Lang Resources & Evaluation (2015) 49:721–751
DOI 10.1007/s10579-014-9284-1

ORIGINAL PAPER

A 700M+ Arabic corpus: KACST Arabic corpus design and construction

Abdulmohsen O. Al-Thubaity

Published online: 16 October 2014

Abstract Compared with English, Arabic is a poorly-resourced language within the field of corpus linguistics. A lack of sufficient data and research has negatively affected Arabic corpus-based researchers and natural language processing practitioners. Although a number of Arabic corpora have been developed in recent years, the overall situation has improved little. The aim of this paper is twofold. First, it reviews 14 Arabic corpora categorized by their designated purpose, target language, mode of text, size, text date, location, text type/medium, text domain, representativeness, and balance. The review also describes the availability of the reviewed corpora, the presence of tokenization, lemmatization and tagging, and whether there are any tools available to search and explore them. Second, it introduces the King Abdulaziz City for Science and Technology (KACST) Arabic corpus, which was designed and created to overcome the limitations of existing Arabic corpora. The KACST Arabic corpus is a large and diverse Arabic corpus with clearly defined design criteria. It is carefully sampled, and its contents are classified based on time, region, medium, domain, and topic, and it can be searched and explored using these classifications. The KACST Arabic corpus comprises more than 700 million words from the pre-Islamic era to the present day (a period covering more than 1,500 years), collected from 10 diverse mediums. Each text has been further classified more specifically into domains and topics. The KACST Arabic corpus is freely available to explore on the Internet (http://www.kacstac.org.sa) using a variety of tools.

Keywords Arabic corpora · Corpus evaluation · Corpus design · Corpus compilation · Arabic language resources · Natural language processing (NLP)

A. O. Al-Thubaity (✉)
King Abdulaziz City for Science and Technology, P O Box 6086, Riyadh 11442, Saudi Arabia
e-mail: aalthubaity@kacst.edu.sa

1 Introduction

While the use of corpora for language studies is well-established for many of the world's languages, such as English, the case of Arabic is different. There is a relative lack of interest in, and even awareness of, the use of corpora in Arabic language studies. This lack of interest is due to two main reasons. First, there is no publically available corpus of Arabic that is sufficiently large and representative by current standards. Second, the idea of using the empirical evidence that corpora can provide to test and evaluate linguistic hypotheses about the Arabic language is still relatively new and not widely considered by the Arabic language research community. As for Arabic language processing, while there is a degree of interest and awareness, there is still no standard benchmark dataset or corpus through which developed techniques can be challenged and applied to illustrate their effectiveness.

In Sect. 2, several existing Arabic corpora will be reviewed. This involves classifying the corpora based on their designated purpose, target language, mode of text, size, text date, location, text type/medium, text domain, representativeness, and balance. The availability of the corpora, the presence of tokenization, lemmatization and tagging, and whether there are any tools available to search and explore them, is also examined. Based on this review, the efforts that have been taken to design and build a corpus that will provide a valuable source of data for linguistic research are illustrated. The resulting KACST Arabic Corpus comprises more than 700 million words from different and diverse Arabic written genres spanning over 1,500 years, from the pre-Islamic period to the present day. The processes of the design and construction of the KACST Arabic Corpus are discussed in Sects. 3 and 4, respectively. Finally, the paper's conclusions and future research tasks are described in Sect. 5.

2 Review of available Arabic corpora

Currently, several Arabic corpora are available commercially and freely for different purposes. There are reports that as many as 19 Arabic language corpora have been developed, ranging in size from 1 million to 3 billion words (Al-Sulaiti and Atwell 2006). These corpora were, by and large, extracted from newspapers, and they were developed almost exclusively for the use of the researchers, and thus they were neither freely available as a data source, nor could they be searched on the web. From 2005 to 2013, however, this situation changed because at least 14 Arabic corpora were developed and made available for free download or search online.

Our review of these available corpora will focus on two main aspects. The first is criteria that define the choice of the corpus content, namely its intended purpose, target language, text mode, size, text date, text location, text type/medium, text domain, corpus representativeness, and corpus balance. The second aspect relates to the features that enhance corpus usability, namely free availability, the presence of tokenization, lemmatization and part of speech (POS) tagging, and, if there is tagging, whether there are any tools available to search and analyze the tags. Tables 1, 2, 3 and 4 list the 14 Arabic corpora released between 2005 and 2013

together with their corresponding evaluation criteria. The corpora are arranged based on the date of their published references. Those with no available publication references are at the end. Here, it should be mentioned that there are some language resources available for Arabic that are not considered to be corpora, such as the Quranic Arabic Corpus by Kais Dukes from the University of Leeds (Dukes et al. 2013) because it contains only one text, the Quran.

2.1 Purpose of a corpus

Before researchers begin designing and compiling corpus materials, they must first explicitly and exactly define the purpose of the corpus, that is, what research question(s) the corpus intends to answer. The Arabic corpora illustrated in Tables 1, 2, 3 and 4 can be generally grouped into two categories.

The first category comprises the corpora designed and built to satisfy a specific research purpose, such as the modern standard Arabic (MSA) corpus (Abdelali et al. 2005), Alwatan-2004 (Abbas et al. 2011), the KACST text classification corpus (Khorsheed and Al-Thubaity 2013), and the Arabic Gigaword Corpus, Fifth Edition (Parker 2011), which was built for Arabic NLP research. The Corpus of Contemporary Arabic and the Arabic Learner Corpus v2 (Alfaifi and Atwell 2013), which were designed and built for language teaching and learning research, can also be included in this category. Any of these corpora can be used for other purposes as long as the relevant research question(s) can be answered.

The second category comprises corpora designed and built to be broadly useful for numerous purposes, such as the International Corpus of Arabic (Alansary et al. 2007), the arTenTen12 corpus (Belinkov et al. 2013), and the Leeds Internet Arabic Corpora by Serge Sharoff, University of Leeds.

2.2 Language and mode

While all the corpora introduced here are monolingual Arabic corpora, there are three main varieties of Arabic: Classical Arabic, MSA, and Arabic Dialects. Moreover, there are different sub-varieties within each of these main varieties. Most Arabic corpora focus on MSA, which is to be expected because most of the available resources used to collect text data are written in MSA. MSA is the language of most Arabic newspapers, and more importantly, it is the official language that can be understood by all Arabs. The King Saud University Corpus of Classical Arabic (Alrabiah et al. 2013) is the only corpus solely devoted to Classical Arabic; while a part of the arabiCorpus contains Classical Arabic, this fact is not stated clearly in the corpus documentation. The arTenTen12 corpus may also include Classical Arabic texts because of its large size and the fact that it was collected from a variety of sources across the Internet. The arabiCorpus is the only corpus that has a section devoted solely to one of the Arabic dialects, colloquial Egyptian. The arTenTen12 corpus may also contain some Arabic Dialects texts because of how its data were collected, as stated above.

Table 1 Arabic corpora released between 2005 and 2013

Corpus name	Reference	Source	Purpose
1. Modern standard Arabic corpus	Abdelali et al. (2005)	Ahmad Abdelali http://aracorpus.e3rab.com/index.php?content=english	Specific/Arabic NLP
2. Akhbar Al Khaleej (2004)	Abbas and Smaili (2005)	Mourad Abbas http://sourceforge.net/projects/arabiccorpus/files	Specific/Arabic NLP
3. Corpus of contemporary Arabic	Al-Sulaiti and Atwell (2006)	University of Leeds http://www.comp.leeds.ac.uk/eric/latifa/research.htm	Specific/teaching Arabic as a foreign language and Arabic language engineering
4. International corpus of Arabic (ICA)	Alansary et al. (2007)	Bibliotheca Alexandrina http://www.bibalex.org/ica/en/	General
5. Open source Arabic corpus (OSAC)	Saad and Ashour (2010)	Motaz Saad https://sites.google.com/site/motazsite/arabic/osac	Specific/Arabic NLP
6. Alwatan-2004	Abbas et al. (2011)	Mourad Abbas http://sourceforge.net/projects/arabiccorpus/files	Specific/Arabic NLP
7. Arabic Gigaword corpus fifth edition	Parker (2011)	LDC https://catalog.ldc.upenn.edu/LDC2011T11	Specific/Arabic NLP
8. King Saud University Corpus of Classical Arabic (KSUCCA)	Alrabiah et al. (2013)	King Saud University http://ksucorpus.ksu.edu.sa/	General
9. KACST text classification corpus	Khorsheed and Al-Thubaity (2013)	KACST contact authors	Specific/Arabic NLP
10. 2012 Arabic newspapers corpus	Al-Thubaity et al. (2013)	Abdulmohsen Al-Thubaity http://sourceforge.net/projects/kacst-acptool/	Specific/Arabic NLP
11. Arabic learner corpus v2	Alfaifi and Atwell (2013)	University of Leeds http://www.arabiclearnercorpus.com/	Specific/language teaching and learning research

