Algerian Journal of Human and Social Sciences
المجلة الجزائرية للعلوم الإنسانية والاجتماعية
ISSN : 2588–1558 / EISSN : 2710–8635

المجلد: 07/العدد: 02
(ديسمبر 2023) ص: 46–64

منهجيّة بناء مدوّنة لغويّة متوازية لغرض تعليم اللغة العربيّة للناطقين بغيرها

The Methodology of Building a Parallel Corpus for Teaching Arabic to Non-Native Speakers

[1*]إبـــراهيـــــم أكـــوريـــدي حـــميـــد

[*1] Ibrahim Akorede Hammed

[1*] معهد الدوحة للدراسات العليا، (دولة قطر)، activeib@yahoo.com

تاريخ الإرسال: 2022/10/05 تاريخ القبول:2023/12/19 تاريخ النشر: 2024/01/26

ملخص:

تسعى الدّراسة إلى تقديم منهجيّة واضحة لبناء مدوّنة لغويّة متوازية لنصوص اللغة العربيّة وغيرها من اللغات البشريّة–بالتركيز على لغة يوربا نموذجا–لغرض تيسير تعليم وتعلّم اللغة العربيّة للناطقين بغيرها. وعلى ذلك، تطرّق الدراسة إلى أهميّة استخدام المدوّنة اللغويّة المتوازية بين اللغتين في مجال تعليم اللغة للناطقين بغيرها ومميّزات مثل تلك المدوّنة.

الكلمات المفتاحية: مدوّنة متوازية، تعليم اللغة، الناطقين بغير العربية، التوازن، وسم.

Abstract :

The study seeks to present a clear methodology for building a parallel corpus between Arabic and one other natural language- with focus on Yoruba as case study texts for teaching Arabic to non-native Arabic speakers. In addition to that, the study discussed the importance of using parallel corpus in teaching the language to Non-native speakers as well as features peculiar to this kind of corpus.

Keywords: Parallel Corpus; Teaching; Arabic Non-Native Speakers; Balance; Annotation

* المؤلف المرسل.

1. مقدمة

إنّ المدوّنة اللغويّة المتوازية نوع من المدوّنات ثنائيّة اللغة فأكثر، وهي عبارة عن مجموعة من النصوص، يتمّ ترجمة كلّ منها إلى لغة أخرى فأكثر (Sinclair, 1995: 19)، وكان هذا النوع من المدوّنات محصور في دراسات الترجمة وبحوث اللغة المقارنة، ولم يطبّق على عمليّة تعليم اللغة الأجنبيّة حتى أواخر الثمانينات من القرن العشرين. ويُعتبَر تيم جون Tim John أوّل من استخدم مصطلح "تعلّم اللغة بالاعتماد على المدوّنة المُحَوسَبة Driven-Data Learning" في سنة 1990 (Boulton & Vyatkina, 2021: 66)، حيث يَقصد به منهجا تعليميّا يشجّع به معلّم اللغة طلابَه في البحث عن القواعد والأنماط اللغويّة المنشودة في المدوّنات، ولكن قبل ذلك، لا بدّ للمعلّم من تحديد الإشكاليّة التي ينوي البحث عنها وتقرير نوع الأنماط التي تمثّل بنية اللغة، كما يحاول إنشاء الأنماط وصياغة القواعد على أساس نصوص المدوّنة التي قدّمها للمتعلّمين (Alotaibi, 2017: 320).

وقد انعكس تطوّر التكنولوجيا على لسانيات المدوّنة، وشهدتْ الآونة الأخيرة تطوّرا واضحا في مستوى استخدام المدوّنات في عمليّة تعليم اللغة الأجنبيّة وتعلّمها، ويشهد على ذلك عدد من الدراسات التي أُجريتْ بالاعتماد على المدوّنة المتوازية وأَثبتتْ نجاحها في عمليّة التعليم والتعلّم. على سبيل المثال، أجرى تساي وتشوي Tsai and Choi (2005) دراسة عن التطوّر المعجميّ عند متعلّمي اللغة الصينيّة كلغة أجنبيّة من الناطقين باللغة الإنجليزيّة باستخدام مدوّنة متوازية، وأَثبتتْ الدراسة أنّ المتعلّمين عن طريق المدوّنة اللغويّة المتوازية أفضل من ناحية اكتسابهم للموادّ المعجميّة المختبرة والاحتفاظ بها من نظرائهم الذين درسوها على الطريق التقليديّ.

والمدوّنات تختلف باختلاف الغرض من بنائها. على ذلك، من المهم معرفةُ طريقة بناء المدوّنة المناسبة لتحقيق غرض معيّن في عمليات التعليم والتعلّم، وذلك يدلّ على أهميّة ما تناولته هذه الدّراسة من تقديم منهجيّة واضحة لبناء المدوّنة بين اللغة العربية ولغة اليوربا نموذجا.

