغة الفصحى، ناهيك عن التكلفة المادية الباهظة لذلك والاحترازات اللازمة لعدم خرق الخصوصية الشخصية.

3. 3 تاريخ النصوص

تختلف الفترات الزمنية التي يجب أن تغطيها المدونة باختلاف الأغراض التي تبنى من أجلها، فالمدونات التي تسعى لدراسة اللغة الحديثة أو المتخصصة عادة ما تتميز بقصر الفترات التي تغطيها، وتتراوح هذه الفترات من سنة إلى عدة سنوات. وعلى النقيض من ذلك المدونات التي تسعى لدراسة تطور اللغة إذ تتضمن نصوصاً من فترات زمنية طويلة تتراوح بين عشرات السنين إلى المئات منها. وبالطبع فإن طول الفترة الزمنية التي تغطيها المدونة يميزها عن غيرها من المدونات التي تغطي فترات أقصر وخصوصا فيما يتعلق بالمدونات المرجعية Reference Corpora سواء كانت حديثة أو تاريخية. ويجب مراعاة أن تتضمن المدونة نصوصا تغطي جميع أجزاء الفترة الزمنية لا أجزاء متفرقة منها قدر الإمكان، وتسعى بعض المدونات التي تهتم بأغراض مقارنة اللغة بين فترة زمنية وأخرى إلى تضمين نصوص من فترات زمنية مختلفة ومتباعدة مثل أن تحوي المدونة نصوصًا من السنوات العشر الأولى من القرنين العشرين والواحد والعشرين الميلاديين.

3. 4 المنطقة الجغرافية

يقصد بالمنطقة الجغرافية هنا البلد أو البلدان التي صدرت فيها النصوص، وقد تستدعي الحاجة تحديد المناطق المختلفة داخل البلد الواحد نفسه، فعلى سبيل المثال لو كان الغرض إنشاء مدونة لغوية تختص بالمملكة العربية السعودية فإنه من المفترض الأخذ بعين الاعتبار جميع مناطق المملكة.

وتزداد قيمة المدونة عند تضمينها نصوصاً من بلدان ومناطق مختلفة تتحدث نفس اللغة بحيث تكشف الاختلافات اللغوية والثقافية بين هذه البلدان أو المناطق، فعلى سبيل المثال لا يمكن لمدونة أن تكشف عن الأنماط اللغوية المشتركة أو المختلفة في اللغة العربية ما لم تُضمن نصوصا من جميع البلدان العربية، ولكن التنوع في البلدان أو المناطق ليس شرطًا واجبًا بل إن الحاجة لهذا التنوع يمليه الغرض من المدونة فحسب.

3. 5 الوعاء

تظهر النصوص في أوعية مختلفة مثل الصحف والمجلات والكتب والرسائل الجامعية أو الشابكة، ولكل من هذه الأوعية سماته اللغوية العامة وخصائصه التي يمكن أن تميزه عن غيره على المستوى اللفظي والتركيبي، فلغة العلم في الرسائل الجامعية والدوريات المحكمة تتميز بالدقة والوضوح واستخدام المصطلحات العلمية، ولا تستخدم المجاز، فلا يمكن التعبير عن فكرة أو معنى معين إلا بمصطلح واحد وثابت متفق عليه في الأغلب؛ بينما الحال مختلف في لغة الصحافة الأدب مثلاً.

إن تحديد أوعية متنوعة لتُضمّن في المدونة يزيد من قيمتها وفائدتها لدراسات وتطبيقات مختلفة بخلاف المدونات المنحصرة في وعاء واحد، وعلى العموم فإن تعدد الأوعية يُعد من علامات المدونات المرجعية في الأغلب.

3. 6 المجال

لكل وعاء من الأوعية مجالات تختص به وقد تظهر فيه فقط ولا تظهر في غيره، فعلى سبيل المثال نجد الأخبار والمقالات والتقارير بوصفها مجالات مختلفة يمكن أن تكون في الصحف والمجلات ولا يمكن أن توجد في الكتب أو الرسائل الجامعية. وتتنوع اللغة المستخدمة داخل الوعاء الواحد باختلاف مجالاته، فالدوريات العلمية المحكمة على سبيل المثال تعد وعاء يجمع عدة مجالات مختلفة مثل أصول الفقه والطب والهندسة، وعلى الرغم من أن هذه المجالات قد تتسم بطابع واحد عام وهو استخدام لغة العلم إلا أنها تتفاوت في طرق التعبير والمصطلحات المستخدمة. إن التباين في الأفكار بين هذه المجالات والطرق المستخدمة في التعبير عنها يثري المدونة ويسمح بظهور الأنماط العامة للغة العلم وكذلك الأنماط الخاصة بكل مجال، كما أن معرفة هذه المجالات وتحديد نسبتها من كل وعاء يساعد في التخطيط المبكر لجمع نصوص المدونة وتيسيره فيما بعد.

3. 7 حجم العينة

في هذا المعيار يتم تحديد ما إذا كانت المدونة سوف تتضمن نصوصًا كاملة أم أجزاء من النصوص، ويوجد رأيان مختلفان بهذا الخصوص: فالرأي الأول يرى تضمين النصوص كاملة ما أمكن (سنكلير 2005)؛ لأن وجود النص كاملا يزيد من فرصة ظهور خواص لغوية متعددة ومختلفة ومترابطة في سياقها الطبيعي، وفي المقابل يرى الرأي الثاني أن اختيار جزء من النص يزيد من تنوع مواد المدونة، ناهيك عن العوائق المتعلقة بحقوق الملكية الفكرية وصعوبة الحصول على النصوص كاملة (ماير 2002). وعلى الرغم من ذلك فيجب عدم الاكتفاء بجزء واحد من النص، بل بأجزاء متعددة

من أوله وأوسطه وآخره (ماكنري وآخرون 2006). وبطبيعة الحال فإن مثل هذا القرار له علاقة بالنصوص ذات الحجم الكبير مثل الكتب والرسائل الجامعية وليس النصوص ذات الحجم الصغير مثل نصوص الصحف.

ومهما يكن فإن تحديد حجم عينة النصوص مرتبط بحجم المدونة، فقد يكون من المناسب أن تكون عينة النصوص أجزاء من النص عندما تكون المدونة صغيرة الحجم ومن الممكن الوصول للحجم بهذه الطريقة، أما إن كانت المدونة كبيرة الحجم فالأولى تضمين النص كاملاً، مع التأكيد على وجوب الحرص على استخدام نفس حجم العينة في جميع النصوص ما أمكن.

