A New English–Arabic Parallel Text Corpus for Lexicographic Applications

Hashan Al-Ajmi, *Department of English, Kuwait University, Kuwait, (hashan98@yahoo.com)*

Abstract: Bilingual lexicographers, translation specialists and English teachers in the Arab world do not have access to computerized corpora of parallel texts for the English–Arabic language pair. This project has been carried out to meet this requirement by establishing the first general parallel corpus of English texts and their Arabic translations. The first phase of the project involved the selection of general source texts having appropriate lexical and stylistic features. The chosen source texts deal with a variety of topics such as the environment, globalization, psychology, history, politics, drama, etc. Their Arabic translations were taken from *The World of Knowledge* series published by the National Council for Culture, Arts and Letters (NCCAL) in Kuwait.

Keywords: PARALLEL CORPUS, LEXICOGRAPHY, TRANSLATION, BILINGUAL DICTIONARY, COLLOCATIONS, ALIGNMENT, SYNONYMS, DERIVATIVES, ANTONYMS, GLOSSARY, FREQUENCY

Opsomming: 'n Nuwe Engels–Arabiese parallelletekskorpus vir leksikografiese toepassings Tweetalige leksikograwe, vertaalkundiges en Engelsonderwysers in die Arabiese wêreld het nie toegang tot gerekenariseerde korpusse van parallelle tekste vir die Engels–Arabiese taalpaar nie. Hierdie projek is onderneem om in dié behoefte te voorsien deur die eerste algemene parallelle korpus van Engelse tekste en hul Arabiese vertalings tot stand te bring. Die eerste fase van die projek het die keuse van algemene brontekste behels wat geskikte leksikale en stilistiese eienskappe besit. Die gekose brontekste handel oor 'n verskeidenheid onderwerpe soos die omgewing, globalisering, psigologie, geskiedenis, politiek, drama, ens. Hul Arabiese vertalings is geneem uit *The World of Knowledge*-reeks gepubliseer deur die National Council for Culture, Arts and Letters (NCCAL) in Koeweit.

Sleutelwoorde: PARALLELLE KORPUS, LEKSIKOGRAFIE, VERTALING, TWEETALIGE WOORDEBOEK, KOLLOKASIES, OOREENSTEMMING, SINONIEME, AFLEIDINGS, ANTONIEME, GLOSSARIUM, FREKWENSIE

1. Introduction

Parallel corpora ought to be treated as essential tools in bilingual lexicographic activities in the Arab world where translation plays a key role in the transfer of knowledge, especially from English sources. Yet, Arab translation specialists and training programs lack corpus-based bilingual dictionaries and lexical data

bases. Unfortunately, and despite the growing awareness of their importance, no parallel corpora for the English–Arabic language pair have yet been developed. This is partly due to the lack of the necessary programs to compile such resources and the funding authorities' doubts and uncertainty regarding the effectiveness of parallel corpora. There are also those who confuse this type of corpus with translation memories which are widely used by professional translators and by developers of machine translation systems and term banks. The available literature on parallel corpora outlines several applications of this type of corpus including:

— Development of bilingual dictionaries:

  English–Arabic dictionaries are considered to be inefficient lexical sources as they rely on the provision of non-contextual translation equivalents. They force the user to undertake most tasks such as the selection of the appropriate equivalent and ordering the words in the target language phrases or sentences. Here, the parallel corpus can play a vital role by providing several appropriate context-based equivalents that can be included in a bilingual dictionary (cf. Dickens and Salkie 1996).

— Helpful information for users:

  Users of the corpus can see both source and target words in their contexts and be their own judges when selecting the appropriate synonym — an advantage typically unavailable in the use of bilingual dictionaries.

— Language learning:

  Many language teachers around the world have already started to integrate and utilize parallel corpora in the foreign-language curriculum. Many of them have noticed that the parallel corpus contributed in making learners more independent, e.g. by making comparisons, understanding how different contexts lead to differences in meaning.

  Also, parallel corpora have been used in a variety of ways such as learning by discovery, enhancement of vocabulary, preparation of teaching materials, and understanding of translation problems and techniques (Baker 1993, Barlow 1996, St. John 2001, Zanettin 1994).

In the following sections, the English–Arabic parallel text corpus will be described in terms of its sources, the programs used in compiling and searching it, its components, and its uses in the improvement of bilingual dictionaries and the development of collocations dictionaries.

2. Sources

The aim was to find original English texts characterized by both currency and

variety of topics to ensure that the corpus will provide new and modern vocabulary and usage. Therefore, the focus has been on those texts written quite recently, e.g. in the nineties. Publications of the Kuwait National Council for Culture, Arts and Letters were considered to be the appropriate resource. These publications enjoy great popularity among educated readers in the Arab world, especially the *World of Knowledge* series of translated books that explore a wide variety of topics and are written by the best translators in the region. It is believed that Arabic translations in the higher stylistic level of this diglossic language (cf. Ferguson 1972) are necessary in this corpus because most translations from English are of scientific, legal, political or economic texts with similar stylistic levels. Accordingly, these translations would make the corpus a practical tool in translator training.

3. Means of access: Restricted web site

To ensure that a larger number of users can access it, the corpus has been allocated a URL on the Kuwait University server. But due to copyright restrictions online users of the corpus will need to enter their user names and passwords. Students taking lexicography and translation courses at the Department of English use this web site in various course activities such as the extraction of bilingual word lists and the study of translator equivalents versus dictionary equivalents.

