Corpus Pragmatics (2023) 7:103–119
https://doi.org/10.1007/s41701-023-00138-x

ORIGINAL PAPER

Building the Leeds Monolingual and Parallel Legal Corpora of Arabic and English Countries' Constitutions: Methods, Challenges and Solutions

Hanem El-Farahaty[1,2] · Nouran Khallaf[1,3] · Amani Alonayzan[1,3]

Received: 21 February 2023 / Accepted: 28 February 2023 / Published online: 7 April 2023

Abstract

Arabic corpora have existed since the last decade of the past century. Although they are constantly increasing, more advanced tools and morpho-syntactically annotated Arabic corpora are still needed for research and teaching. Likewise, parallel and specialised corpora are rare despite the growing need to use them in empirical linguistic investigations of authentic Arabic texts and for language and translation teaching. Therefore, building legal corpora will pave the way for more research in Arabic legal translation, an area which is under-researched worldwide. This paper aims to discuss the building of a collection of specialised parallel and monolingual legal corpora. In particular, it will discuss the building of diachronic corpora, which include all available constitutions of 22 Arabic countries. The aim of building all available versions of these constitutions is two-fold: (1) interdisciplinary corpus-based and socio-cultural investigations and (2) research-led and blended-learning pedagogical approaches to translation teaching and learning. Thus, these corpora are of great value to translation trainers and researchers, law academics and professionals, and governmental, non-governmental and international organisations. The paper will demonstrate the process of building these specialised complex corpora and the challenges encountered throughout this process. Among the challenges faced during the data collection and processing phases are (1) limitations of finding the original constitutions for each Arabic country since some of them date back to 1922; (2) file conversion and the difficulty of choosing one Optical Character Recognition (OCR) tool to rely on for the Arabic language since many lack accuracy, efficiency as well as encoding issues in Arabic.

Keywords Arabic corpora · Arabic/English parallel and monolingual legal corpora · Corpus Linguistics · Corpus-based pragmatics · Corpus-based translation

✉ Hanem El-Farahaty
h.el-farahaty@leeds.ac.uk

Extended author information available on the last page of the article

Introduction

The technology revolution in the 20th century has significantly facilitated the development and growth of corpora and computer tools in the fields of descriptive and applied linguistics. Corpus Linguistics has changed the way language is interpreted and has led to conducting of numerous empirical studies investigating every aspect of language (McCarthy & O'Keeffe, 2012: 4). For example, corpora and pragmatics (Rühlemann, 2019) is one of the important fields contributing to the analysis of legal language, and construction of legal meaning, e.g. analysis of speech acts (Austin, 1962; Searle, 1969, 1979) such as directives (Cao, 2007; Visconti, 2009 and Solum, 2021) and the analysis of modality of obligation, permission and prohibition (Biel, 2014; El-Farahaty & Elewa, 2020; Palmer, 2001).

It quickly became evident that corpus linguistics contributes to many fields such as law which depends on language heavily (Goźdź-Roszkowski, 2021:1515; Solan, 2017:1315). Therefore, corpora and corpus-based translation methodologies are increasingly becoming more integral in translation practice and research since the 1990s (see Baker et.al.,1993; Baker, 1995, 2019). These empirical methodologies uncover translation norms, universals (Olohan, 2004), characteristics and rigorous analyses of legal concepts and phraseology in original and translated texts. They focus on 'the interplay of theoretical constructs and hypotheses, variety of data, novel descriptive categories and a rigorous, flexible methodology (Kruger, 2004:1), hence facilitating translation training and enhancing translation competencies. More recent research dedicated to the use of corpora in translation and language teaching has been published (see Zaki, 2020, 2021; Zaki et al., 2021, among others).

Researchers at The University of Leeds have been active in creating open-access Arabic corpora and promoting projects in Arabic corpus linguistics (see Alfaifi & Atwel, 2016; Sharoff, 2006). The need for advanced tools and morpho-syntactically annotated Arabic corpora to be used in teaching and research is increasing (see section on Arabic Corpora: A Brief Review). Likewise, parallel corpora and specialised legal corpora are rare in spite of the growing need to use them in empirical linguistic investigations of authentic Arabic texts and for language and translation teaching. Therefore, this paper fills an existing gap in the field of Arabic corpus linguistics and legal translation. Building these corpora and sub-corpora is needed to address the possible inconsistencies in the legal drafting and translation of specific linguistic and system-bound areas and will pave the way for more research in the field of Arabic legal translation, an area which is under-researched worldwide. Examples of these research applications using the current corpora include a comparative analysis of performative verbs in the diachronic parallel corpus of Arabic constitutions. Other corpus-based pragmatic investigations include the analysis of directive speech acts, their translations in the parallel corpus of Arabic constitutions and comparative diachronic investigations of deontic modals of obligation, permission and prohibitions in the above parallel corpus.

This paper aims to discuss the building of a collection of specialised parallel and monolingual legal corpora. In particular, it will discuss the building of diachronic corpora, which include all available constitutions of 22 Arabic countries: Algeria, Bahrain, Comoros, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, the United Arab Emirates, and Yemen.

The paper answers the following overarching research question: What is the process of building diachronic corpora of Arabic countries' constitutions and what are the challenges of building these corpora? To answer this research question, we will demonstrate the process of building the specialised corpora for all available constitutions and discuss the solutions/suggestions for challenges encountered throughout this process. It will discuss: (a) researching and collecting the extensive set of data (b) pre-processing the data, i.e. corpora cleaning; alignment of the corpora and the semi-automatic verification of the articles using Lf-aligner (c) uploading and organising the corpora on Sketch Engine and testing them. Sketch Engine provides a bespoke corpus management system and offers several useful functionalities for Arabic language data processing (Kilgarriff et al., 2004; Kilgarriff, 2014). (See Method). Below is a list of the corpora we aim to build:

- ***The Leeds Parallel Corpus of Arabic Countries' Constitutions (LPCACC)***

The final version of this corpus includes the constitutions of 20 Arabic countries and their English translations, from 1922 to 2022; Arabic version (407,633 words) and English version (489,448 words). The LPCACC will include sub-corpora for each country separately but it excludes Comoros and Somalia.