3. 8 حجم المدونة

يقاس حجم المدونة بعدد كلماتها. والكلمة هنا تعني أي مجموعة متتابعة من الرموز لا يفصل بينها فراغ، وبالتالي فإن بعض الكلمات –حسب هذا التعريف– قد تكون كلمة معروفة وصحيحة مثل "كتاب"، أو قد تكون أرقاما "9730" أو كلمات ليس لها معنى "عمخعج" أو كلمات تحوي أخطاء طباعية "كعنبوت" أو كلمات تمت إضافة الكشيدة[8] في وسطها "بســـم". وبالتالي فإن "بســـم" و"بسم" تعدان كلمتين مختلفتين بالنسبة لأدوات معالجة المدونات لاختلاف شكليهما على الرغم من كونهما كلمة واحدة. مثل هذه الأمثلة موجودة في أغلب المدونات خصوصاً الكبيرة منها، ونسبتها إلى إجمالي حجم المدونة لا يذكر ولكن الإشارة إليها لازمة لفهم معنى الكلمة التي بناء عليها يتم قياس حجم المدونة.

ولا بد من التنبيه على أن الآراء تتفاوت حول حجم المدونة، فهناك من يرى أنه كلما ازداد حجم المدونة كان ذلك أفضل (سنكلير 1991، Sinclair)، والسبب في ذلك أن أغلب كلمات المدونة لا تتكرر بالشكل الكافي، ولذا فإنه كلما زاد الحجم، سنحت الفرصة لظهور أنماط أكثر ثباتاً وقبولاً. ومع ذلك يرى عدد من الباحثين أن مدونة بحجم مليون كلمة كافية للأبحاث اللغوية التي تبحث عن الظواهر العامة في اللغة (انظر مناقشة هذا الموضوع في السليطي وأتول، 2006).

كما أن حجم المدونة يتعلق بنوعية العمل أو الدراسات المطلوبة، فالمدونات المتعلقة بالدراسات المعجمية تتطلب حجما أكبر مقارنة بالمدونات المتعلقة بالتراكيب النحوية، وقد ساعد التطور الحاصل في تقنية المعلومات والتوفر المتزايد للنصوص الإلكترونية على بناء مدونات لغوية كبيرة بصورة أسهل من ذي قبل.

وبشكل عام، فإن التحديد الدقيق للحجم الكافي لتحقيق الغرض الذي من أجله تبنى المدونة أمرٌ صعب للغاية ولا توجد معادلة رياضية لذلك؛ بل يخضع بشكل كبير للخبرات السابقة لمصمم المدونة بالإضافة إلى الخبرات المنشورة في هذا المجال مما هو مقارب لغرض المدونة ويتضح بصورة أكبر بعد استخدام المدونة.

ويمكن تحديد نسبة أو عدد الكلمات التي تمثل كل معيار من معايير تصميم المدونة بطريقتين: الطريقة الأولى تعتمد على التوزيع من الأعلى إلى الأسفل، حيث يتم تحديد الحجم الكلي للمدونة اللغوية ثم تحديد نسبة الجزء المكتوب والمنطوق منها، يتبع ذلك تحديد النسبة المناسبة لكل وعاء من أوعيتها، ثم يتم تقسيم هذه النسبة على مجالات الوعاء ثم موضوعاته، يتبع ذلك تحديد نسبة كل فترة زمنية ومنطقة جغرافية إن كانت المدونة تغطي عدة بلدان وفترات زمنية مختلفة، ويمكن بدلاً من ذلك توزيع الحجم الكلي للمدونة على الفترات الزمنية ثم البلدان ثم الأوعية والمجالات والموضوعات.

أما الطريقة الأخرى فتعتمد على التوزيع من الأسفل إلى الأعلى، فيتم فيها تحديد عدد الكلمات المناسب لكل موضوع ونسبة كل بلد وفترة زمنية من هذا الموضوع، وبالتالي فإن مجموع عدد كلمات كل موضوع يعطي العدد الإجمالي للمجال، ويكون مجموع عدد كلمات كل وعاء هو مجموع عدد كلمات كل مجالاته، وبذلك يمكن تحديد الحجم الكلي للمدونة اللغوية بجمع عدد كلمات كل وعاء فيها.

3. 9 معايير أخرى

إن المعايير السابقة تُعد أهم المعايير التي يجب وضعها في الحسبان عند تصميم المدونة، وهناك معايير أخرى يجب اعتبارها في بعض المدونات حسب الغرض الذي من أجله يتم بنائها. من هذه المعايير تحديد جنس الكاتب ذكراً أو أنثى، وكم النسبة التي يجب أن تحتويها المدونة من نتاج كلٍ من الجنسين. ومثل هذا المعيار مهم عندما يكون غرض المدونة الدراسات المقارنة بين كتابات الرجال والنساء. ومن الأمثلة الأخرى تحديد الفئة العمرية والمستوى التعليمي والاجتماعي أو حتى عدد المتابعين للمغرد إن كانت المدونة تدرس لغة الخطاب في تويتر Twitter. وهذه المعايير التي تم ذكرها إنما هي أمثلة فقط لبعض المعايير الخاصة التي يتطلبها تصميم مدونات ذات مواصفات وأغراض خاصة ولا تعني الحصر.

3. 10 التمثيل والتوازن

المقصود بالتمثيل هو قدرة المدونة على تمثيل اللغة أو صورها المختلفة محل الدراسة، ويرى ماكنري وآخرون (2006) أن ما يميز المدونات عن أي مجموعة عشوائية من النصوص هو قدرتها على تمثيل اللغة ومتغيراتها، ويرتبط التمثيل ارتباطاً وثيقاً بحجم المدونة والتنوعات المختلفة فيها ونسبة كل تنوع، فكلما زاد الحجم والتنوع كانت قدرة المدونة أكبر على تمثيل اللغة والإجابة عن أسئلة الدراسة.

