BUILDING AN ENGLISH - VIETNAMESE PARALLEL CORPUS OF CONTEMPORARY ART TERMS

NGUYEN THI MY NGOC[*]*, PHAN THI THANH THAO*
Hue University, University of Foreign Languages and International Studies
[*]*Email: ngocntm.qh@hue.edu.vn*

Abstract: The study aims at building an English-Vietnamese parallel corpus of contemporary art terms to meet the current pressing demands in Vietnam for the translation of art-related documents from Vietnamese into English and vice versa. To achieve its aim, the study adopts a combination of qualitative and quantitative research method to examine the original English texts and accurately translate them into Vietnamese, as well as generate the corpus linguistic database statistics. The generated English-Vietnamese parallel corpus of contemporary art terms consists of about 130,000 words. Some implications regarding the English-Vietnamese translation of texts related to contemporary art are also suggested for translators as well as teachers and students in Vietnamese art schools.

Keywords: corpus, corpus linguistics, translation, contemporary art.

1. INTRODUCTION

The work of translating specialized texts including technical jargons of a scientific discipline from a language to another one and vice versa is inevitable in the current era of globalization. In this context, whether corpus linguistics is considered a newly-emerging methodological basis for carrying out linguistic investigations or a sub-branch of linguistics with the theoretical disciplines of its own, corpus linguistics has long affirmed its significant position in the translation arena. Within the realm of translation alone, corpora have been exploited in a plethora of specialized subject fields such as legal translation (Klabal, 2019) or medical translation (Delégerab et al., 2009), etc. Corpora can also assist in translating art-related documents which often pose great obstacles for translators since these texts are filled with a large amount of emotional-expressive vocabulary, adjectives, extended metaphors, comparisons, parallelisms or emotional reversals (Boyarkina, 2021). In fact, corpus linguistics has proved to be a powerful weapon for translation studies (Shen, 2010), expanding the research scope and leading to brand-new perspectives on the existing theories.

As a result of Vietnam’s rising international integration in the realm of arts, there is a high demand for the translation of Vietnamese artworks into foreign languages, especially into English- the most prominent language in the globe with more than 1.4 billion internationally users nowadays ("English," 2022). At the same time, when the issue of international cooperation is receiving more attention in Vietnamese art schools, learners and teachers are required to get used to certain activities like writing academic papers in English, writing research reports in English for presentation at international conferences, reading specialized art-related documents written by foreign authors, etc. This setting necessitates the translation of documents, research papers, and articles which contain contemporary art terms from Vietnamese to English and vice

Journal of Science, Hue University of Education
ISSN 1859-1612, No. 4(64)A/2022: pp.5-13
Received: 24/8/2022; Revised: 13/9/2022; Accepted: 14/9/2022

versa. However, in terms of reliable and accurate references, there is a serious shortage of English- Vietnamese documents consisting of contemporary art terms. Although there are several scientific research and articles on art terms, the contents of these studies only revolve around basic terms of arts, and none of them give a sharper focus on contemporary arts terminology. Almost no research has been done to create electronic versions of reference materials to assist users in translating contemporary art terms from English to Vietnamese. Due to this deficiency, translators are more likely to encounter great challenges when rendering contemporary art materials from English to Vietnamese and vice versa. The teaching and learning processes of lecturers and students in Vietnamese art institutions tend to be hampered as well. The products of this study are thereby hoped to serve as useful and practical references for translators, as well as students and teachers in Vietnamese art schools to deal with art-related texts, helping them to expand their knowledge of this field.

This study attempts to answer the following research question:

What are the procedures for building an English- Vietnamese parallel corpus of contemporary art terms?

2. LITERATURE REVIEW

2.1. Corpus

2.1.1 The definition of "corpus"

The term "corpus" (the singular form of "corpora") is of Latin origin and literally means "body" (of either a human or an animal). Its current meaning, a machine-readable structured collection of naturally occurring language created for specific purposes, can be traced back to the early 18th century. Nowadays, most scholars and researchers all around the world have come to a consensus on this definition of corpus. For example, Tognini-Bonelli (2001) proposes that a corpus is a collection of texts intended to be representative of a given language or a specific language variety and it can be used for linguistic analysis.