2. أهميّة المدوّنة اللغويّة المتوازية في مجال تعليم لغة للناطقين بغيرها

يرى العديد من الباحثين أنّ المدوّنة اللغويّة المتوازية تقدّم مزايا جديدة لكلّ من لغة المصدر والمستهدف التي لا يمكن اكتسابها من خلال المدوّنة أحاديّة اللغة، منها: إيجاد الباحثين والطلاب المنصة لمقارنة اللغتين فأكثر من جوانب مختلفة (Alatoibi, 2017)، مثل جانب الثقافة، والنظام اللغويّ، والدلالة، وغير ذلك. وقد أكّد Meece (2003) من خلال بحثه أنّ مشاركة المتعلّم في عمليّة التعليم يلعب دورا كبيرا في فهمه للدرس وقدرة تذكّره فيما بعد، وتوفّر المدوّنة تلك البيئة الدراسيّة للمتعلّمين. إذن، يمكن القول إنّ لكلّ من المعلّم والمتعلّم إفادة كثيرة في الاستعانة بالمدوّنات اللغويّة المتوازية في عمليّة التعليم والتعلّم. على سبيل المثال، يستطيع المعلّم تقديم السياقات المختلفة المستخرجة من مدوّنة لكلمة معيّنة إلى طلابه، ثم يطلب منهم اكتشاف معنى الكلمة بأنفسهم أو نوع الكلمة من خلال السياقات، كما يُمكنه من خلال الاستعانة بالمدوّنة معرفة مستوى طلابه قبل التعليم وبعده، وكذلك بإمكانه استثمارها في بناء وتصميم الدروس، حيث يقدّم للطلاب الأمثلة والتدريبات الفعالة الجاهزة الملائمة لمستواهم لتعزيز هدف محدّد (الدكروري، 2018). وينبغي للمعلّم عند دراسة الأصوات العربيّة لغير الناطقين بها–مثلا–أن يركّز أكثر على الأصوات غير الموجودة في لغتهم، مثل"ث" و"ض" و"ع" وغيرها لأبناء اليوربا.

ومن أهمّ ما يستفيد منه المعلّم في الاستعانة بالمدوّنة المتوازية أيضا تدريس المتلازمات اللغويّة، أي؛ مجموعة من الكلمات تدلّ على معنى واحد الذي قد لا يُدرك عند

تفكيك الكلمات مثل العبارة "رأسا على عقب"، والمتصاحبات اللغويّة مثل "المدينة المنوّرة"(نجي العاتي، 2019: 26). ومن المهم أن يطّلع متعلّموا اللغة العربيّة من الناطقين بغيرها – على الأقل–على أشهر هذه العبارات كي لا تكون جديدة عليهم فيحاولون فَهمَها بتفكيك الكلمات المكوّنة منها. علاوة على ذلك، إذا اعتمد المعلّم على ذاكرتِه في الإتيان بالأمثلة قد يَنسى العبارة التي يريدها تماما أو جزءا منها، فمن المستحسن أن يستعين بالمدوّنة في تجهيز الأمثلة الكافية والمناسبة في هذه الظاهرة.

وأمّا من ناحية الطلاب، فمثال هذه المقاربة لا تعطيهم فرصة المشاركة الفعّالة في عمليّة التعليم فحسب، بل تتيح لكلّ طالب إمكانيّة التعلّم الذاتيّ والإمداد بالمعلومات الواسعة، مثلا يستطيع إدراك معاني المفردات وكيفيّة استعمال الأدوات اللغويّة وبعض الكلمات المتصاحبة بنفسه من خلال السياقات في المدوّنات المتوازية.

كذلك تساعد المدوّنة المتعلّمين على إتقان أساس اللغة الأجنبيّة في أقرب وقت ممكن. وذلك لأنّ النصوص المقدمّة إليهم جُمعت بطريقة منهجية عِلمية –لا عشوائيّة–تهدف إلى قضاء حاجاتهم اللغويّة، فحيث يتدرّبون على كيفية نطق كلمة مثلا، سيُدركون فيما بعد أنّهم تعلّموا مفردة لغويّة جديدة وكيفية كتابتها في الوقت نفسه دون أن يشعروا.

ويُلاحظ من خلال ما سبق أنّه يمكن تقسيم طريقة الاستفادة من المدوّنة المتوازية في تعليم لغة لغير الناطقين بها إلى قسمين: الطريق المباشر وغير المباشر. والطريق المباشر هو تعامل المعلّمين والمتعلّمين مع المدوّنة تعاملا مباشرا أثناء عمليّة التعليم والتعلّم، وأمّا الطريق غير المباشر فهو تعامل الباحثين ومصمّمي المناهج والموادّ الدراسيّة مع المدوّنة. (Boulton & Vyatkina, 2021: 66)

3. مميّزات المدوّنة اللغويّة المتوازية لغرض تعليم اللغة العربيّة للناطقين بغيرها

إنّ للمدوّنة المتوازية التي بُنيت لغرض تعليم لغة ما للناطقين بغيرها سمات تتميّز بها من غيرها من المدوّنات، ويمكن ذكر بعضها على النحو الآتي:

تضمين عناصر اللغة الأكثر مداولة في البيئة التي يعيش فيها المتعلّمون؛ لأنه من المتوقّع تطويرُ المتعلِّمين في كلا المهارات الاستيعابيّة والإنتاجيّة عندما يدرسون عناصر وأنماط اللغة الأكثر شُيوعا في بِيئتهم، واستبعاد التركيز على الكلمات والتراكيب النادرة استعمالا فيها.

تمثيل المدوّنة للّغة المستهدفة، ويمكن تقييم ذلك من خلال عاملين؛ مدى التوازن في تنوّعات أوعية الموادّ Genres، وطريقة اختيار العيّنة من كل نوع الأدبيّات Sampling (Alotaibi, 2017: 329). ويلاحظ أنّ كثيرا من المدوّنات المتوازية المُنجزة تنقص قيمتها من خلال أحد هذين العاملين، حيث توجد مدوّنات جُمعت موادّها من خلال نصوص محدودة التنوّع، أو عدم التوازن في أنواع أدبياتها، كما توجد بعض المدوّنات المتوازية مغلقة لا سبيل إلى تحديثها فيما بعد، فتصبح–بعد ما كانت تمثيليّة–غير تمثيلية بمرور الوقت (Hunston، 2002:30)، إذن، ينبغي أن تكون المدوّنة بذلك الغرض تمثيليّة ومفتوحة قابلة للتحديث.