أما المقصود بالتوازن فهو أن تمثل نصوص المدونة الواقع اللغوي كما هو في خارجها فلا تكون منحازة لمحتوى أو مستوى دون آخر، وليس بالضرورة أن يعني ذلك التساوي في حجم ما تحويه المدونة من كل وعاء أو مجال إلا إن كان ذلك لازماً مثلما هو الحال في الدراسات المقارنة. فلو نظرنا للنتاج الكتابي خارج المدونات فسنجد أن الغلبة للصحف مقارنة بالرسائل الجامعية، وعلى هذا يجب أن يكون حجم الصحف من أوعية المدونة المرجعية أكبر من الرسائل الجامعية بحسب الواقع فعلاً. وكذلك الحال عند التوزيع الجغرافي لمواد المدونة، فالدول التي تتميز بنتاج أكبر يجب أن يكون لها النصيب الأكبر من الحجم بحسب الواقع. وعلى الرغم من أن هذا الكلام صحيح في مجمله لكننا لا نستطيع أن نحدد بدقة تامة مقدار النسبة لكل جزء من محتوى المدونة.

إن التمثيل والتوازن مفهومان مهمان ومرتبطان لا يمكن فصلهما عن بعضهما، ويتحققان بشكل كبير عند مراعاة جميع المعايير السابقة بحسب الغرض من المدونة. وقد حظي هذان المفهومان باهتمام بالغ –ولا زالا– في كثير من الدراسات التي تناقشهما وتعرض العديد من وجهات النظر حولهما، انظر على سبيل المثال (سنكلير 1991) (أتكنز وآخرون 1992، Atkins et al.) (بايبر 1993، Biber) (بايبر وآخرون 1998، Biber et al.) (ليتش 2006، Leech). وبسبب الاختلاف الكبير حول سبل تحقيق التمثيل والتوازن كانت دقة تمثيل المدونات الكبرى الشهيرة للغة الإنجليزية، كالمدونة الوطنية البريطانية British National Corpus محل تساؤل عند بعض الباحثين (أحمد 2008، Ahmad). وللتغلب على هذا الإشكال وصعوبة تحقيقه فعليا بشكل كامل اقترح تيوبرت وكيرماكوفا (تيوبرت و كيرماكوفا 2007، Teubert and Cermáková) استخدام المدونات السانحة Opportunistic Corpora؛ إذ تمتاز المدونات السانحة بميزتين رئيستين، أولاهما: أنها كبيرة الحجم يجمع فيها كل ما يمكن جمعه من نصوص مختلفة، وثانيهما: أنها توفر معلومات تفصيلية شاملة

عن النصوص التي تحتويها، وعندها يكون بإمكان الباحث اختيار النصوص التي تمثل من وجهة نظره التمثيل والتوازن لدراسته التي يرغب القيام بها.

ومهما تكن وجهة النظر حيال التمثيل أو التوازن فإنني أرى أنه لا توجد مدونة لغوية ممثلة للغة وصورها المختلفة بشكل كامل، وأن ما يمكن الحكم عليه بشكل أدق هو قدرة المدونة على تمثيل موضوع الدراسة أفضل من غيرها من المدونات الأخرى.

3. 11 البيانات الأساسية للنصوص

البيانات الأساسية للنصوص شبيهة إلى حد كبير بالمعلومات الببليوغرافية. وتحديد هذه المعلومات الأساسية وتوفيرها للباحثين مع نصوص المدونة يساعد في إدارة المدونة، وتحديد مدى مناسبتها لغرض الدراسة بشكل كبير، وكذلك يمكن الاستفادة منها في تحديد الأجزاء الأكثر مناسبة للدراسات اللاحقة، والاكتفاء بها، كما تساعد البيانات الأساسية للنصوص في الاستفادة من المدونة في المستقبل في بناء مدونات لغوية أخرى تتضمن هذه المدونة بكاملها أو أجزاء منها.

تشمل البيانات الأساسية للنصوص المعلومات التالية وليست حصرًا عليها:

- عنوان النص.
- اسم المؤلف وجنسيته وجنسه.
- وعاء النص ومجاله وموضوعه.
- تاريخ صدور النص.
- ناشر النص والبلد الذي ينتمي إليه.
- مصدر النص.
- تاريخ إضافة النص للمدونة.

بناء المدونات

إن بناء المدونات لا يعني جمع النصوص حسب المعايير فحسب، بل يتطلب الأخذ بالاعتبار إجراءات عدة تسبق الجمع وتليه، وفيما يلي سوف نوضح هذه الخطوات مع إعطاء نماذج لما يمكن أن تتضمنه هذه الخطوات:

3. أ حقوق الملكية الفكرية

إن حقوق الملكية الفكرية لنصوص المدونة هي أول ما يتوجب النظر إليه بحرص قبل الشروع في بناء المدونة؛ إذ إن انتهاك حقوق الملكية الفكرية لمالكي النصوص –جهاتٍ أو أفرادا– قد يعرّض

الشخص أو الجهة المسؤولة عن بناء المدونة للمساءلة القانونية، وقد يجهض المشروع برمته خصوصاً عند توزيع نصوص المدونة سواء بمقابل أو بالمجان. وعلى الرغم من أهمية هذا الموضوع وحساسيته فإن التغلب على صعوباته ممكن في أغلب الأحيان.

قبل التفكير في الحصول على إذن مسبق لتضمين النصوص المحمية بقوانين الملكية الفكرية في المدونة من المستحسن البحث عن نصوص ليست محمية بهذه القوانين، ويتضمن هذا النوع النصوص التاريخية القديمة، والنصوص التي انتهت مدة حمايتها النظامية بموجب القانون، والنصوص التي ينص أصحابها على مجانية الحصول عليها وتوزيعها، وكذلك ما توفره بعض الجهات الرسمية من إصدارات ونشرات تعريفية وتثقيفية.

وتنص بعض قوانين حماية الملكية الفكرية على عدم خضوع بعض النصوص للحماية. فعلى سبيل المثال فإن نظام حماية حقوق المؤلف ولائحته التنفيذية[9] في المملكة العربية السعودية ينص على استثناء بعض المصنفات من الحماية، كالأنظمة والأحكام القضائية والوثائق الرسمية وما تنشره الصحف والمجلات من الأخبار اليومية أو الحوادث ذات الصبغة الإخبارية، كما تسمح بعض الأنظمة بالاستفادة من النصوص كاملة أو أجزاء منها لأغراض تعليمية أو بحثية. ومع هذا السماح بالاستفادة من النصوص إلا أنها لا تهمل وجوب الإشارة إلى اسم صاحب الحق في ملكية هذا النص. وقد لجأت بعض المدونات كبيرة الحجم إلى الاستفادة من هذا النظام، فلا تسمح بتوزيع نصوص المدونة ولا تعرض نصوصها للاطلاع بل تعرض السياقات التي تظهر فيه الكلمة فقط مع الإشارة الى مؤلف النص وكافة المعلومات المتوفرة عن هذا النص في المدونة.