2.1.2. Applications of corpora

Recent technology advancements have sparked an increase in the use of corpora as reliable sources of empirical data for numerous investigations in a variety of scientific domains. In this context, scholars in around the world have attempted to present countless applications of corpora in their works. Huang and Yao (2015) assert that the first and most widely-known use of corpus is in lexicography and since then, corpus has become increasingly essential in general linguistic study. Although it is difficult to comprehensively quantify the involvement of corpora in the research methodology across all branches of linguistics, let alone all areas of science, Lüdeling and Kytö (2008) suggest an effective way to aggregate corpora's applications based on grouping types of data offered by corpora into three main kinds of information: (1) empirical support, (2) frequency information, and (3) meta-information. Today, modern corpora can be developed for both general descriptive purposes such as to respond to questions at a wide range of linguistic levels on the lexis, grammar, prosody, discourse patterns or pragmatics of the language; and specialized purposes such as discovering which words and their definitions should be contained in a learners' dictionary, etc. (Kennedy, 1998).

2.1.3. Building a corpus

Any choice must be based on a few criteria, so as for the process of compiling corpora. Overall, researchers have a consensus view on the criteria to consider when building a corpus, including:

the corpus size, the corpus data collection, the corpus data annotation and representation. In terms of the corpus size, while there are some assertions that bigger corpora are better for the linguistic hypothesis testing (Sampson, 2001), the entire issue of corpus size and what is judged an adequate size for a corpus vary depending heavily on the linguistic features being studied and the purpose of corpus inquiry. Regarding the data collection phase for the corpus, it entails both selecting and digitizing texts thanks to recent technological advancements. The sequence of locating and choosing suitable texts, converting them to appropriate electronic format, reviewing and annotating files result in the researcher gaining a far more profound understanding of the data's peculiarities.

2.1.4. Parallel corpus and its applications in the translation domain

The acceleration of parallel corpora has produced invaluable tools for cross-linguistic studies. In fact, they have been long exploited for a broad range of applications. In bilingual lexicography, parallel corpora serve as a source of representative examples of the real usage of a specific term from a huge body of text, as well as data on collocation frequency and quantification, etc. Regarding language pedagogy, they facilitate the processes of curriculum designing and language testing including test development, compilation and selection, test organization, response capture, test scoring, and calculation and delivery of results (as explained by Alderson, 1996). Moreover, several researchers have pointed out the potential applications of parallel corpora in training second language teachers (Conrad, 1999; O'Keeffe and Farr, 2003). With respect to the translation domain, it is believed that parallel corpora have an essential role in translator training, acting as a basis in studies on cross-linguistic terminology, computer-assisted translation, machine translation, and other cross-lingual information retrieval activities.

2.2. Issues surrounding contemporary art terms' translation

With respect to art terms translation, several scholars have proposed problems facing translators while rendering art-specialized discourses. First of all, these types of publications blend theoretical and critical writings, with both descriptive and informative contributions meant to influence the audience's reaction. As a result, the composite, hybrid structure of these documents creates big issues for the translation of their semantic and communicative contents (Pireddu, 2020). Because contemporary art employs a variety of techniques and media forms, frequently it can combine body language, emotional language, etc. which cannot be conveyed while translating. The language of art is often multi-sense and specialized, so the translation of contemporary art terms requires background knowledge of not only contemporary art history but also an understanding of sub-branches such as sculpture, painting, music, etc. To conclude, there is no doubt that contemporary art terms' translation can impose significant challenges on translators.