Table 1 continued

Corpus name	Reference	Source	Purpose
12. arTenTen12	Belinkov et al. (2013)	Sketch Engine https://www.sketchengine.co.uk/	General
13. arabiCorpus	Dilworth Parkinson, Brigham Young University	Brigham Young University http://arabicorpus.byu.edu/	General
14. Leeds Internet Arabic Corpora	Serge Sharoff, University of Leeds	University of Leeds http://smlc09.leeds.ac.uk/query-ar.html	General

Criteria reference, source, and purpose

Table 2 Arabic corpora released between 2005 and 2013

Corpus name	Mode and language	Size	Availability	Date	Location
1. Modern standard Arabic corpus	Written MSA	113M	Free to download	2002	11 Arab countries
2. Akhbar Al Khaleej (2004)	Written MSA	3M	Free to download	2004	1 Arabic country (Bahrain)
3. Corpus of contemporary Arabic	Written and Spoken MSA	1M	Free to download	1990s up to 2005	Different Arab countries (no distribution given)
4. International Corpus of Arabic (ICA)	Written MSA	100M	Free to explore	Not available	Distribution given for newspapers and magazine; 7 countries)
5. Open Source Arabic Corpus (OSAC)	Written CA and MSA	18M	Free to download	Not available	No distribution given
6. Alwatan-2004	Written MSA	10M	Free to download	2004	1 Arabic country (Oman)
7. Arabic Gigaword Corpus Fifth Edition	Written MSA	1,077M	Fees to download	2002–2010	6 Countries (UK, France, China, Egypt, Tunisia, Lebanon)
8. King Saud University Corpus of Classical Arabic (KSUCCA)	Written CA	50M	Free to download Fees to explore	From the pre-Islamic era until the end of the fourth *Hijri* century	Not available
9. KACST text classification corpus	Written MSA	11.55M	Free to download permission required	2008	Mostly from Saudi Arabia
10. 2012 Arabic newspapers corpus	Written MSA	2.5M	Free to download	2012	All Arab countries (evenly distributed)
11. Arabic learner corpus v2	Written and Spoken MSA	282K	Free to download	2012 and 2013	Saudi Arabia
12. arTenTen12	Written CA, MSA and dialects	6.6G	Fees to explore	2012	No distribution is given
13. arabiCorpus	Written, CA, MSA and dialects	173M	Free to explore	No distribution given	Distribution given for newspapers, 7 countries
14. Leeds internet arabic corpora	Written MSA	317M	Free to explore	No distribution given	No distribution given

Criteria mode and language, size, availability, date, location

All materials included in any corpus must be in written electronic format for ease of processing, regardless of whether they were originally produced through speech or written modes of communication. Notably, all of the reviewed Arabic corpora were collected from written modes of communication, except for the Corpus of Contemporary Arabic, which includes a small portion of spoken Arabic data.

2.3 Size

The size of a particular corpus is usually defined by a specific research question or purpose. Sinclair (1991:18) has suggested that "a corpus should be as large as possible, and should keep on growing." A large corpus of many millions of words is required to obtain useful empirical evidence regarding word use and collocation behavior. Conversely, several researchers believe that a corpus of one million words is sufficient for general linguistic research. (See Al-Sulaiti and Atwell 2006 for more details). As a general rule of thumb, however, larger corpora are better because many words and collocations occur with low frequencies.

The sizes of the reviewed Arabic corpora range from several thousand words to several billion. Most, however, contain less than 20 million words, such as Akhbar Al Khaleej 2004 (Abbas and Smaili 2005), the Open Source Arabic Corpus, and Alwatan-2004. While most of these corpora were collected for Arabic NLP, mainly for the purpose of text classification, the corpora collected for linguistics-related research are relatively large (i.e., more than 100 million words), such as the arabiCorpus (173M), the Leeds Internet Arabic Corpora (317M), the Arabic Gigaword Corpus Fifth Edition (1,077M), and the arTenTen12 corpus (6.6G). Clearly the average size of Arabic corpora has increased with time. This may be due to new technological capabilities that allow the gathering and storage of larger amounts of data in addition to the increased availability of electronic Arabic text.

2.4 Text dates

The text dates indicate the time period the corpus materials cover. This text date information can indicate a great deal about the particular language or language varieties included in the corpus. For example, texts collected more recently are more likely to be written materials comprising MSA, while texts collected from the Medieval period comprise classical Arabic texts. Some of the Arabic corpora cover only 1 year, such as Alwatan-2004 and the 2012 Arabic Newspapers corpus (Al-Thubaity et al. 2013), while others cover lengthier periods of time, such as the King Saud University Corpus of Classical Arabic, which covers more than 300 years of written Classical Arabic from the pre-Islamic era until the end of the fourth *Hijri* century (1010 Gregorian). The Corpus of Contemporary Arabic covers a shorter period of MSA, from the 1990s through 2005.

Unfortunately, some corpora do not provide any information regarding the time period the texts cover; moreover, all such corpora are large in size, such as the arTenTen12 corpus, the arabiCorpus, and the Leeds Internet Arabic Corpora. Additionally, it is noteworthy that for all corpora except those that cover only 1 year

Table 3 Arabic corpora released between 2005 and 2013

Corpus name	Medium	Domain
1. Modern standard Arabic corpus	Newspapers	Not available
2. Akhbar Al Khaleej (2004)	Newspapers	Economy, local news, international news, and sports
3. Corpus of contemporary Arabic	Magazines, radio, websites, newspapers, and emails	Natural sciences, applied sciences, social sciences, politics, commerce, life, arts, and leisure
4. International corpus of Arabic (ICA)	Newspapers, books, magazines, electronic press and internet articles	Strategic sciences, social sciences, sports, religion, literature, humanities, natural sciences, applied sciences, art, biography, and miscellaneous
5. Open source Arabic corpus (OSAC)	Websites	10 Categories: economics, history, entertainments, education and family, religious and fatwas, sports, health, astronomy, law, stories, and cooking recipes
6. Alwatan-2004	Newspapers	Culture, religion, economy, local news, international news, and sports
7. Arabic Gigaword corpus fifth edition	Newswires	Not available
8. King Saud University Corpus of Classical Arabic (KSUCCA)	Old Manuscripts	Religion, linguistics, literature, science, sociology, and biography
9. KACST text classification corpus	Saudi Press Agency (SPA), Saudi Newspapers (SNP), Websites, Writers, Forums, Islamic Topics, and Arabic Poems	SPA (cultural, sports, social, economic, political, general), SNP (cultural, sports, social, economic, political, general, IT), Websites (IT, economics, religion, medical, cultural, scientific), 10 writers' opinions, Forums (IT, economics, religion, medical, cultural, scientific, sport, general), Islamic Topics (Hadeeth, Aqeedah, Lughah, Tafseer, Feqh), Arabic Poems (love, wisdom, description, praise, bemoaning, lampoons)
10. 2012 Arabic newspapers corpus	Newspapers	Politics, economics, cultural, religion, sports, and science
11. Arabic learner corpus v2	215 texts written by learners of Arabic in Saudi Arabia	Materials produced by Arabic learners
12. arTenTen12	Websites	Not available

Table 3 continued

Corpus name	Medium	Domain
13. arabiCorpus	Newspapers, Books, Old Manuscripts	Newspapers, modern literature, nonfiction, Islamic Discourse, Egyptian Colloquial, Pre-modern, Arab Literature, Grammarians, Medieval Philosophy/Science, Hadith Literature, Quran, 1001 Nights
14. Leeds internet Arabic corpora	Internet	General, wikipedia, legal, scientific

Criteria medium, domain, tagged

Table 4 Arabic corpora released between 2005 and 2013

Corpus name	Tokenized	Lemmatized	Tagged
1. Modern standard Arabic corpus	No	No	No
2. Akhbar Al Khaleej (2004)	No	No	No
3. Corpus of Contemporary Arabic	No	No	No
4. International Corpus of Arabic (ICA)	No	No	Yes
5. Open Source Arabic Corpus (OSAC)	No	No	No
6. Alwatan-2004	No	No	Yes, KALIMAT corpus
7. Arabic Gigaword Corpus Fifth Edition	Partially yes (ATB)	Partially yes (ATB)	Partially yes (ATB)
8. King Saud University Corpus of Classical Arabic (KSUCCA)	No	Yes, via sketch engine	Yes, via sketch engine
9. KACST text classification corpus	No	No	No
10. 2012 Arabic newspapers corpus	No	No	No
11. Arabic learner corpus v2	No	No	Partially annotated for errors
12. arTenTen12	Partially yes	Partially yes	Partially yes
13. arabiCorpus	No	No	No
14. Leeds internet arabic corpora	Yes	No	Yes, via sketch engine

Criteria tokenized, lemmatized and tagged

of time, the texts are not divided or classified according to their dates or the time period to which they belong. Such a situation limits the usability of the corpus and makes it difficult to compare the languages used in different time periods, or to monitor how the Arabic language has evolved.