وعندما تدعو الحاجة للحصول على إذن بتضمين نصوص محمية بموجب النظام، كحقوق الملكية الفكرية أو نصوص قد تنتهك الخصوصية الشخصية، كحوارات المرضى مع أطبائهم، يجب أن يفرق القائمون على المدونة بين موضوعين رئيسين، والحصول على الإذن بهما، وهما حق تضمين النصوص في المدونة، وحق توزيع أو نشر هذه النصوص مجاناً أو بمقابل مادي.

3. ب تحديد المصادر

إن تحديد المصادر التي يمكن أن تجمع منها نصوص المدونة يوفر الكثير من الوقت والجهد عند القيام بالخطوة التالية وهي جمع النصوص، ويمكننا تقسيم هذه المصادر إلى قسمين رئيسين، الأول: المصادر التي توفر نصوصاً إلكترونية للجمع وللمعالجة المباشرة، مثل: مواقع الصحف والمجلات والمكتبات الإلكترونية التي توفر نصوصا بصيغة نصية (DOC, DOCX, TXT). والآخر: المصادر

التي توفر نصوصا قابلة للجمع ولكنها غير قابلة للمعالجة الإلكترونية مباشرة، بل تتطلب جهدا إضافيا لتحويلها لصيغة نصية، كأن تكون نصوصها على صيغة صور أو PDF أو أن تكون ورقية، ويمكن في هذه الحالة استخدام برامج التعرف الضوئي على الحروف لتحويلها إلى صيغة نصية إلكترونية أو كتابة النص وحفظه بصيغة نصية.

ويتطلب تحديد المصادر إعداد قائمة بها وبروابطها الإلكترونية على الشابكة أو طريقة الحصول عليها إن كانت ورقية، مع تسجيل أي ملاحظات تخص رخص الحصول عليها، وتضمينها في المدونة. وتعتبر هذه القائمة، قائمة أولية يجب اختبارها وتجريبها وجمع نصوص متعددة منها؛ لمعرفة الأفضل والأسهل في جمع النصوص، كما يجب الإضافة إليها وتنويعها بقدر الإمكان وحذف ما لا يصلح منها لأي سبب، كعدم توفر معلومات دقيقة عن النصوص مثل نسبة نص لغير صاحبه.

3. ج الجمع

اعتماداً على قائمة المصادر التي تم تحديدها مسبقا يبدأ العمل في جمع نصوص المدونة. وتوفيراً للوقت والجهد والتكلفة يجب التركيز على القسم الأول من المصادر (المصادر التي توفر نصوصاً إلكترونية للجمع وللمعالجة المباشرة) وعدم العمل على القسم الثاني (المصادر التي توفر نصوصا قابلة للجمع ولكنها غير قابلة للمعالجة الإلكترونية المباشرة) إلا في أضيق الحالات. ومن الممكن في بعض الحالات أن يتم جمع النصوص آلياً حيث توجد عدة برامج يمكن أن تقوم بهذا العمل كبرنامج بوت كات BootCat[10] على سبيل المثال لا الحصر. وقد استُخدمت هذه الطريقة بالكامل لبناء بعض المدونات، انظر على سبيل المثال (الزهراني 2013، Alzahrani) (خوجه 2009، Khoja) (جاكوبيكيك وآخرون 2013، Jakubí□ek et al.). ومع ذلك، يجب التنبه إلى أن نتاج الجمع الآلي للنصوص بحاجة إلى مراجعة دقيقة؛ فبعض هذه النصوص قد يكون مليئاً بالأخطاء، وبعضها قد يكون مكرراً أو يحتوي على بيانات ليس لها علاقة بالنص الأصلي، كروابط لصفحات أخرى أو إعلانات تجارية أو بيانات لها علاقة بموقع الشابكة. وعلى الرغم من فائدة الطريقة الآلية في الجمع السريع لنصوص المدونة إلا أنه من الصعب تحديد الأوعية الخاصة بالنصوص ومجالاتها والتواريخ التي ظهرت فيها ومن هم مؤلفوها، لذا فإن تعيين المواقع التي تجمع منها النصوص لتحديد الأوعية والموضوعات يجب أن يكون محل الاهتمام. ولضمان جودة نواتج عملية جمع النصوص يفضل أن تُجزأ عملية الجمع إلى مراحل يتم تقسيمها بناء على الأوعية مثلا، وكلما انتهى قسم تتم مراجعة نصوصه والتأكد من تحقيقه لمعايير التصميم.

3. د الترميز والتسمية والحفظ

بعد جمع النصوص، ولزيادة الاستفادة من المدونة وتسهيل إدارتها من المستحسن القيام بالتالي:

- توحيد ترميز النصوص، ومن المفضل تحويلها كاملة إلى ترميز واحد يكون مقبولاً من أغلب أنظمة التشغيل وبرامج معالجة المدونات، مثل:UTF8 و UTF16، فاستخدام ترميز الوندوز Windows مثلاً قد لا يساعد على الاستفادة من المدونة في حال معالجتها بأنظمة مختلفة عن الوندوز.
- توحيد طريقة التسمية، ومن المفضل استخدام طريقة تعتمد على الأرقام والأحرف اللاتينية، فمثلاً يمكن تقسيم اسم الملف إلى عدة خانات يفصل بينها "-" حيث تعبر كل خانة منها عن أحد معايير التصنيف، فيكون لكل بلد ووعاء ومجال وموضوع وفترة زمنية رمزه الخاص وتكون آخر خانة للرقم التسلسلي للملف.
- حفظ الملفات في مجلدات منفصلة حسب وعائها أو فترتها الزمنية أو البلد الذي صدرت فيه مع حفظ القائمة المشتملة على اسم النص ومعلوماته حسب معايير التصميم. وعندما تكون المدونة كبيرة الحجم وتحوي عدداً كبيراً من النصوص فقد يكون من الأفضل حفظ النصوص وبياناتها الأساسية في قاعدة بيانات، ليمكن عند ذلك إدارتها والتحكم فيها بسهولة، كما يمكن بهذه الطريقة تصديرها بصيغ متعددة حسب الرغبة متى ما دعت الحاجة إلى ذلك.