- ***The Parallel Corpus of Preambles of Arabic Countries' Constitutions (PCPACC)***

This version contains all available preambles of the Arabic countries' constitutions (32.660 words).

- ***Monolingual Corpora of Arabic and English Constitutions (MCAEC)***

These corpora contain two separate versions of constitutions in Arabic (788,477 words) and in English (343,582 words).

- ***The Leeds Monolingual Corpus of English Countries' Constitutions (LMCECC) (Comparative Corpus)***

This corpus consists of the constitutions of 8 countries of which English is the official language, from 1985 to 2016. It consists of 677,056 words. (For more information, see sections on Method and Corpora)

The aims of building all available versions of Arabic constitutions and the comparative English corpus are (1) interdisciplinary corpus-based, socio-cultural investigations, comparative Arabic-English legal translation and advanced comparative linguistic analysis (cf. Biel, 2014). (2) research-led and blended-learning pedagogical approaches to translation teaching and learning. The research of building corpora starts with collecting original texts and translated texts in their various forms (e.g.txt, pdfs and images), storing, documenting, processing, aligning them automatically or manually, and uploading them to the Sketch Engine platform (Kilgarriff, 2004; 2014).

The paper is structured as follows. Section two provides a review of the current literature on Arabic corpora and Arabic legal corpora. Section three presents the detailed methodology used to build the data set and prepare it before uploading and publishing it on Sketch Engine. It will give full details about the data set used in building the corpora and sub-corpora. The challenges encountered in building the data set and the tools used to solve them will also be discussed in this section. Finally, value propositions, as well as future projects, will be highlighted in section four.

Arabic Corpora: A Brief Review

Corpus and computational linguists across the world have been working in the past three decades on building Arabic corpora (Abbas & Smaili, 2005; Al-Sulaiti & Atwell, 2006; Brierley & El-Farahaty, 2019; El-Farahaty & Elewa, 2020; El-Haj & Koulali, 2013; Goweder & De Roeck, 2001); databases (Boudelaa & Marslen-Wilson, 2010; Khwaileh et al., 2018); online interfaces (Dukes & Atwell, 2012; Sharoff, 2006) and developing NLP tools (Habash, 2010; Al-Jawfi, 2009, among others). Although they are constantly increasing (Alfaifi & Atwell, 2016), Arabic is understudied by corpus-based methodologies compared to its demographic and societal relevance (McEnery et al., 2019:1). It is worth mentioning that creating Arabic corpora comes with challenges, such as creating NLP algorithms and techniques for Arabic and developing tools specific to the language (Al-Thubaity et al., 2013). Other challenges include the ambiguity and difficulty of Arabic, the employment of several dialects of Arabic, each having distinct features and the shortage of freely accessible databases that may be utilised in the study and development of Arabic information extraction and processing (Al-Thubaity et al., 2013). Some academics started constructing electronically searchable corpora of Arabic literature not long after such intrinsic problems in Arabic computing started to be overcome.

This section will review the current literature on Arabic monolingual and parallel corpora, including those built by researchers and scholars at the University of Leeds. We will then focus on Arabic legal corpora and will end up by highlighting the research gap that this paper addresses.

Arabic Monolingual and Parallel Corpora

Corpus linguists and researchers built various Arabic corpora in different fields and genres for pedagogical or research-based purposes. For example, in 1992, the Linguistic Data Consortium (LDC) at the University of Pennsylvania, produced many corpora in twenty languages, including Arabic. Although The LDC individually licensed many of the corpora in its catalogues, users need to have a membership or have access through libraries and universities to be able to use the corpora (Cieri, et al., 2022).

One of the first projects on monolingual Arabic corpora, which contain Arabic texts only, is the Al-Hayat newspaper corpus (Goweder & De Roeck, 2001). This accessible Arabic language corpus includes 18.5 million words. After that, other web-based corpora were built using specialised technology to collect specialised linguistic content covering different genres and text types from the internet. For example, ArTenTen (Belinkov et al., 2013) is a general Arabic corpus from the family of linguistic corpora known as (TenTen corpora) that includes many languages such as English, Japanese, Russian, and Chinese, among others. The ArTenTen corpus, used as a reference corpus, contains more than 10 billion words in Arabic and is accessible via Sketch Engine.

Researchers have worked over the past decade to create Arabic corpora (Alfaifi & Atwell, 2016; Zeroual & Lakhouaja, 2018). Several Arabic corpora were mainly derived from newspapers and designed primarily for researchers' projects but could not be accessed online (Al-Thubaity et al., 2013). However, between 2005 and 2013, many Arabic corpora were created and made available on an online platform or could be downloaded, which led to a shift in the status of Arabic corpus linguistics. Examples of these include the International Corpus of Arabic (ICA) (Alansary & Nagi, 2014), and King Saud University Corpus of Classical Arabic (KSUCCA) (Alrabiah et al., 2013) and arabiCorpus (Parkinson, 2012). Significant efforts have been made at the University of Leeds to establish different types of Arabic corpora including the Quranic Arabic Corpus (Dukes et al., 2013) and Corpus of Contemporary Arabic (CCA), for teaching Arabic as a foreign language (Al-Sulaiti & Atwell, 2006) in addition to the web interface introduced by Sharoff (2006). It is also important to mention the work done by Lancaster University Centre for Computer Corpus Research on Language (e.g. El-Haj et al., 2015; McEnery et al., 2019), which significantly impacted the field. Zaghouani, (2017) presented a Critical survey discussing the freely available Arabic corpora, and more recently, Ahmed et al. (2022) founded 48 free and accessible Arabic corpora by searching the most popular information technology (IT) resources[1].