2.5 Text locations

Text location refers to where the text was originally published, and more specifically, the country of the publisher. In most cases, the location of the text is the same as the home country of the writer. The value of a corpus increases when it contains a variety of texts from writers of different countries who speak the same language. The language varieties that exist across the Arab countries at the lexical or semantic level can be revealed only when a corpus includes them. For example, in MSA the plural of the word "Bank" (**بنك** *Bank*) is written (أبناك *Abnak*) in Morocco but (بنوك *Bounok*) in the Gulf region. The reviewed Arabic corpora vary with respect to the information provided about the locations of the texts, with information on exact locations being frequently patchy.

Three of the corpora include texts from only one country. The texts in Akhbar Al Khaleej 2004 are from Bahrain, those in Alwatan-2004 are from Oman, and those in the KACST text classification corpus are from Saudi Arabia. The texts of two other corpora span multiple Arab countries; the MSA corpus (Abdelali et al. 2005) covers 11 Arab countries, and the 2012 Arabic Newspapers corpus covers all 19 Arab countries.

Another set of corpora shows only the locations of the newspapers that comprise one part of the corpus, such as the International Corpus of Arabic, which contains newspaper articles from seven Arab countries, and the arabiCorpus, which also contains newspapers from seven countries. Notably, the large corpora in our list do not provide any information regarding the locations of their texts, all of which were collected for linguistic research. The exclusion of such information limits the possible benefits of such corpora, such as the possibility of performing linguistic comparisons and contrasts based on geographic location.

2.6 Text type/medium and domain

Text type, or medium, refers to whether a written text is, for example, from a newspaper, magazine, book, or refereed journal. Related to this category is the text domain, which refers to whether a text is news or reportage (in the case of newspapers or magazines), or whether it is related to theoretical or applied linguistics (in the case of refereed journals, for example). A corpus that has diverse text domains spanning different text types will better represent multiple language varieties and should be better able to answer a variety of research questions. Moreover, the intended purpose of a corpus decides what text types are included.

There is great variation among the reviewed Arabic corpora with respect to their coverage of text mediums and domains. Some corpora are based on the single medium of newspapers, such as the MSA Corpus, Alwatan-2004, and the 2012 Arabic Newspapers corpus. The King Saud University Corpus of Classical Arabic is also a single-medium corpus, with all texts originating from old manuscripts. As for the Corpus of Contemporary Arabic, the International Corpus of Arabic, and the arabiCorpus, the texts are drawn from multiple mediums, one of which is newspapers. Additionally, while some corpora clearly specify the text mediums, others such as the corpora collected automatically from the Internet, do not provide any such information; these include the Leeds Internet Arabic Corpora and the arTenTen12 corpus.

As for the text domains of the 14 reviewed corpora, the available information is similar to that for the text mediums. On the one hand, some of the corpora provide specific information regarding their text domains, such as the 2012 Arabic Newspapers corpus, the King Saud University Corpus of Classical Arabic, and the Corpus of Contemporary Arabic. On the other hand, there is a lack, or total absence, of text domain information for other corpora, such as the MSA corpus and the arTenTen12 corpus.

2.7 Corpus representativeness and balance

The problems of corpus representativeness and balance are arguably two of the most important issues of corpus design. A number of studies have analyzed these issues (Sinclair 1991; Atkins et al. 1992; Biber 1993; Biber et al. 2002; Sinclair 2005), and the representativeness of well-known corpora such as the British National Corpus (BNC) has been questioned (Ahmad 2008). Generally speaking, the design criteria

used to sample and collect corpora data define the extent to which they are representative and balanced for the designated research question(s). As a result, the representativeness of corpora can vary greatly depending on their design criteria, and this can affect their relative importance in a particular field of study. It is beyond the scope of this paper to discuss or argue these concepts further, but it is nonetheless worth noting the meaning of these concepts and how they relate to Arabic corpora.

The purpose of a corpus governs its design, and hence its representativeness and balance. A representative corpus must be able to answer the research question(s) under investigation. Biber (2002:246) argues that "a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research." In this sense, the researcher must be aware of the design and contents of a corpus before deciding to use it.

The balance of a corpus refers to the range of text genres it contains and how those genres are sampled. However, the definition of "balanced corpus" varies from researcher to researcher, making it difficult if not impossible to create a truly balanced one. For this reason, Teubert and Cermáková (2007) suggest the "opportunistic" (or cannibalistic) corpus as an alternative to the reference corpus. The opportunistic corpus has two main characteristics, an effort to collect as many texts as possible, and the comprehensive documentation of those texts. These characteristics allow us to tackle the problem of corpus representativeness and balance from a different perspective according to our research interests because they provide the information necessary to select texts that sufficiently match the scope of our research.

For a similar reason, it is difficult to evaluate the representativeness and balance of the Arabic corpora reviewed here without first specifying a particular research question that is related to either linguistics or NLP. It is the author's view that there is no fully representative corpus for any language, but rather corpora that are more representative than other corpora for specific research question(s). In general, all the corpora assembled for the purpose of NLP, such as the 2012 Arabic Newspapers Corpus and Alwatan-2004, can be used for text classification and clustering. Moreover, because most such corpora comprise newspaper content, they can also be used to study the MSA used in Arabic newspapers. Meanwhile, large corpora, such as the arTenTen12 corpus and the Leeds Internet Arabic Corpora, can be used for general linguistics research or for inferring the general patterns of the Arabic language for NLP applications. Unfortunately, however, these corpora cannot be used for constructive studies of Arabic language varieties across time, regions, or mediums because they do not provide the required information or tools to conduct such studies.

2.8 Availability

The availability of any language resource to its research community will increase its value to the community while also revealing its strengths and weaknesses. It is

notable that all of the corpora reviewed in this study are available to download for exploration either free of charge or for a fee. Previously, most Arabic corpora were not available or only available for a fee (Al-Sulaiti and Atwell 2006).

The 14 reviewed corpora can be classified into four categories. The first category comprises corpora that are free (i.e., at no charge) to download, such as the Corpus of Contemporary Arabic and the 2012 Arabic Newspapers Corpus. Note that most of the reviewed corpora are within this category, enabling users to use them as they wish. The second category contains corpora that are only free (i.e., at no charge) to explore via the websites through which they are available. This category comprises three corpora, namely, the arabiCorpus, the International Corpus of Arabic, and the Leeds Internet Arabic Corpora.

The third category contains corpora that require a fee to download. Only one corpus falls within this category, the Arabic Gigaword Corpus Fifth Edition. The fourth category contains corpora that can only be explored by paying a fee via the Sketch Engine website,[1] namely, the arTenTen12 corpus and the King Saud University Corpus of Classical Arabic. Note, however, that the latter is also available as a free download.

2.9 Tagging, tokenization and lemmatization

Different kinds of annotation can be applied to corpora at different levels to enrich their contents and leverage their value. Grammatical annotation, i.e., POS tagging, is one of the most important annotation types. Six out of the 14 reviewed corpora are either fully or partially POS tagged. The four fully POS tagged corpora are the International Corpus of Arabic, which was tagged based on the Tim Buckwalter Arabic morphology analyzer tag set, the Alwatan-2004 corpus, which is tagged via the KALIMAT Corpus (El-Haj and Koulali 2013) using the Stanford Arabic POS tagger (33 POS tags), the King Saud University Corpus of Classical Arabic, which is tagged using the MADA system (Morphological Analysis and Disambiguation for Arabic) (34 POS tags) (Roth et al. 2008), and the Leeds Internet Arabic Corpora which is tagged using AMIRA system (23 POS tags) (Diab 2007).