3. هـ التحشية

من المستحسن –وقد يكون لازماً في بعض الحالات– أن يتم إثراء نصوص المدونة بمعلومات تزيد من فائدتها، وتساعد في إجراء المزيد من الدراسات المتعمقة. وتنقسم هذه المعلومات إلى ثلاثة أقسام:

القسم الأول: يتضمن معلومات عن النص نفسه وهي التي أطلقنا عليها من قبل مسمى البيانات الأساسية للنصوص، ويتم ذلك باتباع عدة منهجيات وأساليب معيارية تساعد في تبادل المدونات وسهولة استخدامها بين الباحثين. ومن أبرز هذه المعايير مبادرة ترميز النصوص[11] Text Encoding Initiative (TEI)، والمرجع القياسي لترميز المدونات[12] Corpus Encoding Standard (CES). وهناك من يرى أن هذا القسم ليس له علاقة بالتحشية فيضع له مصطلحاً آخر وهو التعليم (من العلامة) Mark-up (ماكنري وآخرون، 2006).

القسم الثاني: معلومات تتعلق ببنية النص وتركيبه الظاهريين، مثل تحديد نهاية وبداية الفقرات والجمل والعبارات داخل النص، وهذه التحشية تساعد في الدراسات التي تنظر في العلاقات بين الجمل في الفقرات الواحدة وبين الفقرات في النص ذاته، كما تكون أساساً لبناء البنوك الشجرية Treebanks والتحليل النحوي Parsing. ويمكن أيضا كما في حالة الأوراق العلمية مثلاً تحديد بداية ونهاية الأجزاء المكونة للورقة العلمية مثل العنوان والكلمات المفتاحية والمقدمة والدراسات السابقة والتجارب ومناقشة النتائج والخاتمة، ومثل هذه التحشية تفيد في الدراسات التي تستهدف معرفة الخواص اللغوية لمثل هذا النوع من الكتابات وكيف تترابط أجزاءها.

القسم الثالث: ما يتعلق بإضافة نتائج التحليل اللغوي للنصوص كإضافة الوسوم النحوية والوسوم الدلالية وإضافة معلومات الإحالة بين الضمائر والأسماء وكذلك معلومات التحليل النحوي. وعلاوة على ذلك يمكن للمدونات اللغوية العربية أن تشمل تحشيات من نوع آخر، كالتحشية الخاصة بالجذور وأنواعها والتوسيم الصرفي. ويجب التنويه إلى أن هذا النوع والأنواع الأخرى من التحشية يجب أن تتبع منهجا واحداً ودقيقاً في تحديد معلومات التحشية وذلك بوضع قائمة محددة سلفاً فلا يستخدم غيرها ولا تهمل إن وجدت.

وبطبيعة الحال، فإن عمليات التحشية في القسمين الثاني والثالث يمكن أن تتم بصورة آلية، وهذا هو المتبع في غالب المدونات الإنجليزية بسبب دقة مثل هذه الأنظمة الآلية في اللغة الإنجليزية؛ ولكن الحال في العربية مختلف، فالأبحاث في هذا المجال محدودة والموجود منها لا يراعي ما هو مستقر في النحو العربي إجمالاً (الشبيتي 2014). وعلى الرغم من ذلك فالاستفادة من أنظمة التوسيم النحوي المتوفرة حاليا ممكنة وقد تكون ذات فائدة متى ما أدرك الباحث جوانب النقص فيها، وكان مطلعاً على قائمة الوسوم النحوية فيها، وعالماً بمدلولاتها، وكانت كافية ومحققة لغرضه البحثي.

الأدوات

مهما كانت الجهود والأوقات والأموال التي تصرف في تصميم المدونة وجمع نصوصها لتوائم التصميم فإنها ستكون بيانات بلا فائدة ما لم يكن هناك أدوات قادرة على البحث فيها واستعراض نتائج هذا البحث والمساعدة في تحليله كماً ونوعاً للنظر أولاً في مدى تحقيق المدونة لغرض الدراسة، ولإجراء التحليلات اللغوية المختلفة على بياناتها ثانياً. وسوف نستعرض باختصار أهم الوظائف اللازمة في أنظمة معالجة المدونات وهي كما يلي:

أ- إنتاج بيانات إحصائية عامة عن المدونة، كحجمها، وعدد كلماتها دون تكرار، وعدد نصوصها.

ب- إنتاج قوائم التكرار والتكرار النسبي للمدونة للغوية كاملة، والتوزيع الإحصائي لتكرار الكلمات على أقسام المدونة، مثل: الأوعية والفترات الزمنية والمناطق إن كان مثل هذا التقسيم موجوداً. ويقصد بالتكرار عدد مرات ظهور الكلمة في المدونة، ويقصد بالتكرار النسبي نسبة ظهور الكلمة في المدونة مقارنة ببقية كلماتها. ويتم احتساب التكرار النسبي بقسمة تكرار الكلمة في المدونة على كامل حجم المدونة. ويستفاد من التكرار النسبي عند مقارنة التوزيع الإحصائي للكلمة بين مدونات مختلفة الأحجام.

ج- استخراج الكلمات الدليلية أو المميزة للمدونة اللغوية وذلك بمقارنة قوائم التكرار لكلمات لمدونة اللغوية مع قوائم التكرار لمدونة لغوية أخرى تسمى اصطلاحا بالمدونة المرجعية. وبالإمكان إجراء هذه المقارنة أيضاً على مستوى الوسوم النحوية والدلالية والصرفية متى ما كانت متوفرة والوسوم المستخدمة في المدونتين متطابقة.

د- إنتاج قوائم الكشاف السياقي للكلمة مناط البحث، والقصد من الكشاف السياقي هو استعراض جميع السياقات التي وردت فيها الكلمة داخل المدونة للكشف عن معانيها المختلفة، والكلمات التي تظهر بصحبتها في السياق، واختلاف المعنى من سياق لآخر، باختلاف الأوعية والفترات. ومن المناسب توفر إمكانية تحديد عدد الكلمات التي يتضمنها السياق قبل الكلمة مدار البحث وبعدها.