A significant project in (Arabic) Parallel corpora is the English-Arabic Political Parallel Corpus (EAPPC) by Ahmad et al. (2017). It is specialised since it focuses on contemporary political issues in Jordan but it is not available to the public. As Alotaibi (2016) reported, there is a rise in interest among scholars in studying parallel corpora. For instance, the first general parallel corpus of English texts /Arabic translations was published by the National Council for Culture, Arts and Letters (NCCAL) in Kuwait (Al-Ajmi, 2004), the open parallel corpus known as OPUS (Tiedemann, 2012) and Arabic/English Parallel Corpus (AEPC) by Alotaibi (2016). The AEPC is not available, and its website dedicated to it is not working.

[1] For more information about Arabic corpora, see Awdeh et al. (2019), Al-Saif and Markert (2010), Atwell (2018), Sharaf and Atwell (2012a, b), Sharaf et al. (2010).

Arabic Legal Corpora

There is a growing interest in building specialised corpora. However, due to the fact that certain legal papers are private by their very nature or produced inside institutional frameworks, there may be issues with the accessibility of legal materials. There are several studies about current legal corpora (Goźdź-Roszkowski, 2021; Vogel et al., 2018), and the Sources of Language and Law (SOULL, 2020)[2], an open online platform consistently updated to offer a wealth of knowledge on current data and corpora of legal language.

International organisations like the UN, the EU and the WTO have been responsible for the creation of publicly accessible legal corpora. Examples of these corpora include the United Nations Parallel Corpus, which is significant and vital for Arabic legal translation research and pedagogy (Ziemski et al., 2016). The corpora, however, are displayed as separate files, requiring a well-trained user who knows how to parse the files to get them into an accessible, easy-to-use version. The JRC-Acquis Parallel Corpus contains a collection of written legislative texts in 23 official languages of European Union countries (Steinberger et. al., 2006), and the Arab-Acquis contains over 600,000 words altogether has been professionally translated from both English and French. In addition to these major corpora, the Digital Corpus of the European Parliament (DCEP) consists of different document types produced between 2001 and 2012 in various subject areas and the majority of the material is taken from the official website of the European Parliament (Hajlaoui. et. al., 2014).

Salhi (2013) built the English-Arabic Parallel Corpus of United Nations Texts (EAPCOUNT). It comprised 341 English-Arabic bitexts aligned on paragraph level, from data (mostly resolutions and annual reports) covering the period between 1996 and 2009 collected from different international organisations, mainly the UN (e.g. UNICEF, IMF, UNESCO), among others. Unfortunately, in spite of its importance in research and translation training, this source is not available online.

The Corpus of Arabic Legal Documents (CALD) (Müller, 2021)., a database developed by the ERC-AdG-project Islamic Law Materialised (ILM), contains a tool for studying and comparing legal Arabic documents in Islam from the 7th to the 16th centuries from various regions of the Muslim world aiming to facilitate the study of Islamic law from a historical perspective. Furthermore, it is a handy interface and tool for teaching and researching Arabic/English legal translation.

The Women's Learning Partnership, a body of non-profit, non-governmental organisations, launched a database/Corpus of Laws in 2012. It is a freely available collection of constitutions, civil family laws, gender-based violence-related penal codes, and victim-protection legislation in different countries and elsewhere around the globe. It is just a large dataset of legal documents in this genre, and the files are available in Arabic, English and many other languages. The corpus is not in a parallel format (i.e., aligned) and does provide online tools for teaching legal language or translation.

[2] SOULL (2020). An open web platform offers collected data on past and contemporary legal linguistics study worldwide. https://legal-linguistics.net/.

From this short review, it is clear that the variety of legal corpora is expanding due to the ongoing addition of new legal genres and legal languages. However, many of these corpora reviewed in this paper are not freely available. The available ones are not easy to use and require specialist computational tools that may not be accessible by many users (e.g. translation tutors and PGR researchers). This review also revealed that there are no available legal corpora of constitutions. The current paper seeks to fill an existing research gap and responds to an important need in the field of Arabic/English corpus-based legal translation by building a series of open-access monolingual and parallel corpora of Arabic and English constitutions (See Introduction and Method).

Method

This research adopts an applied methodology to build Arabic legal corpora. This methodology enables researchers to benefit from the procedures and tools used in building the corpora in future projects and research in many fields. In the following, we will present the stages of building legal corpora in detail and present the types of corpora that we have complied. Finally, we will discuss the main research question of methods, challenges, and solutions.

Corpus Building Procedures

In developing the Arabic constitutions corpora, we collected all available versions of the constitutions from 22 Arabic countries: Algeria, Bahrain, Comoros, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, the United Arab Emirates, and Yemen. The oldest version of the constitution dates back to 1923 (in Egypt), and the most recent version dates back to 2021 (in Oman).

Corpora building is divided into three main steps: (i) data collection, (ii) preprocessing the set of data, and (iii) corpora alignment. The following subsections discuss these steps in more detail.

(1) *Data Collection*

We collected the original Arabic versions of the constitutions, along with the English translation of those versions, from the Arab government's official websites as well as websites specialised in publishing state constitutions. We used the Constitute Project (Zachary et al., 2012) which provides a chronological compilation of constitutions from around the world, but many old versions of the Arabic constitutions were unavailable on this site. In addition, some versions were available in Arabic or English only. Therefore, we manually selected the relevant documents from the Constitute Project and downloaded all files separately. We then searched the World

Intellectual Property Organisation (WIPO)[3] website to find translations of unavailable constitutions.