تضمين المواقف اللغويّة الأكثر إشكالا لدى المتعلّمين سواء على المستوى الصوتيّ أو الصرفيّ أو التركيبيّ أو المعجميّ. وعند مقارنة نظام لغتين من اللغات البشريّة، فمن الطبيعيّ أن تُدرك أوجه الاتفاق والاختلاف بينهما، وأنّ أهمّ التحدّيات عند متعلّم اللغة الثانية تكمُن عادةً في أوجه الاختلاف، حيث يجد المتعلّم في ناحية أَوجُه الاتّفاق سهلا (العصيلي،

2002: 193)، ومما يوضّح ذلك على المستوى الصوتيّ مثلا هو التركيز على حرف "ض" لأبناء اليوربا؛ لأنه لا نظير له في لغتهم، وكذلك حرف"غ" المشبّهة بحرف"G" في لغتهم اليوربا؛ لذلك ثمّة احتمال كبير بأنّهم سيواجهون صعوبة في إتقان نطقهما. وكذلك الأمر في التمكّن من التفريق بين الحروف العربيّة بأشكالها المختلفة؛ لأنّ الحروف في لغتهم ليس لها أشكال متنوّعة، إنما لكل حرفٍ شكلٌ واحدٌ يبقى عليه بغضّ النظر عن موقفه في الكلمة. على ذلك، لعلهم سيجدون صعوبة في تعليم اللواصق والمتصاحبات والمتلازمات وبناء الجملة الفعليّة وغير ذلك في المستويات الأخرى.

استخدام نصوص المدوّنة باللغة المعاصرة، المكتوبة أو المحكيّة، باعتبارها تمثيلا للواقع اللغويّ عند أبنائها، وتضمّ نسبة هائلة من مجالات المعرفة كالدين والآداب، والعلوم والفنون والإعلام. وينبغي استبعاد المفردات والأخبار التي تخصّ مجتمع العرب دون مجتمع المتعلّمين في المدوّنة المصنوعة بين اللغتين خاصة لتعليم المبتدئين. مثال ذلك كلمة النقال والريال اللتان تخصان بيئة العرب؛ لأنّ تعلّم ذلك يتطلب عمليّة حفظ الكلمة دون معرفة ماهيتها، وقد يطول على المتعلّم الأمَدُ دون أن يحتاج إليهما في تعبيره، فيمكن يَجِد المتعلّم الصعوبة في تذكّرها عند حاجته إليها، ويلاحظ عكس ذلك عند تقديم المفردات والعبارات التي تخصّ بيئة المتعلّمين، سرعان ما يتذكرونها عند الحاجة، ومثال ذلك كلمة "النيرة" – العُمْلَة المستخدمة في بلاد اليوربا، فمن المتوقّع أن يحتاج المتعلّم إليها عدة مرات يوميّا؛ لأنّها كثيرة الاستخدام في بيئته.

استخراج المدوّنة بالجودة العالية الدقيقة. نظرا إلى أنّ الغرض من بناء المدوّنة هو تعليم اللغة، ينبغي أن تكون خالية من الأخطاء، ولكن يصعب بناء مدونة لغويّة متوازية بهذه السمة في عصرنا الراهن، خاصة بين اللغة العربيّة واليوربا؛ لوجود عدة تحديات عند معالجة آلية لكلٍّ من نصوص اللغتين، لكن بغضّ النظر عن تلك التحدّيات، ينبغي ألا

تتجاوز نسبة الأخطاء في المدوّنة خمسة بالمائة (السعيد، 2015: 22–23)؛ لأن وجود خطأ فيه قد يؤدي إلى فقدان الثقة في المدوّنة بكاملها من قبل مستهدفيها.

تضمين النصوص كبيرة الحجم في المدوّنة. يعتمد حجم المدوّنة اللغويّة بشكل أساسيّ على نوع الأسئلة التي تُسهم المدوّنة اللغويّة في الإجابة عنها إضافة إلى الأهداف التي تسعى تحليلات المدوّنة اللغويّة إلى تحقيقها، وحجم المدوّنة يقاس بعدد كلماتها، والكلمة– عند المعالجة الآلية–هي مجموعة متتابعة من الرموز لا يفصل بينها فراغ. وبعبارة أخرى، سيُعتبر الاسم أو الفعل المرتبط بضمير كلمة واحدة، وليس كلمتين مثلا: "كتب"، و"كتبت"، و"كتبتها" – كلٌّ من تلك المفردات تعدّ كلمة واحدة.

ومن سمات المدوّنة المتوازية لهدف تعليم اللغة الأجنبيّة اشتمالها عددا هائلا من الكلمات، فكلّما ازداد عددها كان أفضل؛ لأنّ ذلك عادةً يزيد احتمال وجود مفردة أو تركيب أو نمط لغويّ آخر في المدوّنة. وقد رأى بعض الباحثين أنّ مدوّنة بمليون كلمة كافية للبحث عن الظواهر العامة في اللغة (الثبيتي، 2015: 157). على ذلك، يمكن القول إنه ينبغي ألّا يقلّ عدد كلمات المدوّنة اللغويّة بين اللغتين عن مليون كلمة، وكلما زادت كان أفضل.