هـ- حسابات التصاحب اللفظي لكلمة معينة من خلال عدة معاملات إحصائية مثل مربع كاي Chi-Squired ومعامل كسب المعلومات Information Gain ومعامل المعلومات المتبادلة Mutual Information ومعامل الاحتمالية اللوغاريثمي Log Likelihood على سبيل المثال لا الحصر. ولا يشترط أن يكون الباحث على معرفة بالخلفية الرياضية لهذه المعاملات الإحصائية، ويكفيه معرفة متى يستخدم هذه المعاملات وحدود استخدامها. وتسعى هذه المعاملات الإحصائية إلى الكشف عن مدى ارتباط الكلمة مع الكلمات الأخرى التي ظهرت معها في السياق، ولا يشترط في هذا الحساب أن تتوالى الكلمتان، بل أن تظهرا في سياق واحد حسبما يحدد الباحث حدود هذا السياق. ويمكن استخدام نفس المعاملات المذكورة سابقاً لحساب

التلازم اللفظي بين كلمتين (أن تظهر الكلمة الأولى قبل الثانية مباشرة) مع تغيير طفيف في طريقة الحساب. ومتى ما كانت المدونة موسومة نحوياً أو دلالياً أو صرفياً فإنه يمكن حساب التصاحب أو التلازم النحوي أو الدلالي أو الصرفي.

تتوفر عدة برامج مجانية وأخرى بمقابل مادي للقيام بهذه الوظائف وغيرها، ومن أشهر البرامج المجانية لمعالجة المدونات برنامج أنتْ كونكْ[13] AntConc ولكنه لا يراعي اتجاه الكتابة العربية (من اليمين لليسار) أثناء عرض الكشاف السياقي. ولم يصمم نظام لمعالجة المدونات العربية يراعي خواصها ويحتوي كل ما ذكر أعلاه من وظائف سوى أداة معالجة المدونات العربية "غواص"[14] (الثبيتي وآخرون، 2013 .Al-Thubaity et al) الذي تم تطويره في مدينة الملك عبدالعزيز للعلوم والتقنية. وعلى الرغم من أهمية ما توفره هذه الأدوات والبرامج من معلومات وبيانات إلا أن هذه المعلومات والبيانات ليست هي نهاية البحث اللغوي بل بدايته، ومهمة الباحث بعد ذلك هي الكشف عما تعنيه هذه البيانات من خلال تفحصه ودراسته لها.

مثال تطبيقي

يعرض هذا القسم مثالا تطبيقياً بسيطاً للخطوات التي سبق شرحها الخاصة بتصميم المدونات وبنائها. وقد تختلف معي عزيزي القارئ في التفاصيل التي سوف أضعها لمعايير التصميم أو لبناء المدونة، وهذا أمر طبيعي. ومرد هذا الاختلاف – إن وجد – لسببين: الأول اختلاف الفهم للغرض من المدونة، والآخر هو اختلاف الخبرات والتجارب في هذا المجال، وعلى كل حال فإننا سوف نتفق على الكثير.

بداية سوف أحدد سؤال البحث أو الغرض الذي لأجله سوف أستخدم المدونة ثم أبحث عن مدى توفر مدونة لغوية بالإمكان استخدامها للإجابة عن سؤال البحث.

إن الغرض الذي حددته لاستخدام المدونة هو الكشف عن الفروقات في استخدام اللغة بين الكتّاب السعوديين والكاتبات السعوديات من خلال ما ينشر في الصحافة السعودية. وللأسف لم أجد أي مدونة لغوية يمكن أن تساعد في تحقيق هذا الغرض، ولاهتمامي بهذا الموضوع ومعرفتي باهتمام الكثيرين به سوف أضطر إلى بناء مدونة لغوية يمكنها الإجابة عن استفساراتي بهذا الخصوص. فيما يلي سوف أعرض للمعايير التي اخترتها لتصميم هذه المدونة:

معايير التصميم

أ- لغة المدونة: العربية الفصحى.

ب- طبيعة النصوص: النصوص المكتوبة.

ج- تاريخ النصوص: 2014م.

د- المنطقة الجغرافية: المملكة العربية السعودية.

هـ- الوعاء: الصحف.

و- المجال: المقالات.

ز- الموضوعات: المقالات الاجتماعية، والثقافية، والدينية. وقد اقتصرت هنا على هذه المجالات الثلاثة لأنه من الممكن وجود كتابات لكلا الجنسين فيها بينما يصعب هذا في مجالات أخرى مثل السياسية، والرياضية، والعلوم والتقنية.

ح- حجم العينة: النص كاملاً.

ط- جنس الكاتب: الذكور والإناث.

ي- حجم المدونة: ثلاثة ملايين كلمة، توزع بالتساوي بين المجالات وبين الجنسين مثلما هو موضح في الجدول (1)

المجال	الذكور	الإناث	المجموع
المقالات الاجتماعية	500،000	500،000	1،000،000
المقالات الثقافية	500،000	500،000	1،000،000
المقالات الدينية	500،000	500،000	1،000،000
المجموع	1،500،000	1،500،000	3،000،000

جدول (1) توزيع كلمات المدونة على المجالات وجنس الكاتب

ك- معايير أخرى: قد يكون من المفيد تحديد معايير أخرى مثل المستوى التعليمي والفئة العمرية؛ لكن الوصول لهذه المعلومات وتحديدها بدقة صعب جداً.

ل- التمثيل والتوازن: كما تم توضيحه سابقاً فإن التمثيل والتوازن مفهومان مهمان ومترابطان، وقد حققت المدونةُ التوازنَ من خلال التوزيع المتساوي لعدد الكلمات بين المجالات وبين الجنسين. وعلى الرغم من غلبة عدد الكتّاب على الكاتبات في الصحافة السعودية وفي الصحف أيضاً إلا أن التساوي في توزيع عدد الكلمات والمجالات مهم جداً لأن القصد

هو المقارنة بين كتابات الجنسين وليس أي نوع آخر من الكتابات، كما تحقق التوازن من خلال حصر المجالات واقتصارها على المجالات التي من الممكن وجود كتابات صحفية للجنسين فيها بسهولة، وتحقق التمثيل من خلال تنوع المجالات ومن خلال اعتبار الصحف السعودية جميعها.

م- البيانات الأساسية: سوف أذكر هنا جميع البيانات الأساسية حتى ما هو بدهي منها، والغرض من هذا هو الاستفادة الكاملة من المدونة لاحقا فيما لو تم إضافتها لتكوين مدونات أخرى لنفس الغرض لكن تغطي فترات زمنية أخرى أو لتغطي بلداناً عربية أخرى. والمعلومات الأساسية المقترحة لنصوص لمدونتنا هذه هي: عنوان النص، اسم المؤلف، جنسية المؤلف، جنس المؤلف، وعاء النص، مجال النص، موضوع النص، تاريخ صدور النص باليوم والشهر والسنة، اسم الصحيفة، الرابط الإلكتروني للنص، تاريخ إضافة النص للمدونة اللغوية.