2.3. Previous studies

Since the advent of modern computers, corpus linguistics has returned to its peak with the advent of corpora of all sizes. Studies on the field of building corpora have been conducted throughout the globe, varying in the corpus language, size, and purposes. For example, while Caseli et al. (2009)'s research centers on building a Brazilian Portuguese parallel corpus of original and simplified texts, Tadic (2001) studies the procedures involved in building the Croatian-English Parallel Corpus. However, the quantity seems to be more limited regarding the number of English- Vietnamese parallel corpora. Some studies that include the construction of English- Vietnamese parallel corpora can be mentioned as Ngo and Dien (2004)'s study on building English- Vietnamese named entity corpus with aligned bilingual news articles, Dien

(2005)'s research regarding building an annotated English-Vietnamese parallel corpus, Dang and Ho (2007)'s investigation on the automatic construction of English-Vietnamese parallel corpus through Web mining, etc. However, none of these studies mined data related to the field of art, specifically contemporary art.

Despite the current pressing demand in the field of art for the translation of documents, research papers and articles from Vietnamese to English and vice versa, there is a severe lack of reliable and accurate references consisting of English- Vietnamese contemporary art terms. Certain related works proposed in Vietnam over the past few decades are Từ điển Mỹ thuật (Dictionary of Fine Arts) by Le Thanh Loc (1998), Từ điển thuật ngữ Mỹ thuật phổ thông (Dictionary of common Fine Art terms) by Dang Bich Ngan (2002), the Từ điển bách khoa Việt Nam (Literally Encyclopedic Dictionary of Vietnam) (Vietnam National Council, 2005), etc. Although several scientific research and articles on art terms can be found on certain websites, none of them place a stronger emphasis on contemporary arts terminology and instead just revolve around basic terms of arts. Moreover, little research has been done to create the electronic version of reference materials like an online English- Vietnamese dictionary of contemporary arts terms, or a machine-readable English- Vietnamese glossary of contemporary arts terms, etc.

All things consider, in order to address the present learning and research demands in Vietnam, it is imperative to create an English-Vietnamese parallel corpus of contemporary art terms.

3. METHODOLOGY

The study incorporates a combination of both quantitative and qualitative approach, given the fact that a mixed-method is ideal to seek answers to the research questions.

Data sources

Overall, a total number of 35 websites serve as reliable data sources for the corpus. English documents are extracted from 130 original articles, forming 130 English texts whose lengths range from 104 words to 2333 words. The majority of articles feature several pictures with captions, which is why only extractions rather than the entire original English texts are taken. In certain situations, the unnecessary details and image captions are deleted. Although some of the excerpts are modest in length, their quality is assured with a high density of contemporary art terms.

Data collection procedure

Step 1: At the beginning of this process, after selecting reputable sources for data collection, the researcher first selects articles that are relevant to the research topic. Next, she copies the entire content of the original article into the word file, including the accompanying images.

Step 2: The next step involves deleting redundant information in the file, which can be called word elimination. Based on the corpus' goals and the target users' needs, certain parts such as the publication date of the article, the contact information of the artist mentioned in the article, the image caption, etc. were omitted.

Step 3: Finally, the author saves the data by pressing *Ctrl + S*, then name the file and click *Save*.

Translating English source texts into Vietnamese using Smartcat

In this phase, the researchers translate the English source materials into Vietnamese with the help of Smartcat, a robust cloud-based CAT (computer-aided translation) tool. It helps to translate the source-language texts by segmenting the content into separate parts and storing them in the database. Smartcat also includes several quality control capabilities for auto-correction, spelling and grammatical checks, improving sentence structure, etc. Once a translation is complete, the pairs of source and translation text will be then stored in the translation memory. Compared to conventional desktop systems, this cloud-based translation memory system is more practical and user-friendly.

First, in order to use Smartcat, the researchers create an account in Smartcat. After that, the researchers clicks MY TASKS to create a new project, then uploads all the 130 existing files containing English source language texts from her computer into Smartcat. At this step, if the users have an existing translation memory, they can click Add to select the translation memory to help improve the quality of the machine translation. Next, the authors name the project, choose the source and target language, deadline (optional) and comments (optional). After the project has been successfully created, all the English source language texts will be split into separate segments. The authors select the available translation provided by the machine in the right part of the interface and edit it for accuracy.

