MPRC: A Multilingual Parallel Reference Corpus

The last decades have seen a growing interest in use of translation corpora for language learning, translation training e.g. Bowker (2003) and corpus-based translation studies e.g. Baker (1996). Multilingual parallel corpora "consist of original texts (i.e. texts originally written in a given language), and/or translations from other languages" (Mikhailov & Cooper 2016). Most available translation corpora dealing with (Arabic, English, Chinese) languages are bitexts focusing on the alignment of two languages only (i.e. English-Arabic, English-Chinese, or Arabic-Chinese) e.g. Wang (2005). None of the available multilingual corpora deals with aligning multi-targets of parallel Arabic, English and Chinese data e.g. Tiedemann (2012). Thus, this paper introduces a new resource for use in translator training and for research on contrastive and comparative linguistics. It describes the current status of the ongoing project called Multilingual Parallel Reference Corpus (MPRC). The resource itself is a subcorpus in a conglomerate of corpora, most of them learner translator corpora, some of them reference corpora. The paper aims to present the information about the corpus design and compilation process.

One of the aims of this project is to provide comprehensive standardized metadata (i.e. author, translators, source text, target texts, and the translator's preface). It comes in response to Saudi Arabia decision to include the Chinese language in the curriculum in schools and universities (2019). It should inform textbooks compilers of Chinese in Saudi Arabia and provide them with interesting naturalistic examples as the languages involved in this project (Arabic, English, Chinese) are not cognate and belong to different language families. MPRC is constructed to be a reference corpus to a parallel learner corpus that intends to include future productions of the learners of Chinses in Saudi universities and schools. The corpus consists of sample published texts from different genres and translation directions. Texts have been classified according to the source language and genre in a balanced way. The current corpus consists of 5 million words and manually aligned trilingual sentences. The following table indicates information about the available data in the corpus.

Table 1. MPRC preprocessing corpus

Number of Files	Source Language	Source Texts Tokens (Approx.)	First Target Tokens (Approx.)	Second Target Tokens (Approx.)
15	Arabic	600,000	455,000	365,000
15	English	590,000	555,000	490,000
15	Chinese	400,000	600,000	500,000

MPRC developed a user-friendly interface and easy to use corpus analysis tools. MRPC concordancing tool allows for retrieval of all the available examples of any query in context. The concordancing tool allows for the retrieval of the extended context of any query up to 10 sentences from the text.

