https://kaica.org.sa/site/page/30

مجلة اللسانيات العربية
The Arabic Linguistics Journal
مجلة علمية محكمة نصف سنوية تصدر عن مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية
ISSN: 1658-9858

مجمع الملك سلمان العالمي للغة العربية
King Salman Global Academy for Arabic Language

مجلة اللسانيات العربية، العدد *18*، جمادى الآخرة 1445/ January 2024

بناء المدونات اللغوية القانونية الخاصة بجامعة ليدز: منهجيات وتحديات

Building the University of Leeds Legal Corpora: Methodologies and Challenges

[1]هانم الفرحاتي [2]أماني سيف آل عنيزان* [3]نوران خلّاف

[3،2،1] مدرسة اللغات، والثقافات، ودراسات المجتمع، كلية الآداب والعلوم الإنسانية، جامعة ليدز، ليدز، المملكة المتحدة

توثيق البحث APA Citation:

الفرحاتي، هانم.، آل عنيزان، أماني.، خلّاف، نوران. (2024). بناء المدونات اللغوية القانونية الخاصة بجامعة ليدز: منهجيات وتحديات. *مجلة اللسانيات العربية*، *18*، 96-116.

استقبِل في: 28-10-1444/**رُوجع في:** 14-11-1444/ **قُبل في** 28-01-1445/ **نُشر في:** 19-06-1445

Received on: 2023-05-18 / **Revised on:** 2023-06-03 / **Accepted on:** 2023-08-15 / **Published on:** 2024-01-01

الملخص

رغم التطور في بناء المدونات العربية بمختلف أنواعها، لا تزل هنالك حاجة لبناء مدونات مترجمة متوازية. وما زالت المدونات المتخصصة المتوازية، كالمدونات القانونية، نادرة رغم الحاجة المتزايدة لاستخدامها. لذا فإن بناء المدونات القانونية المترجمة والمتوازية سيمهد الطريق لمزيد من البحوث، كما تعد أيضا مرجعا للقانونيين. يهدف هذا البحث إلى أ. بناء مدونات أُحادية اللغة ومتوازية لدساتير الدول العربية، تشمل جميع نسخ الدساتير المتاحة لجميع الدول العربية منذ عام 1922. ب. بناء مدونة مقارنة لدساتير ثمان دول لغتها الرسمية هي اللغة الإنجليزية (انظر منهجية البحث). ويناقش البحث منهجية مفصلة لبناء هذه المدونات والتحديات التي واجهتها الباحثات خلال جمع البيانات وبناء هذه المدونات وكيفية التغلب على تلك التحديات. يتيح بناء المدونات الفرصة إلى: (1) إجراء بحوث لغوية متعددة التخصصات؛ (2) استخدام مناهج تربوية قائمة على البحث والتعلم المختلط لتعليم الترجمة وتعلمها. وستكون هذه المدونات ذات قيمة كبيرة لباحثي الترجمة وللأكاديميين في الترجمة والقانون والمترجمين والمنظمات الحكومية وغير الحكومية وكذلك المنظمات الدولية.

الكلمات المفتاحية: مدونات متوازية. المدونات اللغوية العربية، ترجمة قانونية.

Abstract

Arabic corpora have been in use since the last decade of the past century and have been constantly increasing. However, there is still a need for more advanced tools and morpho-syntactically annotated Arabic corpora for research and teaching. Despite this growing need, parallel and specialized corpora are still rare. Building the legal Arabic/English corpora can lay the foundation for more research in Arabic legal translation, which is an area currently insufficiently studied worldwide. This paper presents the compilation of a collection of specialized parallel and monolingual legal corpora, including diachronic corpora that include all available constitutions of 22 Arab countries. The goal is to enhance :1. interdisciplinary corpus-based and socio-cultural investigations; and 2. research-led and blended-learning pedagogical approaches to translation teaching and learning. The paper will provide an overview of the methods used to create these specialized complex corpora as well as the obstacles experienced.

Keywords: Parallel corpora; Arabic corpus linguistic; Legal translation.

*المؤلف المراسل: *Corresponding author*

Email:*Amani.onayzan@windowslive.com*

1. المقدمة

بدأت المدونات العربية في الظهور في العقدين الأخيرين من القرن العشرين. وبرغم التطور الكبير في بناء المدونات العربية بكافة أنواعها (الفيفي وأتويل، 2016)، كمدونة arTenTen، ومدونة جامعة الملك سعود للغة العربية الفصحى، والمدونة اللغوية لتدريس معلومات عن الإسلام، لا تزال هنالك حاجة لبناء مدونات عربية متخصصة ومدونات مترجمة متوازية في مجالات مختلفة ليتم توظيفها، باستخدام الأدوات الحاسوبية اللغوية المتطورة، في مجالات البحث العلمي والتدريس وتدريب المترجمين والترجمة بكافة أنواعها. وتعرف المدونات المتوازية (Parallel Corpora) بأنها مجموعتان من النصوص المترجمة بين لغتين مختلفتين، حيث تتوازى النصوص في اللغة المصدر وترجمتها إلى اللغة الهدف (بيكر Baker، 1993). ومع أن هذا النوع من المدونات لا يزال في بدايته، فإنه يحمل إمكانيات كبيرة لتطوير تقنيات الترجمة الآلية والتحليل اللغوي في اللغة العربية (المجيول،2015). وما زالت المدونات المتخصصة، كالمدونات القانونية، نادرة برغم الحاجة المتزايدة لاستخدامها في البحوث اللغوية التجريبية للنصوص العربية، وتحليل لغة القانون، وتدريب المترجمين. ومن ثم جاءت فكرة بناء هذه المدونات القانونية لتمهد الطريق لمزيد من البحوث في مجال الترجمة القانونية ولتكون مرجعا للقانونيين المهتمين بالصياغة القانونية.

يهدف البحث إلى بناء مدونات مختلفة (أُحادية اللغة ومتوازية) لدساتير الدول العربية (انظر الفرحاتي وعليوة-El Farahaty & Elewa، 2020; برايرلي والفرحاتي Brierley & El-Farahaty ، 2019). تشمل هذه المدونات جميع نسخ الدساتير المتاحة لجميع الدول العربية، منذ عام 1922.كما يهدف البحث لبناء مدونة مقارنة لدساتير 8 دول لغتها الإنجليزية هي اللغة الرسمية (انظر منهجية البحث). ومن شأن بناء جميع الإصدارات المتاحة من هذه الدساتير أن يسهّل إجراء بحوث لغوية متعددة التخصصات، تضم لسانيات المدونات اللغوية، واللسانيات القضائية كما تشمل بحوث الترجمة القانونية المقارنة من العربية إلى الإنجليزية، وتقدم بعض تقنيات التحليل اللغوي المقارن المتقدم (مقتبس من بيل Biel ، 2014) ، فضلًا عن تدريب المترجمين وتطوير برامج الترجمة الآلية. كما تسهل للأكاديميين استخدام مناهج تربوية جديدة تواكب التغير المستمر وترضي احتياجات سوق العمل، حيث تستخدم هذه المناهج التكنولوجيا وأدوات التعلم المختلط في تعليم الترجمة وتعلمها.

تتمثل إسهامات البحث في بناء المدونات التالية وجعلها متاحة على منصة إسكتش إنجن Sketch Engine :

1. **المدونة المتوازية الكاملة لدساتير الدول العربية:** وتشمل النسخة النهائية من هذه المدونة دساتير 20 دولة عربية وترجماتها الإنجليزية، ما عدا دولتيْ جزر القمر والصومال، في الفترة ما بين عام 1922 وعام 2022.
2. **المدونات الأحادية العربية أو الإنجليزية:** وتحوي مدونتين منفصلتين لإصدارات الدساتير باللغة العربية واللغة الإنجليزية، كل على حدة، ويبلغ العدد الإجمالي لكلمات مدونة اللغة العربية 788,477 كلمة، فيما بلغ عدد كلمات مدونة اللغة الانجليزية 343,582 كلمة.
3. **مدونة مقدمات (ديباجات) الدساتير العربية:** وتضم هذه النسخة الديباجات المتاحة للدساتير العربية.
4. **مدونة أحادية اللغة للدساتير باللغة الإنجليزية (مدونة مقارنة):** وتضم دساتير **8** دول لغتها الإنجليزية هي اللغة الرسمية، من عام 1985 إلى عام 2016. وتتألف من 677,056 كلمة.

يركز البحث على منهجية بناء المدونات بدءًا من جمع النصوص الأصلية والنصوص المترجمة بأشكالها المختلفة ثم تخزينها وتوثيقها وتجهيزها ومحاذاتها آليا أو يدويا وتحميلها على منصة إسكتش إنجن (Sketch Engine) (كيلغاريف وآخرون (Kilgarriff et al., 2004, 2014).

يناقش الجزء الأول من هذا البحث مقدمة البحث وأهدافه وأسئلة البحث. ويتناول الجزء الثاني مراجعة للدراسات والمشاريع السابقة في مجال المدونات اللغوية العربية، مع التركيز على الفجوة البحثية. ويعرض الجزء الثالث منهجية البحث، وإجراءات جمع النصوص ومصادرها، وخطوات بناء المدونات، والتحديات التي واجهها الباحثون في بنائها، وكيفية التغلب عليها. ويناقش الجزء الأخير النتائج المتوقعة والجهات المستفيدة من البحث والمشاريع المستقبلية.