4. منهجيّة بناء المدوّنة اللغويّة المتوازية بين لغتين

تنوّعت طرق بناء المدوّنة عند علماء لسانيات المدوّنة، لكن ثمة بعض التساؤلات التي لابدّ أن تكون إجاباتها واضحة عند صانع المدوّنة قبل مبادرته بالمشروع، ولعلّ من أهمّها؛ ما الغرض مِن إنشاء المدوّنة؟ ومَا الفئة المستهدفة للمدوّنة؟ وهل هي للاستعمال الشخصيّ أو العامّ؟ وهل الموارد المناسبة لتجميع النصوص متاحة ويمكن الحصول عليها باعتبار قواعد حقوق المِلكيّة الفكريّة؟ ولكلّ هذه التساؤلات دور كبير في نجاح بناء

المدوّنة المنشودة، ومن أدوارها إرشاد صانع المدوّنة إلى جمع أنسب النصوص لتحقيق غرضه، وتحديد نسبة النوع لكلّ نصّ، واختيار نوع المدوّنة الذي يخدم غرضه، واتّخاذ القرار ببناء مدوّنة –مغلقة أو مفتوحة، وغير ذلك.

رغم وجود التنوّع في طرق بناء المدوّنة إلّا أنّ هناك ثلاث مراحل أساسيّة في إنجاز المدوّنات اللغويّة من جديد؛ بداية من تصميم نوع موادّ المدوّنة، مرورا بجمع نصوصها، إلى القيام برمزها ووسمها(Annotation (Biber & Reppen, 2015: 35.

1.4 مرحلة تصميم نوع موادّ المدوّنة

من أهمّ ما يقوم به صانعو المدوّنات المتوازية في مرحلة تصميم نوع موادّ المدوّنة هو تحديد أوعية الموادّ المختلفة من لغة المصدر التي تؤدّي إلى تحقيق أهداف بناء مدوّناتهم؛ حتى يكون المشروع عملا منهجيّا مخطّطا ذا أهداف واضحة من البداية. وإنّ الغاية في التصميم هي الحصول على مدوّنة تمثيليّة للغة من كلّ الظواهر الملائمة مع أهداف بناء المدوّنة، يعني لا يكتمل التصميم دون مراعاة عنصر التمثيل في المدوّنة المنشودة؛ لأنّ العنصر يتماشى مع أهداف بناء المدوّنة، وقد تتناسب درجة التمثيل في المدوّنة مع صحة النتائج التي توصّل إليها صانع المدوّنة من خلال تحليل المدوّنة.

ولكنّ السؤال الذي يُطرح هو: على أيّ معيار يتمّ اختيار موادّ المدوّنة حتى تكون المدوّنة تمثيليّة؟ هناك صِنفان من المعايير للاختيار؛ المعايير الداخليّة، وهي تتعلّق بالبنية اللغويّة الداخليّة للنص ذاته–مثل أنواع النص Text Type، والسمات النحويّة للنص؛ والمعايير الخارجيّة، وهي تعتمد على الوظيفة التواصليّة للنص في البيئة التي ظهر فيها– مثل وعاء النص، والسجلات، وجنس الكاتب، ومستواه الاجتماعي، وجنسيته. ومن الصعب اعتبار المعايير اللغويّة معيارا أساسيّا للاختيار؛ لأنّه تُبنى المدوّنة أساسا بشكل تمثيليّ

لدراسة التوزيعات اللغويّة، وإذا كان توزيع اللغويّ–مثلا– محددا مُسبقا عند تصميم موادّ المدوّنة، فلا فائدة من تحليل مثل تلك المدوّنة لاكتشاف توزيعات السمات اللغويّة التي تحدث بشكل طبيعيّ. على ذلك، يصحّ القول إنّ النصوص التي سيتمّ تضمينها في المدوّنات ينبغي أن يتمّ اختيارها وفقا لمعايير خارجيّة لا لخصائصها اللغويّة؛ حتى يمكن الاستفادة من نتائج تحليل المدونات في تحسين سمة التمثيل للمدوّنات الأخرى (الثبيتي،2015: 151).

ويمكن تقسيم المدوّنات باعتبار طبيعة نصوصها إلى قسمين؛ المدوّنات العامّة والمتخصّصة (السعيد، 2019: 65). وتمثّل المدوّنات المتخصّصة المدوّنات التي تنتمي نصوصها إلى مجال معيّن، مثل مجال الطبّ، والقانون؛ أو وعاء محدّد، مثل نصوص الصحافة، أو الأدب. وأمّا المدوّنات العامّة فهي عادةً ما تكون نصوصها عامّة شاملة لعدد من الظواهر اللغويّة من حقول معرفيّة مختلفة. ومن المدوّنات التي تُعتبر عامّة المدوّنة الوطنيّة البريطانيّة(British National Corpus (BNC ؛ لأنّ الهدف في بنائها هو تمثيل اللغة الإنجليزيّة البريطانيّة الحديثة كاملة. ومن سمات المدوّنات العامّة هي أن تغطي– بشكل متناسب– أكبر عدد ممكن من أوعية النصوص الموجودة لتكون المدوّنة تمثيليّة إلى أقصى حد للّغة التي تسعى إلى تمثيلها. ويعتمد عنصر التمثيل للمدوّنات العامّة غالبا على أخذ عيّناتها من مجموعة واسعة من أوعية النصوص المختلفة.

بناء على ذلك، يلاحظ أنّ المدوّنات المتوازية التي بُنيت لغرض تعليم مهارات اللغة العربيّة الأربع للناطقين بغيرها تندرج تحت المدوّنات العامّة. ويُعرف مدى تمثيل المدوّنات لتعليم اللغة من خلال عنصري المدوّنة: التوازنBalance، وطريقة أخذ العيّنات Sampling (McEnery, & Et. Al; 2006: 13-14).

