Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES

Alia Al-Sayed Ahmad[1], Bassam Hammo[2] and Sane Yagi[3]

(Received: 20-Jun.-2017, Revised: 19-Aug.-2017 and 04-Oct.-2017, Accepted: 07-Oct.-2017)

ABSTRACT

This study reports on the construction of a one million word English-Arabic Political Parallel Corpus (EAPPC), which will be a useful resource for research in translation studies, language learning and teaching, bilingual lexicography, contrastive studies, political science studies and cross-language information retrieval. It describes the phases of corpus compilation and explores the corpus, by way of illustration, to discover the translation strategies used in rendering the Arabic and Islamic culture-specific terms takfīr and takfīrī from Arabic into English and from English into Arabic. The Corpus consists of 351 English and Arabic original documents and their translations. A total of 189 speeches, 80 interviews and 68 letters, translated by anonymous translators in the Royal Hashemite Court, were selected and culled from King Abdullah II's official website, in addition to the textual material of the English and Arabic versions of His Majesty's book, Our Last Best Chance: The Pursuit of Peace in a Time of Peril (2011). The texts were meta-annotated, segmented, tokenized, English-Arabic aligned, stemmed and POS-tagged. Furthermore, a parallel (bilingual) concordance was built in order to facilitate exploration of the parallel corpus. The challenges encountered in corpus compilation were found to be the scarcity of freely available machine-readable Arabic-English translated texts and the deficiency of tools that process Arabic texts.

KEYWORDS

Parallel corpus, Political, English-Arabic translation, Corpus compilation, Challenges.

1. INTRODUCTION

A parallel corpus is a collection of original texts and their translations in a target language. These texts can be aligned at different levels, such as paragraph, sentence, phrase or word level. Bilingual concordances are used to display all the occurrences of a search term in the source language (SL) together with their equivalents in the target language (TL). This type of corpora plays a crucial role in research that involves two or more languages, such as machine translation, cross-language information processing, contrastive studies, language research, language learning and teaching and bilingual lexicography [1]-[6].

1.1 Parallel Corpora and Translation Studies

The benefits of parallel corpora to translation studies and translation activity are well recognized. Parallel corpora are important educational tools when they are used in translation training programs, as they provide learners with authentic examples and with a flow of language data, from which they can discover and learn the strategies employed by translators [6].

Parallel corpora play an important role in the development of machine translation (MT) systems. There are three approaches to machine translation: linguistic knowledge, statistical and computer-assisted [7].

The first approach is rule-based and depends on such linguistic knowledge as morphological, syntactic, semantic and idiomatic knowledge of the source and target languages. In this approach, the source sentence is translated into the target sentence using a parser. This parser analyses the source sentence into its components, such as NP, VP, AdvP and PP, then replaces them with their TL equivalents with

This paper is an extended version of a short paper that was presented at the international conference "New Trends in Information Technology (NTIT) 2017", 25-27 April 2017, Amman. Jordan.

1. A. Al-Sayed Ahmad is with the English Department, University of Jordan, Amman, Jordan. Email: aliaahmadsh@gmail.com.
2. B. Hammo is with the Department of CIS, University of Jordan, Amman, Jordan. Email: b.hammo@ju.edu.jo.
3. Sane Yagi is with the Department of Linguistics, University of Jordan, Amman, Jordan. E-mail: saneyagi@yahoo.com.

the help of a dictionary. Then, the output sentence is reorganized in accordance with the linguistic rules of the target language.

The second approach is statistical machine translation. It analyses a parallel corpus, selects the SL-TL patterns that coincide most frequently and uses them in the translation. For example, in such statistically-based systems as Google Translate, parallel corpora, monolingual corpora and statistical models of their data are heavily used to automatically render texts from one language into another [8].

The third approach is computer-assisted translation, which involves an interactive process between the machine and the translator. Many types of computer-assisted translation software (e.g., electronic dictionaries and translation memories) were developed to facilitate and automate the process of translation. Bilingual and multilingual electronic dictionaries contain information about SL and TL words, such as part of speech, pronunciation and collocations. These dictionaries can be found in different forms: special devices (e.g. Atlas Modern Dictionary English-Arabic), computer software (e.g. Golden Al-Wafi), smart phone applications (e.g. Britannica Dictionary), CD-ROMs and DVD-ROMs usually sold with the printed version (e.g. Oxford Elementary Learner's Dictionary with CD-ROM) and online dictionaries (e.g. https://www.merriam-webster.com/). These dictionaries have the advantage of swiftly finding a query term.

One of the most precious computer-assisted translation resources is a translation memory that a company can create for its translators. Translation memory (TM) is a repository of translated phrases, where SL text segments are aligned with their TL equivalents. When the translator activates the translation memory and starts to translate a new document, the TM would quickly offer him/her a suggested translation for any SL segment that matches a previously translated one in its database. Building a translation memory is a "process of comparing a source text and its translation, matching the corresponding segments and binding them together as translation units" [9]. TMs save the translator's time and effort, particularly in translating documents of highly repeated texts, such as legal contracts. Some websites, such as Glosbe (https://glosbe.com/), have numerous dictionaries and translation memories that offer the user access to parallel texts in different languages including Arabic and English.

Parallel corpora have proven to be useful in developing machine translation systems. They spare time and effort and contribute to the resolution of some translation challenges. Even though machine-translation output is occasionally dull and literal, parallel corpora can capture subtle meanings of the source text (ST), idiomatic expressions and metaphors [10].

1.2 Parallel and Comparable Corpora

Parallel and comparable corpora are two types of multilingual corpora. Both parallel and comparable corpora consist of texts in two or more languages, but the first requires that there be a source language and a translation version of the same texts. The latter, on the other hand, makes no such requirement. The texts that it contains are merely of the same sampling frame (i.e., the same text size from the same genres and published in the same period of time). Parallel and comparable corpora are invaluable to translation and contrastive studies [11].

1.3 Challenges of Compiling Parallel Corpora

Developing a parallel corpus is not a straightforward task. This is due to technical and linguistic challenges encountered at most stages of construction: text selection, conversion, segmentation, stemming, alignment and annotation [12]-[15].

In the text selection process, it is challenging to find a large number of translated open-access texts that are available in the desired language combination. These texts should be machine-readable, accessible and representative samples of the use of the specific language combination [12]-[13], [16].