2. تساؤلات البحث

يحاول البحث الإجابة عن هذين السؤالين: ما منهجية بناء مدونات الدساتير العربية والإنجليزية؟ وما التحديات التي واجهتها الباحثات خلال بناء هذه المدونات؟

3. الدراسات السابقة

سنعرض في هذا الجزء بعض الدراسات التي نُشرت في مجال بناء المدونات العربية وسنقدم لمحة عن هذه المدونات وخاصة ما قام به الباحثون في جامعة ليدز، وسنتطرق للمدونات المتوازية وخاصة المتاح منها في مجال الترجمة القانونية، ونختم بمناقشة الفجوة البحثية لنبين أهمية بناء المزيد من المدونات اللغوية القانونية.

تعد المدونات اللغوية في وقتنا الحالي حجر الأساس لتحليل اللغات المختلفة، ودراسة مفرداتها وتراكيبها، بالإضافة إلى استخدامها في تعليم اللغة وتعلمها، وأبحاث اللغة والترجمة وغيرها. وقد تم بناء وتطوير العديد من المدونات اللغوية الحاسوبية في لغات عدّة، كالمدونة الوطنية البريطانية بي. أن. سي (BNC)، التي تعد من أُولى المدونات الإنجليزية التي أُنشئت في العقد الأخير من القرن العشرين. أما في اللغة العربية، التي سنركز عليها في هذا البحث، فيوجد العديد من المدونات المتنوعة، ويأتي هذا التنوع حَسَبَ اختلاف المواضيع، والتصاميم، والغرض، وحجم المدونات، الذي يعد من المعايير الهامة في بنائها.

في البداية قام الباحثون والمختصون ببناء مدونات متنوعة تُستخدم للإجابة عن مختلف أنواع الأسئلة لاحتوائها على أوعية مختلفة، وتسمى بالمدونات المرجعية، ومن أمثلتها مدونة أر تن تن (arTenTen) (Belinkov et. al., 2013)، وهي مدونة عربية من عائلة المدونات اللغوية المعروفة بـ (TenTen corpora)، وتشمل العديد من اللغات كالإنجليزية، واليابانية، والروسية، والصينية، وغيرها. وقد تم بناؤها باستخدام تقنية متخصصة تقوم بجمع محتوى لُغَوي متخصص من الإنترنت. وتشمل المدونة أكثر من 10 بليون كلمة باللغة العربية. ويمكن استخدامها عبر منصة إسكتش إنجن (Sketch Engine). ومن أوائل الدراسات في بناء المدونات، مدونة نصوص جريدة الحياة (قويدر و رويك Goweder & Roeck، 2001) وهي مدونة لغوية عربية متاحة للاستخدام، تضم 18.5 مليون كلمة.

وقامت مكتبة الإسكندرية بدعم بناء المدونة اللغوية العربية العالمية المعاصرة، وهي منصة بحثية تحتوي على 100 مليون كلمة، تم تحليلها صرفيا، وتمثل إقليما كبيرا من الدول الناطقة باللغة العربية المعاصرة، واعتمد في بنائها على أربعة أوعية نشر أساسية هي: الكتب، والصحف، والمقالات الإلكترونية، والدراسات الأكاديمية (الأنصاري وناجي Alansary & Nagi، 2014).

ويعد مشروع (المدونة اللغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية) من أكبر المدونات المفتوحة وأكثرها تنوعا من حيث النصوص والمصادر. وتحتوي على أوعية نشر متنوعة ما بين الصحف، والمجلات، الكتب، والرسائل الجامعية، والدوريات المُحَكَّمة، والإصدارات الرسمية، ووكالات الأنباء، والإنترنت، والمناهج الدراسية في مجالات علمية وفكرية مختلفة.

وتتميز بأنها تغطي حقبة زمنية طويلة نسبيا تمتد من العصر الجاهلي وحتى عصرنا هذا. وتضم حاليا ما يزيد عن 700 مليون كلمة. وعلى الرغم من أنها تعدّ من أهم المدونات العربية الموجودة وأكبرها، فإنها تصنف كمدونة لغوية عامة ولا تركز على نوع معين من النصوص (الثبيتيAl-Thubaity ، 2015). وهناك أيضا مدونة arabiCorpus (انظر: Parkinson باركنسون، 2012).

وهنالك مدونات لغوية عربية صوتية، ومن أهم المشروعات من هذا القبيل قاعدة بيانات الصوتيات العربية KACST التي أنشأها (الغامدي Alghmadi ، 2003)، وأصدرتها مدينة الملك عبد العزيز للعلوم والتقنية، إذ تضم تفاصيل دقيقة عن نطق الأصوات في اللغة العربية من حيث مخارج الحروف.

ومن المهم ذكر موقع CLARIN الذي يجمع أهم المدونات المتوازية لمختلف اللغات، مع إمكانية تحميلها إذا كانت مجانية. ومن بين المدونات المتوازية، مدونة TED-Parallel-Corpus التي أنشأها Ajinkya Kulkarni وجمع فيها الخطابات المقدمة في فعاليات Ted مع ترجماتها. ويهدف المشروع إلى تطوير معالجة برامج محاذاة الجملة وأنظمة الترجمة الآلية. وتمت جميع عمليات المعالجة ألكترونيا بدون أي تصحيحات يدوية. وتشمل مجموعة النصوص المتوازية متعددة اللغات: اللغة العربية، والصينية، والهولندية، والفرنسية، والألمانية، والعبرية، والإيطالية، واليابانية، والكورية، والروسية، والإسبانية.

3. 1 .مدونات ومنصات جامعة ليدز

قام عدد من الباحثين في جامعة ليدز البريطانية ببناء العديد من المدونات اللغوية باللغة العربية ليتسنى لهم دراسة مختلف العناصر اللغوية والبحث في خصائص اللغة العربية. ويعدّ مشروع مدونة تعليم العربية بواسطة الحاسب الآلي (Arabic by Computer ABC) أول أبحاث المدونات اللغوية في جامعة ليدز، ويهدف إلى إنشاء مصدر يحتوي على قاعدة بيانات للنصوص العربية، ومعجم لمتعلمي اللغة العربية. ولكن اقتصر عرض المدونة على الحواسيب الآلية المشغلة بأنظمة (بروكيت وآخرون Brockett et. al.، 1989).

ومن أهم المشروعات الحاسوبية في اللغة العربية والقرآن الكريم في جامعة ليدز، حسب شرف وآخرون (Sharaf et al., 2010)، موقع المدونة العربية لنصوص القرآن الكريم (The Quranic Arabic Corpus) الذي يُعدّ مصدرا لُغويا مفتوحا ومجانيا، يشرح معاني كلمات القرآن، بالإضافة إلى كونه موسوما بمعلومات لغوية تفصيلية لكل كلمة، ويشمل أيضا ترجمة معاني القرآن الكريم إلى اللغة الإنجليزية (ديوكس وآخرون Dukes et. al.،2013). ومن المدونات الأخرى التي صدرت عن جامعة ليدز مدونة الترابط الدلالي بين آيات القرآن الكريم (QurSim: Quran Verse Similarity Corpus)، وتمثل إضافة جديدة في مجال التحشية اللغوية للنص القرآني (شرف وأتويل Sharaf & Atwell، 2012b)، إضافة إلى مدونة الإحالة الثنائية لضمائر القرآن الكريم (Quran Pronoun Anaphoric Co-Reference Corpus: QurAna) (شرف وأتويل Sharaf & Atwell، 2012a)، وهي أول مدونة للغة العربية الفصحى يمكن تحميلها مجانًا، وقد قام الباحثون بوسم ما يزيد عن ٢٤,٥٠٠ ضميرا.

وأصدر العديد من الباحثين والمختصين في الجامعة مدونات أخرى عربية، كالمدونة اللغوية للعربية المعاصرة (CCA: the Corpus of Contemporary Arabic) التي صممت لتقابل مدونة الإنجليزية البريطانية المعاصرة (Lancaster-Oslo- Bergen Corpus (LOB)) ومدونة الإنجليزية الأمريكية المعاصرة (Brown Corpus)، وكلتاهما بحجم مليون كلمة (ليتش وآخرون Leech et. al. ، 1983 a) (السليطي وأتويل Al-Sulaiti & Atwell ، 2006).

واعتمد شاروف (Sharoff, 2006) منهجية استخدام النصوص وجمعها من الإنترنت كمدونة لغوية لبناء مدونات ضخمة للغات مختلفة منها الروسية، والعربية، والصينية، وإتاحتها للجميع من خلال واجهة ألكترونية مع إضافة خاصية البحث بالسياقات والمتصاحبات اللفظية (Arabic Internet Corpus). ثم أضاف صوالحة وأتويل (Sawalha & Atwell، 2013) Lemmatization إلى كلمات المدونة لاحقا.

وأنشأت السيف وماركرت (Al-Saif & Markert, 2010) البنك الشجري للخطاب العربي، ويشمل 537 نصا أخباريا بالإضافة إلى تطوير أداة للوسم، وإنشاء موقع على الويب لنشر المدونة. وقد طور صوالحة وآخرون (Sawalha et. al. 2013) أداة جديدة لتحليل الوحدات الصرفية الصغرى في اللغة العربية مع وسمها Arabic Corpus Part-of-Speech Tagging and Morphological Analysis، وأطلق عليها اسم SALMA.