1.1.4 التوازن Balance:

إنّ التوازن في عمليّة بناء المدوّنة ظاهرة تتعلّق بنطاق أوعية الموادّ المتضمّنة في المدوّنة، والطريق الذي يتمّ اختيار أوعية نصوصها. ويعتمد التوازن في المدوّنات بشكل كبير على مستخدميها بغضّ النظر عن نوع نصوصها. على سبيل المثال، تُعدّ المدوّنات العامّة التي تحتوي على كلٍّ من النصوص المكتوبة والمنطوقة–مثل المدوّنة الوطنيّة البريطانيّةBNC – متوازنةً كما تعتبر المدوّنات التي تحتوي على النصوص المكتوبة فقط– مثل مدوّنة براون Brown Corpus ومدوّنة لانكستر أوسلو بيرغن Lancaster–Oslo-Bergen(LOB)، أو المنطوقة فقط –مثل مدونة كامبريدج ونوتنجهام للخطاب باللغة الإنجليزية CANCODE– متوازنةً؛ لأنّ نصوص كلّ منها تغطي نطاقا واسعا من أنواع النصوص التي يُفترض أن تكون ممثّلة للظواهر اللغويّة التي في قيد الدراسة. ويتمّ أخذ عيّنات عادةً من أنواع النصوص المختلفة بشكل متناسب لإدراجها في المدوّنة بحيث تقدّم نموذجا صغيرا لموادّ اللغة التي يرغب صانع المدوّنة دراستها.

وتُجدر الإشارة إلى عدم وجود مقياس عِلْميّ لمعرفة مستوى التوازن للمدوّنات بشكل عامّ، إنّما تعتمد الفكرة غالبا على الحدس والتقديرات. ويرى أتكينز Atkins أنّ مدى توازن المدوّنات ليس أمرا يُمكن معرفته إلّا بعد بناء مدوّنة أوّليّة (McEnery, Et. Al; 2006: 16). ولعلّ النهج الأكثر نموذجيّةً لتوازن المدوّنات هو أن يعتمد صانعو المدوّنات على نموذج المدوّنات المنجزة مُسبقا وهي معروفة بالتوازن لبناء مدوّناتهم الخاصة على افتراض أنّ التوازن سيتحقّق من النموذج المعتمد عليه. وقد بُني العديد من المدوّنات على هذا النهج. على سبيل المثال، قد تمّ قبول المدوّنة الوطنيّة البريطانيّة عموما على أنّها مدوّنة متوازنة، فأُخذتْ نموذجا لبناء عدد من المدوّنات، منها المدوّنات الآتية:

المدوّنة الوطنيّة الأمريكيّة The American National Corpus.

المدوّنة الوطنيّة الكوريّة The 21st Century Sejong Corpus.

المدوّنة الوطنيّة البولنديّة The National Corpus of Polish.

المدوّنة المرجعيّة الروسيّة The Russian Reference Corpus (Sharoff, 2006: 66).

ولعلّ إلقاء نظرة فاحصة على معايير التصميم المستخدم في بناء المدوّنة الوطنيّة البريطانيّة يساعد على إعطاء فكرة عامّة عمّا يفترض أن يكون عليه تصميم المدوّنات المتوازنة. إنها تحتوي على مائة مليون كلمة تقريبا، وتسعون بالمائة منها نصوص مكتوبة وعشر بالمائة نصوص منطوقة، وقد تمّ اختيار النصوص المكتوبة باستخدام ثلاثة معايير: مجال النص، وفترة إنتاج النص، ونوع النشر النصيّ، مثل كتب ودوريات أو مخطوطات غير منشورة (Aston & Burnard, 2002: 29–30). أمّا النصوص المنطوقة فتمّ جمعها على أساس معيارين: ديموغرافيّة ومحكومة بالسياق. ويتمثّل المكوّن الديموغرافي في المدوّنة من لقاءات غير رسميّة سجّلها 124 متطوّعا، وتمّ اختيارهم حسب الفئة العمريّة والجنس والطبقة الاجتماعيّة والمنطقة الجغرافيّة، بينما تتكوّن المحكومة بالسياق من لقاءات رسميّة، مثل الاجتماعات، والمحاضرات، والبرامج الإذاعيّة المسجّلة في أربع تصنيفات سياقيّة. ويُكمّل نوعا النصوص المنطوقة بعضهما البعض، فقد لا تغطي المدوّنة العديد من سياقات الكلام إذا اعتمد على تقنيات العيّنات الديموغرافية وحدها في جمع النصوص (McEnery, & Et. Al, 2006 : 17).

ويجدر الذكر أنّ تَدْوين المعلومات الأساسيّة عن تصميم المدوّنات يلعب دورا كبيرا في صناعة المدوّنات؛ إنه يساعد في إدارة المدوّنة وتحديد مدى مناسبتها لغرض البناء، وقد يُستفاد منه في بناء مدوّنات أُخر. وأهمّ هذه البيانات معلومات المؤلف، وعنوان النص، ونوع النص ومجاله، وموضوعه، وتاريخ إصدار النص، وناشر النص مع مكان

النشر. وكلّ تلك البيانات، في الوقت نفسه، قد تساعد في تقييم المدوّنة بما أنها توفّر للباحثين معرفة معلومات النصوص التي تكوّن منها المدوّنة.

وينبغي الحذر عند اتّخاذ القرار بناء على نتائج الدراسات المعتمدة على المدونات، لاسيما التي لم توفّر البيانات المستخدمة لبنائها؛ لأنه صعبٌ للمستخدم معرفة مدى تلاؤم المدوّنة لتحقيق أهدافه. ومع ذلك، لا يعني أن تُفقد الثقة تماما في نتائج المدوّنات، ويقول اتكينز والآخرون (McEnery, et. Al, 2006: 19) في هذه المناسبة:

"إنه من قصر النظر حقا الاعتقاد بعدم الاستفادة من المدوّنات إلّا بعد التمكّن من تأكيد وجود التوازن فيها عِلميّا، والتسرّع في رفض نتائج تحليل المدوّنات باعتبارها غير موثوقة، أو غير مفيدة؛ لأنّ المدوّنات المستخدمة لا يمكن الإثبات أنّها متوازنة".

