Open Source Corpus as a Tool for Translation Training

Taj Rijal Muhamad Romli

Department of Modern Languages, Faculty of Languages and Communication, UPSI

35900 Tg. Malim, Perak

Department of Foreign Languages, Faculty of Modern Languages of Communications, UPM,

43400 Serdang, Selangor

taj.rijal@fbk.upsi.edu.my

Muhamad Fauzi Jumingan

Department of Foreign Languages, Faculty of Modern Languages of Communications, UPM,

43400 Serdang, Selangor

uzi@fbmk.upm.edu.my

Abstract

Building a sentence into Arabic is rather difficult for amateur translators. Similarly,is the case for Malay students who particularly learn how to build sentences in writing. Usage of dictionaries also is not enough to convey the translation, especially in translating phrases and sentences from the Malay language into Arabic. Students are incapable of building sentences in Arabic because of lack of exposure to the structure of Arabic sentences. This weakness is discovered by most schools and universities in their writing exercises (Rosni, 2012), Ab. Halim Mohamad (2009), Che Radiah (2009). Generally, the dictionary is very suitable to be used in the search for meaning in the words but not the meaning of the sentence. This paper proposes a method of comparing comparable text of both languages through comparable corpora of both. It can also be called as a tool for translators. In addition to using the dictionary, students are guided to understand the structure of the original Arabic sentences with the comparative method, then apply it in the form of a writing exercise. In this process, teachers, students and amateur translators need to use the computer as a tool and open access data corpus in websites as the ingredient. Translated texts or guide texts for writing exercises are based on Aker and colleagues (2012) method of selection. Text is filtered using Webcorp open corpus engine http://www.webcorp.org.uk/live/ and also through Google open database https://www.google.com. Through this method, the search for similarities between the first and the second language can be exploited. Any text that is identified as having the closest comparable will be used in the classroom. It helps students and translators to build sentences into Arabic by comparison and evaluation of the original text in the corpus. At the same time students are also able to understand and recognize indirectly the structure of the original Arabic sentences. Hopefully this method will help amateur translators and students improve their quality of translation and writing in Arabic.

Keywords: corpus, comparable, databases

Introduction

This method was introduced since the widespread use of bilingual corpus. It is not a method of translation, but it is a method for finding comparable texts between two languages. The aim was to find meaning equivalence that can be a model translation that is near to original level of language usage. Mona Baker's theory of translation is used as measurement in determining comparability between the two texts. It is expected to be used as a method of learning in class for translation courses and also as a plan and structure to a software translation tool that is complemented with Malay and Arabic corpus data. The software will display examples of comparable Malay-Arab sentences as a guide to students studying correct Arabic sentence structures.

Background

One of the problems in schools and universities that need solution is the students' weakness in mastering the Arabic language, especially in writing and translation skills. Arabic language students experienced this problem since the school years and weakness was further brought to the university level. The problem becomes more apparent when focused on the weaknesses in the building of phrases and sentences which is the main mean for effective communication as found by studies such as Rosni (2012), Ab. Halim Mohamad (2009), Che Radiah (2009), Noor Anida binti Awang, Norhayati Binti Che Hat and Nurazan binti Mohmad Rouyan (2014) and Ghazali Yusri and Ahmad Bin Salleh (2006).

Based on the studies above, among the factors that lead to this weakness is due to the fact that the students are affected by the structure of their mother tongue. Other reasons are low mastery of Arabic vocabulary, negligency and no high motivation in learning. This weakness can be seen through significant mistakes in their writing and also in their translation texts from Malay into Arabic language. The usage of dictionary is not enough help to convey the translation's meaning, because the dictionary only translate words and phrases. The example given is also limited. Even if the students were able to find the meaning of each word but they still have trouble in structuring sentences into Arabic.

This method is as a proposal to the development of a software which displays example sentences in both languages. The texts selected are appropriate to the needs of their essay. Students only need to search the sentences needed on the topic of their essay by keyword root word or phrase. Based on the limited ability of the dictionary, thus this process is intended to help students get to know and understand the Arabic sentence structure more easily through studying and comparing with the sentences they construct.