بناء المدونة

أ- حقوق الملكية الفكرية: كما أشرت سابقاً إلى أن المحافظة على حقوق الملكية الفكرية من أهم ما يجب التفكير فيه عند جمع نصوص المدونة، كما أن الرغبة في تعميم الاستفادة من المدونة وإتاحتها للباحثين الآخرين كمصدر لدراساتهم أو لضمها لمدونات أخرى يزيد من أهمية استئذان الصحف قبل البدء في جمع نصوص المونة اللغوية. بالطبع قد تأخذ هذه المرحلة بعض الوقت ولكن الإجابة بالقبول هي الغالبة، وعلى الباحث أن يدرك أهمية متابعة الموضوع والاتصال الدائم بالصحف وتوضيح فكرته ليحصل على الموافقة، كما يجب عليه ألا ينسى شكر الصحف وذكر موافقتها على ضم نصوصها لمدونته مع أهمية احتفاظه بنسخة من هذه الموافقة وإرفاقها مع المدونة.

ب- تحديد المصادر: الجدول (2) يوضح مثالا للمصادر التي سيتم جمع نصوص المدونة منها ويمكن السير على منواله لتحديد بقية الصحف، مع العلم أن المعلومات الواردة في خانة موافقة الصحيفة هي معلومات افتراضية وليست حقيقية.

ج- الجمع: سيتم جمع نصوص المدونة يدوياً من مواقع الصحف بسبب أن المدونة صغيرة، كما أن تحديد جنس الكاتب آليا وتجهيز أحد البرامج لجمع نصوص كاتب معين يستغرق وقتاً طويلاً. ويمكن أن يقوم بهذا العمل عدة أشخاص يتولى كل منهم جمع

النصوص من صحيفة معينة، ويجب الحرص على تنوع الكتّاب والكاتبات وعدم الاقتصار على كتّاب وكاتبات بعينهم.

د- الترميز والتسمية والحفظ: يتم حفظ النصوص في ملفات نصية بصيغة TXT بترميز UTF8، ويتم حفظ نصوص الكتاب في مجلد باسم SNP_M[15]، والكاتبات في مجلد باسم SNP_F[16]. بالنسبة لتسمية الملفات سوف نقسم الاسم إلى 6 أجزاء كما هو موضح في الجدول (3)

م	الصحيفة	العنوان	الموقع على الشابكة	موافقة الصحيفة
1	الرياض	المملكة العربية السعودية – الرياض، حي الصحافة أول طريق القصيم ص.ب 851 الرياض 11421 سنترال: 2996000، فاكس: 4871070	http://www.alriyadh.com/	إعداد الخطاب
2	مكة	مكتب صحيفة مكة الرئيسي في مكة المكرمة: صندوق بريد: 5803، الرمز البريدي: 21955 تليفون: 966125206776، فاكس: 966125203054	http://www.makkahnewspaper.com/	تم إرسال الخطاب
3	اليوم	السنترال: 8580800 3 966، الفاكس: 8588777 3 966 الرقم المجاني: 8006121212 ص.ب. 565 الدمام ، 31421، المملكة العربية السعودية mail@alyaum.com	http://www.alyaum.com/	تم التواصل إلكترونياً، في انتظار الرد
4	الوطن	أبها – مدينة سلطان ، طريق المطار التحرير- هاتف مجاني (8007540007)،	http://www.alwatan.com.sa/	تمت الموافقة

م	الصحيفة	العنوان	الموقع على الشابكة	موافقة الصحيفة
		سنترال 2273333 فاكس التحرير 2273756 ، ص.ب. 15155		

جدول (2) مصادر المدونة

القسم	الدلالة	القيم الممكنة
2	السنة	2014
3	جنس الكاتب	1 للكتّاب، 2 للكاتبات
4	الموضوع	CUL الثقافية، REL الدينية، SOC الاجتماعية
5	الصحيفة	يأخذ عدداً صحيحاً من خانتين كل عدد منها يدل على صحيفة معينة. الرياض:01، اليوم:02، مكة:03، الوطن:04 وهكذا لبقية الصحف
6	التسلسل في المدونة	رقم يتكون من أربع خانات يدل على تسلسل النص في المدونة

جدول (3) تسمية ملفات المدونة

لو نظرنا لاسم الملف التالي 2014-2-CUL-04-0091 لعرفنا أنه لكاتبة، وأن موضوعه ثقافي، وأنه من صحيفة الوطن السعودية، وأن تسلسله في المدونة هو 91. وسوف تحفظ جميع المعلومات الأساسية للمدونة اللغوية في قائمة كما هو موضح في الجدول (4)

اسم الملف	العنوان	جنس الكاتب	الكاتب	الموضوع	الصحيفة	رابط النص
2014-2-CUL-04-0001	إلى الساخرين من التراث!!	أنثى	ملحة عبدالله	ثقافي	الوطن	http://www.alwatan.com.sa/Articles/Detail.aspx?ArticleId=22691

اسم الملف	العنوان	جنس الكاتب	الكاتب	الموضوع	الصحيفة	رابط النص
2014-1-SOC-02-0002	تحصين المجتمع الجامعي لمواجهة المتغيرات	ذكر	عادل رشاد غنيم	ديني	اليوم	http://www.alyaum.com/article/4013773

جدول (4) المعلومات الأساسية للمدونة اللغوية

الخاتمة

المدونات ما هي إلا بيانات لا تختلف عن أي بيانات أخرى تجمع لإجراء الدراسات العلمية، وهي عينة من اللغة وليست اللغة كلها، وبحسب وضوح المنهج المتبع وانضباطه في جمع المدونات يمكن الاعتماد والوثوق بنتائج الدراسات القائمة عليها؛ فالهدف من استخدام المدونات في دراسة اللغة هو الكشف عن الأنماط الشائعة في اللغة المستخدمة الذي قد يتطابق أو يختلف عما نعرفه عنها معيارياً.

وقد شرح هذا المبحث بشكل مختصر ومبسط خطوات إجرائية، تصلح إطاراً عاماً يمكن اتباعه عند تصميم المدونات وبنائها، والأدوات المهمة الرئيسية لمعالجة هذه المدونات. حيث وضح القسم الثاني من هذا الفصل أهم ما يجب النظر إليه عند تصميم المدونات مثل لغة المدونة، وطبيعة نصوصها، وتاريخ هذه النصوص، ومكان صدورها، والأوعية التي ظهرت فيها، والمجالات المختلفة التي تغطيها، كما ذكرنا في هذا القسم النقاط التي تراعى في تحديد حجم المدونة وحجم العينات التي تؤخذ للوصول للحجم المطلوب مع الإشارة إلى المعايير الأخرى التي قد يكون الباحث بحاجة لأخذها بالحسبان، كما تطرق القسم الثاني إلى مفهومين مهمين في المدونات يحددان أهميتها، هما التمثيل والتوازن.