فيمكن الاعتماد على نتيجة المدوّنة التي تتلائم بيانات بنائها مع الأهداف المطلوبة.

2.1.4 طريقة أخذ العيّنات:

بما أنّ الهدف هو بناء مدوّنة لغويّة متوازية لتيسير كلّ مهارات اللغة المستهدفة للناطقين بلغة المصدر، فالأمر يتطلّب جمع كلّ النصوص في اللغة المستهدفة، لكن هل يوجد سبيل إلى تحقيق ذلك؟ قد أجاب عن ذلك السؤال الثبيتي (2015: 148) قائلا: "من المستحيل عَمَلِيّا جمع كلّ ما قيل وكُتب للحصول على أحكام دقيقة وفاصلة، إنّنا أمام خيار يبدو أكثر واقعيّة ويستند إلى أساس علميّ، ألا وهو جمع عيّنة متوازية وممثّلة للّغة..."، أي: من الضروريّ الالتفات إلى فِكرة أخذ العيّنات التمثيليّة للغة المصدر. ويبقى على صانع المدوّنة النظر في الطريقة المتّبعة في اختيار العيّنات التمثيليّة للمدوّنة المنشودة. وسعيا لتحقيق ذلك، فإنّ أوّل خطوة هي تحديد الحدود التي تُؤخذ منها العيّنات Population، ثم تحديد وحدة أخذ العيّنات Sampling Frame. وبالنسبة للنصوص

المكتوبة–مثلا– قد تكون الحدود التي يؤخذ منها العيّنات عبارة عن قائمة جميع النصوص المكتوبة التي توجد في إطار كل الوحدات، فحينئذ تتمثّل وحدة أخذ العيّنات في كتب، أو دوريات، أو صحف، وغيرها من النصوص المكتوبة.

على سبيل المثال، إنّ الحدود التي أُخذت منها عيّنات مدوّنة براون الرائدة هي الكتب المكتوبة باللغة الإنجليزيّة التي نُشرت في الولايات المتّحدة في عام 1961 بينما كان إطار أخذ العيّنات عبارة عن قائمة من مجموعة الكتب والدوريات في مكتبة جامعة براون وبروفيدنس أثينيوم Providence Athenaeum. ومن الصعب في عمليّة بناء المدوّنات تحديد وحدة العيّنات، والحدود التي يتم أخذها؛ وذلك لعدم وجود وصفات واضحة تخصّ الموضوع، ولكن تتمّ المهمّتان عادةً بالنظر إلى طريقة إنتاج اللغة التعليميّة –كما اعتُبرت في إنشاء مدوّنة براون– واستقبال اللغة، وطبيعة اللغة بذاتها.

وبمجرّد التمكّن من المهمّتين، يمكن أخذ العيّنات المنشودة من خلال عدّة أساليب، ولعلّ الأبسط منها أسلوب أخذ العيّنات العشوائيّ البسيط Simple Random Sampling، حيث يتمّ ترقيم جميع الوحدات في إطار أخذ العيّنات، ثمّ يختار صانع المدوّنة العيّنة باستخدام جدول الأرقام المرتّبة عشوائيّا. قد يُنتج هذا الأسلوب عيّنة لا تتضمّن العناصر النادرة نسبيّا في المدوّنة، على الرغم من أنّها قد تكون ذات فائدة لبعض صانعي المدونات، إلّا أنّها غير مناسبة للمدونات المتوازية التي بُنيتْ لغرض تعليم كلّ مهارات اللغة المستهدفة للناطقين بلغة المصدر. وتجنّبا لتلك المشكلة، يَجدر استخدام أسلوب أخذ العيّنات العشوائيّ الطبقيّ Stratified Random Sampling، حيث يُقسم أوّلا جميع الحدود التي يتمّ أخذ العيّنات إلى مجموعات متجانسة نسبيّا، ثمّ يتمّ أخذ العيّنات من كلّ الطبقات بشكل عشوائيّ. وفي مدوّنة لانكستر أوسلو بيرغن (LOB)، على سبيل المثال، قُسّمت الحدود التي أُخذت العيّنات منها أوّلا إلى خمس عشرة فئة نصيّة من

التقارير الإخباريّة، والنثر الأكاديميّ، والروايات المختلفة؛ ثمّ سُحب العيّنات من كلّ فئة نصيّة. ويلاحظ أنّ أسلوب أخذ العيّنات العشوائيّ الطبقيّ أيضا لا يخلو من القيود، خاصة في تحديد نسبة العيّنات وعددها لكلّ فئة نصيّة، ذلك لأنه ينبغي أن تكون أعداد العيّنات عبر فئات النص متناسبة مع تكراراتها في الحدود التي أُخذت منها العيّنات حتى يتمّ اعتبار المدوّنة الناتجة على أنّها تمثيليّة، ويُلاحظ أنه من الصعب تحديد هذه النسبة موضوعيّا (Hunston 2002: 28-30).