Objectives

This paper aims to introduce the method of comparing text meaning through comparable corpora of two languages, Malay and Arabic . The results are expected to be used as a tool in teaching and learning translation classes and as the basis for the construction of Malay - Arabic dictionary of sentences .

Research Significance

Through this method, it can be used to search for comparabilities in texts and sentences between the first language and the second. Each text identified as having immediate comparability will be used as teaching and learning tool in the classroom. It helps students to Arabic translators in structuring new sentences into Arabic by comparing and evaluating the original texts in the corpus. At the same time students are able to understand and recognize indirectly the structure of the original Arabic sentences. This method is expected to help the amateur translators and students to improve their translation and writing in Arabic. Among the advantages of this method, it offers greater data links rather than using manual methods and is expected to form the basis for constructing a data software with a collection of selected Malay and Arabic texts, placed at par.

Literature Review

According to Zanettin (1998), Rusli and Norhafizah (2001), and Kruger (2004), there are two types of corpus that can be used as study tools to replace the dictionaries. First, known as parallel corpus (parrallel corpora), a corpus which compares the original text with the translated text. The second, known as comparable bilingual corpus (comparable corpora), the corpus that compares the text in two different languages, but share the same topic. For example, some topics or newspaper headlines of the world reporting an important event in multiple languages (Li Shao and Tou Hwee Ng, 2004).

According to Rusli Abdul Ghani and Nurhafizah Mohamed Husin (2001), the DBP has made an effort to build a database of phrases in Malay whether idiomatic or not based on actual use of the phrases in their translation texts. This database

includes common phrases and regular expressions in the source language (English) with its equivalent in the target language (Malay). Phrases and regular expression with its matches, are all derived from parallel and comparable corpora.

In Europe, comparable corpus studies began in the 1990s. Many studies concerning corpus were carried out. Comparable corpus as has been described is an unparallel bilingual texts but related and deliver a lot of overlap data in the web such as news in various languages released by news agencies such as CNN and BBC. Among the studies that utilize comparable corpus are studies by Munteano and Marcu (2005) and Munteano (2006).

Various techniques have been introduced by researchers such as Rapp (1995), have made the assumption that comparable words that can be translated appear in the same context, even in unrelated text. Rapp took 100 words and their translations representing the context as vector of similar incident (co-occurrence vector). The result is the matrix of the same events become more common when the composition of words in the matrix is the same in both languages.

Aker et. al (2012) in collaboration with Google has shown a simple technique to collect comparable corpus from the web. This is because the techniques introduced by the researchers before, such as Rapp (1999), Monteanu and Marcu (2002), Resnik (1999), Huang et.al. (2010), Talvensaari (2008), and others were time-consuming and requires substantial resources. The objective of his research is to reduce the amount of time and resources. Previously, researchers have to go through three steps to gather and build a comparable corpus, namely:

First: by downloading the document from the list of titles of the two languages. The process of downloading the document takes a long time and have to go through many obstacles.

Second: is the process of matching with comparable data and the third is to extract them. However, with the proposed technique, the first and second steps become easy. This study used English, Greek and Germany corpus. The methodology is by making a search of news articles through webs and RSS feeds without having to download the entire document. Topics headlines that are beneficial to the study from various categories of the selected languages are taken and at the same time, the time and date of the newscast, URL articles and cluster URL Google News are all recorded. From the topic search and URL cluster, a total of 30 articles with headlines are collected and downloaded forming monolingual Google News search. This process is performed in a specified time period, ie within a period of one week so that only the latest news are taken.

Third, is to divide the title into several entities in the source language and named after people, places or organizations. It is then translated via Google translate to the target language. The next phase is the process of aligning the document to compare the titles of the articles from the collected corpora. If it is comparable, then the actual article is being downloaded to obtain the equivalent corpus.