وتطرق القسم الثالث من هذا المبحث إلى النقاط التي ينبغي الاهتمام بها عند جمع نصوص المدونة استرشاداً بمعايير التصميم المشروحة في القسم الثاني، فتناول حقوق الملكية الفكرية، وتحديد مصادر جمع النصوص، وكيفية جمع النصوص مع التطرق إلى بعض المسائل التقنية عند الانتهاء من الجمع وهي ترميز ملفات نصوص المدونة وتسميتها وحفظها وإضافة معلومات مساعدة في تعزيز الفائدة من المدونة من خلال تحشيتها بمعلومات إضافية مختلفة.

وتطرق القسم الرابع باختصار إلى أهم الأدوات التي تساعد في تحليل المدونات ودراستها، فتطرق إلى خمس وظائف يمكن أن تساعد بشكل أساسي في هذا الشأن هي الإحصاءات العامة عن

المدونة بمجملها، وقوائم التكرار، والكلمات المميزة للمدونات، والكشاف السياقي، والتصاحب اللفظي، ثم تطرق القسم السادس إلى مثال تطبيقي لما سبق شرحه.

ومع أهمية اتباع هذه الخطوات والإجراءات لجمع عينة من اللغة تكون ممثلة للغة أو إحدى صورها مجال الدراسة فإن الجهد يجب ألا يتوقف عند هذا الحد؛ فتصميم المدونة ونصوصها يجب أن يخضعا للمراجعة والتقويم المستمر متى ما استدعى الأمر ذلك، مع إتاحتها للباحثين الآخرين متى ما كان ذلك ممكناً.

الحواشي

(7) http://www.sketchengine.co.uk/

(8) المد الذي يضاف في وسط الكلمة بغرض تصفيف الكلمة ومحاذاتها

(9) http://www.info.gov.sa/copyrights/SectionDetails.aspx?id=6

(10) http://bootcat.sslmit.unibo.it/

(11) http://www.tei-c.org/index.xml

(12) http://www.xces.org/

(13) http://www.antlab.sci.waseda.ac.jp/software.html

(14) http://sourceforge.net/projects/ghawwasv4

(15) SNP_M:Saudi NewsPapers_Males الصحف السعودية_الذكور

(16) SNP_F: Saudi NewsPapers_Females الصحف السعودية_الإناث

المراجع

الثبيتي، عبدالمحسن. المدونات العربية، التحليل الصرفي والتوسيم النحوي. محاضرة بمركز اللغويات التطبيقية. جامعة الإمام محمد بن سعود الإسلامية. الرياض. 7 مايو 2014م. (2014. http://www.slideshare.net/althubaity/ss-34381929)

Ahmad, **Khurshid**. "Being in Text and Text in Being: Notes on representative texts." Incorporating Corpora: The linguist and the translator (2008): 60-94.

Alansary, **Sameh**, **Magdy Nagi**, **and Noha Adly**. "Building an International Corpus of Arabic (ICA): progress of compilation stage." 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007. 2007.

Alfaifi, **A**. **Y**. **G**., **Eric Atwell**, **and I**. **Hedaya**. "Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners." (2014).

Alkanhal, **Mohamed I**., **et al**. "Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions." Audio, Speech, and Language Processing, IEEE Transactions on 20.7 (2012): 2111-2122.

Al-Sulaiti, **Latifa**, **and Eric Steven Atwell**. "The design of a corpus of contemporary Arabic." International Journal of Corpus Linguistics 11.2 (2006): 135-171.

Al-Thubaity, **A**. **O**. (2014).A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Language resources and evaluation. DOI 10.1007/s10579-014-9284-1

Al-Thubaity, **Abdulmohsen**, **et al**. "New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool."Asian Language Processing (IALP), 2013 International Conference on. IEEE, 2013.

Alzahrani, **Salha M**. "Building, Profiling, Analysing and Publishing an Arabic News Corpus Based on Google News RSS Feeds." Information Retrieval Technology. Springer Berlin Heidelberg, 2013. 488-499.

Atkins, **Sue**, **Jeremy Clear**, **and Nicholas Ostler**. "Corpus design criteria."Literary and linguistic computing 7.1 (1992): 1-16.

Biber, **Douglas**, **Susan Conrad**, **and Randi Reppen**. Corpus linguistics: Investigating language structure and use. Cambridge University Press, 1998.

Biber, **Douglas**. "Representativeness in corpus design." Literary and linguistic computing 8.4 (1993): 243-257.

Crystal, **David**. An encyclopedic dictionary of language and languages. Middlesex, UK: Blackwell, 1992.

Jakubíček, **Miloš**, **et al**. "The TenTen Corpus Family." Proc. Int. Conf. on Corpus Linguistics. 2013.

Khoja, **Shereen**. "An RSS Feed Analysis Application and Corpus Builder."Proceedings of the Second International Conference on Arabic Language Resources and Tools. 2009.

Khorsheed, **Mohammad S**., **and Abdulmohsen O**. **Al-Thubaity**. "Comparative evaluation of text classification techniques using a large diverse Arabic dataset." Language resources and evaluation 47.2 (2013): 513-538.

Leech, **Geoffrey**. "New resources, or just better old ones? The Holy Grail of representativeness." Language and Computers 59.1 (2006): 133-149.

McEnery, **Tony**, **Richard Xiao**, **and Yukio Tono**. Corpus-based language studies. London: Routledge, 2006.

Meyer, **Charles F**., ed. English corpus linguistics: An introduction. Cambridge University Press, 2002.

Sinclair, **John**. "Corpus and text-basic principles." Developing linguistic corpora: A guide to good practice *(*2005): 1-16.

Sinclair, **John**. Corpus, concordance, collocation. Oxford University Press, 1991.

Teubert, **Wolfgang**, **and Anna Cermáková**. Corpus linguistics: A short introduction. Continuum, 2007.

المبحث الرابع

لسانيات المدونات: نماذج وتطبيقات في لغة الصحافة العربية

عقيل بن حامد الشمري وعبدالمحسن بن عبيد الثبيتي