2.4. جمع نصوص المدوّنة

وبعد التمكّن من تخطيط الطريق المؤدي إلى جمع موادّ المدوّنة المنشودة، ليس على صانع المدوّنة إلّا اتّباعه خطوةً خطوة، فيبدأ بجمع نصوص مدوّنته. ويمكن تقسيم النصوص المتضمّنة في المدوّنة إلى شكلين: إمّا أن تكون مكتوبة أو منطوقة، وقد يُدمج الشكلان معا لتحقيق غرض إنشاء المدوّنة– وهو المختار عادةً في بناء مدوّنات متوازية بين اللغتين لغرض التعليم. وحينما كانت موادّ المدوّنة ممثّلة بنصوص مكتوبة، كان يتمّ إدخالها إلى الحاسوب يدويّا، ولكن مع التطوّر التكنولوجيا الذي يوفّر لنا الوقت والجهد والطاقة البشريّة، صار يُستعمل قارئ المحارف البصرية Optical Character Reader لمسح النصوص المطبوعة وتحويلها إلى نصوص رقميّة، أو يُستعان بالنصوص المتاحة في الشبكة العنكبوتيّة. وأمّا النصوص المنطوقة فإنّها تتطلب تسجيلا نصيا Transcripting، وعلى ذلك يتمّ نسخ النصوص ببرامج الإملاء الآليّ التي تحوّل النصوص المنطوقة إلى المكتوبة Speech-To-Text. وقد تنسخ عن طريق النسخ الصوتيّ، حيث يستعمل رموز للدلالة على منعطفات الحديث من إيماءات، وتداخلات لفظيّة وغير لفظيّة، ووقف، وتردّد، وما شابه ذلك من ظواهر تتّسم بها النصوص المنطوقة والحوارات (الدكروري، 2018: 55–56).

ومن هنا يسعى صانع المدوّنة إلى ترجمة النصوص إلى اللغة المستهدفة للحصول على النسخة باللغة المستهدفة، ثم تتمّ محاذاة النصين مع بعضهما البعض Alignment آليّا أو يدويّا. ومن المعروف أنّ ترجمة النصوص لإنتاج المدوّنة المتوازية ليست أمرا سهلا؛ لأنّها تتطلّب التوازن بين النصين بالدقة، أي: أن يُوجد لكلّ كلمة من النص المصدر دليل في النص المستهدف– وهو أمر غير هيّن في عمليّة الترجمة خاصة في ضوء مدوّنة كبيرة نسبيّا؛ لأنّ الأمر يتطلب تدريبا عميقا على أدوات الترجمة الآلية كي تتمكّن من إنتاج النصوص المناسبة لغرض تعليم اللغة الأجنبيّة، أو لغويّا يجيد اللغتين المصدر والمستهدفة وثقافتهما إن كانت الترجمة بشريّة حتى يعرف متى يزيد أو ينقص كلمةً فأكثر في النص المستهدف لسبب من الأسباب قبل أن يستقيم المعنى(Alotaibi, 2017: 330).

والجدير بالذكر أنّ النصوص بحاجة إلى تَنقية مستمرّة ومراجعة دقيقة بعد كلّ معالجة آلية تقريبا؛ لاحتمال وجود الأخطاء مثل وجود ألفاظ أجنبيّة، أو تكرار النص، أو الفراغات غير المطلوبة، أو احتواء بيانات النص الهامشيّ مثل رقم الصفحات والإعلانات التجاريّة. ويمكن التخلّص من معظم تلك الآثار غير المطلوبة عن طريق عدّة تطبيقات، منها Notepad++، وNote Tab Pro، وjustText demo. وينبغي القيام بمراجعة يدويّة في النهاية؛ للتأكّد من أنّ الأخطاء الإجماليّة في النصوص لم تجاوز النسبة المسموح بها.

3.4. رمز المدوّنة ووسمها

وبعد تجهيز النصوص بشكل موازٍ بين اللغتين تحوّل إلى ترميز واحد قابل للمعالجة عند معظم البرامج المستخدمة لمعالجة المدوّنات، مثل UTF8 وUTF16، ثم تحفظ على شكل نصوص رقميّة بسيطة plain format؛ لتميّزها في التعامل مع برمجيات مختلفة.

أمّا بخصوص وسم المدوّنة المتوازية Parallel Corpus Annotation –أي تحشية المدونّة– فيقصد بها إجراء تحليل لغويّ لنصوص المدوّنة لاستخراج بعض التفسيرات والمعلومات اللغويّة. وبعبارة أخرى، تعني مجموعة من التحليلات والمعالجات اللغويّة التي تتمّ على النصوص بهدف إضافة توصيف دقيق إليها، فتوفّر إمكان استرجاع واستخراج المعلومات وفقا لنوع التحشية الذي تنتمي إليه كلمة أو جملة أو فقرة، مثلا: وَسْم أقسام الكلمPOS Tagging ، والإعراب التركيبي، والأسر اللغويّة Lemmas، وغير ذلك.

وقد قُسم التحشية إلى نوعين؛ التحشية الداخليّة Embedded Annotation وهي التي توضع لتفسيرات في داخل النص، والتحشية الخارجيّة Stand-off Annotation وهي التي توضع لتفسيرات في ملف مستقل ثم تدمج مع نصوصها الأصليّة. وقد مالَ عددٌ من الباحثين إلى تفضيل النوع الثاني على الأوّل؛ استدلالا بأنّ الأوّل يشكّل تشويشا على النصوص المسترجَعة (McEnery & Hardie, 2014:13).

هناك ثلاث طرق يمكن تحشية المدوّنات من خلالها، ويمكن ذكرها على النحو الآتي:

الطريقة الآلية هي الاستفادة من البرمجيات المصنوعة بناء على قواعد وخوارزميات التي تمّ إعدادها مسبقا لإضافة التفسيرات إلى نصوص المدوّنات، وتعتبر هذه الطريقة أسرع في المعاملة مع المدوّنات الكبيرة إن لم يُنظر إلى مدة إعدادها.

والثانية هي التحشية اليدويّة؛ أي: الاعتماد تماما على محلّل بشريّ دون أيّ تدخّل للآلة في عمليّة التحشية. ويُنظر إلى هذه الطريقة بأنها مُكلفة ومستنفدة للوقت، إلّا أنّها تُقدّم على الطريقتين الأخريين عند معظم اللغويّين لاسيّما إذا كانت المدوّنة صغيرة؛ لقلة الأخطاء في إنتاجها.