Getting the statistics of the numbers of tokens, words, and sentences of the corpus

In order to get the details about the numbers of tokens, words, and sentences of each sub-corpus, on the interface of Sketch Engine, the researchers select the corpus name and then click CORPUS INFO.

4. FINDINGS AND DISCUSSIONS

4.1. Statistics of the number of tokens, words, sentences, and the type-token ratio (TTR)

Overall, the total number of words in the English- Vietnamese parallel corpus is 129,639. In particular, the English sub-corpus labeled “Contemporary art terms” contains 53,077 words, while the Vietnamese sub-corpus named “Thuật ngữ Nghệ thuật đương đại” contains 76,562 words in total. The utilities that are accessible on Sketch Engine make it simple to calculate these data. The details about the numbers of tokens, words, and sentences of each sub-corpus are illustrated in the following Table 1.

Table 1. *The total numbers of tokens, words, and sentences of each sub-corpus*

	English sub-corpus	**Vietnamese sub-corpus**
Tokens	61,059	84,878
Words	53,077	76,562
Sentences	2,778	2,830

It is obvious that the statistics for all of the above categories of the Vietnamese sub-corpus are much higher than those of the English sub-corpus. These figures reveal that in general, the Vietnamese translation is longer than the original English text, both in word count and in sentence count. This might result from the fact that the English source materials are specialized discourses written in an elaborate and sophisticated style, with a high density of specialized fine arts vocabulary. Therefore, in most cases, a single English vocabulary is aligned with its multi-word Vietnamese translation. For example, certain English nouns made up of one word such as “diaspora” and “artwork” are rendered into Vietnamese as “cộng đồng hải ngoại” and “tác phẩm nghệ thuật”, respectively, using four separate words. Another example is the one-word-made-up verb “explore”, which is translated into Vietnamese as “khám phá”. Take the following English sentence and its Vietnamese version as an example:

English original sentence:

(1) Agnes Martin was a key figure in the male-dominated field of abstract art in the USA.

Vietnamese translation:

(2) Họa sĩ trừu tượng Agnes Martin là một nhân vật chủ chốt trong lĩnh vực nghệ thuật trừu tượng do nam giới thống trị ở Mỹ.

While sentence (1) contains only 16 words, its Vietnamese equivalent comprises 26 separate words in total. Proper names referring to famous artists in the field of contemporary art, such as "Agnes Martin" in the above example, are often picked up to build the glossary. They are typically translated using a descriptive strategy to help the target readers better understand these characters. Under those circumstances, "Agnes Martin" is thus referred to as "Họa sĩ trừu tượng Agnes Martin" in the Vietnamese version, which is four-word longer than the source language term. Besides, adjectives like "key" or "male-dominated" in sentence (1) are alternatively rendered as "chủ chốt" and "do nam giới thống trị", contributing more words to the target language text.

Additional justifications for the larger size of the Vietnamese sub-corpus in comparison with the source language one pertain to the rule of forming plural form and possessive form of nouns in English, along with the norm of generating past tenses of English verbs. While the general rule in modern English for pluralizing nouns is to insert the suffixes *-s* or *-es* into the stems, Vietnamese requires specific quantifiers like "những", "các", etc. For instance, the plural noun "dreams" is interpreted as "những giấc mơ". Likewise, if English imposes the rule of producing past tenses by adding *-d* or *–ed* to non-finite verbs, Vietnamese language employs words such as "đã", "từng", etc. In terms of possessives, English people add an apostrophe and an *s* to the noun, whereas Vietnamese use the word "của". The following two sentences are an example:

English original sentence:

(3) Martin's work was influenced by Zen Buddhism and Chinese Tao philosophy.

Vietnamese translation:

(4) Tác phẩm của Martin chịu ảnh hưởng bởi Phật giáo Thiền tông và triết học Đạo Lão của Trung Quốc.

It can be observed from those sentences that the English expression for the idea that the work belongs to Martin consists of two words, whilst the Vietnamese translation contains four different words.