وكما بين أتويل (Atwell, 2018)، فقد جُمعت مدونة اللغة العربية حول العالم (World Wide Arabic Corpus) لدراسة اختلاف لهجات اللغة العربية بين البلدان، وتشمل عدة مدونات فرعية من كل دولة، تحتوي كل منها على مئتي ألف كلمة. واستخدم الويب لجمع المدونة اللغوية لتدريس معلومات عن الإسلام (Corpus for Teaching about Islam)، وكان الهدف من هذه المدونة التخصصية هو تأليف موسوعة جامعية تستخدم في تدريس معلومات حول الإسلام والمسلمين.

كما أنشئت المدونة اللغوية لمتعلمي اللغة العربية (Arabic Learner Corpus: ALC) لتكون مصدرا لُغويا لأبحاث تعلم اللغة العربية وتعليمها، ومجالا لمعالجة اللغة الطبيعية (الفيفي وأتويل Alfaifi & Atwell ، 2013). وتضم المدونة نصوصا مكتوبة ومنطوقة شارك بها متعلمو اللغة العربية في المملكة العربية السعودية خلال العامين 2012 و2013، وتضم 1585 نصًّا (282,732 كلمة)، شارك في كتابتها 942 طالبا من 67جنسية، وبـ66 لغة.

وقد تعاونت جامعة ليدز مع جامعة الملك سعود لإنشاء مدونة جامعة الملك سعود للغة العربية الفصحى (KSUCCA King Saud University Corpus of Classical Arabic)، التي تضم 50 مليون كلمة من اللغة العربية الفصحى القريبة من فترة نزول القرآن الكريم، وتعدّ المدونة مدخلا للدراسات اللغوية التاريخية القائمة على المدونات (الربيعة وآخرون Alrabiah et.al.،2013).

هذا وقد أنشأ باحثون في قسم الترجمة بجامعة ليدز بعض المنصات لدراسة العناصر اللغوية العربية، كمقارنة التكرارات، والكشاف السياقي، (Collocation) (شاروف Sharoff، 2006)، ومنصة (Intelitext) التي قام ببنائها ولسن وآخرون (Wilson et. al., 2010) وتستخدم هاتان المنصتان في التدريس والبحث في جامعة ليدز.

2.3. المدونات المتوازية

تختلف المدونات المتوازية (Parallel Corpora) عن المدونات المتقابلة (Comparable Corpora) في الطريقة التي يتم من خلالها عرض النصوص. ففي المدونات المتقابلة، لا تكون النصوص نتيجة ترجمة سابقة، وغالبا ما تكون الترجمة آلية، وتكون النصوص منفصلة في اللغات المصدر والهدف. أما في المدونات المتوازية، فيتم توازي النصوص في اللغة المصدر مع ترجماتها في اللغة الهدف باستخدام برامج المحاذاة التي تقوم بتحديد الترجمات المتوافقة لكل جزء من النصوص. ويعتمد توازي النصوص على نتائج الترجمة الفعلية، وهذا يسمح بتحليل العلاقة بين النصوص في اللغتين واستخدامها في تطوير نظم الترجمة الآلية والأبحاث اللغوية الآلية (المجيول، 2015).

يُعد البحث اللغوي الآلي في المدونة الحاسوبية واللغة العربية مجالًا جديدًا ومتطورًا. ومع أن هذا النوع من المدونات لا يزال في بدايته، فإنه يحمل إمكانيات كبيرة لتطوير تقنيات الترجمة الآلية والتحليل اللغوي في اللغة العربية.

ومن أول مشروعات المدونات المتوازية العربية ما قام به (العجمي Al-Ajmi, 2004) حين جمع نصوصا إنجليزية عامة وحصل على ترجماتها باللغة العربية من المجلس الوطني الكويتي للثقافة والفنون والآداب. لكن هذه المدونة متاحة لجامعة الكويت فقط. ومؤخرا لاقت المدونات العربية المتوازية رواجا في عدة مجالات كاللغويات التطبيقية، وتعليم اللغات، وصارت لها أهمية كبيرة خاصة في مجال بحوث الترجمة وتدريسها وتدريب المترجمين. ويرى العديد من الباحثين أن لهذا النوع من المدونات أهمية كبرى حيث تعدّ مصدرا نفيسا في مجال تعليم اللغات.

وتعدّ مدونة الأمم المتحدة المتوازية المعروفة باسم"المتن المتوازي لوثائق الأمم المتحدة" أول مدونة متوازية لوثائق الأمم المتحدة والوثائق الدولية، وتضم نصوصا مترجمة باللغات الرسمية للأمم المتحدة، منها العربية. وقد كتبت هذه النصوص وترجمت خلال الفترة ما بين (1990- 2014) (زمنسكي وآخرون Ziemski et. al.، 2016).

ومن المدونات اللغوية المتوازية إيبكاونت (EAPCOUNT) الخاصة بنصوص وتقارير منظمة الأمم المتحدة، وتحتوي على 261 نصا باللغة الإنجليزية مع ترجمتها باللغة العربية (انظر صالحي Salhi، 2013). وهناك أيضا مدونة OPUS، وهي مدونة متعددة اللغات تم جمع نصوصها من النصوص المتوفرة على الإنترنت بطريقة آلية، لكنها لم تراجَعْ يدويا، ولذلك تعد أقل دقة (تايدمان Tiedemann ، 2012). ومن بين المدونات المتوازية المتعددة اللغات، مدونة EuroMatrix التي أنشأها الاتحاد الأوروبي، وتضم مجموعة كبيرة من اللغات، من بينها العربية. تحتوي المدونة بشكل عام على 51 مليون كلمة، بما في ذلك مليون ونصف المليون كلمة عربية، وتهدف إلى دعم البحث في مجال الترجمة الآلية. ومن المدونات أيضا مشروع مدونة متوازية متعددة اللغات في ألمانيا تحت اسم MultiUN (انظر إيزيل وتشي Eisele & Chen ، 2010).

وتحتوي مدونة µtopia المتوازية التي أنشأها لينق وآخرون (Ling et. al., 2013) على مجموعة من التغريدات والمدونات باللغات التالية: الإنجليزية-الصينية، الإنجليزية-العربية، الإنجليزية-الروسية، الإنجليزية-الكورية، والإنجليزية-اليابانية.

وبالرغم من أهمية وجود مدونات متوازية باللغة العربية يمكن استثمارها في مجالات متعددة كتدريب المترجمين، فإنه لا زالت هناك حاجة لبناء هذا النوع من المدونات لإثراء المصادر اللغوية العربية. ولخصت العتيبي (2016) بعض التحديات التي يواجهها مصممو المدونات العربية عموما والمتوازية خصوصا ومنها: أن حروف اللغة العربية مترابطة فيما بينها وتستخدم النقاط وعلامات التشكيل وغيرها من العناصر الخارجية، وهذا ممّا يزيد صعوبة معالجة اللغة العربية مقارنة باللغات الأخرى إضافة إلى وجود قيود كحقوق النشر، وقلة التمويل المادي والدعم المعنوي.

واستعرضت العتيبي (2016) ملخصا لأكثر المدونات المتوازية تداولا، وأشارت إلى أن اتحاد البيانات اللغوية في بنسلفانيا يتبنى العديد من المشروعات التي من بينها مدونات باللغة العربية، لكنها غير متاحة للجميع وتبلغ قيمة الاشتراك فيها قيمة عالية. وقد أنشئت أول مدونة متوازية لعدة لهجات عربية مع ترجمتها باللغة الانجليزية. وشملت اللهجات (العربية الفصحى، واللهجات المصرية، والتونسية، والأردنية، والفلسطينية والسورية) وتضم الترجمات ل2000 جملة (بوعمر وآخرون Bouamor et.al., 2014).

قام إزفيني (Izwaini, 2003) بإنشاء مدونات متخصصة متوازية تخدم دراسات الترجمة في مجال تقنية المعلومات، وتضم نصوصا باللغة الإنجليزية مع ترجمتها إلى اللغتين العربية السويدية. وقد تم تجميع النصوص من مواقع ألكترونية

مختصة بالحاسب، ومن الكتب، والدوريات العلمية، مع إضافة بعض أدلة الاستخدام المتوفرة لأنظمة التشغيل. وطورت العتيبي (2016) مدونة AEPC المتوازية (عربية/إنجليزية) التابعة لكلية اللغات والترجمة في جامعة الملك سعود، وتُستخدَم مصدرا لتعليم اللغة وللمساهمة في تدريب المترجمين وإثراء المحتوى العربي. كما قام عبدالعالي وآخرون (Abdelali et. al, 2014) بتطوير مدونة AMARA Corpus لإثراء المحتوى التعليمي، تحت إشراف معهد قطر لبحوث الحوسبة، إلى منصة QCRI Educational Domain Corpus وذلك لكتابة المحتوى التعليمي المرئي والمحاضرات وترجمتها من عدة منصات تعليمية، مثل أكاديمية خان و تيد TED.