We compiled at least one parallel constitution for the 22 countries, except for Comoros and Djibouti, for which only the Arabic versions were available, and there was not a parallel English version for both constitutions. We found several different English translations of some constitutions, and these were added to a monolingual corpus in English (see Types of compiled corpora). We faced another challenge while collecting the constitutions; some Arab government countries' websites did not contain any information about the original constitutions. In addition, some library websites have placed restrictions on constitution files in PDF format, such as the US Library of Congress.

(B) *Pre-processing*

The stage of pre-processing went through three steps: (i) standardisation of file formatting, (ii) cleaning, formatting, and normalising data, and (iii) parallel corpus alignment. To undertake these procedures, we used technical tools that we will show in each of the following steps:

(i) Standardising File Format

Most of the PDF files included texts or images, while the rest of the documents were bundled as (.txt) text files or HTML web pages. In this step, we converted all the bundled files to text (.txt) file format. It was challenging to choose one Optical Character Recognition (OCR) tool to rely on for the Arabic language since many lack accuracy, efficiency, and encoding issues in Arabic. Since these tools work differently depending on the hidden format of PDF files, we used the following free OCR tools:

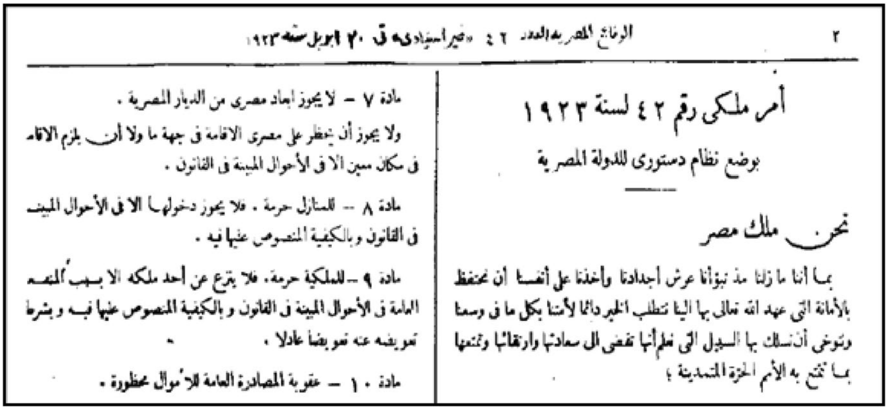
- Sotoor OCR[4]: is an online tool that reliably and accurately recognises Arabic characters, but it is limited to 100 free pages per user;
- i2OCR[5]: is another web-based tool that recognises each page individually;
- Google translate: we used the OCR embedded tool inside the Google Translate mobile application. The use of this application was used only if the previous systems failed to recognise Arabic texts.

In addition to the general challenges faced by Arabic OCRs, some PDF files contained old images (typewritten old Arabic script), as shown in Fig. 1, a sample of the

[3] WIPO LEX DATABASE https://wipolex.wipo.int/ar/main/legislation

[4] Sotoor https://sotoor.ai/en/home.

[5] i2OCR, a free online Optical Character Recognition OCR) https://www.i2ocr.com/free-online-arabic-ocr.



٢ الوقائع المصرية العدد ٤٢ « غير اعتيادى » فى ٢٠ ابريل سنة ١٩٢٣

أمر ملكى رقم ٤٢ لسنة ١٩٢٣
بوضع نظام دستورى للدولة المصرية

نحن ملك مصر

بما أننا ما زلنا منذ تبوأنا عرش أجدادنا وأخذنا على أنفسنا أن نحتفظ
بالأمانة التى عهد الله تعالى بها الينا نتطلب الخير دائما لأمتنا بكل ما فى وسعنا
ونتوخى أن نسلك بها السبيل التى نعلم أنها تفضى الى سعادتها وارتقائها وتمتعها
بما تتمتع به الأمم الحرة المتمدينة ؛

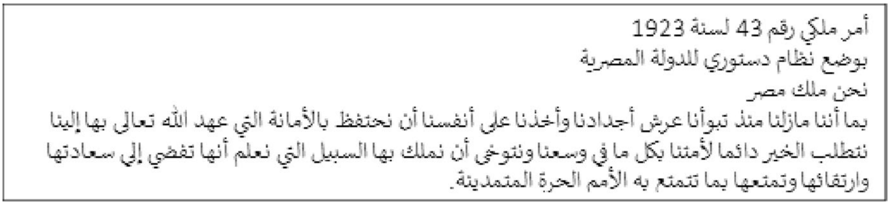
مادة ٧ – لا يجوز ابعاد مصرى من الديار المصرية .
ولا يجوز أن يحظر على مصرى الاقامة فى جهة ما ولا أن يلزم الاقامة
فى مكان معين الا فى الأحوال المبينة فى القانون .

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

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

مادة ١٠ – عقوبة المصادرة العامة للأموال محظورة .

Fig. 1 A snapshot of 1923 Egyptian Constitution



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.