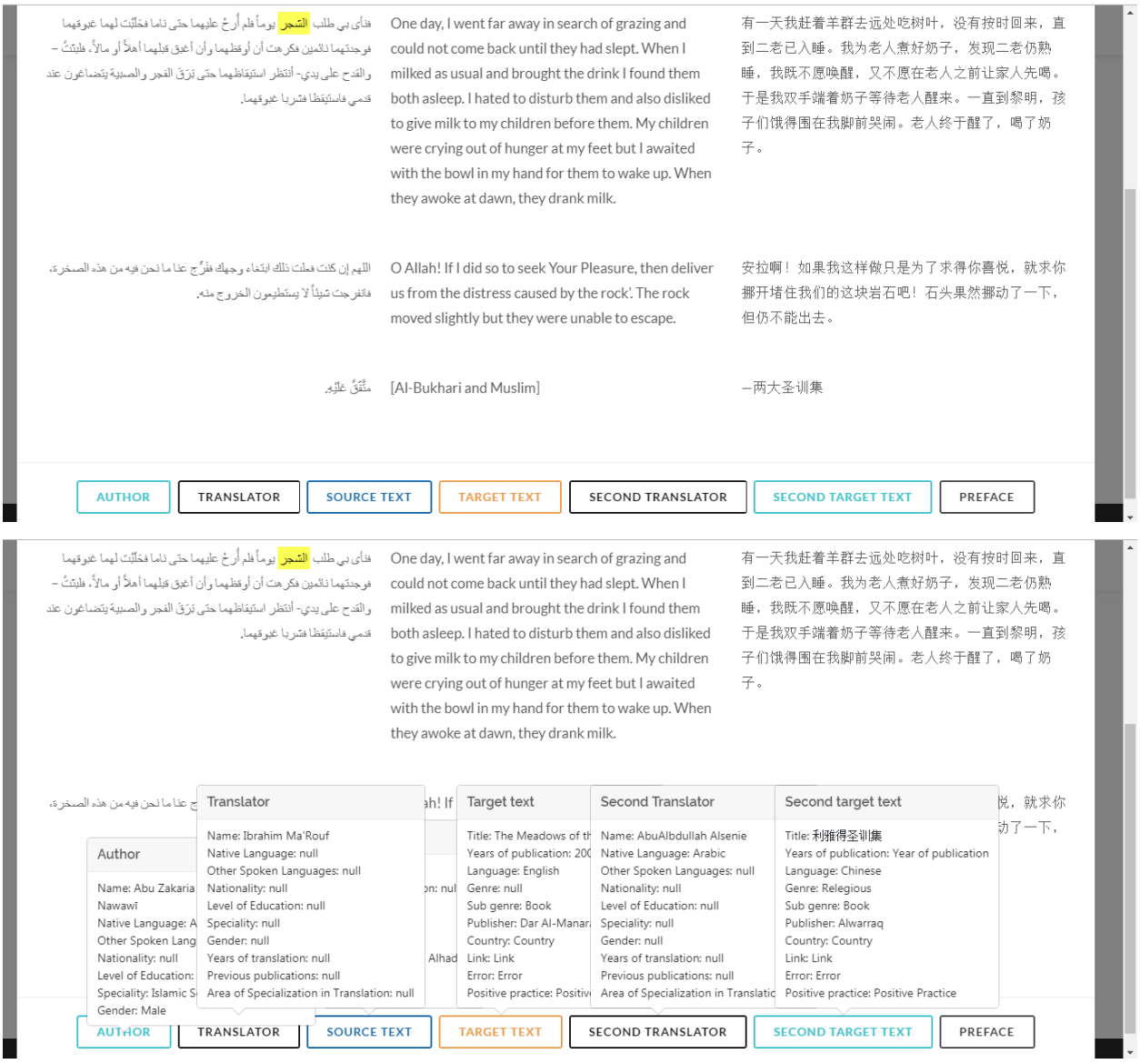

Figure 1. MPRC Annotation Layers and Concordancing tool

Additionally, MRPC developed a multilingual collocation tool that can extract formulaic sequences that can be bigrams or trigrams (see Figure 3).

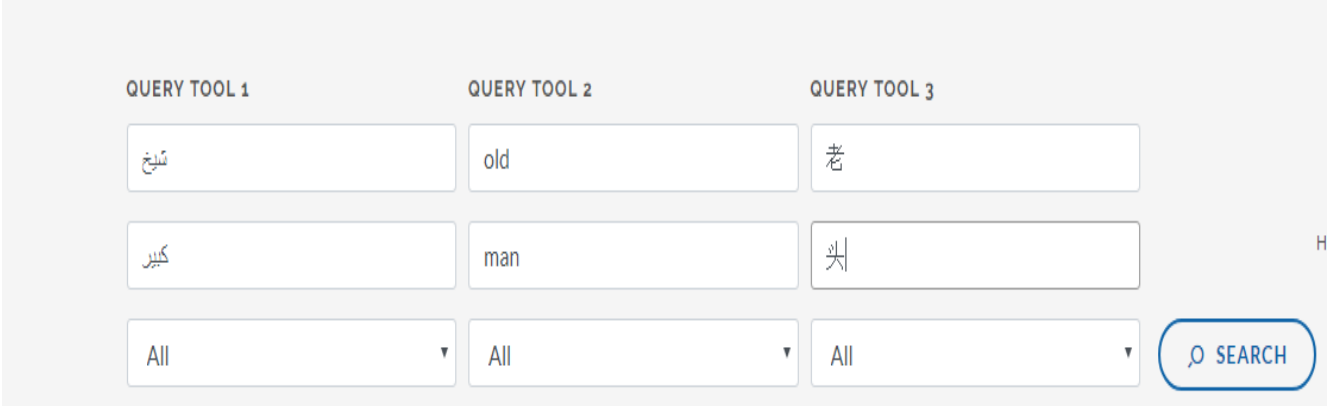

Figure 3. MPRC Multilingual collocation tool

This paper addresses MPRC design criteria and compilation process of data gathered in the first phase of project. The main aim of this project is to make data accessible for teaching and research.

References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Som-ers (Ed.), Terminology, LSP and translation. Studies in language engineering in honour of Juan C. Sager (pp. 175–186). Amsterdam: Benjamins.

Bowker, L., and Peter, B. (2003). "Student Translation Archive and Student Translation Tracking System: Design, Development and Application." In Corpora in Translator Education, edited by Federico Zanettin, Silvia Bernardini, and Dominic Stewart: 103–119. Manchester: St Jerome Publishing.

Mikhailov & Cooper. (2016). Corpus linguistics for translation and contrastive studies: A guide for research. Routledge. Corpus Linguistics Guides. London & New York: Routledge

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12) (pp. 2214–2218). Istanbul: European Language Research Association.

Wang, L. (2005). E-C Concord. Hong Kong Institute of Education. http://ecconcord.ied.edu.hk/paraconc/index.htm.

Dear Reem Alfuraih,

We thank you for your submission to the 15th Teaching and Language Corpora (TaLC) Conference Corpus Linguistics International Conference 2022 which will take place at the University of Limerick in conjunction with Mary Immaculate College.

We are delighted to inform you that your paper (details given below) has been accepted for this year's conference. Comments from the Reviewer Panel can now be found linked to your submission in the original Abstract Submission Portal. This portal will re-open on 11th April 2022 for one week only to facilitate any edits/revisions suggested by reviewers.