Moreover, some languages, such as Arabic, suffers from the scarcity and/or inefficiency of tools that are used for text conversion (e.g., OCR), segmentation, tokenization and part of speech tagging [12]-[13], [17]-[18].

Minority languages and languages of technologically underdeveloped countries have to overcome the formidable challenge of automating texts alignment [17]. It is widely acknowledged that aligning a

large amount of texts is probably prohibitively expensive [19]. That is why it is difficult to find parallel texts of any significant size that are aligned at phrase level for any language combination. Although Arabic is not a minority language, it is lacking in reliable text-alignment and text-annotation tools [12]-[13], [17], [20].

The rest of this paper is organized as follows. The next section provides a brief overview of the existing English and Arabic parallel corpora. Section 3 presents the methodology of building the corpus. Section 4 gives information about the parallel concordance. Corpus experimentation is discussed in section 5. Concluding remarks are presented in Section 6.

2. LITERATURE REVIEW

2.1 English Parallel Corpora

Parallel corpora began to appear in 1988, when Bell Communications Research and the IBM T. J. Watson Research Center compiled the Hansard corpus; the first parallel corpus of French and English [19]. It included 50 million words collected from transcriptions of the Canadian Parliament debates between 1975 and 1988 [19]. Since then, many parallel corpora projects were initiated. The corpus of European Corpus Initiative (ECI) contains about 19 million words from French, English and Spanish texts; the English-Norwegian Parallel Corpus (ENPC) consists of two million tokens that were culled from original fiction and non-fiction English and Norwegian texts and their translations. The original texts were aligned with their translations at sentence level. This corpus was aimed at carrying out comparisons between original texts and their translations, originals in both languages, translations in both languages and originals and translations in one language [21]. Later, the ENPC was joined by the German-Norwegian, French-Norwegian and Russian and Norwegian parallel corpora to form the Oslo Multilingual Corpus [21].

In the wake of ENPC, other parallel corpora that included English have been compiled (e.g., the English-Swedish Parallel Corpus, the English-French corpus, the English-German corpus and the English-Spanish corpus) [21]. These parallel corpora followed the design criteria of ENPC and shared some of its English original texts [21]. The JRC-ACQUIS Multilingual Parallel Corpus included more than one billion tokens from 22 languages [22]. In addition, the Official Journal of European Community multilingual parallel corpus involved English-German, English-Italian, English-Spanish and English-French aligned combinations [22]. The Open Parallel Corpus (OPUS) consists of nearly 352 million tokens in sixty European and Asian languages including Arabic [23]. Perhaps, these EU corpora were geared more towards research in natural language processing (NLP) than in linguistics and translation studies.

The list of parallel corpora that include English is too long to cover here. This is not the case for Arabic however.

2.2 Arabic Parallel Corpora

Arabic parallel corpora began to come into existence only in the late 1990's when the English-Arabic Parallel Egypt Corpus was developed at John Hopkins University in 1999 for the purpose of facilitating machine translation [24]. It consisted of the Qur'an in English and Arabic. Then in 2004, the English-Arabic Parallel Corpus was compiled by Al-Ajmi [17]. It contained three million words that were collected from the Kuwaiti World of Knowledge book series. This is a series of translated books about a variety of topics in history, economics, arts, science and literature. In addition to the aforementioned OPUS, a parallel Spanish-Arabic corpus was built from the annual reports of United Nations institutions for the purpose of experimenting with alignment at sentence level [12]. Samy et al. reused tools made for the Spanish language with Arabic texts. One year later, Samy et al. added English texts from the United Nations documents and developed the Arabic-Spanish-English multilingual corpus [13]. The Quranic Arabic Corpus[1] [25] is excellent for illustrating morphologically annotated classical Arabic. The English-Arabic Parallel Corpus of United Nations Texts (EAPCOUNT) was compiled by Hammouda Salhi in 2013 [26]. Finally, the Linguistic Data Consortium at the University of Pennsylvania (LDC)2 developed several English Arabic parallel corpora from broadcast conversation

[1] http://corpus.quran.com/

(e.g., talk shows), broadcast news and news wires, which amount to around 40 million words.

Although these may appear like numerous parallel corpora, most of them are either proprietary, experimental or restricted to specific text sources (e.g. News, UN documents and the Holy Quran). They are also aligned at either paragraph or sentence level but not at phrase and word levels. Most of them are not POS-tagged, as shown in Table 1. There is a clear need for properly-annotated Arabic resources that translators, learners, educators, researchers and language engineers can use free of charge.

Table 1. Examples of English-Arabic parallel corpora.

	Egypt Corpus	**The English-Arabic Parallel Corpus**	**OPUS**	**Arabic-Spanish-English Parallel Corpus**	**Quranic Arabic Corpus**	**EAPCOUNT**	**LDC**
Size (words)	77,430	3 million	352 million	3 million	77,430	5.5 million	40 million
Availability	×	×	✓	✓	✓	×	×
Medium	Written	Written	Written	Written	Written	Written	Written and spoken
Source	The Holy Quran	The World of Knowledge (a series of translated books)	OpenOffice.org documentation KDE manuals including KDE system messages PHP manuals	UN documents	The Holy Quran	UN documents	Broadcast conversation, traditional broadcast news and newswires
Alignment level	Sentence	Sentence	Sentence	Sentence	Sentence	Paragraph	Sentence and word
POS tagging	×	×	×	✓	✓	×	×
Stemming	×	✓	×	×	✓	×	×

To fill this gap, the present study developed a freely available, human-verified English-Arabic political parallel corpus (EAPPC) that contains more than one million words culled from His Majesty King Abdullah II's written and spoken texts. To the best of our knowledge, this is the first work that aligns English-Arabic parallel texts at multiple levels (i.e., sentence, clause, phrase and word levels). To make this resource even more valuable, the corpus texts have not only been stemmed and POS-tagged automatically, but also manually verified. The present corpus can be a springboard to a Jordanian English-Arabic Parallel corpus. However, the most important limitation of our parallel corpus is that it is restricted to the Arabic and English speeches, interviews, letters and book of one person and that the translation was performed by anonymous translators in the Royal Court. The size of the corpus is also a limiting factor, but the fact that this is a research in progress is a consolation.