Another metric that can be exploited to get a better picture of the parallel corpus is the TTR. Although Sketch Engine does not present the TTR, the platform offers all the information required for the calculation. They can be found on the corpus info page where the LEXICON SIZES table provides the number of unique items in the corpus and the COUNTS table supplies information of the total numbers found in the corpus.

The calculation can be done using the figure for 'word' from LEXICON SIZES and the figure for 'tokens' from COUNTS as following:

$$\boxed{\text{TTR= word} \div \text{tokens} \times 100}$$

Using the above-given formula, the researcher managed to calculate the TTR of English sub-corpus, which is approximately 14.8 %, and that of Vietnamese sub-corpus, which is nearly 5.4%. The ratio is a measure of the corpus's lexical diversity, which means the degree of linguistic variance in the text increases as the TTR rises. According to Mcenery & Hardie (2012), if the ratio approaches 1 (or 100 percent), the vocabulary is considered to be more diverse. At this point, both sub-corpora have low TTR. However, this cannot be interpreted that the parallel corpus provides a negligible amount of contemporary art terms. Instead, it reveals the high frequency of repetition of each contemporary art term throughout the collected texts, implying that the English- Vietnamese parallel corpus provides a variety of contexts in which a particular term appears.

All things consider, the Vietnamese sub-corpus is of more lexical items than the original English version due to certain fundamental differences in the vocabulary construction of the two languages. Moreover, there is clear evidence to suggest that the English- Vietnamese parallel corpus provides a wide range of settings in which a single contemporary art term arises.

4.2. Statistics of Part of Speech (POS) in the English sub-corpus

Table 2 below provides information about 4 main kinds of POS including Noun, Verb, Adjective, and Adverb in the contemporary art terms English sub-corpus. They are displayed based on three categories: the number of different items (Lempos_lc), the total frequency, and the percent of the whole corpus.

On the one hand, Lempos is described by Sketch Engine as a positional attribute. It is a combination of lemma and part of speech (POS) containing the lemma, a hyphen and a one-letter abbreviation of the part of speech, e.g., "paint-v", "market-n". The part of speech abbreviations vary between corpora. Using this case-sensitive function, "art-n" is distinguished from "Art-n". On the other hand, Lempos_lc can be seen as a positional attribute, or a lowercased version of Lempos. All the existing uppercase letters are transformed to lowercase; thus "Art-n" is converted to "art-n".

Table 2. *Statistics of Part of Speech (POS) in the English sub-corpus*

	Number of different items (Lempos_lc)	**Total Frequency**	**Percent of whole corpus**
Noun	4,039	17,320	28.37%
Verb	1,051	8,236	13.49%
Adjective	1,222	4,873	7.981%
Adverb	341	2,112	3.459%

It can be seen that while Noun has the highest number of 4,039 different items and the highest total frequency of 17,320, the lowest number of 341 different items and the lowest total frequency of 2112 belong to Adverbs. Verb possesses a smaller number of different items than the figure for Adjective, which is 1051 and 1222, respectively. However, in terms of the total frequency, Verb ranks second with a figure of 8236, almost two times higher than the total frequency of Adjective, which is 4873. This indicates that there is a significant amount of verb overlap in the text.

5. CONCLUSION AND IMPLICATION

To conclude, an English- Vietnamese parallel corpus of contemporary art terms was successfully built by performing many stages, combining the efforts of the researchers and the effective help of computer tools. Text materials totaling 53,077 words used to create the corpus were carefully collected from 35 reliable websites. The source language documents are 130 news extracts with a variety of themes encompassing seven contemporary art forms, ranging from architecture, film, literature, music, painting, to sculpture and theater. Smartcat, a CAT tool, plays an indispensable role in hastening the translating stage while compiling the corpus. Besides, the system provides a great quantity of corpus essential data such as the number of tokens, words, sentences, extracted terms, etc. for the sake of corpus mining and analyzing.