The two corpora that are partially tagged are the arTenTen12 corpus and the Arabic Gigaword Corpus Fifth Edition. Two samples of the arTenTen12 corpus (29,884,791 words and 115,315,274 words) were tagged using the Stanford Arabic POS tagger and MADA system respectively, while part of the Arabic Gigaword Corpus Fifth Edition is included in the Pennsylvania Arabic Tree Bank (ATB), which is POS tagged based on the Tim Buckwalter Arabic morphology analyzer tag set.

One corpus, the Arabic Learner Corpus v2, is annotated with a different type of information and is partially tagged with learners' errors. With the exception of the ATB, all of these annotated corpora were released in 2013 due to an increased interest in Arabic POS tagging research and the availability of such tools.

The presence of tokenization and lemmatization varies among Arabic corpora. Among the reviewed corpora, three were tokenized. These are the ATB part of the Arabic Gigaword Corpus Fifth Edition, the part of the arTenTen12 corpus

[1] http://www.sketchengine.co.uk/.

(29,884,791 words) that was tagged with the Stanford Arabic POS tagger and the Leeds Internet Arabic Corpora.

Among the reviewed corpora, three were lemmatized. They include the ATB part of the Arabic Gigaword Corpus Fifth Edition, the King Saud University Corpus of Classical Arabic and the part of the arTenTen12 corpus (115,315,274 words) that was tagged with the MADA system.

2.10 Tools and systems

Regardless of their perceived value, corpora are useless without the availability of tools and systems to adequately explore them and provide statistical distributions of their content, such as word and N-gram frequencies and related concordance information. POS tagging information is also of great importance. Additionally, the Arabic language lends itself to special consideration for morphological analysis due to the richness and complexity of its morphology.

In the case of corpora available for free download over the Internet, researchers have the freedom to choose the most suitable tools to manage and manipulate the corpus under consideration. Unfortunately, most currently available standalone corpora processing tools do not take into account the right-to-left writing direction of Arabic, yielding inaccurate concordance results. For this reason, Roberts et al. (2006) released aConCorde[2] as an open source processing tool for Arabic. The aConCorde application provides both frequency lists and concordance values of the target corpus. More recently, the KACST Arabic Corpora Processing Tool "Khawas" ("غواص," or "diver" in English)[3] was released as a Java-based open source processing tool for Arabic corpora. Khawas provides several functions and options for corpus processing that are not available in aConCorde, such as N-gram frequency and concordance, word and N-gram collocations, and corpora comparisons. More details regarding the functions and options available in Khawas can be found in Al-Thubaity et al. (2013).

As for the other corpora available for search and analysis via their respective websites, each case is different, providing its own unique tools and functions. The Leeds Internet Arabic Corpora, for example, enables the user to list the concordance of a single word or a sequence of words with an option to examine the texts in which the search terms appear. Collocations of the search terms are also provided based on different measures, such as Mutual Information, T-Score, and Log-likelihood Score. Additionally, the website for the Leeds Internet Arabic Corpora enables users to explore other corpora as well, namely the Al Hayat corpus, the Wikipedia corpus, the Corpus of Contemporary Arabic, the Arabic Legal Corpus, and the Computer Science Corpus of Arabic.

The arabiCorpus provides the concordances (citations) of single words according to four POS tags (noun, verb, adjective, and adverb) and strings. The corpus itself is not actually tagged, however, because the developers designed the search function to predict these tags based on morphological features. It also provides possible word

[2] http://www.andy-roberts.net/coding/aconcorde.

[3] https://sourceforge.net/projects/kacst-acptool/.

forms and collocations according to the frequency counts of the word forms that appear within four-word windows on both sides of the target word. The users of the arabiCorpus must first be made aware of the tagging and word form errors that occur regularly.

Both the arTenTen12 corpus and the King Saud University Corpus of Classical Arabic can be accessed through the Sketch Engine website. Sketch Engine provides several useful features to access corpora, such as those found in the websites of the Leeds Internet Arabic Corpora and the arabiCorpus, and moreover it provides additional distinct features, such as the option to perform a Word Sketch or Thesaurus function on a chosen word or lemma.

The International Corpus of Arabic offers concordance information of searched words based only on exact search terms, roots, lemmas, and stems. The search results can then be refined based on country, POS tags, stem patterns, and some morphosyntactic features such as number, gender, and definiteness. The corpus website does not provide any other information, such as frequency lists or collocation measures, but its ability to refine search results based on stem patterns is a unique feature of the site.

As illustrated in the above subsections, several positive and negative characteristics can be inferred from the evaluation of the Arabic corpora summarized in Tables 1 and 2. First, with respect to positive characteristics, compared with the situation described in Al-Sulaiti and Atwell (2006), the overall situation changed in 2013, when several corpora became freely available for different purposes. Second, additional options are now available for exploring these corpora, either via standalone software systems or through websites. Third, the sizes of corpora have expanded and should only become larger with time. Fourth, several corpora are entirely or partially POS tagged.

As for the negative characteristics, the website interfaces for all of the reviewed corpora are in English, which is a drawback that limits some Arab linguists from effectively using these corpora. Second, the contents of larger corpora are not classified according to their genres, times, or locations. This lack of information limits the usability of these corpora because contrastive studies cannot be performed, and language change cannot be accurately monitored. Third, and most importantly, the design criteria for many of the corpora are unclear, which in effect limits the user's ability to accurately evaluate the findings of any research based on them, and, moreover, no justification is given for excluding such criteria. Note, however, that the International Corpus of Arabic could be the sole exception. The following sections describe and discuss the design and construction of the KACST Arabic Corpus, which has overcome some of these drawbacks of existing Arabic corpora.

3 KACST Arabic corpus design

Defining the purpose of the corpus is the first step in corpus design; the purpose will guide and outline the design criteria and thereby the corpus construction itself. The stated purpose of the KACST Arabic Corpus project is to develop a free access,

large-sized, and sufficiently diverse Arabic corpus to represent the many varieties of Arabic language across three main dimensions: time, region, and genre. Such a corpus could be used for different research interests, beginning with linguistic studies at various levels and extending to the development of NLP applications.

3.1 KACST Arabic corpus design criteria

The following guidelines summarize the criteria used to create a sample KACST Arabic Corpus based on Sinclair (2005). It should be noted that the following criteria are interrelated and that each complements the others to increase the diversity of the corpus.

a. *Corpus size* 700 million words in the sample phase, with increases planned as required in subsequent phases.
b. *Corpus languages* The Arabic language and its varieties spanning the three main levels: time, region, and genre. Currently, *Fusha* "الفصحى" (both Classical Arabic and MSA) is only considered and not any other Arabic dialects because there is currently no standard system for writing the many dialects of Arabic, and, moreover, most are not used in written form.
c. *Text mode* Written text, because it is easy to capture, and most of the useful knowledge is codified in written texts. The inclusion of spoken materials is also important, however, and thus such materials will be considered in the next phase of the project.
d. *Sampling size* Full texts, because sampling full texts will maximize the chance of capturing the most linguistic features and hence will provide better research results. Additionally, recent advances in computer hardware and software now allow for the storing and processing of large amounts of data without difficulty.
e. *Text dates* The period from before Islam up to the present (more than 1,500 years), so that the many varieties of Arabic, from Classical Arabic to MSA, and the transience between these two main forms, can be represented.
f. *Text locations* Mainly from the Arabic region, but also Arabic publications from other regions because some newspapers and books are published outside of geographically Arabic countries.
g. *Text medium* Ten mediums spanning the time period, namely Old Manuscripts, Books, Newspapers, Magazines, Curricula, University Theses, Websites, Refereed Periodicals, Official Prints, and News Agencies.
h. *Text domains and topics* The appropriate domains and topics for each time period and medium. For example, the topic of sports news within the news domain is appropriate for the medium of newspapers in recent years. More details regarding the distribution of domains and topics are given in the following subsection.