3. BUILDING THE CORPUS

3.1 Data Selection

A preliminary survey of Arabic texts on the World Wide Web (WWW) was conducted to identify the kinds of existing Arabic texts that were translated into English and English texts that were translated into Arabic. The survey results showed that there were different types of texts such as UN documents, news (e.g., Petra News Agency and the British Broadcasting Corporation's news texts), novels (e.g., Najib Mahfouth, Ghassan Kanafani and Agatha Cristy's) and books (e.g., on history and science). However, most of these texts were not freely available. Furthermore, most available Arabic texts were found in a Portable Document Format (PDF) whose conversion into machine-readable text would result in highly corrupted content. In order to obtain high-quality textual material for the present corpus, we considered the following selection criteria:

1. Arabic data should be in Modern Standard Arabic (MSA) and must have an English translation.
2. English data should be in Standard English regardless of geographic origin and must have a translation in Modern Standard Arabic.
3. The translation must not be produced by machine or non-professional translators.
4. The data should be available in machine-readable textual format.
5. The data should be representative of MSA in general or of a particular MSA genre.
6. Texts ought not to have been used in previous parallel corpora. This avoids duplication and opens the way for our corpus to get integrated with previous English-Arabic parallel corpora in the future.
7. The copyright must permit corpus compilation.

Fortunately, the required data were found in His Majesty's official website (https://www.kingabdullah.jo/). His speeches, interviews and letters met all the selection criteria. However, we decided to add His Majesty's book, *Our Last Best Chance: The Pursuit of Peace in a Time of Peril (2011)*, despite the fact that it was not available to the researcher in electronic format. This was done for two reasons: first, to enlarge the corpus size; second, to shed light on the real problems that are often encountered by corpus compilers when they deal with non-availability of texts.

It must be acknowledged that our corpus is limited in size and in representativeness of political language. As it stands, this parallel corpus is a valuable asset; not only to political science and media specialists, but also to translation specialists.

3.2 Data Description

The present corpus consists of 351 Arabic and English original documents that were translated into the opposite direction. These documents fall into four categories: speeches, interviews, letters and one book.

Table 2. The data extracted from His Majesty's speeches, interviews and letters.

Text	**Speeches**	**Interviews**	**Letters**
Time range	1999-2015	1999-2015	1999-2015
Total number of translated texts	189	80	68
English source text	131	45	0
Arabic source text	58	35	68
Translators	Royal Court	Royal Court	Royal Court
Range of length (words)	250-2350	308-6403	200-3050
English words	200,767	210,384	45,385
Arabic words	177,444	163,140	39,730

Table 3. The King's book.

Title	**Our Last Best Chance: The Pursuit of Peace in a Time of Peril**
Publisher	Viking Press
Date of publication	2011
Place of publication	New York
No. of chapters	27
Chapter length (words)	2300-8194
English words	109491
Arabic title	فرصتنا الأخيرة: السعي نحو السلام في وقت الخطر
Publisher	Daralsaqi
Date of publication	2011
Place of publication	Beirut
Translator	Shukri Rahim
No. of chapters	27
Chapter length (words)	2300-8194
Arabic words	107634

The book was written in English and translated and published in Arabic. Tables 2 and 3 below describe the corpus texts.

3.3 Data Extraction

After obtaining permission from the Royal Hashemite Court to use the King's speeches and writings for non-commercial purposes, Arabic and English data were extracted from the official website of His Majesty King Abdullah II and from His book.

The texts on the website were initially saved in Microsoft Word document format and subsequently into utf-8 text only format. Likewise, The English version of the book was scanned and submitted to an optical character recognition (OCR) process to convert picture into text. This process is relatively fast, yet its product is not without inaccuracies. Therefore, the converted text had to be manually checked and edited. OCR was only possible for the English version of the book, but not for the Arabic one. Available OCR tools were not capable of satisfactorily converting the Arabic version of the book into any electronic text format. Thus, the Arabic version was manually retyped, checked and saved in text format to render it ready for machine processing.

3.4 Data Processing

Data processing involved six stages: metadata annotation, text segmentation, tokenization, alignment, stemming and POS tagging. These processes were completed primarily by the researchers with one checking and verifying what the other had done and ensuring consistency. On occasion, verification was sought from experts at the Arabic and English departments at the university of the researchers. As this verification was not methodical, it is the intention of the authors to conduct a systematic inter-annotator agreement study and to convert the current corpus into a gold standard that is thoroughly human-verified and validated.

3.4.1 Metadata Annotation

Metadata include text title, author, year, era, category, occasion, region or place and source language. All texts in the corpus were annotated with these eight metatags.

3.4.2 Text Segmentation

Segmentation is the process of splitting a text into smaller segments, such as paragraphs, sentences and clauses[27]. Segmenting a text allows search terms to find matches and renders texts ready for analysis [22].

Identification of sentence boundaries in the source texts (ST) and their matches in the target texts (TT) is a major challenge for the segmentation process. This is because the boundaries do not always correspond. The relations between text segments and their translations are not always in one to one correspondence. One sentence in one language might be translated into one sentence or two sentences, as illustrated in Table 4; or two sentences in ST might be translated into one in TT, as shown in Table 5. Besides, there are many examples where a clause corresponds to a sentence, as demonstrated in Table 6. Therefore, text segmentation is manually carried out at sentence, clause and phrase levels. This is done in order to obtain the best matching between ST and TT segments.

Sentence length is another issue that was taken into consideration during the segmentation process. Many lengthy sentences in ST corresponded to long ones in TT, so they were segmented into smaller meaningful chunks that would potentially recur in other texts. This was done after satisfying the best matching constraint. These chunks are similar to what Andrew Pawley calls 'conventionalized sentence stems'[28]. This makes our corpus of particular value to language learners, as they can easily recognize and learn authentic instances of sentence stems together with their translations.

3.4.3 Tokenization

Tokenization is carried out manually. Word boundary identification during the tokenization process is also a challenge. Idiomatic expressions are often treated as single dictionary entries; hence they are at the same rank as words. In many cases, two or more words are kept together as one token. This is

Table 4. One ST sentence corresponds to two TT sentences.

Source	Line	Arabic sentence (ST)	English sentence (TT)
Letter of re-designation to Ali Abul Ragheb	7	وقد تلقيت كتاب استقالتك الذي يعبر عما عرفته فيك من ولاء وانتماء وحرص على النهوض بالواجب وتحمل المسؤولية بإخلاص وتميز في الأداء وقدرة على تحقيق الإنجاز في إطار من العمل المؤسسي المستند إلى قواعد المعرفة ومواكبة روح العصر.	I received your letter of resignation, in which you articulate what is well known to us of your loyalty, of your sense of belonging and of responsibility, sincerity and excellence. Such performance has been demonstrated in the execution of your duties.