وثالثها هي الطريقة شبه الآلية، حيث يتمّ استخدام كلٍّ من الآلة والبشر في عمليّة التحشية. وفي هذه الطريقة تُبدأ العمليّة عادةً بالآلة، ثم يتدخّل البشر لِحلّ حالات الخطأ أو اللبس التي تستعصي على البرامج الآلية. وهذه الطريقة تنتج مسترجعة أكثر موثوقيّة من الطريقة الأولى، إلّا أنها تتطلّب وقتا أطول وطاقة بشريّة (الدكروري، 2018: 70).

5. خاتمة

بعد ما كانت الأهداف واضحة لدى صانع المدوّنة، تَبيّن من ضمن هذه الدراسة أنّ الطريقة النموذجيّة لبناء مدوّنة لغويّة متوازية لغرض تعليم اللغة العربيّة للناطقين بغيرها هو تحديد الفئة المستهدفة للمدوّنة، ونوع المدوّنة المناسب لتحقيق أهدافها، وتأكيد إمكانيّة الحصول على النصوص المناسبة للدراسة، وغير ذلك من المعلومات الأساسيّة.

ثمّ ينتقل صانع المدوّنة إلى مرحلة التصميم، حيث يُصمّم منهجا متّبعا للحصول على مدوّنة موثوقة ويحدد حدود أخذ موادّ مدوّنته، ولتحقيق ذلك لابدّ أن تتمتّع المدوّنة بعنصر التمثيل، ولا يتمّ ذلك إلا بالعناية بالتوازن وطريقة أخذ العينات. وليست هناك طريقة عِلميّة تَضْمن لصانع المدوّنة إدراك التوازن والعينات المناسبة تماما لتحقيق أهدافه، إنما يبقى الأمر على التقديرات، والاستفادة من تصاميم المدوّنات المنجزة مُسبقا وتلقّي القبول لدى معظم علماء لسانيات المدوّنة وباحثيها، مثل المدوّنة الوطنيّة البريطانيّة، ومدونة لانكستر أوسلو بيرغن(LOB)، وغيرهما.

وبعد التمكّن من تجهيز التصميم، ليس على صانع المدوّنة إلا اتّباعه خطوة خطوة، فيبدأ بجمع نصوص مدوّنته من المكتوبة والمنطوقة، ثم يقوم بمحاذاة نصوص اللغة المصدر مع نصوص اللغة المستهدفة، ويلي ذلك تحويل كلّ النصوص إلى ترميز واحد قابل للمعالجة– مثل UTF16، ومن هنا يمكنه القيام بتحشية النصوص حسب أغراضه،

والتي قد تكون على مستوى الكلمة، أو شبه الجملة، أو غيرهما. ومن المهم العناية بتنقية النصوص إثر كلّ معالجة لتقليل نسبة الأخطاء في المدوّنة.

6 . الهوامش والمراجع

[1] الثبيتي، عبد المحسن، 2015، لغويات المدونة الحاسوبية: تحليلات تطبيقية على العربية الطبيعية. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

[2] الدكروري، أيمن، 2018، المدوّنات اللغويّة ودورها في معالجة النصوص العربية. الرياض. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، ط1.

[3] السعيد، المعتز بالله 2015، نحوَ مُعجم للّغة العربيّة للنّاطقينَ بغيرها" معالجة حاسوبيّة احصائيّة". مجلة الدوليّة لهندسة اللغة العربيّة واللسانيّات العامّة، المغرب، ع18، صص 1–37.

[4] السعيد، المعتز بالله 2019، الموارد اللغويّة الحاسوبية. تحرير: رشوان، محسن والسعيد المعتز بالله، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، صص51–88.

[5] عصيلي، عبد العزيز2002، أساسيات تعليم اللغة العربية للناطقين بلغات أخرى. مركز بحوث اللغة العربية وآدابها، مكة.

[6] نجي العاتي، أيمن الطيّب2019، ترتيب الوحدات المعجمية المركبة في المعجم العربي المعاصر – معالجة لغوية حاسوبية. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية.، الرياض.

[7] Alotaibi, H. 2017, Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. Arab World English Journal, 8(3), pp. 319-337.

[8] Biber, D. & Reppen, R 2015, The Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press.

[9] Boulton, A. & Vyatkina, N. 2021, Thirty Years of Data-Driven Learning: Taking Stock and Charting New Directions Over Time. Language Learning & Technology, 25(3), pp. 66- 89.

[10] Guy Aston, G. & Lou Burnard, L. 2002.The BNC Handbook. Edinburgh: Edinburgh University Press.

[11] Hunston, S. 2002, Corpora in applied linguistics. Cambridge: Cambridge University Press.

[12] Judith, MEECE 2003, Applying Learner-Centered Principles to Middle School Education, Theory into Practice, (42)2, pp. 109-116.

[13] McEnery, T. & Hardie, A. 2014, Corpus linguistics: Method, theory and practice. Cambridge; Cambridge University Press.

[14] McEnery,T. & Et. Al. 2006. Corpus-Based Language Studies. An Advanced Resource Book (Routledge Applied Linguistics). London, New York

[15] Sharoff, S. 2006, Creating General-Purpose Corpora Using Automated Search Engine Queries. Retrieved on 5-8-2022 from: http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf.

[16] Tsai, C. & Choi, H. 2005, Parallel Corpus and Lexical Acquisition in Chinese Learning, Proceedings of Fourth International Conference on Internet Chinese Education, Taipei, Taiwan.