4. Corpus processing and search tools

The source and target texts were scanned, converted to text files through an OCR (optical character recognition) program and saved as XML (extensible markup language) files. Another software program aligned the converted texts on the sentence level and each textual unit was assigned an alignment type and identification. Header files stored in two separate corpora were used to link the English and Arabic texts. The search tool employed in this project is the al-Idrisi search program developed by Sakhr Software. This program was chosen for its capability of handling Arabic morphology and syntax. Its search abilities cover exact matches, root-based searches as well as derivatives, synonyms, antonyms and recognition of Arabic affixes. It also uses wild card searches, employs relevancy ranking and ignores common Arabic errors, e.g. *hamza*, *ha/ta* endings entered by the user or present in the original text. It searches for all or any of the words or phrases.

5. Components

The corpus is composed of a number of HTML links for a glossary of English words, and a search page as well as a help page.

5.1 The English glossary

This glossary is actually an edited index of all words in the English part of the corpus. It includes word frequencies and is used as a reference tool for looking up words and knowing their frequencies directly. This is done by selecting the first letter of the needed word, then pressing the second letter to show all the words in that section with their frequencies. The user of the glossary is subsequently capable of accessing the results page by clicking on the chosen hypertext word which will highlight its English results.

5.2 Search page

This page contains several options allowing users to find results of a specific nature. For example, users can specify a source text or subject such as economics, arts, psychology, etc. The search options here include the search language, which makes it possible to search for identical words or phrases and for both Arabic and English synonyms as well as derivatives which can be the focus of morphological studies of English and Arabic. Users can utilize the corpus as either a monolingual or a bilingual corpus. They are also able to choose the number of needed results which can range between ten and fifty. They can opt to ignore Arabic common errors such as *hamza* and the *ya* ending. The corpus has search capabilities similar to those available in Google such as searching for all or any of the words or phrases, but it is also able to search for Arabic words with their affixes, e.g. the attached definite *al* or pronouns which characterize Arabic. Users can access more text of any given result by clicking on 'more' to see the whole sentence highlighted in its wider context on an HTML page.

6. Corpus-based exercises

The corpus has already been used in a number of course activities in translation and lexicography at the Department of English. One of these involved comparisons between the collocations provided in the *Oxford Collocations Dictionary for Students of English* (2002) and the collocations in the corpus. It was found that the corpus contains about 25% of the OCD collocations but many corpus collocations have not been listed in the dictionary. Of course, the corpus was useful in providing Arabic equivalents for these collocations and, therefore, it can be a good source for a corpus-based bilingual dictionary of this type. In another exercise the corpus was compared with the *Al-Mawrid English–Arabic Dictionary* (2003) indicating that about 10 000 Arabic equivalents and senses are missing in the dictionary. And in a further exercise comparisons were made between frequency ordering in the corpus and sense ordering in some English–Arabic dictionaries. The comparison has revealed that many entries need to be reordered in order to match users' expectations when dealing with living texts in translation and reading.

The corpus has other capabilities especially in the field of bilingual lexicography, e.g. obtaining illustrative examples in one or both languages.

7. Conclusion

It is hoped that this parallel corpus will meet the needs of researchers, translators and English teachers in the Arab world. However, given its small size (three million words) it should at this stage be regarded as a prototype for a larger corpus which is planned to be compiled in future. The envisaged version would be about ten times the current size. It will be bi-directional with more textual genres and styles, and it will be processed with POS (part of speech) tags to improve the search options in both directions.

Acknowledgement

I wish to thank the Kuwait Foundation for the Advancement of Sciences (KFAS) and Kuwait University for supporting this project (research grant no. 2001-13-01).

References

Dictionaries

Ba'albaki, M. 2003. *Al-Mawrid: A Modern English–Arabic Dictionary*. Dar El-Ilm: Lil-Malayin.

Crowther, J. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Other references

Baker, M. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. Baker, M., G. Francis and E. Tognini-Bonelli (Eds.). 1993. *Text and Technology*: 233-250. Philadelphia: John Benjamins.

Barlow, M. 1996. Parallel Texts in Language Teaching. Botley, S., J. Glass, T. McEnery and A. Wilson (Eds.). 1996. *Proceedings of Teaching and Language Corpora 1996*: 45-56. UCREL Technical Papers 9. Lancaster: University of Lancaster.

Dickens, A. and R. Salkie. 1996. Comparing Bilingual Dictionaries with a Parallel Corpus. Gellerstam, M., J. Järborg, S.G. Malmgren, K. Norén, L. Rogström and C. Röjder Papmehl (Eds.). *EURALEX '96 Proceedings I–II*: 551-559. Gothenburg: Department of Swedish, Göteborg University.

Ferguson, C. 1972. Diglossia. Cashdan, A. et al. (Eds.). *Language in Education: A Course Book*: 38-45. London: Routledge and Kegan Paul.

St. John, E. 2001. A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language. *Language Learning and Technology* 5(3): 185-203.

Zanettin, F. 1994. Parallel Words: Designing a Bilingual Database for Translation Activities. Wilson, A. and T. McEnery (Eds.). 1994. *Corpora in Language Education and Research: A Selection of Papers from Talc94*: 99-111. UCREL Technical Papers 4. Lancaster: University of Lancaster.