ومن المشاريع في مجال تطوير نظم الترجمة الآلية، ما قام به الباحثان القحطاني وتيهان (Alkahtani & Teahan 2015) حين أنشأ مدونة متوازية تحتوي على نصوص عربية مختلفة مترجمة إلى اللغة الإنجليزية، وذلك من مدونتين: مدونة جريدة الحياة، والمدونة المتوازية المفتوحة OPUS، وتهدف إلى توفير بيانات ودعم الأبحاث في هذا المجال, وتحتوي المدونة على 27.8 مليون كلمة باللغة العربية، و 30.8 مليون كلمة باللغة الإنجليزية.

وهناك المدونة المتوازية للمقالات الصحفية للغتين العربية واليابانية، وهي من أوائل المدونات المتوازية، وتضم مقالات صحفية مع ترجماتها، وكان جمعها من قبل جامعة طوكيو للدراسات الأجنبية، وهي متاحة للجميع بصيغة XML على مستوى الوثيقة، وصيغة plain text على مستوى الجملة (Inoue et. al. 2018).

وهناك أيضا المدونة المتوازية للحديث النبوي، وهي من أحدث المدونات المتوازية في جامعة ليدز، إذ جُمِعت فيها الأحاديث النبوية باللغة العربية من كتب الأحاديث الستة مع ترجماتها الإنجليزية، وذلك من قبل التمامي وآخرون (Altammami et. al.,2020). ومشروع حسان وأتويل (Hassan & Atwell,2016) الذي يعدّ من أهم مشاريع تصميم المدونات المتوازية العربية. فقد صممما مدونة الأحاديث النبوية الشريفة المتوازية ومتعددة اللغات، وتضم نصوص الأحاديث الواردة في صحيح البخاري مع ترجماتها باللغات: الإنجليزية، والفرنسية، والروسية.

من جهته، قدّم رياش وحيدر (Rayyash & Haider ،2022) مدونة لغوية جديدة لترجمة الأفلام، وتتألف من 1،254،278 كلمة من الإنجليزية إلى العربية (EAMSC). وتتضمن نصوص أفلام إنجليزية تم اختيارها بناءً على تقييمات عالية على الإنترنت (IMDB)، مع ترجماتها المستخرجة من Netflix و (OSN) Orbit Showtime Network. ويمكن استخدام هذه المدونة في تدريس اللغة وتدريب المترجمين، وفي أبحاث الترجمة السمعية البصرية، مع بيان الاستراتيجيات المستخدمة في ترجمة هذه الأفلام.

ومن بين المشاريع الحديثة، أنشأ حمدي وآخرون (Hamdy et. al. 2020) مدونة متوازية للتغريدات العربية /الإنجليزية. حيث جمعوا تغريدات لمغردين ينشرون المحتوى ذاته باللغتين العربية والإنجليزية، بالاضافة إلى قائمة بحسابات تويتر Twitter التي تنشر التغريدات باللغتين العربية والإنجليزية، وتعد هذه المدونة مصدرا مهما لتطوير أنظمة الترجمة الآلية.

وفي المجال القانوني، تم تطوير قاعدة بيانات أطلق عليها اسم مجموعة الوثائق القانونية العربية (CALD)، وذلك من قبل مشروع (ILM) ERC-AdG-project Islamic Law Materialized، وتحوي أداة لدراسة ومقارنة الوثائق العربية القانونية في الإسلام من مختلف مناطق العالم الإسلامي، ومن القرن السابع حتى القرن السادس عشر. ويهدف المشروع إلى تسهيل دراسة الشريعة الإسلامية من منظور تاريخي. ويعدّ أداة وواجهة مفيدة للتعليم والبحث في مجال الترجمة القانونية بين العربية والإنجليزية (مولر Müller، 2021).

3. 3. الفجوة البحثية

في الجزء السابق، استعرضنا أهم الدراسات والمشروعات الخاصة بالمدونات العربية. وقد تبيّن من خلاله أنه بالرغم من وجود العديد من المدونات التي يختص معظمها بلغة الصحافة أو القرآن الكريم أو تعليم اللغة العربية، فإن أغلبها أحادي اللغة أو غير متاح، أو هو متاح مقابل اشتراك باهظ التكلفة. وعلى حدّ علم الباحثات، فإن المدونات العربية المتخصصة في مجال القانون قليلة جدا، وهنالك حاجة لجمع وإنشاء مدونات أحادية اللغة ومتوازية في هذا المجال. لذلك كان لابد من العمل على مشروع جمع وبناء مدونات لغوية قانونية باللغة العربية. وتتميز هذه المدونات عن غيرها بأنها تقوم على جميع النصوص المتاحة من دساتير الدول العربية وترجماتها باللغة الإنجليزية، وعند بنائها أخذنا بعين الاعتبار الكثير من التخصصات ومجالات الاستخدام. كما قمنا في المدونة المتوازية بمحاذاة النصوص على مستوى المادة بحيث تُعرض كل مادة باللغة العربية وما يوازيها باللغة الإنجليزية، كما تتيح منصة إسكتش إنجن الفرصة لمحاذاة المواد بالجمل أيضا.

4. منهجية البحث ومناقشة الأسئلة البحثية

يتبنى هذا البحث منهجية تطبيقية بينية لبناء المدونات القانونية، تمكن الباحثين من استخدام إجراءاتها وأدواتها في مشروعات وأبحاث مستقبلية تشمل مجالات عديدة. وفيما يلي عرْض مفصّل لمراحل بناء المدونات القانونية، ولأنواع المدونات التي قمنا ببنائها، مع مناقشة، من خلال هذه المنهجية، لتساؤلات البحث الرئيسة، من طرق وتحديات وحلول.

4. 1. إجراءات بناء المدونات اللغوية

من أجل بناء المدونات اللغوية لدساتير الدول العربية، قمنا بجمع كل النسخ المتاحة من دساتير 22 دولة عربية هي: الأردن، والإمارات، والبحرين، وتونس، والجزائر، وجزر القمر، وجيبوتي، والسعودية، والسودان، وسوريا، والصومال، والعراق، وعمان، وفلسطين، وقطر، والكويت، ولبنان، وليبيا، ومصر، والمغرب، وموريتانيا، واليمن. وينقسم بناء المدونات إلى ثلاث خطوات إجرائية رئيسية هي: (1) جمع النصوص؛ (2) تجهيز النصوص وتنسيقها؛ و(3) معالجة المدونات. وسنعرض هذه الخطوات ونناقشها بالتفصيل في الأقسام التالية.

4. 1.1. جمع النصوص

قمنا بجمع النسخ العربية الأصلية المتاحة من الدساتير وترجماتها الإنجليزية من شبكة الإنترنت، وبالأخص من المواقع المتخصصة في نشر دساتير الدول والمواقع الحكومية للدول العربية. على هذا النحو، اعتمدنا موقع Constitute Project الذي يوفر مجموعة من الدساتير من جميع أنحاء العالم، تم تصنيفها زمنيا. ولأن التركيز هنا كان فقط على البلدان العربية، وجدنا أن العديد من الدساتير القديمة لم تكن متوفرة في هذا الموقع. إضافة إلى ذلك، كانت بعض الإصدارات متاحة إمّا باللغة العربية أو الإنجليزية فقط. لذلك، قررنا تحديد الوثائق ذات الصلة يدويًّا من موقع Constitute Project وتنزيل جميع ملفات

الدساتير كل على حِدَة. بعد ذلك، حددنا المواقع الإلكترونية للمكتبات الحكومية بالإضافة إلى مواقع أخرى، مثل موقع المنظمة العالمية للملكية الفكرية [1](WIPO)، لاستكمال ترجمة الدساتير غير المتاحة. وبهذا تمكّنّا من جمع دستور متوازٍ واحد على الأقل لعشرين دولة، باستثناء دولتي جزر القمر وجيبوتي؛ وعثرنا على نسخ عربية لدستور جزر القمر ترجع إلى عام 2003 ولدستور جيبوتي، تعود لعام 1977، وعدة دساتير مختلفة باللغة الإنجليزية أضفناها إلى مدونة لغوية فرعية أحادية، باللغة الإنجليزية (انظر أنواع المدونات). ومن التحديات الأخرى التي واجهناها في أثناء جمع الدساتير عدم احتواء بعض مواقع المنظمات العربية على أية معلومات حول الدساتير الأصلية، أو تصنيفها بشكل خاطئ. في حين وضعت بعض مواقع الإنترنت الخاصة بالمكتبات قيودا على ملفات الدساتير من تنسيق PDF، مثل موقع مكتبة الكونجرس الأمريكي.