According to Aker et. al (2012), to measure the equivalence of corpus, two titles were tested with various heuristic techniques. The best 'heuristic' technique is TS (similar title), HS (time difference), and TLD (title-length difference), when used in combination TS-HS-TLD. It was then assessed according to 'Kendall's rank order' and also through human judgment based on Braschler's comparison (1998), ie five categories: same story, related story, same aspect, similar terms or unrelated. The hypothesis is that when the two articles contain the same story.

Some of the findings resulting from this comparison showed that parallel and comparable corpus can be used to build a database of phrases. However, due to the small size of the corpus leads only to few findings. From parallel corpus, some examples of phrases and expressions have been quoted, while from the comparable corpus only terms are available with no results of idiomatic phrase. Comparable phrases from different texts (and different translator) gives an indication that there is a consensus that assures that it is the most suitable match. In case of only having one phrase source with multi-matching, researchers can make their own choices based on compatibility.

Methodology

This study will collect comparable data in Malay and Arabic. Data search and collection is through open corpus online using Webcorp search engine corpus http://www.webcorp.org.uk/live/ and Google https://www.google.com database. The scope of this study is focused on general materials which are appropriate to the writing skills course beginning from level two of primary schools, all levels of secondary schools and Arabic writing skills courses in local university.

The selected topics are topics that dominate the debate of every major world newspaper which will open up a wider debate, as explained by Maia (2003), thus triggering the stages of new language usages and arising many terms related to this topic.

Data samples taken for this method is from sports genre under the topic of the World Cup Championship. This topic was chosen because of the importance of these topics covering the headlines, front and back pages of the newspapers. The probability score to achieve comparable text is greater. Topics for important matches especially the final always received wide coverage as it relates to the world's biggest hit in sports favoured by many.

Related data, evaluated by Aker's 'heuristic' technique (2012) is TS (title similarity), HS (time difference), and TLD (title-length difference) when used in combination TS-HS-TLD and Braschler and Schäuble (1998) category which is same story, related story and same aspect. Each category is then measured of its strong comparability of three levels, as recommended by Guidere (2002) i.e. strong, medium and weak comparability.

Figure 1 shows an overview of the whole methodology of the study:

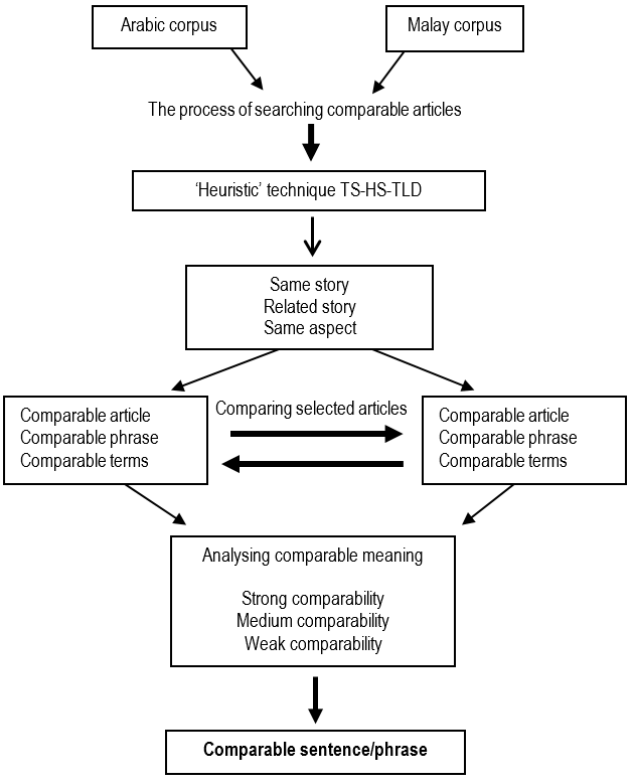

Figure 1

RESEARCH METHOD

More specific title in the 2010 World Cup Championship event have been selected for large probability of comparative findings between the texts.