Table 5. Two ST sentences correspond to one TT sentence.

Source	Line	English sentence (ST)	Arabic sentence (TT)
The King's book Chapter 5	110	I went to study international relations at Pembroke College, Oxford. I spent a year among the grassy quads and honey-colored stone buildings of that venerable institution, studying Middle Eastern politics.	انتسبت إلى كلية "بمبروك" في جامعة أوكسفورد حيث أمضيت سنة أدرس العلاقات الدولية وسياسات الشرق الأوسط وسط تلك المربعات الخضراء الواسعة التي تحيط بها الأبنية التراثية ذات اللون العسلي، في ذلك الصرح العلمي والثقافي المهيب.

Table 6. One ST clause corresponds to one TT sentence.

Source	Line	Arabic clause (ST)	English sentence (TT)
Letter of re-designation to Ali Abul Ragheb	8	وإنني إذ أعرب عن عميق اعتزازي بما حققته هذه الحكومة من إنجازات وما تصدت له من تحديات على الصعيد الداخلي أو على الصعيد الإقليمي والدولي،	We herewith express our deep pride in what this government has accomplished and the challenges it has faced, whether locally, regionally or internationally.
	9	فإنني أتوجه بالشكر بشكل خاص لكل من عمل وأسهم في إنجاز الانتخابات البرلمانية التي أردناها غاية في النزاهة والموضوعية والشفافية.	In particular, we extend special gratitude to all those who worked and contributed to the successful completion of the parliamentary elections, which we wanted to be conducted with utmost integrity and transparency.

because a single token in Arabic might correspond to a phrase in English (e.g.,يُكَوّن corresponds to the phrase "*make up"*). On the other hand, a single token in English might correspond to two or more words in Arabic (e.g., "*cousins*" corresponds to أبناء العمومة).

3.4.4 Alignment

Accurate alignment is crucial for extracting information out of a parallel corpus. It enables users to easily and swiftly find equivalents of search terms or phrases. For example, when the user types a search term in one language, the concordance displays all occurrences of this word in that language. It also displays all the aligned equivalents in the target language. The results can be then extracted and analyzed [29].

To automatically align the corpus texts, SDL Trados WinAlign 2011 was used as an alignment tool. However, the output was unsatisfactory. Furthermore, WinAlign altered the predesignated text segmentation. Given the particular importance of alignment to the parallel corpus, it is regarded essential to manually verify its accuracy.

The text alignment was carried out at sentence, clause, phrase and word levels. However, it was not a straightforward process either. There were in the data some instances where lines in ST were left untranslated (see Table 7). Similarly, some texts in TT were inserted without any correspondence texts in ST (see Table 7). These phenomena created alignment problems.

Table 7. Examples of untranslated lines.

Line	**ST** (His Majesty's Speech at the Opening Session of the World Economic Forum, 20-May-05)	**TT**
1	No English equivalent	السلام عليكم، وأهلاً بكم في الأردن
2	Thank you Professor Schwab.	No Arabic equivalent
3	And thank you all	والشكر لكم جميعاً

To solve these problems, the lines with no equivalents were either deleted or attached to a neighboring line if they had significant contribution to the text.

The words in ST were manually matched to their equivalent words or phrases in TT to create a bilingual terminology list which would be useful for compiling bilingual dictionaries. Alignment at word level was more complex due to word order differences between Arabic and English. Words with no equivalents were left unaligned.

3.4.5 Stemming and Lemmatization

Stemming and lemmatization are beneficial for information retrieval. They reduce multiple word forms to a single word type. This multiplies the chance that the corpus users find more words that are morphologically related to a search term.

The Arabic light stems, roots and lemmas were automatically extracted using MADAMIRA. Its accuracy was reported in the literature to be respectable [18]. Then, the lemmatized texts were manually verified. English words were stemmed using the Porter stemmer [30].

3.4.6 Part of Speech Tagging

POS is of paramount concern to the linguist who wants to know how language works and how it is used. Hence, corpora are often annotated with POS. Hunston (2002) argued that POS tagging is a fundamental step in corpus exploration [5]. She stated four reasons for this. First, POS tags allow the search to be restricted to specific POS instances of a given word (e.g. searching the corpus for the noun instances of the word *play*). Second, they help identify the frequent collocates of a search term (e.g. *outdoor, creative, imaginative and pretend* usually collocate with the noun *play*, while *brilliantly, excellently, superbly, well and badly* collocate with the verb *play*). Third, POS tags allow the frequency comparison between words in different categories or genres within the corpus. Finally, POS tags enable researchers to discover the common word-class in each corpus category.

In the present corpus, Arabic words were automatically POS-tagged using MADAMIRA[18] and then manually verified.

3.4.7 Corpus Structure

EAPPC consists of Arabic and English sub-corpora, each of which is further divided into two sub-corpora that contain ST and TT in each language, as illustrated in Figure1.

4. BUILDING A PARALLEL CONCORDANCER

We built a parallel concordancer which consisted of two parts: the application through which the end-

user interacts with the corpus as shown in Figure 2 and the database which stores the parallel corpus (See Figure 3).

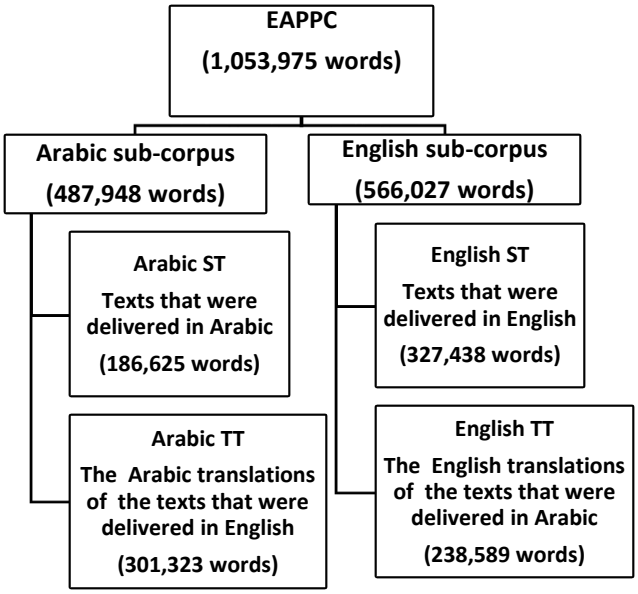


Figure 1. EAPPC structure.