4. 2.1. تجهيز النصوص وتنسقها

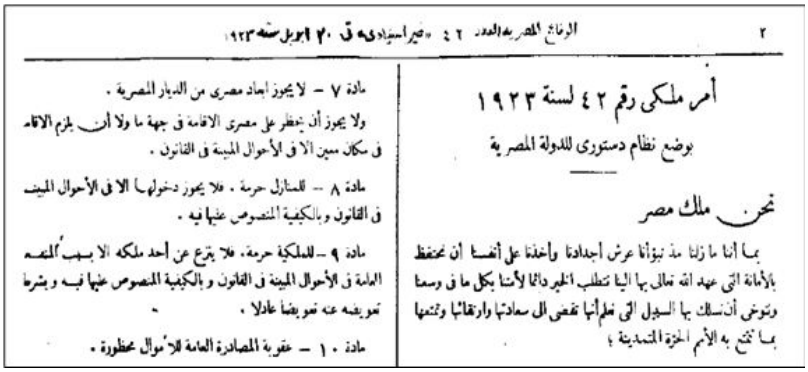
مرت مرحلة تجهيز النصوص وتنسيقها بثلاث خطوات: أولا، توحيد تنسيق الملفات. وثانيا، تنظيف البيانات وتنسيقها وتطبيعها. وثالثا، محاذاة النصوص. وللقيام بهذه الإجراءات، استخدمنا أدوات تقنية لغوية مختلفة سنعرضها في كل خطوة على حدة.

أ. توحيد تنسيق الملفات: حيث شملت أغلب الملفات المجمعة بتنسيق PDF نصوصا أو صورا، في حين تم تجميع بقية الوثائق كملفات نصية بتنسيق txt أو صفحات إنترنت بتنسيق HTML، قمنا في هذه الخطوة بتحويل كل الملفات المجمّعة إلى تنسيق ملف نصي (.txt). نظرًا إلى وجود عدد قليل من قوارئ المحارف البصرية OCR للغة العربية، ودقتها المحدودة وكفاءتها المتواضعة، فكان من الصعب اختيار أداة واحدة. لذلك، لم نعتمد أداةً بعينها. فضلًا عن ذلك، هناك مشكلات ترميز باللغة العربية. وحيث تعمل هذه الأدوات بشكل مختلف وفقًا للتنسيق المخفي لملفات PDF، استخدمنا القوارئ المجانية التالية: (1) سطور Sootor OCR[2]، وهي أداة متاحة على الإنترنت تتعرف على الحروف العربية بشكل موثوق ودقيق، لكنها تقتصر على 100 صفحة مجانية لكل مستخدم. (2) i2OCR[3] وهي أداة أخرى متاحة على الإنترنت تُستخدم لتحويل ملفات PDF، لكنها تتعرف على كل صفحة على حدة. (3) مترجم جوجل Google Translate : استخدمنا القارئة البصرية داخل تطبيق ترجمة جوجل للهاتف المحمول، واقتصر استخدام ذلك التطبيق في حال فشلت الأنظمة السابقة في التعرف على النصوص العربية.

وإلى جانب التحديات العامة التي واجهتنا من قوارئ المحارف البصرية OCR للغة العربية، احتوت بعض ملفات تنسيق PDF الخاصة بالدساتير على صور قديمة لكتابات بالآلة الكاتبة، بخط عربي قديم، كما هو موضح في الشكل 1 الذي يمثل عينة من الدستور المصري تعود لسنة 1923، في حين يوضح الشكل 1 النص المماثل للعمود الأول الذي تم التعرف عليه بقارئ سطور Sootor. علاوة على ذلك، قمنا بتوحيد تسمية الملفات لتسهيل طريقة البحث في ملفات المدونة على هذا النحو: [سنة النشر] _ [اسم الدولة] _ [عربي] أو [إنجليزي] أو [Ar_En] على سبيل المثال،" [1923_Egypt_Ar-En] للإشارة إلى ملف الدستور المصري المتوازي الذي تم نشره عام 1923.

شكل 1

جزء من ملف دستور مصر عام 1923



٢ الوقائع المصرية العدد ٤٢ «غير اعتيادي» في ٢٠ ابريل سنة ١٩٢٣

أمر ملكي رقم ٤٢ لسنة ١٩٢٣

بوضع نظام دستوري للدولة المصرية

نحن ملك مصر

بما أننا ما زلنا منذ تبوأنا عرش أجدادنا وأخذنا على أنفسنا أن نحتفظ بالأمانة التي عهد الله تعالى بها إلينا نتطلب الخير دائما لأمتنا بكل ما في وسعنا ونتوخى أن نسلك بها السبل التي نعلم أنها تفضي الى سعادتها وارتقائها وتمتعها بما تتمتع به الأمم الحرة المتمدينة ؛

مادة ٧ — لا يجوز إبعاد مصري من الديار المصرية .

ولا يجوز أن يحظر على مصري الاقامة في جهة ما ولا أن يلزم الاقامة في مكان معين الا في الأحوال المبينة في القانون .

مادة ٨ — للمنازل حرمة . فلا يجوز دخولها الا في الأحوال المبينة في القانون وبالكيفية المنصوص عليها فيه .

مادة ٩ — للملكية حرمة . فلا ينزع عن أحد ملكه الا بسبب المنفعة العامة في الأحوال المبينة في القانون وبالكيفية المنصوص عليها فيه وبشرط تعويضه عنه تعويضا عادلا .

مادة ١٠ — عقوبة المصادرة العامة للأموال محظورة .

شكل 2

نتيجة القارئ سطور Sootor للعمود الأول من الشكل 1

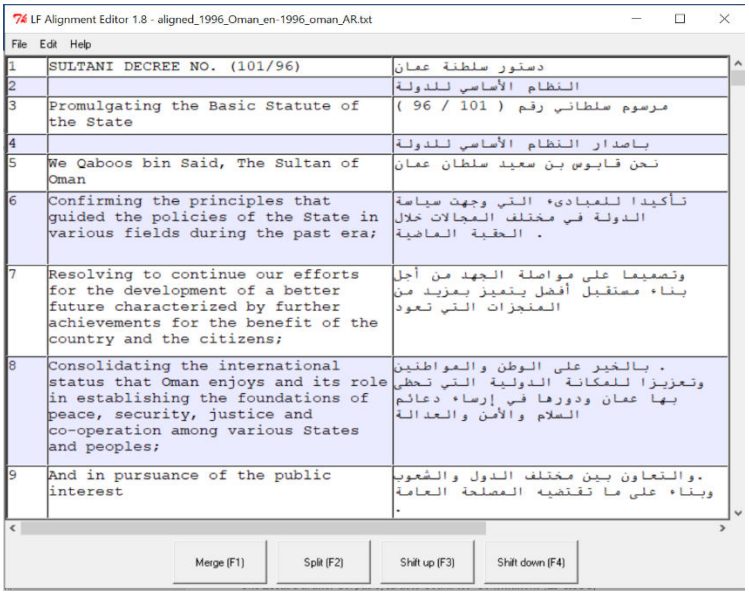
أمر ملكي رقم 43 لسنة 1923
بوضع نظام دستوري للدولة المصرية
نحن ملك مصر
بما أننا مازلنا منذ تبوأنا عرش أجدادنا وأخذنا على أنفسنا أن نحتفظ بالأمانة التي عهد الله تعالى بها إلينا نتطلب الخير دائما لأمتنا بكل ما في وسعنا ونتوخى أن نملك بها السبيل التي نعلم أنها تفضي إلى سعادتها وارتقائها وتمتعها بما تتمتع به الأمم الحرة المتمدينة.

ب. **تنظيف وتنسيق الملفات**: تتضمن هذه الخطوة تنظيف الملفات وتنسيقها وتطبيعها (مثل إزالة علامات التشكيل والرموز). ويتم ذلك باتباع إجراءات عامة لتوحيد الكتابة العربية في الملفات. على سبيل المثال، استخدمت بعض الدساتير رقمًا فقط أو شكلًا مختصرًا للإشارة إلى المادة؛ في هذه الحالة، قمنا بتغيير جميع الاختصارات والأشكال لكلمة "مادة" حتى يتم توحيد جميع الملفات على نسق واحد. وكذلك استبدلنا بكلمة "فصل" كلمة "مادة" في بعض الدساتير العربية، كدستور دولة المغرب، حيث يشار إلى كلمة "مادة" بكلمة "فصل".

ج. **محاذاة النصوص**: هي الخطوة الأخيرة في تجهيز النصوص، وتتعلق بمحاذاة المواد داخل الملفات العربية والإنجليزية. تعد المحاذاة اليدوية عملية معقدة وتستغرق وقتًا طويلا. وبذلك، فإن اعتماد تطبيق LF-aligner[4]، الذي يعمل على محاذاة الملفات المترجمة، يعطي خيارًا لمراجعة المحاذاة يدويًا قبل الانتهاء من محاذاة المواد داخل الملفات. لذا، فقد سهل هذا التطبيق عملية محاذاة شبه آلية لمواد الدساتير. وتجدرالإشارة هنا إلى أن المحاذاة تمت على مستوى المادة بالكامل وليس على مستوى الجملة. وهذا من أهم ما يميز هذه المدونة، وذلك لتوسيع استخدامها في مختلف المجالات، مثل الدراسات السياسية واللغوية.