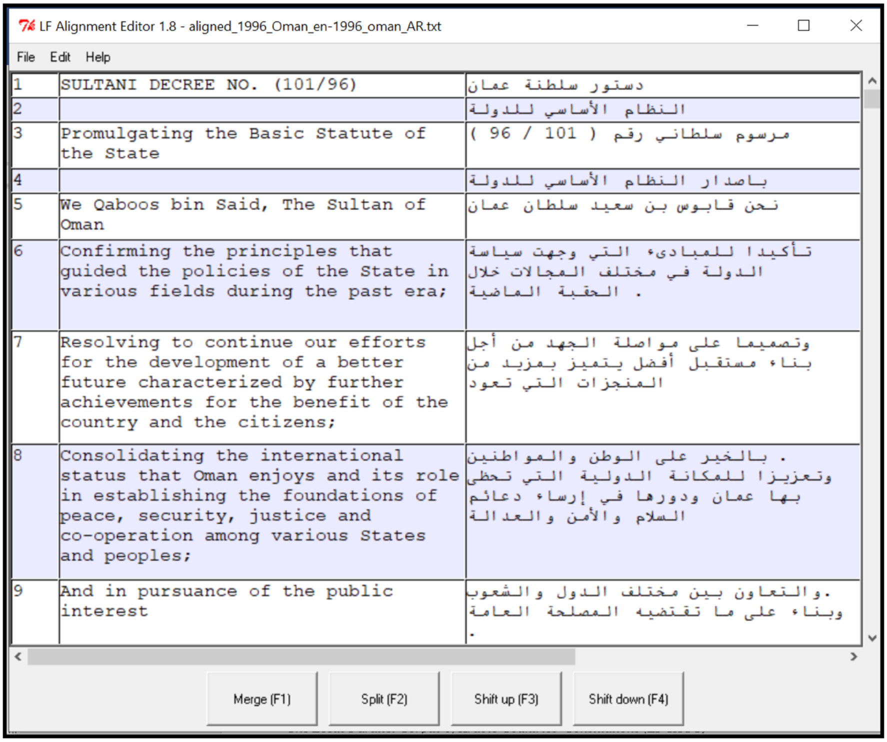
Fig. 2 The OCR result of the first column in figure one

1923 Egyptian constitution, while Fig. 2 shows the text similar to the first column (in figure two) that was recognised by the OCR tool, Sotoor.

Furthermore, to make it easier to look for a certain file, we standardised file naming as [publication year] _ [country name] _ [Arabic] or [English] or [Ar_En] describing the language of the constitution. For example, "1923_Egypt_Ar_En" refers to the parallel constitution file of Egyptian that was published in 1923.

(b) Data Cleaning

This step involves cleaning, formatting, and normalising files (such as removing Arabic diacritics and symbols). This is done by following general procedures to standardise Arabic writing in files. For example, some constitutions used only a number or an abbreviated form, as in "art." to refer to the number of articles and in those cases, we changed all the abbreviations to the word "Article" so that all the files are unified into a single format. Likewise, in the Arabic files, we replaced the word "فصل" with the word "article" in some Arab constitutions, as it was used to refer to the word "المادة" as in the Constitution of Morocco.


LF Alignment Editor 1.8 - aligned_1996_Oman_en-1996_oman_AR.txt
File
Edit
Help
Merge (F1)
Split (F2)
Shift up (F3)
Shift down (F4)

1	SULTANI DECREE NO. (101/96)	دستور سلطنة عمان
2		النظام الأساسي للدولة
3	Promulgating the Basic Statute of the State	مرسوم سلطاني رقم (101 / 96)
4		باصدار النظام الأساسي للدولة
5	We Qaboos bin Said, The Sultan of Oman	نحن قابوس بن سعيد سلطان عمان
6	Confirming the principles that guided the policies of the State in various fields during the past era;	تأكيدا للمبادىء التي وجهت سياسة الدولة في مختلف المجالات خلال الحقبة الماضية .
7	Resolving to continue our efforts for the development of a better future characterized by further achievements for the benefit of the country and the citizens;	وتصميما على مواصلة الجهد من أجل بناء مستقبل أفضل يتميز بمزيد من المنجزات التي تعود
8	Consolidating the international status that Oman enjoys and its role in establishing the foundations of peace, security, justice and co-operation among various States and peoples;	. بالخير على الوطن والمواطنين وتعزيزا للمكانة الدولية التي تحظى بها عمان ودورها في إرساء دعائم السلام والأمن والعدالة
9	And in pursuance of the public interest	.والتعاون بين مختلف الدول والشعوب وبناء على ما تقتضيه المصلحة العامة .

Fig. 3 LF-aligner automatic alignment for Oman 1996 constitution

(iii) Parallel Corpus Alignment

It is the last step in text processing, which is related to the alignment of the Arabic and English files. Manual alignment is a laborious and time-consuming process. Thus, the adoption of the LF-aligner[6] application, which aligns the translated compiled files, gives the option to manually review the alignment before it is completed, as shown in Fig. 3. Therefore, this tool facilitated a semi-automated alignment process for all articles of the collected constitutions. It is worth mentioning that the alignment was done at the level of the entire article and not at the sentence level to allow users specialised in other disciplines (e.g. politics, law) to use the corpus.

Parallel Corpus Compilation

All constitutions are collected, then aligned in excel sheets, and finally saved as text files. We used Sketch Engine to test and automate the corpora. This platform can be used for processing users' data, collecting data from the web, and exploring a vast

[6] LF Aligner https://sourceforge.net/projects/aligner/.

number of open access/available corpora for many languages. It allows fundamental features (Arabic Word Sketches; Arabic Concordance or keyword in context (KWIC); Arabic Thesaurus; Arabic Word Lists and Arabic N-grams which identify patterns relating to multi-word units (MWU) in Arabic (Kilgarriff, 2004; 2014).

Types of Compiled Corpora

The following corpora have been built and made available on the Sketch Engine platform:

The Leeds Parallel Corpus of Arabic Countries' Constitutions (LPCACC)

The final version of this corpus includes the constitutions of 20 Arabic countries and their English translations from 1922 to 2022, Arabic version (407,633 words) and English version (489,448 words). The LPCACC excludes Comoros and Somalia since there are no parallel versions of these constitutions, and it excludes the preambles, which are compiled in a separate corpus. Figure 3 shows the number of parallel constitutions in each country and indicates that Egypt represents the most significant number of constitutions in the parallel corpus with eight parallel files. The United Arab Emirates, Lebanon, Qatar, and Saudi Arabia are all represented in the constitution by one parallel constitution. The other 15 states are represented by an average of three parallel constitutions for.