Figure 2. Parallel concordancer interface.

The first step in building EAPPC was the preparation of the annotated texts. All the required data were manually arranged in three Excel sheets and then exported to a relational database as shown in Figure 3.

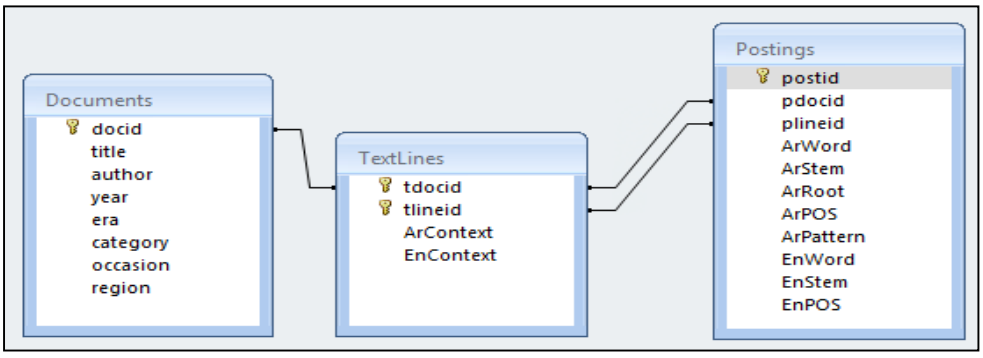


Figure 3. EAPPC relational database.

This relational database consists of three tables:

1. Documents Table, which stores information about each collected document (1 entry per document). A Document can have many text lines.
2. Text Lines Table, which stores the words of each text line in a document. Each entry (1 entry per text line) contains an ST line and its aligned equivalent TT line. A text line of each document can have many postings.
3. Postings Table, which stores information about every word in each text line (i.e., SL and TL,

English and Arabic). This includes the word itself, its stem, root, lemma and part of speech tag.

The concordance was developed using Eclipse as an integrated development environment running the Java 1.8 programming language. Both Eclipse and Java are open source software packages available for free.

5. CORPUS EXPERIMENTATION

Parallel corpora can be used to study different aspects of language, such as the features of source and target languages, the influence of SL on TL, the translation strategies used and the ideology and style of individual translators [31]-[34]. Furthermore, translators may learn strategies from parallel corpora and use them in their translation tasks [35].

Translating political texts has been labeled a complex activity [36]-[38]. Translators of such texts attempt to maintain the ideological and cultural aspects of the ST during the translation process [36], [38]. Hence, translators need to use the translation strategies and techniques that would enable them to preserve the ideology of the SL text and to cope with translation problems that surface during the translation process [38]-[39]. One of the problematic issues that often encounter translators is the translation of non-equivalence, particularly when a given concept is either unknown in the TL (a culture-specific concept) or known but is unlexicalized. Baker (1992) listed seven strategies that translators use to render non-equivalence [39]:

1. Using a hypernym.
2. Using a more neutral/ less expressive term.
3. Cultural substitution.
4. Using a loan word or a loan word plus explanation.
5. Paraphrasing.
6. Illustration and exemplification.
7. Omission of the problematic concept.

In order to demonstrate the EAPPC's utility in translation studies, the corpus was used to investigate how different Royal Court translators rendered the Arabic and Islamic culture-specific terms تكفير (takfīr)and تكفيري (takfīrī) in the speeches, interviews, letters and book of His Majesty King Abdullah II, since this is an Arabic term that illustrates non-equivalence in English.

5.1 Methodology

This research uses EAPPC, classical and modern Arabic dictionaries, *such as*, Al-ʿayn العين*(786* CE*)*, Muʿjam Maqāyīs Al-luġah(1004 CE), Al- mufradāt fī Ġarīb Al-qurʾān(1109 CE), *Lisān al-ʿarab*(1311 CE) andMuʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah(2003 CE)[2].

First, the Arabic dictionaries were consulted in order to determine the meaning of the word تكفير (takfīr) and تكفيري (takfīrī). Next, the EAPPC corpus was explored using the Arabic root كفر (kfr) as a query term. The parallel corpus concordance displayed, in the keyword in context (KWIC) style, all occurrences of the term along with their translations. However, the query terms are not highlighted so that the user can easily and swiftly identify them on the screen, but we are working on a better version and it will be highlighted. Moreover, information about the text, if it is a source or a target one, is not explicitly provided. The displayed data were then exported to an Excel spreadsheet and analyzed. Finally, the study used Baker's taxonomy of translation strategies to analyze the data and to discover the adopted translation strategies.

5.2 Findings and Discussion

The term تكفير (takfīr) is an abstract noun derived from the verb of intensification, kaffara, while the term تكفيري (takfīrī), often used as a substantive adjective (i.e., a noun), is derived from the nounتكفير (takfīr). *Takfīr* has multiple senses in dictionaries. See Table 8.

[2]http://lisaan.net.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 3, No. 3, December 2017.

The majority of these senses have changed over time and only one sense has survived. Table 8 shows تكفير (takfīr) to have had the meanings of abasement, submissiveness, obeisance, wearing armor, expiation of sins, nodding and enthronement. Only expiation, however, has survived the ravages of time. Additionally, *Muʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah* reflects our modern conception of تكفيري (takfīrī) as the "attribute of ascribing apostasy to others".

Table 8. Senses of the term تكفير (takfīr).

Sense	Dictionary
إيْمَاءُ الذِّمي برأسِه "Nodding"	Al-ʿayn العين
تَتْوِيْجُ المَلِكِ بتاج "Enthronement"	Al-ʿayn العين
الذّلّ والخضوع "Abasement, Submissiveness"	Lisān al-ʿarab
الانحناءُ الشَّدِيد "Bowing"	Lisān al-ʿarab
ستر الشيء وتغطيته "Covering"	Al-mufradāt fī Ġarīb Al-qurʾān
أن يَتَكَفَّرَ المُحاربُ في سلاحه "Covering the body with weapons"	Lisān al-ʿarab
تُكَفِّر الخطيئة أي تمحوها "Expiation"	Lisān al-ʿarab
التَّكْفير: أن يَخْضَعَ الإنسانُ لغَيْره وينحَني ويُطَأطِئ رَأسَه قَرِيبا من الرُّكوع "Offering obeisance"	Lisān al-ʿarab
جماعة تكفيرية: جماعة متشدِّدة تنسب العصاة والمذنبين إلى الكُفر، أو عدم الإيمان بالله، أو الزّندقة "Takfīrī group: extremists who call people apostates"	Muʿjam Al-luġa Al-ʿarabiyah Al-muʿāsirah

Using EAPPC,we searched for the Arabic root (كفر) "kfr" in order to retrieve all the occurrences of the terms تكفير and تكفيري. Seventy-seven instances of derivatives of this root have been used by His Majesty. Their distribution in the corpus is shown in Figure 4 below.