شكل 3

محاذاة النصوص وترجماتها في تطبيق LF-aligner



4. 1. 3. معالجة المدونات

توجد كل الدساتير التي تم جمعها ثم محاذاتها في أوراق إكسل excel sheets، وتمّ حفظها في شكل ملفات نصية. واستخدمنا منصة إسكتش إنجن Sketch Engine لاختبار الدساتير ومعالجتها. وتتميز هذه المنصة بتوفير بعض الأدوات المفيدة على النحو التالي:

- اسكتشات الكلمات العربية: يعطي Word Sketch نظرة عامة على سلوك الكلمة، مع عرض الكلمات المتلازمة collocations في قوائم مصنفة لتحديد المتلازمات القوية والضعيفة. علاوة على ذلك، يسمح بعرض الرسم التخطيطي للكلمات تبعا لظهورها في النص مع مقارنتها بكلمات أخرى.
- التوافق العربي: يمكن أن يعرض إسكتش إنجن Sketch Engine جميع مثيلات كلمة أو عبارة كما تظهر في سياقات مختلفة في المدونات النصية العربية.
- قاموس المترادفات العربية: يقوم تلقائيًا بإنشاء قائمة من الكلمات المتشاركة في المعنى مع كلمة البحث التي يحددها المستخدم.
- قوائم الكلمات العربية: تعمل مِيزة قائمة الكلمات في منصة Sketch Engine على إنشاء قائمة لعدد تكرارات الكلمات في نص أو مدونة، وهناك أيضا بعض الخيارات التي تمكّن من الحصول على عدد التكرارات للتراكيب النحوية أو أنواع الكلمات.
- N-grams العربية: يمكن من تحديد الأنماط المتعلقة بالوحدات متعددة الكلمات (MWU) في اللغة العربية (كيلغاريف وآخرون، 2004، 2014).

2.4. أنواع المدونات

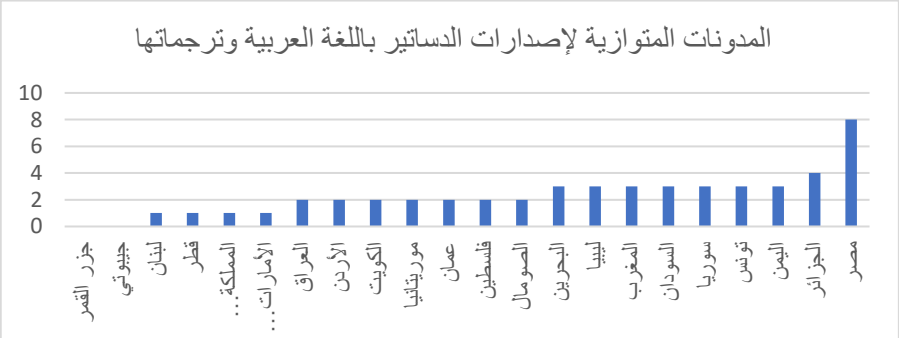
تتمثّل إسهامات البحث في بناء المدونات التالية وجعلها متاحة على منصة إسكتش إنجن Sketch Engine :

4. 2. 1. المدونة المتوازية الكاملة لدساتير الدول العربية

تشمل النسخة النهائية من هذه المدونة دساتير 22 دولة عربية مع ترجماتها الإنجليزية، ما عدا دولتي جزر القمر والصومال، في الفترة ما بين عام 1922 حتى عام 2022 . كما ذكرنا سابقًا، لا توجد نسخ متوازية لدساتير جزر القمر والصومال، فلم يتسنَّ إضافتها إلى هذه النسخة الكاملة. ويوضح الشكل 4 عدد الدساتير المتوازية في كل بلد، ويبين أن مصر لها أكبر عدد من الدساتير في المدونة المتوازية، بثمانية ملفات متوازية. في الوقت نفسه، فإن كلًّا من الإمارات العربية المتحدة، ولبنان، وقطر، والمملكة العربية السعودية ممثلة بدستور واحد موازٍ. وأما الدول الخمس عشرة الأخرى فممثلة بمعدّل ثلاث دساتير متوازية في كل دستور. وتم إضافة جدولين (جدول 1 وجدول 2) لإيضاح تفاصيل الملفات وأنواعها وأعداد الكلمات في كل من الدساتير بحسب الدولة وتاريخ الدستور. وقد تم جمع 51 من الدساتير العربية ومقابلاتها الإنجليزية، وبلغ عدد الجمل في الملفات العربية 17,967 جملة، في حين احتوت النسخة الإنجليزية على 22,569 جملة. وتم فصل ديباجات الدساتير عن المدونات الأساسية وجمعها في مدونة لمقدمات (أو ديباجات) الدساتير العربية، ومجموع كلماتها (32.660 كلمة) وهي مفصله كالآتي (12,360) كلمة باللغة العربية و (20,300) كلمة باللغة الإنجليزية.

شكل 4

توزيع الدساتير المتوازية لإصدارات الدساتير باللغة العربية وترجماتها بحسب الدولة



ملفات الدساتير العربية وعدد الكلمات وأنواع الملفات

م.	الدول	نوع الملفات	عدد الكلمات
1	1923_Egypt_AR.txt	txt	5,692
2	1964_Egypt_AR.txt	txt	4,964
3	1971_Egypt_AR.txt	txt	8,763
4	2007_Egypt_AR.txt	txt	3,996
5	2011_Egypt_AR.txt	txt	13,270
6	2013_Egypt_AR.txt	txt	16,023
7	2014_Egypt_AR.txt	txt	16,015
8	2019_Egypt_AR.txt	txt	16,700
9	1963_Algeria_AR.txt	txt	2,064

7,134	txt	1989_Algeria_AR.txt	10
7,129	txt	2008_Algeria_AR.txt	11
10,159	txt	2016_Algeria_AR.txt	12
8,762	txt	1953_Syria_AR.txt	13
5,544	txt	1973_Syria_AR.txt	14
5,940	txt	2012_Syria_AR.txt	15
6,685	txt	1973_Bahrain_AR.txt	16
8,228	txt	2002_Bahrain_AR.txt	17
8,252	txt	2018_Bahrain_AR.txt	18
7,582	txt	1951_Libya_AR.txt	19
2,297	txt	2012_Libya_AR.txt	20
13,170	txt	2016_Libya_AR.txt	21
9,977	txt	1998_Sudan_AR.txt	22
21,178	txt	2005_Sudan_AR.txt	23
6,363	txt	2019_Sudan_AR.txt	24
3,944	txt	1992_Morocco_AR.txt	25
5,110	txt	1996_Morocco_AR.txt	26
11,666	txt	2011_Morocco_AR.txt	27
4,200	txt	1959_Tunisia_AR.txt	28
5,155	txt	2008_Tunisia_AR.txt	29
9,029	txt	2014_Tunisia_AR.txt	30
6,050	txt	2003_Palestine_AR.txt	31
6,525	txt	2005_Palestine_AR.txt	32
6,301	txt	1925_Iraq_AR.txt	33
8,652	txt	2005_Iraq_AR.txt	34
6,436	txt	1962_kuwait_AR.txt	35
6,425	txt	1992_kuwait_AR.txt	36
7,823	txt	2011_Jordan_AR.txt	37
7,979	txt	2016_Jordan_AR.txt	38
6,842	txt	1960_Somalia_AR.txt	39
16,974	txt	2012_Somalia_AR.txt	40
5,205	xls	1991_Mauritania_AR_EN.xlsx	41
5,204	txt	2012_Mauritania_AR.txt	42
3,833	txt	1996_Oman_AR.txt	43
4,884	txt	2021_Oman_AR.txt	44
9,744	txt	2001_Yeman_AR.txt	45
9,744	txt	2015_Yeman_AR.txt	46
8,198	txt	1978_Yamen_AR.txt	47
8,822	txt	2009_UAE_AR.txt	48
5,166	txt	2004_Lebanon_AR.txt	49
5,608	txt	2004_Qatar_AR.txt	50
2,582	txt	1992_KSA_AR.txt	51

جدول 2

ملفات الدساتير الإنجليزية وعدد الكلمات وأنواع الملفات

م.	الدول	نوع الملفات	عدد الكلمات
1	1923_Egypt_EN.txt	txt	6,340
2	1964_Egypt_EN.txt	txt	6,281
3	1971_Egypt_EN.txt	txt	10,149
4	2007_Egypt_EN.txt	txt	4,800
5	2011_Egypt_EN.txt	txt	15,008
6	2013_Egypt_EN.txt	txt	18,921
7	2014_Egypt_EN.txt	txt	18,880
8	2019_Egypt_EN.txt	txt	19,649
9	1963_Algeria_EN.txt	txt	2,747
10	1989_Algeria_EN.txt	txt	8,747
11	2008_Algeria_EN.txt	txt	10,364
12	2016_algeria_EN.txt	txt	12,915
13	1953_Syria_EN.txt	txt	11,804
14	1973_Syria_EN.txt	txt	6,350
15	2012_Syria_EN.txt	txt	7,395
16	1973_Bahrain_EN.txt	txt	8,403
17	2002_Bahrain_EN.txt	txt	9,892
18	2018_Bahrain_EN.txt	txt	10,195
19	1951_Libya_EN.txt	txt	8,129
20	2012_Libya_EN.txt	txt	2,743
21	2016_Libya_EN.txt	txt	15,668
22	1998_Sudan_EN.txt	txt	12,242
23	2005_Sudan_EN.txt	txt	23,975
24	2019_Sudan_EN.txt	txt	7,419
25	1992_Morocco_EN.txt	txt	4,454
26	1996_Morocco_EN.txt	txt	5,803
27	2011_Morocco_EN.txt	txt	14,887
28	1959_Tunisia_EN.txt	txt	4,200
29	2008_Tunisia_EN.txt	txt	6,006
30	2014_Tunisia_EN.txt	txt	6,287
31	2003_Palestine_EN.txt	txt	7,551
32	2005_Palestine_EN.txt	txt	8,055
33	1925_Iraq_EN.txt	txt	6,774
34	2005_Iraq_EN.txt	txt	10,786
35	1962_Kuwait_EN.txt	txt	8,046
36	1992_Kuwait_EN.txt	txt	7,873
37	2011_Jordan_EN.txt	txt	9,896
38	2016_Jordan_EN.txt	txt	10,790
39	1960_Somalia_EN.txt	txt	8,416
40	2012_Somalia_EN.txt	txt	20,335
41	1991_Mauritania_EN.xlsx	xls	6,872