The first challenge in the construction of such a wide-ranging corpus is the issue of copyright. The corpus content can be classified into two classes. The first class of content has no copyright and contains texts from the era before Islam up to 1950; as such, these texts are provided freely to download. This content includes a large quantity of Books, Curricula, University Theses, Refereed Periodicals, and Official

Table 5 The designed distribution of corpus content in words across the time periods

Time period (Gregorian)	Time period (Hijri)	Number of words	%
0–600	Before Islam	700,000	0.1
600–700	0–99	1,400,000	0.2
700–800	100–199	1,400,000	0.2
800–900	200–299	7,000,000	1.0
900–1000	300–399	7,000,000	1.0
1000–1100	400–499	7,000,000	1.0
1100–1200	500–599	7,000,000	1.0
1200–1300	600–699	7,000,000	1.0
1300–1400	700–799	7,000,000	1.0
1400–1500	800–899	7,000,000	1.0
1500–1600	900–999	7,000,000	1.0
1600–1700	1000–1099	7,000,000	1.0
1700–1800	1100–1199	7,000,000	1.0
1800–1900	1200–1299	7,000,000	1.0
1900–1980	1300–1399	28,000,000	4.0
1980–1990	1400–1409	52,500,000	7.5
1990–2000	1410–1419	98,000,000	14.0
2000–2010	1420–1430	164,500,000	23.5
2011–2013	1431–1434	276,500,000	39.5
Total		700,000,000	100

Prints. The second class contains texts that may have active copyrights, namely Newspapers, News Agencies, Magazines and Websites. For all of these text mediums, the following courses of action have been decided: (1) to not distribute the collected texts; (2) to not allow downloads of the texts; (3) to not allow previews of the full texts, but only the contexts of individual words, i.e., the 15 words before and 15 words after the node word; and (4) to provide bibliographic information of the corpus content. This is because the corpus was intended for research purposes, and using these materials according to the above restrictions is consistent with current Saudi copyright law.[4,5]

3.2 Designed text distribution of the KACST Arabic corpus

The design of the distribution of the corpus materials across all time periods, mediums, domains, and topics is influenced by the information and knowledge production available across those time periods as well as the available text mediums in each time period. Taking into consideration the corpus design criteria, the number

[4] http://www.mci.gov.sa/LawsRegulations/SystemsAndRegulations/IntellectualPropertySystem/Pages/default.aspx.

[5] http://fikratech.kacst.edu.sa/Invention-World/Copyright.aspx.

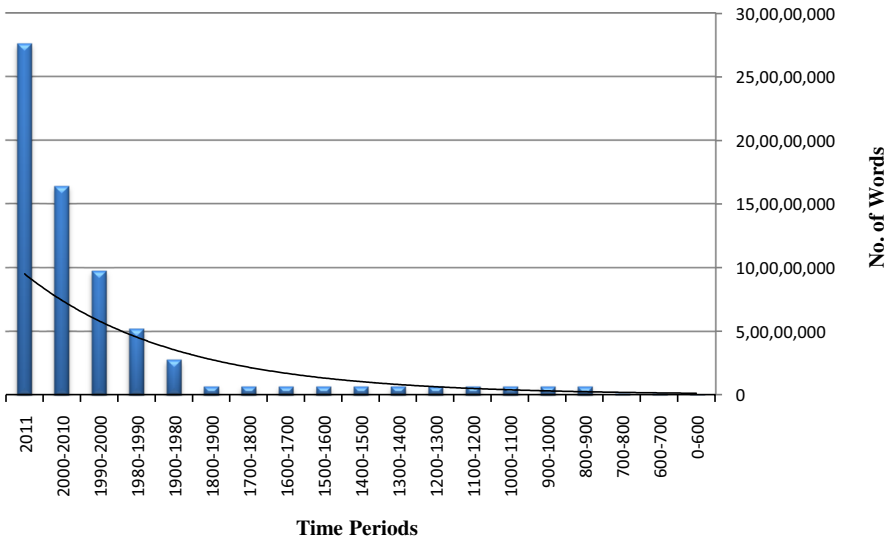

Fig. 1 The designed distribution of corpus content in words across the time periods

of words to be included in each time period, medium, domain, and topic were calculated.

The calculation of content distribution across time periods was inspired by the fact that the amount of information and codified knowledge gradually expands over time. The exponential growth rate is hence used to reflect this fact, whereby the amount of content increases with time across the corpus time periods. This distribution of content (words) across the time periods is shown in Table 5 as well as graphically illustrated in Fig. 1.

For each time period, the appropriate mediums were chosen. Old Manuscripts, for example, was the appropriate medium for earlier time periods, whereas mediums such as Newspapers, Magazines, and Websites were appropriate for the more recent time periods. The content distribution across the 10 medium types is shown in Table 6 and graphically illustrated in Fig. 2. Furthermore, the content distribution across the time periods and the 10 medium types is shown in Table 7.

The texts of each of the 10 medium types are first classified into domains related to the medium, and these texts are then further classified into topics under each domain. The medium of Old Manuscripts, for example, is classified into 16 domains, one of which is Islamic Jurisprudence "الفقه الإسلامي." Islamic Jurisprudence is further classified into eight topics, namely, Maliki Jurisprudence "الفقه المالكي", Hanafi Jurisprudence "الفقه الحنفي", Shafi'i Jurisprudence "الفقه الشافعي", Hanbali Jurisprudence "الفقه الحنبلي", Dhaheri Jurisprudence "الفقه الظاهري", Zaidi Jurisprudence "الفقه الزيدي", Ithna Ashariyyah Jurisprudence "الفقه الاثنا عشري" and Abadhi Jurisprudence "الفقه الإباضي". Table 8 through 15 illustrate the domains and topics for each of the 10 medium types in the corpus. The number of words assigned to each time period and medium is distributed evenly over both the domains and topics. It is worth noting that texts for each domain and topic are available for all assigned time periods. The medium of Old

Table 6 The designed distribution of corpus content in words across the 10 types of mediums

Medium	Number of words	%
Newspapers	241,500,000	34.5
Magazines	140,000,000	20
Books	126,000,000	18
Old manuscripts	80,500,000	11.5
Official prints	42,000,000	6
University Theses	21,000,000	3
Refereed periodicals	21,000,000	3
Websites	14,000,000	2
Curricula	7,000,000	1
News agencies	7,000,000	1
Total	700,000,000	100

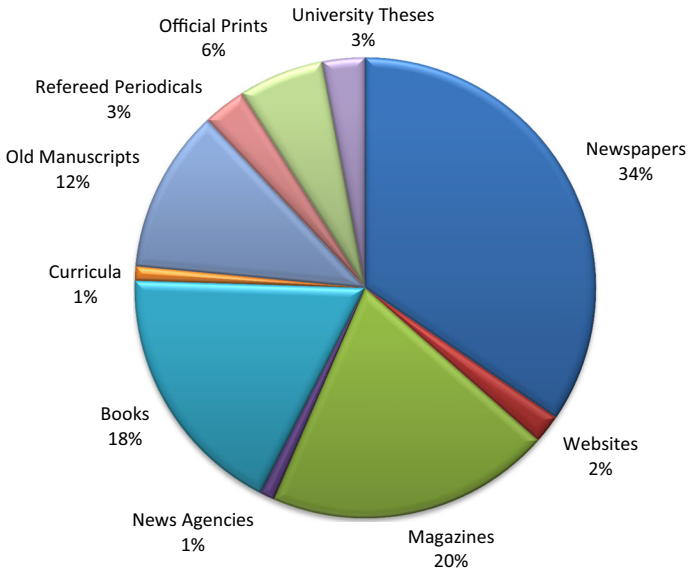

Fig. 2 The designed distribution of corpus content across the ten types of media

Manuscripts, for example, covers the time period of pre-Islam, but not all domains and topics are covered in that period; texts were available from the literature domain only, and specifically only from the topic of poetry from within that domain.

The distribution of texts across regions was difficult to specify because no data are available regarding the publishing activity of each country over time, and the availability of freely downloadable materials varies dramatically from country to country. While it was easy to find freely downloadable text from Saudi Arabia, it was difficult to do so from Mauritania and Sudan, for example.