The Parallel Corpus of Preambles of Arabic Countries' Constitutions (PCPACC)

This version includes all available preambles of the Arab countries' constitutions (32.660 words), Arabic (12,360) and English (20,300).

Monolingual Corpora of Arabic and English Constitutions (MCAEC)

Some countries publish constitutions either in Arabic or in English only, and in some cases, the constitutions are published in French and Arabic with no English translations. Therefore, we created two separate corpora for these versions of constitutions, one in Arabic and another in English (see Fig. 4). The total number of words for the Arabic language corpus is (788,477), and for the English language, the corpus is (343,582). Figure 4 shows a comparison of the number of constitutions issued in each language from 1922 to 2022. In the early years, more versions of constitutions were available in Arabic without English translations. However, after 1990, English translations of the constitutions were available and published on the web, which facilitated access to them, and they were included in the English language corpus (Fig. 5).

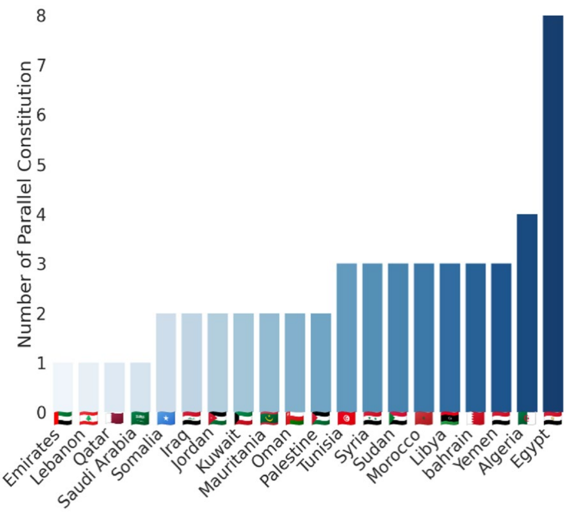


Fig. 4 The total number of parallel constitutions in each Arabic country

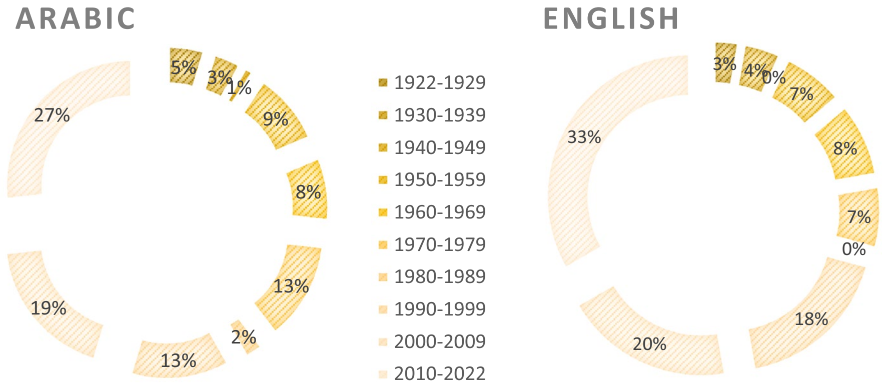


Fig. 5 Monolingual corpora of Arabic and English constitutions

The Leeds Monolingual Corpus of English Countries' Constitutions (LMCECC) (Comparative Corpus)

This comparative corpus was created from the revised and updated versions of the constitutions of eight countries whose official language is English, which consists of Australia, Canada, Ireland, New Zealand, Singapore, South Africa, USA and UK. The oldest version of this corpus is the 1985 Australia, and the latest revised version is for the USA, 2016. This corpus consists of 677,056 words and 557,086 words (See El-Farahaty & Elewa, 2020)[7].

[7] The authors wish to thank Abdulrahman Alosaimy for collecting the comparative corpus and annotating it in 2018.

Concluding Remarks

In this paper, we have presented the methods, procedures and challenges of building sustainable open access diachronic parallel and monolingual corpora of all available versions of Arabic countries' constitutions and their preambles as well as a comparative corpus of constitutions of English-speaking countries.

Throughout the process of building the corpora, we faced different challenges. In the data collection phase, it was challenging to find all the versions of the original constitutions for each Arabic country or their translations since some of the files date back to 1922. This resulted in searching several online websites in addition to searching the translations of the constitutions manually because not all of them were available on each country's government website. Although many of them were available in the Library of Congress, they were in PDF files. Therefore, file conversion was one major challenge and it was not enough to choose one OCR tool to rely on for the Arabic language since many lack accuracy, efficiency as well as encoding issues in Arabic. We used more than one OCR tool to deal with the challenges faced during the data processing phase. For example, PDF files contained old images (typewritten old Arabic script) for which we used the Sotoor OCR.

For the parallel corpus file alignment, it was not possible to fully automate the alignment of the Arabic and English files. Therefore, we used the LF-aligner[8] application, to undertake a semi-automated alignment process for all articles of the collected constitutions. As some of the Arabic files were distorted, this application gives the option to manually review the alignment before it is completed. This semi-automation saved time and effort of manually aligning the parallel corpus which is such a laborious process although we used Excel sheets to align some files. Not all of the constitutions have a parallel Arabic-English version, e.g. Comoros and Djibouti for which only the Arabic versions were available and there was not a parallel English version for both constitutions. All versions of the constitutions which did not have a parallel version were included as part of a monolingual corpus for each language.

The corpora offer different values to different end users. They offer sustainable pedagogical tools for law students, and postgraduate translation and interpreting students enrolled in Arabic and English translation UG and MA programmes all over the world. They are important research tools/data sets for postgraduate researchers, academics, law professionals across the world. The corpora will be used in mapping specialised terminology and phraseology and will be of big value nationally for (law drafters/professionals, NGOs). The corpora will be used to undertake empirical investigations of a range of topics, using corpus-based tools, e.g. a diachronic corpus-based pragmatic investigation of directives (for more examples of such investigations, see El-Farahaty & Elewa, 2020; Brierley & El-Farahaty, 2019). We aim to scale up the project and extend the building of these corpora to include other MSA and Classical Islamic legal genres and text types.