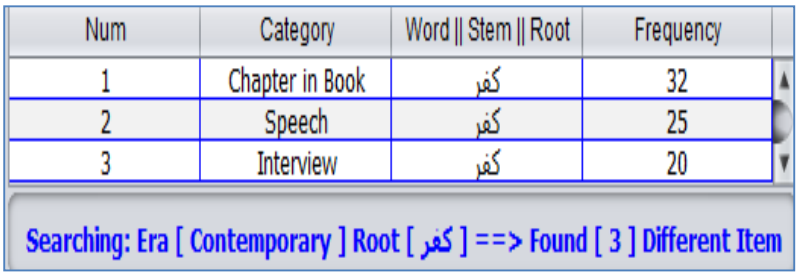


Figure 4. Search results of the root كفر (kfr) in the parallel corpus.

Seventy instances of these relate to the terms تكفير and تكفيري as shown in Table 9.

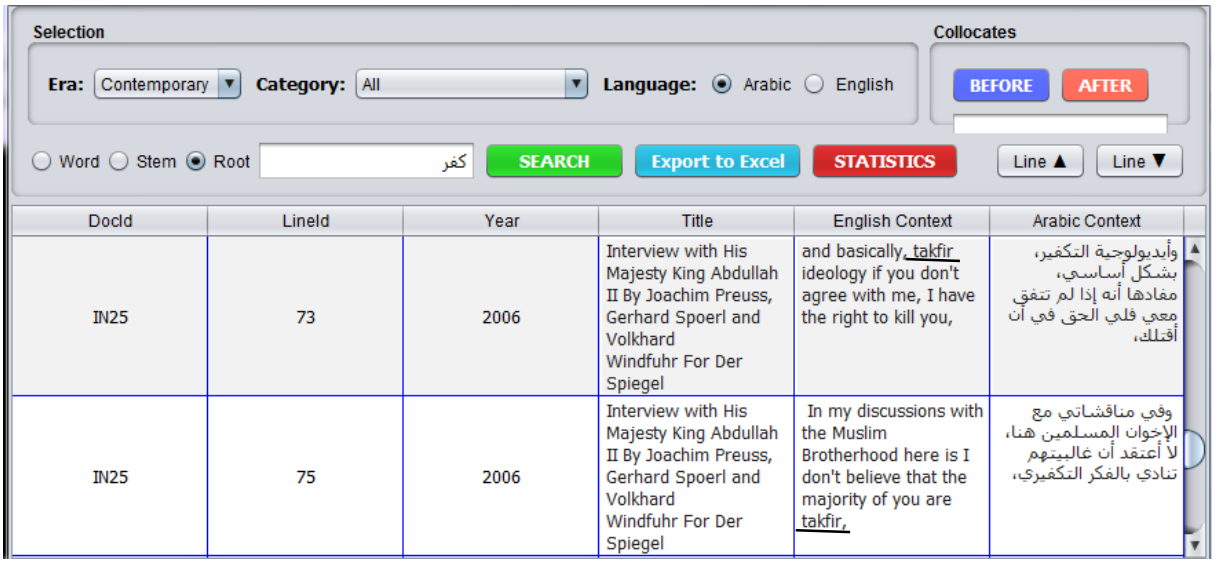
Table 9. Distribution of تكفير (takfīr) and تكفيري (takfīrī) instances in the parallel corpus.

Term	Speeches	Interviews	Book	Total
تكفير takfīr	21	6	3	**30**
تكفيريين takfīrīyīn/تكفيريون takfīrīyūn	-	3	24	**27**
تكفيري takfīrī	1	8	-	**9**
تكفيرية takfīrīyīah	-	-	4	**4**
Total	**22**	**17**	**31**	**70**

In order to examine how each instance of the terms تكفير (takfīr) and تكفيري (takfīrī) were rendered in Arabic and English, it is important to discover the translation strategies that were used by the King's translators. In English ST, the terms تكفير (takfīr) and تكفيري (takfīrī) were used as loanwords as illustrated in Figure 5.

The term تكفير (takfīr) in the Arabic ST component of the EAPPC occurs three times in one interview and 15 times in five speeches. In the English ST component of this corpus, the loan word *takfīr* occurs three times in one interview, six times in six speeches and three times in one chapter in the book.

"English-Arabic Political Parallel Corpus: Construction, Analysis and a Case Study in Translation Strategies", Alia Al-Sayed Ahmad, Bassam Hammo and Sane Yagi.



DocId	LineId	Year	Title	English Context	Arabic Context
IN25	73	2006	Interview with His Majesty King Abdullah II By Joachim Preuss, Gerhard Spoerl and Volkhard Windfuhr For Der Spiegel	and basically, takfir ideology if you don't agree with me, I have the right to kill you,	وأيديولوجية التكفير، بشكل أساسي، مفادها أنه إذا لم تتفق معي فلي الحق في أن أقتلك،
IN25	75	2006	Interview with His Majesty King Abdullah II By Joachim Preuss, Gerhard Spoerl and Volkhard Windfuhr For Der Spiegel	In my discussions with the Muslim Brotherhood here is I don't believe that the majority of you are takfir,	وفي مناقشاتي مع الإخوان المسلمين هنا، لا أعتقد أن غالبيتهم تنادي بالفكر التكفيري،

Figure 5. Examples of the term تكفير (takfīr) and تكفيري (takfīrī) in His Majesty's English ST.