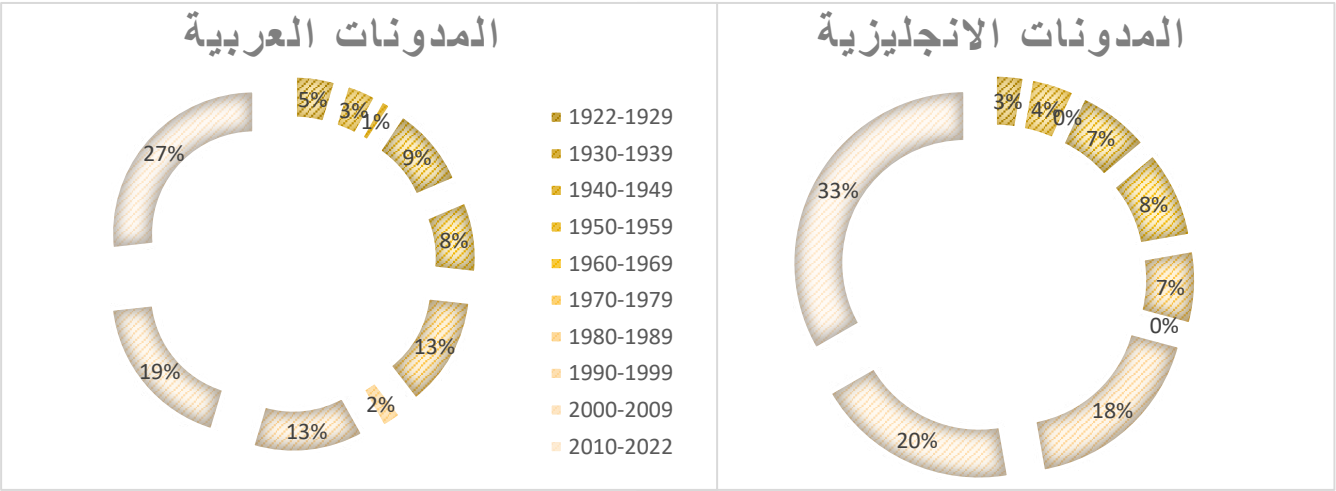
42	2012_Mauritania_EN.txt	txt	7,111
43	1996_Oman_EN.txt	txt	4,507
44	2021_Oman_EN.txt	txt	5,747
45	2001_Yeman_EN.txt	txt	11,673
46	2015_Yeman_EN.txt	txt	11,673
47	1978_Yamen_EN.txt	txt	10,461
48	2009_UAE_EN.txt	txt	10,246
49	2004_Lebanon_EN.txt	txt	6,171
50	2004_Qater_EN.txt	txt	7,188
51	1992_ksa_EN.txt	txt	2,901

4. 2. 2. المدونات الأحادية العربية أو الإنجليزية

كانت بعض الدساتير المنشورة لبعض الدول إمّا باللغة العربية أو باللغة الإنجليزية فقط، وفي بعض الدول الأخرى، جاءت الدساتير باللغتين الفرنسية والعربية، ولا توجد ترجمة إنجليزية لهذه الإصدارات. لذلك قمنا بإنشاء مدونتين منفصلتين لإصدارات الدساتير باللغة العربية واللغة الإنجليزية، كل على حدة (انظر الشكل 4)، وكان العدد الإجمالي لكلمات مدونة اللغة العربية 788,477 كلمة، ولكلمات مدونة اللغة الانجليزية 343,582 كلمة. ويوضح الشكل 5 مقارنة بين أعداد الدساتير الصادرة بكل لغة على حدة، خلال الفترة من عام 1922 وحتى عام 2022. وفي السنوات الأولى، كانت إصدارات الدساتير المتاحة باللغة العربية فقط دون الترجمات الإنجليزية أكثر من النسخ المتاحة بتلك الترجمات. ومنذ سنة 1990، صارت إصدارات الدساتير باللغة الإنجليزية متاحة على الإنترنت، مما أتاح الوصول إليها بسهولة وضمّها إلى المدونة الإنجليزية الأحادية.

شكل 5

المدونات الأحادية لإصدارات الدساتير باللغة العربية واللغة الإنجليزية



4. 2. 3. مدونة أحادية للدساتير باللغة الإنجليزية (مدونة مقارنة)

بعد إنشاء المدونات السابقة الفرعية الأحادية، أنشأنا مدونة مقارنة من النسخ المنقحة المحدثة من دساتير 8 دول لغتها الإنجليزية هي اللغة الرسمية وتضم: أستراليا وكندا وأيرلندا ونيوزيلندا وسنغافورة وجنوب إفريقيا والولايات المتحدة الأمريكية والمملكة المتحدة. وكانت أقدم نسخة في هذه المدونة لأستراليا في عام1985، وأحدث نسخة منقحة كانت للولايات المتحدة الأمريكية في عام 2016. وتتألف هذه المدونة من 677,056 كلمة و557,086 كلمة.

5. الجهات المستفيدة والمشروعات المستقبلية

تناول البحث بناء المدونات القانونية المتوازية وأحادية اللغة لدساتير البلدان العربية ومدونة لدساتير الدول لغتها هي الإنجليزية. وناقش البحث منهجية بناء هذه المدونات والتحديات التي واجهتها الباحثات خلال بنائها والحلول المقترحة للتغلب عليها.

تقدم المدونات المستدامة فوائد مختلفة للعديد من الجهات والمنظمات التعليمية وغير التعليمية. فهي تعدّ أولا أداة تربوية لطلاب القانون، والترجمة التحريرية والترجمة الفورية وطلاب الدراسات العليا المسجلين في برامج الترجمة العربية والإنجليزية في شتى جامعات العالم. وتمد هذه المدونات الباحثين في مجالات الترجمة عامة والترجمة القانونية خاصة بأدوات بحث مهمة للدارسين والأكاديميين والعاملين في مجال القانون في جميع أنحاء العالم. ويمكن أن تُستخدم هذه المدونات كأداة لبناء خرائط للمصطلحات المتخصصة والمفاهيم القانونية، أو لإجراء دراسات تجريبية. فعلى سبيل المثال لا الحصر، يمكن استخدامها في دراسة مقارنة للمصطلحات والمفاهيم الخاصة بحقوق الإنسان في الدول العربية. ونظرا إلى قيامنا بمحاذاة المدونات على مستوى المادة وليس على مستوى الجملة، بالإضافة إلى بناء مدونة فرعية خاصة بالديباجات التي توضح السياق الثقافي والمجتمعي والسياسي، سيتيح مشروعنا البحثي الفرصة للمستخدمين من تخصصات علمية أخرى، كالباحثين والأكاديميين في مجال السياسة والدبلوماسية والقانون الدولي. والمدونات حاليا متاحة على GitHub، وستكون متاحة للباحثين على منصة إسكتش إنجن Sketch Engine قريبا.

الهوامش

[1] wipolex.wipo.int/ar/main/legislation
[2] sotoor.ai/en/home
[3] www.i2ocr.com/free-online-arabic-ocr

[4] https://sourceforge.net/projects/aligner/

المراجع العربية

المجيول ، سلطان . (2015). البحث اللغوي في المدونات العربية الحاسوبية بين الممكن والمحتمل والمأمول. في صالح العصيمي (محرر)، *المدونات اللغوية العربية: بناؤها وطرق الإفادة منها*. (ص ص 236-279). مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

العتيبي، هند. (2016). المعالجة الآلية للمدونات المتوازية واستخداماتها في تعليم اللغات وتدريب المترجمين. في سلطان المجيول (محرر.)، *لغويات المدونة الحاسوبية: تطبيقات تحليلية على العربية الطبيعية*. (ص ص 170-196). مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، الرياض.

المراجع الأجنبية

Abdelali, A., Guzman, F., Sajjad, H., & Vogel, S. (2014). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. *LREC*,*14*, 1044-1054.

Al-Ajmi, H. (2004). A New English–Arabic Parallel Text Corpus for Lexicographic Applications. *Lexikos*. *14*(1), 326–330.

Al-Saif, A., & Markert, K. (2010, May). The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the seventh international conference on language resources and evaluation* (LREC'10).

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, *11*(2), 135-171.

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013, August). New language resources for arabic: corpus containing more than two million words and a corpus processing tool. *In 2013 International Conference on Asian Language Processing* (pp. 67-70). IEEE. doi: 10.1109/IALP.2013.21.