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

Acknowledgements This research is funded by Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia, which is part of the Arab Effort Coordination Program in the field of translation.

Author Contributions All authors whose names appear on the submission: (1) made substantial contributions to the design of the work; collection of data, the creation of new corpora and writing the paper; (2) approved the version to be sent for publication; and (3) agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding This research is funded by Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia, which is part of the Arab Effort Coordination Program in the field of translation.

Data Availability Data is safely stored on the Authors' University Repository. Data will be made available on Sketch Engine and GitHub.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Approval Ethical approval is not required for this research

Consent for Publication No consent to publish form is required for this study

References

Abbas, M., & Smaili, K. (2005). 'Comparison of topic identification methods for the Arabic language'. I*n Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP* pp 14-17.

Ahmad, A. A. S., Hammo, B., & Yagi, S. (2017). 'Construction of an English-Arabic Political Parallel Corpus' *New Trends in Information Technology (NTIT)–2017*, *2*, 93. pp 157-171.

Ahmed, A., Ali, N, Alzubaidi, M. Zaghouani, W. Abd-alrazaq, A., Househ, M. (2022). 'Free and Accessible Arabic Corpora: A Scoping Review', *Computer Methods and Programs in Biomedicine Update*, 100049. Available from https://www.sciencedirect.com/science/article/pii/S2666990022000015 [Accessed 8 February 2023]

Al-Ajmi, H. (2004). A new english–arabic parallel text corpus for lexicographic applications. *Lexikos, 14*, 326–330.

Alansary, S., & Nagi, M. (2014). '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, 192*, 347–357.

Al-Jawfi, R. (2009). Handwriting arabic character recognition LeNet using neural network. *Int. Arab J. Inf. Technol., 63*, 304–309.

Alotaibi, H. M. (2016). 'AEPC: Designing an arabic/english parallel corpus', *Research in Corpus Linguistics*, pp 1-7.

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*, The University of Leeds, pp 5-8.

Al-Saif, A., & Markert, K. (2010). 'The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic' In *Proceedings of the seventh international conference on language resources and evaluation LREC'10)*. pp 2046-2053.

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics, 112*, 135–171.

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). '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.

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

Austin, J. L. (1962). *How to do things with words*. Harvard University Press.

Awdeh, H., Abdallah, A., Bernard, G., Hajjar, M., & El-Sayed, M. (2019). 'A silver standard Arabic corpus for segmentation and validation', *BDCSIntell.*

Baker, M. (2019). 'Corpus Linguistics and Translation Studies: Implications and applications' In: Kim, K.H., & Zhu, Y. (eds.), *Researching Translation in the Age of Technology and Global Conflict*. (pp 9-24). Routledge.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies, 7*(2), 223–243.

Baker, M., Francis, G., & Tognini-Bonelli, E. (eds.) (1993). *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.

Belinkov, Y., Habash, N. Kilgarriff, A., Ordan, N., Roth. R., and Suchomel, V. (2013). ArTenTen: A new, vast corpus for Arabic. Retrieved from: https://www.sketchengine.eu/wp-content/uploads/arTenTen_corpus_for_Arabic_2013.pdf [Accessed February 20 2023].

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

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for modern standard Arabic. *Behavior Research Methods, 422*, 481–487.

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. *The Journal of Specialised Translation (JoSTrans), 32*, 121–145.

Cao, D. (2007). Legal speech acts as intersubjective communicative action. In: *Interpretation, Law and the Construction of Meaning.* Springer, Dordrecht. Available here. [Accessed February 26 2023]

Cieri, C. et al. (2022). 'Reflections on 30 Years of Language Resource Development and Sharing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* (pp. 543-550).

Dukes, K., & Atwell, E. (2012). 'LAMP: A multimodal web platform for collaborative linguistic analysis'. In *Proceedings of the Eight International Conference on Language Resources and Evaluation LREC'12)* (pp. 3268-3275). (European Language Resources Association ELRA).

Dukes, K., Atwell, E., & Habash, N. (2013). Supervised collaboration for syntactic annotation of quranic Arabic. *Language Resources and Evaluation, 471*, 33–62.

El-Farahaty, H., & Elewa, A. (2020). A Corpus-based analysis of deontic modality of obligation in Arabic–English constitutions'. *Estudios De Traducción, 10*, 107–136.

El-Haj, M., & Koulali, R. (2013). 'KALIMAT a multipurpose Arabic Corpus'. In *The Second Workshop on Arabic Corpus Linguistics WACL-2,* pp. 22-25.

El-Haj, M., Kruschwitz, U., & Fox, C. (2015). Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation, 493*, 549–580.

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

Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue Internationale De Sémiotique Juridique, 345*, 1515–1540.

Habash, N., Zalmout, N., Taji, D., Hoang, H., & Alzate, M. (2017). ‘A parallel corpus for evaluating machine translation between Arabic and European languages’. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers:* (pp. 235-241).

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies, 31*, 1–187.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., & Varga, D. (2014). ’Dcep-digital corpus of the european parliament’. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)* (pp. 3164-3171).

Khwaileh, T., Mustafawi, E., Herbert, R., & Howard, D. (2018). Gulf Arabic nouns and verbs: A standardised set of 319 object pictures and 141 action pictures, with predictors of naming latencies. *Behavior Research Methods, 506*, 2408–2425.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, (pp. 105-116).

Kilgarriff, A., et al. (2014). The sketch engine: Ten years on. *Lexicography, 1*, 7–36.