EAPPC evidence shows that translators adopted these strategies when rendering تكفير (takfīr) from Arabic into English: the use of loan words, loan words plus explanation, English equivalents and English equivalents with the TL terms between brackets. In many instances, translators would introduce the loan word *takfīr* with an explanation (e.g. calling others apostates) and then use it without explanation in subsequent occurrences, as shown in the following example from an interview given by His Majesty on 22 April 2006 to *Al Sabah Al Jadid Newspaper:*

ST (Arabic):

- "كما حظي بتوافق إجماعي يدين ممارسات **التكفير** التي يلجأ إليها المتطرفون لتبرير العنف."
- "ولأننا نقف ضد التطرف **والتكفير**، فقد أصبحنا مستهدفين من الجماعات الإرهابية في العراق".

TT (English)

- "This declaration condemned the practice of **takfīr (calling others apostates)** that extremists use to justify violence."
- "And because we stand against extremism and **takfīr**, we have become targets of terrorist groups in Iraq,"

Another strategy for rendering تكفير (takfīr) from Arabic into English is translation by TL equivalents, as illustrated in the following examples from His Majesty's speech at *the opening session of the International Islamic Conference* on 4 July 2005:

ST (Arabic):

"وعدم جواز **تكفير** أي مسلم من أتباعها."

TT (English):

"and that **declaring any one of them an apostate** is unacceptable.

Another strategy is using a TL equivalent with the loan word between brackets. For example, translators paraphrased تكفير (takfīr) as *apostasy* and used the loan word *takfīr* between brackets, as demonstrated in the following example from His Majesty's speech at the opening of the third extraordinary session of the *Islamic Summit* on 7 December 2005:

ST (Arabic):

"لأن عدم الاتفاق على هاتين المسألتين هو سبب الفرقة والاختلاف وتبادل تهم **التكفير**والاقتتال بين أبناء الدين الواحد"

TT (English):

"The absence of consensus on these two issues has led to divisions and differences, accusations of **apostasy (takfir)** and internecine fighting.".

Analysis shows that translators tended to use English equivalents strategy most often when rendering the term تكفير (takfīr) from Arabic into English as shown in Table 10.

Table 10. Frequencies of تكفير (takfīr) translation strategies from Arabic into English.

Translation Strategy	Frequency
The use of loan words	2
Loan words plus explanation	2
English equivalents	12
English equivalents with the TL terms between brackets	2
Total	**18**

The term *takfīrī* is the adjectival form of *takfīr* that is often used as a noun. The corpus offers 40 such instances. Twenty-eight of them occur as loan words in His Majesty's book and one in an English interview. In the subcorpus of Arabic STs, on the other hand, *takfīrī occurs* 10 times in four interviews and once in a speech. Moreover, the corpus evidence shows that thirteen occurrences of the term تكفيري (takfīrī) in the Arabic subcorpus are adjectives and twenty-seven are nouns.

The loan word *takfīrī* in English ST texts was rendered as تكفيري (takfīrī), تكفيرية (takfīrīyah), تكفيريين (takfīrīyīn) or (takfīrīyūn) تكفيريون in Arabic in accordance with the requirements of syntactic inflection.

In some instances, the word *takfīrī* is not found in the ST text but the translator understood that it was intended. In such a case, the translator made it explicit by using *takfīrī* in the TL text, as shown in the following example from His Majesty's book, *Our Last Best Chance (2011):*

ST (English):
"and we helped the Americans understand **what** to look for"
TT (Arabic):
"وقد ساعدنا الأمريكيين في التعرف على **التكفيريين**."
'and we helped the Americans recognize **takfīrī s**'

In other cases, the author referred to the word *takfīrī* by using an anaphoric pronoun. In this case, the translators explicitly used the equivalent word تكفيري (takfīrī), as illustrated in the following example from His Majesty's book, *Our Last Best Chance(2011):*

ST (English):
"Islam celebrates life; **they** seek to destroy it."
TT (Arabic):
"فإذ يحترم الإسلام الحياة الإنسانية ويصونها، لا يتردد **التكفيريون** في تدميرها والقضاء عليها.".

'Even though Islam respects and protects human life, **takfiris** do not hesitate to destroy it and quell it.'

Although the term تكفير (takfīr) has been translated into English using multiple strategies, تكفيري (takfīrī) has only been rendered using the loan word strategy, as shown in His Majesty's interview on 22 April 2006 given to *Al Sabah Al Jadid Newspaper*:

ST (Arabic):
"وجد الفكر **التكفيري** ما يغذي أهدافه البعيدة كل البعد عن قيم الإسلام الحقيقية".
TT (English):
"Takfiri thought found feeding ground for its aims that are alien to true Islamic ethics and values."

To sum up, translators tended to render the loan words *takfīr* and *takfīrī* from English into Arabic by using the same terms as they are Arabic in the first place. On the other hand, when they translated into English they employed several strategies: translation using loan words, loan words plus explanation, translation by TL equivalence and translation by equivalents with the loan word between brackets.

6. CONCLUSION

This study has described the construction of EAPPC. The ultimate aim of EAPPC is to provide translators, learners, educators, researchers and language engineers with a freely available tagged parallel corpus whose annotation has been manually verified. To illustrate its utility, we have carried out an experiment that examined the translation strategies used in rendering a culture-specific term. The

results demonstrated the ease with which knowledge about translation strategies can be gained from this parallel corpus.

REFERENCES

[1] J. Sinclair, Corpus, Concordance, Collocation, Oxford University Press, 1991.

[2] M. Baker, "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research," Target, vol. 7, pp. 223-243, 1995.

[3] D. Biber, S. Conrad and R. Reppen, Corpus linguistics: Investigating Language Structure and Use, Cambridge University Press, 1998.

[4] L. Bowker, "Towards a Methodology for a Corpus-based Approach to Translation Evaluation," Meta: Journal des traducteurs Meta:/Translators' Journal, vol. 46, pp. 345-364, 2001.

[5] S. Hunston, Corpora in Applied Linguistics, Cambridge: Cambridge University Press, 2002.

[6] F. Zanettin, S. Bernardini and D. Stewart, Corpora in Translator Education, London, 2014.

[7] I. Ulitkin, "Computer-assisted Translation Tools: A Brief Review," Translation Journal, vol. 15, 2011.

[8] F. J. Och, "Statistical Machine Translation: Foundations and Recent Advances," presented at the Tutorial at MT Summit, Phuket, Thailand, 2005.

[9] L. Bowker and J. Pearson, Working with Specialized Language: A Practical Guide to Using Corpora, London, NY: Routledge, 2002.

[10] S. Goodman and K. O'Halloran, The Art of English: Literary Creativity, Basingstoke, UK: Palgrave Macmillan, 2006.