Examples of topics are:

1. The 2010 World Cup Final
2. Matches between world big teams or well-known teams .
3. Final World Cup 2010

Search method in Webcorp search engine and Google is almost the same. But Google has the advantage of making easier option at the beginning of the search. Google engine offers the 'Any Country' and 'Any time' keys so that the search can be limited to time and place.

Advantage of Webcorp is that it has a filter engine 'Word Filter'. It is able to filter out the requested words and remove unnecessary words by placing minos symbol (-) before the word. Filter of time given for six to seven days until results come out.

Based on first time general search for the three general topics as mentioned above. More specific headlines were made keywords for searching comparable texts. The topics are extracted from the general title as follows:

1. General Title: The Final World Cup 2010

   Specific topic: *The 2010 World Cup Final between the Netherlands and Spain*

2. General Title: Match between big team or leading team.

   Specific topic: *Round 16 Match of the 2010 World Cup Championship (German vs England).*

3. General Title: The Final World Cup 2010

   Specific topic: *The Final World Cup 2014 between Argentina and Germany*

Results

Table 1 below provides an example of the analysis and the conclusion of a number of comparable text taken from the first title of the 2010 World Cup Final match between Netherlands and Spain after rejecting the difference of phrase and word's levels in both texts. The following result can be concluded as comparable sentences.

Table 1

Arabic Data	Malay Data
DS1 **(بعدما كسر مصيدة التسلل**AA1-1) **ضرب مصيدة التسلل الذي**(AA1-3) DS2 **اثر عرضية متقنة من فابريغاس**(AA1-1) **عقب نشاط مثمر من فابريغاس**(AA1-2) **زميله فابريجاس من قاتلة بينية تمريرة تلقى**(AA1-4) **ومرر فابريجاس كرة هدف المجد لانيستا أمام المرمى** (AA1-4)	(MA1-1) Beliau yang tidak berada dalam posisi 'offside' (MA1-1) hasil daripada umpanan Cesc Fabregas. (MA1-2) mendapat lambungan bola ke dalam kotak penalty (MA1-4) bola dihantar Cesc Fabregas (MA1-7) mengutip hantaran pemain gantian Cesc Fabregas