Table 7 The designed content distribution across the time periods and 10 types of mediums

	Time periods										
	Pre-Islam	600–700	700–800	800–900	900–1000	1000–1100	1100–1200	1200–1300	1300–1400	1400–1500	1500–1600
Medium											
Old manuscripts	0.1	0.2	0.2	1	1	1	1	1	1	1	1
Newspapers	0	0	0	0	0	0	0	0	0	0	0
Magazines	0	0	0	0	0	0	0	0	0	0	0
Books	0	0	0	0	0	0	0	0	0	0	0
Curricula	0	0	0	0	0	0	0	0	0	0	0
University Theses	0	0	0	0	0	0	0	0	0	0	0
Refereed periodicals	0	0	0	0	0	0	0	0	0	0	0
Official prints	0	0	0	0	0	0	0	0	0	0	0
News agencies	0	0	0	0	0	0	0	0	0	0	0
Websites	0	0	0	0	0	0	0	0	0	0	0
Total percentage	0.1	0.2	0.2	1	1	1	1	1	1	1	1
Total number of words	700,000	1,400,000	1,400,000	7,000,000	7,000,000	7,000,000	7,000,000	7,000,000	7,000,000	7,000,000	7,000,000

	Time periods									
	1600–1700	1700–1800	1800–1900	1900–1980	1980–1990	1990–2000	2000–2010	2011–2013	% of Medium	No. of words
Medium										
Old manuscripts	1	1	1	0	0	0	0	0	11.5	80,500,000
Newspapers	0	0	0	1.5	3	5	10	15	34.5	241,500,000
Magazines	0	0	0	1	2	4	5	8	20	140,000,000
Books	0	0	0	1	2	3	5	7	18	126,000,000
Curricula	0	0	0	0	0	0	0	1	1	7,000,000

Table 7 continued

	Time periods									
	1600–1700	1700–1800	1800–1900	1900–1980	1980–1990	1990–2000	2000–2010	2011–2013	% of Medium	No. of words
University Theses	0	0	0	0	0	0.5	1	1.5	3	21,000,000
Refereed periodicals	0	0	0	0	0	0.5	1	1.5	3	21,000,000
Official prints	0	0	0	0.5	0.5	1	1.5	2.5	6	42,000,000
News agencies	0	0	0	0	0	0	0	1	1	7,000,000
Websites	0	0	0	0	0	0	0	2	2	14,000,000
Total percentage	1	1	1	4	7.5	14	23.5	39.5	100	700,000,000
Total number of words	7,000,000	7,000,000	7,000,000	28,000,000	52,500,000	98,000,000	164,500,000	276,500,000		700,000,000

Table 8 Domains and topics of old manuscripts

Domains	Topics
Principles of Islamic Jurisprudence	Maliki, Hanafi, Shafi'i, Hanbali, Zaidi, Ithna Ashariyyah, Abadhi, Dhaheri
Islamic Jurisprudence	Maliki, Hanafi, Shafi'i, Hanbali, Zaidi, Ithna Ashariyyah, Abadhi, Dhaheri
Islamic Faith	Sunni, Shi'I, Zaidi, Abadhi
Non-Islamic Faith	Christianity, Judaism, Sabean
Hadeeth (Prophetic Traditions)	Sunni, Shi'I, Zaidi, Abadhi
Science of Hadeeth	Sunni, Shi'I, Zaidi, Abadhi
Interpretation of Quran	Sunni, Shi'I, Zaidi, Abadhi
Science of Quran	Sunni, Shi'I, Zaidi, Abadhi
Politics and judiciary	Islamic politics, politics, judiciary
Islamic morals	Asceticism, Morals, Sermons
Biography	Biography of the Prophet Mohammad, Biography
Arabic linguistics	Grammar, morphology, semantics, rhetoric, dictionaries, alarowdh
Social sciences	History, geography, genealogy
Natural sciences	Physics, chemistry, mathematics, astronomy, medicine
Culture and literature	Poetry, literature, travel literature
Philosophy	Islamic philosophy, Greek philosophy

3.3 KACST Arabic corpus text metadata

The basic component of any corpus is the text. It is important that as much information as possible is available about the text contents, so that researchers can study the language and its many varieties in both general and specific ways, and across many different levels, thereby allowing for more accurate language models to be constructed. For the current KACST Arabic Corpus, the following metadata were assigned to each text: title, year of publication, time period, author name and gender, region, medium, domain, and topic.

The availability of such metadata is what distinguishes the KACST Arabic Corpus from other large Arabic corpora such as the arTenTen12 corpus, the Leeds Internet Arabic Corpora, and the arabiCorpus. This metadata allows the corpus user to restrict her/his study to specific time periods, regions, mediums, domains, or topics, or to work on all of these in a variety of possible combinations.

Additionally, large files are divided into smaller parts having a maximum of 5,000 words for ease of access and processing in the database, and also to allow future users to build their own corpus based on either the KACST Arabic Corpus full text, or only selected parts of the text.

The design criteria illustrated in the above subsections serves to reinforce the representativeness and balance of the KACST Arabic Corpus through the following interrelated factors:

a. Large corpus size (700M words);

b. Time span covered (from the period before Islam up to almost present day), and how the corpus materials are distributed across time (and also showing exponential growth);
c. Wide diversity of texts content covering 10 mediums, 80 domains, and 481 topics;
Inclusion of corpus texts from all Arab countries.

In conclusion, it is important to emphasize that the design criteria given above are subjective because no strict guidelines are currently available for creating a fully representative and balanced corpus. However, general guidelines for constructing such a corpus are available. Obviously, the availability of clear design criteria benefits not only the process of corpus compilation, but also the tasks of accurately validating and evaluating the results obtained through the corpus by its users.

4 KACST Arabic corpus construction

The given design criteria for the KACST Arabic Corpus posed a difficult challenge because of its sheer size, the long time period covered, and the inclusion of three classification levels of text (medium, domain, and topic). As far as it can be ascertained, no current Arabic corpus explicitly has these design criteria. The following subsections illustrate in more detail the sources of texts and the challenges posed by the process of compiling the KACST Arabic Corpus; specifically, the following describes what has already been achieved in relation to the proposed design goals mentioned above, and a brief description of the technology currently being used to store and manage the corpus.

4.1 Text sources

The main source of the KACST Arabic Corpus texts is the Internet. Several Arabic web sites provide free-to-download, machine readable format texts covering numerous mediums. The corpus texts have been collected from numerous sources. The resources mentioned hereafter are examples of the main sources only.

Most of the Old Manuscripts texts were downloaded from the *Alshamilah*[6] web site. The Books texts were mainly gathered from *Alshamilah, Saaid Alfawaeed*,[7] and the Arab Writers Union in Syria.[8] The Refereed Periodicals texts were mainly collected from several Arab university journal websites, such as those of Umm Alqura University,[9] and King Faisal University,[10] and also from the Arab Writers Union in Syria. The University Theses texts were gathered from different Arab university websites and *Alshamilah.*

[6] http://shamela.ws/.

[7] http://saaid.net/.

[8] http://www.awu.sy/.

[9] https://uqu.edu.sa/page/ar/518.

[10] http://www.kfu.edu.sa/ar/departments/sjournal/Pages/Home.aspx.

The Official Prints texts were mainly collected from sites that specialize in law and regulation, such as United Nations agencies and governmental websites such as the Bureau of Experts in Saudi Arabia[11] and the Arab Legal Portal.[12] The Curricula texts were mainly collected from university websites, the websites of the Arab Ministries of Education, and educational websites.

The Newspapers, Magazines, and News Agencies texts were gathered from the websites of the designated websites of those mediums, such as the Saudi newspaper *Alwatan*,[13] the Egyptian magazine *Rosa Alyousof*,[14] and the Saudi Press Agency.[15]

Most of the corpus texts were classified according to their given classification on the websites from which they were downloaded; using this method, 75 % of the corpus was classified into Old Manuscripts, Refereed Periodicals, Official Prints, Newspapers, Magazines, and News Agencies. The rest (25 %) of the texts were then classified manually. The tasks of text collection and classification were outsourced to an external party.

4.2 Difficulties and challenges

Collecting texts according to the corpus design criteria revealed several difficulties and challenges. The four main challenges were as follows:

a. It was difficult to locate machine-readable texts from certain time periods for some of the domains and topics, such as the time period from 1980 to 2000, for which it was difficult to find newspaper articles;
b. It was difficult to find a reasonable amount of texts from some Arabic geographical regions, such as Mauritania and Sudan;
c. Some of the domains and topics do not exist for all time periods, such as the natural sciences and philosophy with respect to the medium of Old Manuscripts, or the domains and topics of the medium of Official Prints;
d. We were unable to include newspapers from all Arab countries, either because of:
    1. The absence of a means of classifying the newspaper content;
    2. The lack of permission by some newspapers to automatically collect text from their websites; or
    3. Because of recent civil unrest, i.e., what is commonly referred to as the "Arabic Spring," some newspaper websites for countries such as Libya and Syria are no longer available.