Regarding possible applications of the research product, the most obvious and direct application lies in the field of translation. First of all, for translators, the output of this study is expected to serve as a reliable reference source, especially when dealing with art-related materials. To be more specific, the parallel corpus can be easily accessed and mined using the tools provided by Sketch Engine. For example, the system permits the generation of bilingual concordances, which can be used to identify equivalent expressions and examine linguistic differences. In other word, the aligned parallel corpus is hoped to act as a source of translation suggestions.

As for teachers and students, the study is expected to help expedite teaching and learning progress in Vietnamese art schools. First, because the corpus's contents span a wide range of art-related issues, users can expand their understanding of contemporary art by exploiting the English- Vietnamese parallel corpus of contemporary art terms as well as the documents carrying aligned English-Vietnamese sentences. Second, art-majored students tend to be exposed to a great deal of specialized material in English, which might cause a considerable delay in reading comprehension for those who are not proficient in the language. In this context, the product of the present study is hoped to act as a bilingual source of reference, leading to an adequate interpretation of English contemporary art texts. Last but not least, the corpus machine-readable value enables English teachers to instantly extract all authentic, typical examples of a lexical item's usage from a large body of text.

REFERENCES

[1] Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. Longman.

[2] Boyarkina, A. (2021). Translating terminology of media texts dealing with art and culture (in German-Russian texts). *Translation Studies: Theory and Practice*, *1*(1).

[3] Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., & Aluisio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. 10th Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

[4] Conrad, S. (1999). The importance of corpus-based research for language teachers. *System*, *27*, 1–18.

[5] Đặng, B. N. (2002). *Từ điển thuật ngữ Mỹ thuật phổ thông*. NXB Giáo Dục.

[6] Dang, V. B., & Ho, B. (2007). Automatic Construction of English-Vietnamese Parallel Corpus through Web Mining. *IEEE International Conference on Research, Innovation and Vision for the Future*, 261-266.

[7] Delégerab, L., Merkelc, M., & Zweigenbaum, P. (2009). Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, *42*(4).

[8] Dien, D. (2005). Building an annotated English-Vietnamese parallel corpus. *The Mon-Khmer Studies Journal*, *35*, 21-36.

[9] English. (2022). In D. M. Eberhard, G. F. Simons, & C. D. Fennig (Eds.), *Ethnologue: Languages of the World* (25th ed.). Dallas, Texas: SIL International.

[10] Huang, C.-R., & Yao, Y. (2015). Corpus linguistics. *International Encyclopedia of the Social & Behavioral Sciences*, *4*, 949-953.

[11] Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman.

[12] Klabal, O. (2019). Corpora in legal translation: Overcoming terminological and phraseological assymetries between Czech and English. *CLINA*, *5*(2), 165-186.

[13] Lê, T. L. (1998). *Từ điển Mỹ thuật*. Công ty sách Thời Đại & NXB VHTT.

[14] Lüdeling, A., & Kytö, M. (2008). Introduction. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*. Walter de Gruyter.

[15] Ngo, Q. H., & Dien, D. (2004). Building English-Vietnamese named entity corpus with aligned bilingual news articles. 25th International Conference on Computational Linguistics, Dublin, Ireland.

[16] O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: pedagogic issues and practical applications. *TESOL Quarterly*, *37*(3), 389–418.

[17] Pireddu, S. (2020). Translating art catalogues: theoretical and practical issues. *Imaginations: Journal of Cross-Cultural Image Studies*, *11*(3).

[18] Sampson, G. (2001). *Empirical linguistics*. Continuum.

[19] Shen, G.-r. (2010). Corpus-based approaches to Translation Studies. *Cross-Cultural Communication*, *6*(4).

[20] Tadic, M. (2001). Procedures in building the Croatian-English parallel corpus. *International Journal of Corpus Linguistics*, *6*, 107-123.

[21] Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins Publishing Company.

[22] Vietnam National Council. (2005). *Từ điển bách khoa Việt Nam (Literally Encyclopaedic Dictionary of Vietnam).* Vietnam's Encyclopedia Publishing House.

[23] Mcenery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.