الهولندي	(MA1-2) merembat bola tersebut pada minit ke-116. (MA1-4) Iniesta menyempurnakan bola dihantar Cesc Fabregas untuk menewaskan penjaga gol Maarten Stekelenburg (MA1-7) merembat bola melepasi penjaga gol Maarten Stekelenburg.
DS3 **عندما هز انييستا الشباك الهولندية**)AA1-1(**ان يسدد الكرة في الزاوية اليمنى للمرمى الهولندي**)AA1-1(**عندما أسكن إنييستا الكرة في شباك هولندا**)AA1-2(**وسدد كرة قوية في مرمى الحارس الهولندي في الدقيقة 116**)AA1-3(**هولندا شباك في وسددها**)AA1-4(**ليسددا بيمناه في المرمى الهولندي**)AA1-5(**لم يتوانى الأخير في تسديدها في الشباك في الوقت القاتل مع الدقيقة 116**)AA1-6(**في إنييستا أندريس اللاعب عبر الثمين الفوز هدف وسجل الدقيقة 116**.)AA1-7(	(MA1-3) Kemenangan ini mengurniakan gelaran Piala Dunia pertama buat Sepanyol. (MA1-5) SEPANYOL muncul Juara Piala Dunia pertama di bumi Afrika
DS4 **وكان هذا الهدف كافيا لمنح بلاده المجد**.)AA1-1(**ليتوج المنتخب الإسباني بطلا للعالم للمرة الأولى في تاريخها.**)AA1-2(**احرز المنتخب الاسباني كأس العالم في كرة القدم للمرة الاولى في تاريخه**)AA1-8(	(MA1-3) Tatkala kedua-dua pasukan dilihat bakal berdepan penentuan penalti selepas tanpa jaringan dalam permainan 90 minit dalam perlawanan
DS5 **أي قبل نحو أربع دقائق من الاحتكام إلى ضربات الترجيح**)AA1-3(**اي قبل دقائق قليلة من انتهاء الوقت الاضافي الذي يسبق اللجوء الى ضربات الجزاء الترجيحية**.)AA1-8(	**(**MA2-1)Sepanyol bagaimanapun terlebih dahulu berpeluang.
DS6 **وجاءت أول فرصة في اللقاء لصالح المنتخب الأسباني**)AA2-1(**وحصلوا على فرصة لافتتاح التسجيل**)AA2-2(**وكاد المدافع سيرجيو راموس أن يفتتح باب التسجيل في الدقيقة الثالثة**)AA2-3(	(MA2-1)…sepakan sudut dihadiahkan kepada Sepanyol diambil oleh Xavi.
DS7 **اثر ضربة حرة لعبها تشافي هيرنانديز في الدقيقة الخامسة وقابلها سيرخيو راموس.**)AA2-1(**عندما انبرى تشافي لكرة حرة من الجهة اليمنى وصلت الى سيرجيو راموس.**)AA2-2(**المتخصص تشافي الذي نفذها بدقة على رأس راموس**)AA2-3(	(MA2-1)…menerusi tandukan Ramos (MA2-2)…yang kemudiannya ditanduk oleh Sergio Ramos.
DS8 **وقابلها سيرخيو راموس بضربة رأس قوية**)AA2-1(**وصلت الى سيرجيو راموس الذي سددها من مسافة قريبة بضربة رأسية**.)AA2-2(**نفذها بدقة على رأس راموس ليحولها برأسة قوية…**)AA2-4(	(MA2-1) …menguji Maarten Stekelenburg menerusi tandukan Ramos namun sempat ditepis oleh penjaga gol itu pada minit ke-5
DS9 **ولكن الحارس الهولندي مارتن ستيكلنبرج تصدى لها ببراعة فائقة ثم شتتها الدفاع قبل جيرارد بيكيه المتحفز.**)AA2-1(**صدها الحارس ستيكلنبورغ…**)AA2-3(**أنقذها الحارس الهولندي مارتن ببراعة.**)AA2-4(	(MA3-1) Dirk Kuyt, pada minit ke-7 berjaya melepaskan satu sepakan kencang dari jarak 25 meter. (MA3-2) Mujur cubaan jauh Dirk Kuyt**…**
DS10 **محاولاته الهجومية بتسديدة قوية أطلقها ديرك كاوت من مسافة بعيدة في الدقيقة الثامنة.**)AA3-1(**وسدد كاوت كرة ضعيفة كان لها كاسياس**)AA3-2(	

) ورد الهولنديون بتسديدة بعيدة لديرك كاوت(AA3-2	

The results of the comparison of the text in all comparable data (DS) as above. The number one and two data (DS1 and DS2) may be summarized as follows:

DS1

Arabic	Comparability	Malay
(بعدما كسر مصيدة التسلل(AA1-1 ضرب مصيدة التسلل الذي(AA1-3)	Medium comparability (MC) Both data share the same meaning, that is-not offside position. The difference is in the use of term and sentence structure.	(MA1-1) Beliau yang tidak berada dalam posisi 'offside'

Text level in comparable data 1 (DS1), are all medium-class of comparability (MC). Textual level means the same thing that it is not in an 'offside' position.

DS2

Arabic	Comparability	Malay
اثر عرضية متقنة من (AA1-1) فابريغاس	Strong Comparability (SC) Additional adjectives in AA1-1 *Mutqinah* does not affect the purpose of delivering the meaning	(MA1-1) hasil daripada umpanan Cesc Fabregas.
عقب نشاط مثمر من (AA1-2) فابريغاس	Medium comparability (MC) Although Fabregas name is mentioned in AA1-2 but not in 1-2 it still means the passing of ball from Fabregas.	(MA1-2) mendapat lambungan bola ke dalam kotak penalty
من قاتلة بينية تمريرة تلقى(AA1-4) زميله فابريجاس	Medium comparability (MC) The difference is in AA1-4 as there is an adjective *Qatilah* which means deadly passing while in MA 1-4 is not mentioned	(MA1-4) bola dihantar Cesc Fabregas

Text level in comparable data 2 (DS2). Data AA1-1 has strong comparability (SC). AA1-2 data and AA1-4 medium comparability (MC). The overall data means the same thing that is ball passing done by Fabregas.