Because of these and other challenges, the proposed design criteria with respect to the regions of coverage and some of the topics were not fully matched.

[11] http://www.boe.gov.sa/MainLaws.aspx?lang=en.

[12] http://www.arablegalportal.org/.

[13] http://www.alwatan.com.sa.

[14] http://rosa-magazine.com/.

[15] http://www.spa.gov.sa/.

4.3 Distribution of collected texts across time and mediums

The KACST Arabic Corpus contains more than 731 million words from 869,800 texts. The available number of words exceeds the original corpus criteria by more than 32 million. Showing the actual distribution of the KACST Arabic Corpus content across time periods, regions, mediums, domains, and topics will yield a more detailed illustration of its content that should be of interest to a wide variety of researchers. This subsection will describe the KACST Arabic Corpus content based on word counts across the three main dimensions of time period, region, and medium.

The actual distribution of texts and words of the KACST Arabic Corpus across the time periods is shown in Table 16. The table shows that the number of texts belonging to the time period from pre-Islam until the year 1800 is less than that of the subsequent time periods. This is due to using the full texts as the default sample text size, and because all of the texts from this time period are Old Manuscripts that are relatively lengthy compared with more recent texts. The texts from the other time periods are mostly from Newspapers and Magazines, and thus they are short in comparison to the texts of other mediums, such as Old Manuscripts and Books.

The data show that the actual distribution of words across the time periods was achieved for most time periods, and even exceeded the targeted number of words. However, the targeted number of words was not achieved for some of the time periods, namely the periods from pre-Islam until 800, from 1980 until 2000, and the most recent period from 2011 to 2013.

It was difficult to find a sufficient number of texts for the first three time periods: pre-Islam, 600–700, and 700–800. Regarding the pre-Islam time period, the only available documents were old Arabic poetry texts, which are limited in length. As for the other two periods, the main texts consist of documentation of the sayings of the prophet Mohammad and his *sahabies'* interpretations of these sayings. These documents are also limited in their length. As for the time periods of 1980–1990 and 1990–2000, the largest amount of content was based on Newspapers and Magazines that met the design criteria; it was difficult to find downloadable texts in these two mediums for these two time periods.

For the time period 2011–2013, however, the case is different. The shortage of words is because of the unavailability of texts from the mediums of Books, University Theses, Refereed Periodicals, and Curricula. Securing the available texts from these mediums for this time period required more time than expected.

The actual distributions of the texts and words of the KACST Arabic Corpus across mediums are shown in Table 17. The data show that the desired number of words according to the original design criteria was achieved for all mediums except for Newspapers, Magazines, and Official Prints. As mentioned above, it was difficult to find downloadable texts for the mediums of Newspapers and Magazines for the time period 1980–2000. The difference between the original design criteria and the actual content for the Official Prints domain is large; most of the texts are unfortunately in PDF format, making it very difficult to convert the text into easily accessible formats in Arabic.

Efforts to search for more texts to bridge these gaps between the original design criteria and the actual content of the KACST Arabic Corpus with respect to the abovementioned time periods and mediums will continue, and additional texts will be added accordingly.

4.4 KACST Arabic corpus website tools

A relational database was modeled and designed for the KACST Arabic Corpus, which is currently freely available to explore via the Internet (http://www.kacstac.org.sa). The tasks of website design and programming were outsourced to an external party.

The website provides several basic tools, such as:

a. Frequency distribution of words across time periods and mediums in both tabular and graphical formats;
b. Concordance of a given word with flexibility to change the concordance window to be any value between 1 and 15 words from both left and right. The user can also use different filters to restrict the concordance results based on time period, region, medium, domain, and topic. Citation information about the concordance results is also provided;
c. Text title search using metadata information about the texts and their frequency lists. The same filters used for the concordance tools can also be used for this text search;
d. Frequency list of the 200,000 most frequent words of the corpus, with the ability to specify the frequency range;
e. All displayed information can be saved to external files in either TXT or PDF format.

5 Conclusion

The design and construction of corpora is not only costly but also time consuming, and thus any efforts to make corpora more freely available to the research community are both useful and worthwhile. This paper briefly reviewed 14 Arabic corpora based on a number of criteria, including purpose, language, text mode, size, text date, location, text type/medium, text domain, representativeness, and balance. Additionally, the review included information regarding the current availability of the reviewed corpora, tagging, and whether there are any tools available to explore them in detail. The review also showed that there is a need for a freely available Arabic corpus constructed using clearly defined design criteria and associated information related to its content, and that is sufficiently large and diverse for current as well as future research needs. The KACST Arabic Corpus was thus designed and constructed to fulfill these needs.

Using the review criteria used to assess the 14 corpora reviewed in this paper, the KACST Arabic Corpus assessment can be summarized as follows:

Table 9 Domains and topics of books, university theses, and referred periodicals

Domains	Topics
Principles of Islamic Jurisprudence	Maliki, Hanafi, Shafi'i, Hanbali, Zaidi, Ithna Ashariyyah, Abadhi, Dhaheri
Islamic Jurisprudence	Maliki, Hanafi, Shafi'i, Hanbali, Zaidi, Ithna Ashariyyah, Abadhi, Dhaheri
Islamic Faith	Sunni, Shi'i, Zaidi, Abadhi
Non-Islamic Faith	Christianity, Judaism, Sabean
Hadeeth (Prophetic Traditions)	Sunni, Shi'i, Zaidi, Abadhi
Science of Hadeeth	Sunni, Shi'i, Zaidi, Abadhi
Interpretation of Quran	Sunni, Shi'i, Zaidi, Abadhi
Science of Quran	Sunni, Shi'i, Zaidi, Abadhi
Biography	Biography of the Prophet Mohammad, Biography, …Diary
Arabic linguistics	Grammar, morphology, semantics, rhetoric, dictionaries, alarowdh, modern linguistics
Social sciences	History, geography, economy, psychology, sociology, media, management, politics, law, education
Natural sciences	Physics, chemistry, mathematics, biology, geology, astronomy, medicine, environment
Applied sciences	Civil engineering, mechanical engineering, electrical engineering, chemical engineering, aeronautical engineering, industrial engineering, nuclear engineering, biomedical engineering, information technology and communications technology
Culture and literature	Poetry, literary criticism, the story, the short story, the novel, theater, cinema and plastic arts
Philosophy	Islamic philosophy, Philosophy of science, Philosophy of mind, Epistemology, Philosophy of language
Law	Political, economic, social, security, educational and commercial

Table 10 Domains and topics of newspapers

Domains	Topics
News	Social, sports, economic, technical, cultural, political, general, scientific
Articles	Opinion, editorials, sports, social, economic, cultural, political, scientific, technical, general, hobbies, religious
Reportage	Social, sport, economic, cultural, political, scientific, technical, general, hobbies, religious

Corpus purpose General-purpose Arabic corpus that can be useful for linguistic studies at various levels and the development of NLP applications;
Language and mode A monolingual Arabic corpus for written *Fusha* "الفصحى" including both Classical Arabic and MSA, but not any Arabic dialects;
Size 700 million words;
Text period More than 1,500 years, covering the period from before Islam up to the present;

Table 11 Domains and topics of magazines

Domains	Topics
News	Social, sport, economic, technical, cultural, political, general, scientific
Articles	Opinion, editorials, sports, social, economic, cultural, political, scientific, technical, general, hobbies, religious
Reportage	Opinion, editorials, sports, social, economic, cultural, political, scientific, technical, general, hobbies, religious, security
Studies	Linguistics, social sciences, religious, literature, culture, economic, natural sciences, applied sciences, legal, philosophy, security, military

Table 12 Domains and topics of official prints

Domains	Topics
Laws and regulations	Political, economic, social, administrative, health, security, military, education and business
Reports	Political, economic, social, administrative, health, security, military, education and business

Table 13 Domains and topics of websites

Domains	Topics
Official	Social, sport, economic, technical, cultural, political, scientific, medical, religious
Personal	Social, sport, economic, technical, cultural, political, scientific, medical, religious

Table 14 Domain and topics of news agencies

Domains	Topics
News	Social, sport, economic, technical, cultural, political, general, scientific

Location Mainly from the Arabic region, but also Arabic publications originating in other regions;
Text mediums and domains Ten mediums total; the appropriate domains (and topics) for each time period and medium are considered. (See Tables 8, 9, 10, 11, 12, 13, 14, 15);
Representativeness and balance Efforts to achieve corpus representativeness and balance are realized in the overall corpus size, which is relatively large, the time span it covers, the regions covered, and the diversity of mediums, domains, and topics included;
Availability The corpus is free to explore and search only via its designated website, and it is not available for download or distribution by any means;

Table 15 Domains and topics of Curricula

Domains	Topics
Islamic Topics	Islamic Faith, Islamic Jurisprudence, Hadeeth, Quran interpretation
Arabic language	Grammar, morphology, rhetoric, literature
Social sciences	History, geography, psychology, sociology
Natural sciences	Physics, chemistry, mathematics, biology, geology

Tagging The corpus has not been tagged, tokenized or lemmatized yet;
Tools and systems The corpus website provides the basic analysis tools as described in Sect. 4.3.