[11] K. Aijmer, B. Altenberg and M. Johansson, Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund 4-5 March 1994, vol. 88, Lund Studies in English, 1996.

[12] D. Samy, A. Moreno Sandoval and J. M. Guirao, "An Alignment Experiment of a Spanish-Arabic Parallel Corpus," Proceedings of the International Conference on Arabic Language Resources and Tools (NEMLAR 2004), pp. 85-89, 2004.

[13] D. Samy, A. M. Sandoval, J. M. Guirao and E. Alfonseca, "Building a Parallel Multilingual Corpus (Arabic-Spanish-English)," Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations (LREC), 2006.

[14] M. Tadić, "Building the Croatian-English Parallel Corpus," The 5th International Conference on Language Resources and Evaluation (LREC'2006), pp.523-530, Athens, Greece, 2000.

[15] S. Singh, T. McEnery and P. Baker, "Building a Parallel Corpus of English/Panjabi," Parallel Text Processing, ed: Springer, pp. 335-346, 2000.

[16] L. Rura, W. Vandeweghe and M. Montero Perez, "Designing a Parallel Corpus as a Multifunctional Translator's Aid," in XVIII FIT World Congress= XVIIIe Congrès mondial de la FIT, 2008.

[17] H. Al-Ajmi, "A New English–Arabic Parallel Text Corpus for Lexicographic Applications," Lexikos, vol. 14, pp. 326-330, 2004.

[18] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," LREC, pp. 1094-1101, 2014.

[19] J. Véronis, "From the Rosetta Stone to the Information Society," Parallel Text Processing, ed: Springer, pp. 1-24, 2000.

[20] L. Al-Sulaiti and E. Atwell, Designing and Developing a Corpus of Contemporary Arabic, MA Thesis, School of Computing, University of Leeds, UK, 2004.

[21] H. Hasselgård, "Contrastive Analysis / Contrastive Linguistics," The Routledge Linguistics Encyclopedia, K. Malmkjær, Ed., Third Edition, London, NY: Routledge, pp. 98-101, 2010.

[22] G. R. Yepes, "Parallel Corpora in Translator Education," Redit: Revista Electrónica de Didáctica de la Traducción y la Interpretación, pp. 65-80, 2011.

[23] J. Tiedemann and L. Nygaard, "The OPUS Corpus-Parallel & Free," Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 2004.

[24] M. S. S. Sawalha, Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora, PhD Thesis, University of Leeds, 2011.

[25] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," in LREC'10, Malta, 2010.

[26] H. Salhi, "Investigating the Complementary Polysemy and the Arabic Translations of the Noun Destruction in EAPCOUNT," Meta: Journal des traducteurs Meta:/Translators' Journal, vol. 58, pp. 227-246, 2013.

[27] P. Baker, A. Hardie and T. McEnery, A Glossary of Corpus Linguistics: Edinburgh Uni. Press, 2006.

[28] A. Pawley and F. H. Syder, "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency," Language and Communication, vol. 191, p. 225, 1983.

[29] M. Ghadessy, A. Henry and R. L. Roseberry, Small Corpus Studies and ELT: Theory and Practice, SCL 5, Philadelphia, USA: John Benjamins Publishing Co., 2001.

[30] M. F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, pp. 130-137, 1980.

[31] M. Baker, "Corpus Linguistics and Translation Studies: Implications and Applications," Text and Technology: In honour of John Sinclair, M. Baker, G. Francis and E. Tognini Bonelli, Eds., Philadelphia, USA: John Benjamins Publishing Co., 1993.

[32] S. Laviosa, Corpus-based Translation Studies: Theory, Findings, Applications, Amsterdam – New York, NY: Rodopi B.V, 2002.

[33] M. Olohan, Introducing Corpora in Translation Studies, London, NY: Routledge, 2004.

[34] C. Fantinuoli and F. Zanettin, New Directions in Corpus-based Translation Studies, Berlin: Language Science Press, 2015.

[35] G. Shen, "Corpus-based Approaches to Translation Studies," Cross-Cultural Communication, vol. 6, pp. 181-187, 2011.

[36] A. Shunnaq, "Arabic-English Translation of Political Speeches," Perspectives: Studies in Translatology, vol. 8, pp. 207-228, 2000.

[37] G. Quentel, "Translating a Crucial Political Speech," ed: Retrieved October, 2006.

[38] K. Sárosi-Márdirosz, "Problems Related to the Translation of Political Texts," Acta Universitatis Sapientiae Philologica, pp. 159-180, 2014.

[39] M. Baker, In Other Words: A Coursebook on Translation, London, NY: Routledge, 1992.

ملخص البحث:

تتناول هذه الدراسة إنشاء مجموعة كاملة من مليون كلمة بالعربية والإنجليزية في حقل السياسة، من شأنها أن تكون مصدراً مفيداً للبحث في دراسات الترجمة، وتعلُّم اللغة وتعليمها، وعلم المعاجم ثنائية اللغة، والدراسات المقارِنة، ودراسات العلوم السياسية، واسترجاع المعلومات المتعلقة بتقاطع اللغات. وتصف الدراسة مراحل إنشاء المجموعة، وشرحها عن طريق الأمثلة، من أجل اكتشاف استراتيجيات الترجمة المستخدمة في ترجمة المصطلحين "تكفير" و"تكفيري" المتعلقين بالثقافة العربية والإسلامية من العربية الى الإنجليزية وبالعكس. تتكون المجموعة من 351 وثيقة أصلية بالعربية والإنجليزية وترجماتها. فقد تم اختيار 189 خطاباً و80 مقابلة و68 رسالة، ترجمها مترجمون في الديوان الملكي الهاشمي، من الموقع الإلكتروني الرّسمي لجلالة الملك عبدالله الثاني، إضافة الى نصّ كل من النسختين الإنجليزية والعربية لكتاب جلالته المعنون: "فرصتنا الأخيرة: السعي نحو السلام في زمن الخطر" الصادر عام 2011. وبعد استكمال مراحل إنشاء المجموعة، جرى إعداد فهرس أبجدي لتسهيل عملية تحرّي المجموعة. أما أبرز التحديات فقد تمثلت في ندرة النصوص المترجمة من العربية الى الإنجليزية وبالعكس القابلة للقراءة بواسطة الآلة، إضافة الى نقص الأدوات القادرة على معالجة النصوص العربية.