Conclusion

This methodology is expected to produce a result of analysis that can be used as proof of for searching comparable meaning by using comparable corpus. In addition it can also be used as a specific method for learning aided by corpus using a specially designed software online in the hope of helping to develop a method of translation in teaching and learning translation and as translation tools.

Reference

[1] Aker, A.; Kanoulas, E. and Gaizauskas, R. (2012). A light way to collect comparable corpora from the Web. *In Proceedings of LREC* 2012, 21-27 May, Istanbul, Turkey.

[2] Belinda Maia. (2003). What are comparable corpora? *In Proceedings of the Workshop on Multilingual Corpora: Linguistic requirements and technical perspectives, at the Corpus Linguistics* 2003, pages 27–34, Lancaster, UK, March.

[3] Braschler M and Schäuble P. (1998). Multilingual information retrieval based on document alignment techniques. *In Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, pp. 183-197.

[4] Che Radiah Mezah. (2009). *Kesilapan Leksikal dalam Pembelajaran Bahasa Arab*. (Cetakan Pertama). Selangor: Penerbit Universiti Putra Malaysia.

[5] Ghazali Yusri Bin Abd Rahman & Ahmad Bin Salleh. (2006). Kemahiran Menulis Tulisan Arab Di Kalangan Pelajar-Pelajar UiTM: Kelemahan Dan Cara Mengatasinya. *Laporan Penyelidikan Akademi Pengajian Bahasa*, Universiti Teknologi MARA, Shah Alam.

[6] Guidere, Mathieu. (2002). *Toward Corpus Based Machine Translation for Standard Arabic*. Translation Journal. Vol.6. no. 1, January 2002.

[7] Kruger. (2004). *Corpus-Based Translation Research: Its Development And Implications For General*, Literary And Bible Translation. http://www.ajol.info/index.php/actat/article/viewFile/5455/29593.

[8] Li Shao and Hwee Tou Ng. (2004). In Mining New Word Translations from Comparable Corpora, *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*. http://www.mt-archive.info/Coling-2004-Shao.pdf.

[9] Munteanu, D.S. & Marcu, Daniel (2005). *Improving Machine Translation Performance by Exploiting Non-Parallel Corpora*. Association for Computational Linguistics.

[10] Noor Anida binti Awang, Norhayati binti Che Hat, Nurazan binti Mohmad Rouyan. (2014). Analisa Kelemahan Kemahiran Menulis Bahasa Arab Dalam Kalangan Pelajar Unisza. *Prosiding Seminar Pengajaran & Pembelajaran Bahasa Arab 2014*. Fakulti Pengajian Islam, UKM & Fakulti Kontemporari Islam, UniSZA

[11] Olohan, Maeve. (2004). *Introducing Corpora in Translation Studies*. New York: Routledge.

[12] Rosni bin Samah. (2012). *Pembinaan Ayat Bahasa Arab Dalam Kalangan Lepasan Sekolah Menengah Agama*. GEMA Online™ Journal of Language Studies . 12(2), 555-569.

[13] Rusli Abdul Ghani dan Norhafizah Mohamed Husin. (2001, 3 Sept.). Yang Selari dan Yang Setanding: Peranan Korpus dalam Penterjemahan, *Dalam Kertas Kerja Persidangan Penterjemahan Antarabangsa Ke-8*, Langkawi, Kedah.

[14] Tuomas Talvensaari. (2008). *Corpus-based cross-language information retrieval*. Department of Computer Science: Tampere.

[15] Zanettin, Federico. (1998). *Bilingual Comparable Corpora and the Training of Translators*. Meta: Translators' Journal, vol. 43, no. 4, p. 616-630. http://www.erudit.org/erudit/meta/v43n04/zanettin/zanettin.html