In addition to its large size, three main features distinguish the KACST Arabic Corpus from other Arabic corpora: clearly stated design criteria; diversity across a very long period of time (around 1,500 years) as well as across geographical regions, mediums, domains, and topics; and the fact that each text in the corpus is classified according to its time period, region, medium, domain, and topic, so that

Table 16 KACST Arabic corpus content distribution across time periods

Time period (Gregorian)	Time period (Hijri)	No. of texts	Number of words	%		
				Original criteria	Actual	Difference
0–600	Before Islam	51	366,688	0.1	0.05	−0.05
600–700	0–99	20	596,416	0.2	0.08	−0.12
700–800	100–199	31	939,466	0.2	0.13	−0.07
800–900	200—299	54	8,242,722	1.0	1.12	0.12
900–1000	300–399	41	7,800,788	1.0	1.06	0.06
1000–1100	400–499	47	7,235,942	1.0	0.98	−0.02
1100–1200	500–599	56	9,881,847	1.0	1.34	0.34
1200–1300	600–699	41	12,012,654	1.0	1.63	0.63
1300–1400	700–799	50	11,987,918	1.0	1.62	0.62
1400–1,500	800–899	47	10,682,463	1.0	1.45	0.45
1,500–1600	900–999	16	12,005,838	1.0	1.62	0.62
1600–1700	1000–1099	15	7,788,684	1.0	1.05	0.05
1700–1800	1100–1199	57	12,366,525	1.0	1.67	0.67
1800–1900	1200–1299	299	30,851,745	1.0	4.17	3.17
1900–1980	1300–1399	1,818	37,885,968	4.0	5.13	1.13
1980–1990	1400–1409	142	15,237,940	7.5	2.06	−5.44
1990–2000	1410–1419	1,776	28,326,364	14.0	3.83	−10.17
2000–2010	1420–1430	570,290	392,001,661	23.5	53.69	30.19
2011–2013	1431–1434	294,949	125,464,722	39.5	17.33	−22.17
Total		869,800	731,676,351			

Table 17 KACST Arabic corpus content distribution across mediums

Medium	No. of texts	Number of words	%		
			Original criteria	Actual	Difference
Newspapers	587,058	243,740,488	34.50	34.08	−0.42
Magazines	208,501	132,812,342	20.00	18.25	−1.75
Books	592	129,132,457	18.00	18.03	0.03
Old manuscripts	1,323	132,561,082	11.50	17.76	6.26
University Theses	1,817	27,972,420	3.00	3.75	0.75
Refereed periodicals	3,564	23,400,458	3.00	3.22	0.22
Websites	39,794	15,717,433	2.00	2.15	0.15
Curricula	1,260	10,801,115	1.00	1.47	0.47
News agencies	25,336	8,681,402	1.00	1.21	0.21
Official prints	555	6,857,154	6.00	0.93	−5.07
Total	869,800	731,676,351			

users can use the corpus website tools to easily restrict their search queries based on all or any combination of these classifications.

Such availability and usability of this corpus will provide researchers with an unprecedented ability to decide if the corpus is suitable for their research needs, and to judge their findings while using possibilities offered by the corpus to explore new research topics, especially with respect to language contrasts and comparisons that have previously been virtually impossible using available Arabic corpora (Tables 16, 17).

This is not the final state of the KACST Arabic Corpus. Actual use of the corpus may reveal a need for further design modifications. Moreover, as part of the future work, it is planned to increase the corpus size to 1 billion words, and all necessary efforts will be taken to bridge the gap between the design and actual collection of texts across time periods, mediums, domains, and topics. Finally, new tools will be added to the corpus website, such as tools to allow the consideration of word N-grams and the many different forms of Arabic words owing to the productive nature of Arabic morphology.

Acknowledgments This project was fully funded by the King Abdulaziz City for Science and Technology via Grants Number (531-31) and (33-824). The author would like to thank the three anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

Abbas, M., & Smaili, K. (2005). Comparison of topic identification methods for Arabic language. In *International conference RANLP05: Recent advances in natural language processing*, 21–23 September 2005, Borovets, Bulgaria.

Abbas, M., Smaili, K., & Berkani, D. (2011). Evaluation of topic identification methods on Arabic corpora. *Journal of Digital Information Management, 9*(5), 185–192.

Abdelali, A., Cowie, J., & Soliman, H. (2005). Building a modern standard Arabic corpus. In *Workshop on computational modeling of lexical acquisition*. The Split Meeting. Croatia, July 25–28.

Ahmad, K. (2008). Being in text and text in being: Notes on representative texts. In G. Andeman, and M. Rogers (Eds.). *Incorporating corpora*. Clevedon: Multilingual Matters, pp. 60–91 (Chapter 5).

Alansary, S., Nagi, M., & Adly, N. (2007). Building an international corpus of Arabic. In *7th International conference on language engineering*, Cairo, Egypt, December 5–6.

Alfaifi, A., & Atwell, E. (2013). Arabic learner corpus v1: A new resource for arabic language research. In *Second workshop on Arabic Corpus Linguistics* (WACL-2), July 22.

Alrabiah, M., Al-Salman, A., & Atwell, E. (2013). The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. In *Second workshop on Arabic corpus linguistics* (WACL-2), July 22.

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. O., 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 *International conference on Asian Language Processing 2013* (IALP 2013), pp. 67–70.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1–16.

Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., & Suchomel, V. (2013). arTenTen12: A new, vast corpus for Arabic. In *Second workshop on Arabic Corpus Linguistics* (WACL-2), July 22.

Biber, D. (1993). Representativeness in Corpus design. *Literary and Linguistic Computing, 8*(4), 243–257.

Biber, D., Conrad, S., & Reppen, R. (2002). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Diab, M. (2007). Towards an optimal POS tag set for Modern Standard Arabic processing. In *Proceedings of recent advances in natural language processing* (*RANLP*), pp. 91–96.

Dukes, K., Atwell, E., & Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. In *Language resources and evaluation journal (LREJ). Special issue on collaboratively constructed language resources.*

El-Haj, M., & Koulali, R. (2013). KALIMAT a multipurpose Arabic Corpus. In *Second workshop on Arabic Corpus linguistics* (WACL-2), July 22.

Khorsheed, M. S., & Al-Thubaity, A. O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language Resources and Evaluation, 47*(2), 513–538.

Parker, R., et al. (2011). *Arabic Gigaword fifth edition LDC2011T11. Web Download.* Philadelphia: Linguistic Data Consortium.

Roberts, A., Al-Sulaiti, L., & Atwell, E. (2006). aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora, 1*(1), 39–60.

Roth, R., Rambow, O., Habash, N., Diab, M. & Rudin, C. (2008). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of the conference of American association for computational linguistics (ACL08).*

Saad, M. K., & Ashour, W. (2010). *OSAC*: Open source Arabic Corpora. 6th ArchEng international symposiums. In *EEECS'10 the 6th international symposium on electrical and electronics engineering and computer science*, pp. 118–123, European University of Lefke, Cyprus.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2005). Corpus and text—basic principles. In Wynne M. (Ed.) *Developing linguistic corpora: A guide to good practice*, pp. 1–16. Oxford: Oxbow Books. http://ahds.ac.uk/linguistic-corpora/. Accessed August 28, 2013.

Teubert, W., & Čermáková, A. (2007). *Corpus linguistics: A short introduction*. London: Continuum.