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. Language Resources and Evaluation, 49, 721-751. https://doi.org/10.1007/s10579-014-9284-1

Alansary, S., & Nagi, M. (2014, October). The international corpus of Arabic: Compilation, analysis and evaluation. In Proceedings *of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 8-17).

Alfaifi, A., & Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, *19*(2), 347-357.

Alghmadi, M. (2003, August). KACST Arabic phonetic database. In *the Fifteenth International Congress of Phonetics Science*, Barcelona (pp. 3109-3112).

Alkahtani, S., Liu, W., & Teahan, W. J. (2015). A new hybrid metric for verifying parallel corpora of Arabic-English. *arXiv preprint arXiv:1502.03752*. https://doi.org/10.48550/arXiv.1502.03752

Alotaibi, H.M. (2016). Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. Retrieved from http://dx.doi.org/10.2139/ssrn.3053572.

Alrabiah, M., Al-Salman, A., & Atwell, E. S. (2013). The design and construction of the 50 million words KSUCCA. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics* (pp. 5-8). The University of Leeds.

Altammami, S., Atwell, E., & Alsalka, A. (2020). The Arabic-English parallel corpus of authentic hadith. International Journal on Islamic Applications in Computer Science And Technology, 8(2).

Atwell, E. (2018). Classical and modern Arabic corpora: Genre and language change. In: Whitt, RJ, (ed.) Diachronic Corpora, Genre, and Language Change. *Studies in Corpus Linguistics*, 85. John Benjamins, pp. 65-91. ISBN 9789027201485

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In *M. Baker, G. Francis, and E. Tognini- Bonelli (Eds.), Text and technology: In honour of John Sinclair. Amsterdam and Philadelphia:* John Benjamins, pp. 233-250.

Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., & Suchomel, V. (2013). ArTenTen: a new, vast corpus for Arabic. *Proceedings of WACL*, 20.

Biel, Ł. (2014). The textual fit of translated EU law: a corpus-based study of deontic modality. *The Translator*, 20(3), 332-355.

Bouamor, H., Habash, N., & Oflazer, K. (2014, May). A Multidialectal Parallel Corpus of Arabic. In *LREC* (pp. 1240-1245).

Brierley, C., & El-Farahaty, H. (2019). An interdisciplinary corpus-based analysis of the translation of كرامة (karāma, 'dignity') and its collocates in Arabic-English constitutions. *JoSTrans: the Journal of Specialised Translation*, (32), 121-145.

Brockett A, & Atwell, E, & Taylor, O, & Page, M. 1989. An Arabic text database and glossary system for students. *Proceedings of the Seminar on Bilingual Computing in Arabic and English*.

Dukes, K., Atwell, E. & Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. *Lang Resources & Evaluation* 47, 33–62. https://doi.org/10.1007/s10579-011-9167-7

El-Farahaty, H., & Elewa, A. (2020). A Corpus-Based Analysis of Deontic Modality of Obligation and Prohibition in Arabic/English Constitutions (Un análisis de corpus de la modalidad deóntica de obligación y prohibición en las constituciones árabes/inglesas). Estudios de Traducción, 10, 107-136.

Eisele, A., and Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA), pp. 2868–2872.

Inoue, G., Habash, N., Matsumoto, Y., & Aoyama, H. (2018, May). A parallel corpus of Arabic-Japanese news articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Izwaini, S. (2003, March). Building specialised corpora for translation studies. In *Workshop on multilingual corpora: Linguistic requirements and technical perspectives, corpus linguistics.*

Goweder, A., & De Roeck, A. (2001). *Assessment of a significant Arabic corpus. In Arabic NLP Workshop at ACL/EACL.*

Hassan, S., & Atwell, E. S. (2016). Design and implementing of multilingual Hadith corpus. *International Journal of Recent Research in Social Sciences and Humanities*, 3(2), 100-104.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 *the sketch engine. Information Technology,* 105(116), 105-116.

Kilgarriff, A., Baisa, V., Bušta, J. et al. (2014). The Sketch Engine: ten years on. Lexicography ASIALEX 1, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Leech, G., Garside, R., & Atwell, E. (1983). Recent developments in the use of computer corpora in English language reليsearch. *Transactions of the Philological Society*, 81(1), 23-40.

Ling, W., Xiang, G., Dyer, C., Black, A. W., & Trancoso, I. (2013, August). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 176-186).

Mubarak, H., Hassan, S., & Abdelali, A. (2020, May). Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora* (pp. 14-21).

Müller, C. (2021). 'Cald: A very short introduction', *The Documents of Islamic Law in History. Studies on Arabic Legal Documents*. https://dilih.hypotheses.org/763

Parkinson, D. B. (2012). ArabiCorpus. Online. https://arabicorpus.byu.edu/

Rayyash, H. A., & Haider, A. (2022). Construction of a parallel corpus of English-Arabic movie subtitles: a genuine source for audiovisual translators. *The International Journal of Humanities Education,* 21(1), 21.

Salhi, H. (2013). Investigating the complementary polysemy and the Arabic translations of the noun Destruction in EAPCOUNT. *Meta: Translators' Journal, 58(1), 227–246.*

Sawalha, M., Atwell, E., & Abushariah, M. A. (2013, February). *SALMA: standard Arabic language morphological analysis.* In 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE.

Sawalha, M and Atwell, ES. (2013). Accelerating the processing of large corpora: using Grid Computing for lemmatizing the 176 million words Arabic Internet Corpus. In: Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2. The 2nd Workshop of Arabic Corpus Linguistics WACL-2, 22-26 Jul 2013, Lancaster University, UK. UCREL.

Sharaf, A. B., & Atwell, E. (2012a, May). QurAna: Corpus of the Quran annotated with pronominal anaphora. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12) (pp. 130-137).

Sharaf, A. B., & Atwell, E. (2012b, May). QurSim: A corpus for evaluation of relatedness in short texts. n Proceedings of the Eighth International Conference on Language Resources and

Evaluation (LREC'12), (pp. 2295–2302), Istanbul, Turkey. European Language Resources Association (ELRA).

Sharaf, A. B., & Atwell, E & Kais, Dukes & Sawalha, M & Al-Saif, A & Sharoff, S & Markert, K & Al-sulaiti, L & Shawar, B & Abbas, N. (2010). المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز *"Arabic and Quranic Computational Linguistics Projects at the University of Leeds".*

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus linguistics*, *11*(4), 435-462.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Wilson, J., Hartley, A., Sharoff, S., & Stephenson, P. (2010, November). Advanced corpus solutions for humanities researchers. *In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* (pp. 769-778).

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016). The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), (pp. 3530–3534), Portorož, Slovenia. European Language Resources Association (ELRA).

AUTHORS BIODATA

Dr Hanem El-Farahaty is an Associate Professor of Arabic Language, Translation and Interpreting in the School of Languages, Cultures and Societies, and the University of Leeds. She received her PhD degree in Arabic Language, Translation and Interpreting (2011) from the University of Leeds. Her research interests include English-Arabic legal translation, corpus-based legal translation, media and political translation, political satire, and critical discourse analysis.

بيانات الباحثات

د. هانم الفرحاتي، أستاذ مشارك في (اللغويات والترجمة والترجمة الفورية) في جامعة المنصورة بجمهورية مصر العربية، وجامعة ليدز بالمملكة المتحدة. حاصلة على درجة الدكتوراه في الترجمة من جامعة ليدز عام 2011. تتركز اهتماماتها البحثية في الترجمة، والترجمة القانونية بين الإنجليزية والعربية، والمدونات القانونية، والترجمة القانونية باستخدام المدونات، والمدونات المتوازية، والترجمة الإعلامية والسياسية والهجاء السياسي، وتحليل الخطاب النقدي.

معرف أوركيد **(ORCID)** : 4497-2104-0002-0000

Email: h.el-farahaty@leeds.ac.uk

أ. أماني سيف آل عنيزان، باحثة دكتوراه في اللغويات التطبيقية والمدونات الحاسوبية، في جامعة ليدز، بالمملكة المتحدة. تتركز اهتماماتها البحثية في المدونات اللغوية العربية والمدونات اللغوية لمتعلمي اللغة الإنجليزية والكتابة الأكاديمية باللغة الإنجليزية.

معرف أوركيد (ORCID) : 6418-9228-0002-0000

Email: Amani.onayzan@windowslive.com

أ. نوران خلّاف، محاضر في تكنولوجيا اللغة، قسم الصوتيات واللسانيات، بكلية الآداب، جامعة الأسكندرية (جمهورية مصر العربية). حاصلة على درجة الدكتوراه في معالجة اللغات الطبيعية من جامعة ليدز عام2023. تتركز اهتماماتها البحثية في المدونات اللغوية، والتطبيقات اللغوية، ومعالجة اللغات الطبيعية، واللغة العربية.

معرف أوركيد (ORCID) : 0000-0002-5422-5585

Email: mlnak@leeds.ac.uk

Amani Alonayzan is a PhD candidate at the School of Languages, Cultures and Societies, University of Leeds. Her research interests include corpus linguistics, learner corpora, Arabic corpus linguistics and Academic writing.

Nouran Khallaf is a lecturer of Language Technology at the Phonetics Department, Faculty of Arts, University of Alexandria. She obtained her PhD in Arabic NLP (2023) from the University of Leeds. Her research interests include: NLP, corpus linguistics, and Arabic.