Kruger, A. (2004). ‘Corpus-based translation research comes to Africa. *Language Matters: Studies in the Languages of Southern Africa, 35*, 1–5.

McCarthy, M. &. O’Keeffe, A. (2012). ‘Analysing Spoken Corpora’. In C. A. Chappelle (eds.). The Encyclopedia of Applied Linguistics. DOI: https://doi.org/10.1002/9781405198431. Online at: http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0028/full.

McEnery, T., Hardie, A., & Younis, N. (2019). ‘Introducing Arabic Corpus Linguistics’. In T. McEnery, A. Hardie, & N. Younis (eds.), *Arabic Corpus Linguistics,* (pp. 1–16). Edinburgh University Press. Available from http://www.jstor.org/stable/10.3366/j.ctvcwndq8.4 [Accessed February 25 2022]

Müller, C. (2021). ‘Cald: A very short introduction’, *The Documents of Islamic Law in History. Studies on Arabic Legal Documents*. Available from https://dilih.hypotheses.org/763 [Accessed February 25 2022]

Olohan, M. (2004). *Introducing corpora in translation studies*. Routledge.

Palmer, F. R. (2001). *Mood and modality*. Cambridge University Press.

Parkinson, D. B. (2012). ArabiCorpus. Online. Available from: https://arabicorpus.byu.edu/ [Accessed February 20 2022]

Rühlemann, C. (2019). *Corpus linguistics for pragmatics: A GUIDE FOR RESEARCH*. Routledge.

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

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Searle, J. R. (1979). *Expression and meaning: Essays in the theory of speech acts*. Cambridge University Press.

Sharaf, A., Atwell, E. S., Dukes, K., Sawalha, M., Al-Saif, A., Sharoff, S. & Roberts, A. (2010). ‘Arabic and Quranic computational linguistics projects at the University of Leeds’ المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز./Almashārī^{c} Al-hāsūbiyyah ^{c}ala Al-lughah Al-^{c}rabiyyah fī jāmi^{c}at Leeds’. In *Proceedings of the workshop of Increasing Arabic Contents on the Web, Organised by Arab League Educational, Cultural and Scientific Organization (ALECSO)*.

Sharaf, A. B., & Atwell, E. (2012a). ‘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). ‘QurSim: A corpus for evaluation of relatedness in short texts’. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation. (LREC’12)*: (pp. 2295-2302).

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics, 114*, 435–462.

Solan, L. M., & Gales, T. (2017). ‘Corpus linguistics as a tool in legal interpretation’, *BYU L. Rev.*, pp.1311-1358, Available from https://digitalcommons.law.byu.edu/lawreview/vol2017/iss6/5 [Accessed November 10 2022]

Solum, Lawrence. B. (2021). ‘Legal Theory Lexicon 021: Speech Acts’. Available from https://lsolum.typepad.com/legal_theory_lexicon/2004/02/legal_theory_le_4.html [Accessed March 11 2023]

SOULL Sources of Language and Law (2020). Available from https://legal-linguistics.net/data-collections [Accessed November 10 2022]

Steinberger, R. , Pouliquen, B. , Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). ‘The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages’, *arXiv preprint cs/0609058.*

Available from https://publications.jrc.ec.europa.eu/repository/handle/JRC32786 [Accessed November 11 2022]

Tiedemann, J. (2012). ‘Parallel data, tools and interfaces in OPUS’. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, (LREC’12) (pp. 2214-2218).

Visconti, J. (2009). Speech acts in legal language: Introduction. *Journal of Pragmatics, 41*, 393–400.

Vogel, F., Hamann, H., & Gauer, I. (2018). Computer assisted legal linguistics: Corpora and empirical methods as a new instrument in the legal toolbox. *Law & Social Inquiry. Journal of the American Bar Foundation ABF, 434*, 1340–1363.

Women’s Learning Partnership: About Our Corpus of Laws (2012). Available from https://learningpartnership.org/learning-center/learning-center-overview/corpus-laws [Accessed 9 February 2023]

Zachary, E., Ginsburg, T., & Melton, J. (2012) ‘Constitute: The World’s Constitutions to Read, Search, and Compare’. Available from: https://www.constituteproject.org/content/about?lang=en [Accessed March 10 2022]

Zaghouani, W. (2017). ‘Critical survey of the freely available Arabic corpora’, *Available from* https://arxiv.org/abs/1702.07835 *[Accessed November 12 2022]*

Zaki, M. (2020). ‘Corpus-based language teaching and learning: Applications and implications’, *International Journal of Applied Linguistics*, 6 October 4th Quarter/Autumn.

Zaki, M., Wilmsen, D., & Abdulrahim, D. (2021). ‘The Utility of Arabic Corpus Linguistics’, *The Cambridge Handbook of Arabic Linguistics*, pp 473-503.

Zaki, M. (2021). ‘Corpora and translation teaching in the Arab world’. In Said M. Shiyab (eds.), *Research into Translation and Training in Arab Academic Institutions*, (pp. 21-40).

Zeroual, I., & Lakhouaja, A. (2018). ‘Arabic corpus linguistics: major progress, but still a long way to go. In Shaalan, K., Hassanien, A. E., & Tolba, F. (eds.), *Intelligent Natural Language Processing: Trends and Applications*:(pp. 613-636).

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). ‘The United Nations parallel corpus, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, (LREC’16)* (pp. 3530-3534).

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Hanem El-Farahaty[1,2] · **Nouran Khallaf**[1,3] · **Amani Alonayzan**[1,3]

Nouran Khallaf
mlnak@leeds.ac.uk

Amani Alonayzan
mlasao@leeds.ac.uk

[1] University of Leeds, Leeds, UK

[2] Mansoura University, Mansoura, Egypt

[3] PhD Researcher at the University of Leeds, Leeds, UK