A REVIEW OF TRENDS IN BUILDING VIETNAMESE BILINGUAL CORPUS FOR TRANSLATION TEACHING AND LEARNING

LE MINH THANH
Industrial University of Ho Chi Minh City
Corresponding author: leminhthanh@iuh.edu.vn

Abstract. The utilization of Vietnamese Bilingual Corpus (VBC) for translation teaching and learning, particularly in Vietnam, has been inconsistent. While there are abundant gold standard corpora available for English, the same cannot be said for less-resourced languages like Vietnamese (Ngoc Tan Le & Fatiha Sadat, 2015). Therefore, the purpose of this review is to present a comprehensive overview of the existing research in this field, providing a valuable resource for researchers and educators. The review aims to achieve the following objectives: (a) Identify current concerns, trends, and research gaps that can guide future investigations in the teaching and learning of foreign language translation; (b) Explore innovative and effective approaches to integrate Vietnamese computational linguistics into translation teaching and learning practices. The utilization of VBC is rapidly evolving to address various aspects, including maintaining academic integrity, enhancing foreign language teaching and learning, and incorporating artificial intelligence in translation.

Keywords. Vietnamese, bilingual corpus, teaching, learning, translation.

1 INTRODUCTION

The term “corpus” or “corpora” refers to a collection of documents or parts of an electronic document selected and arranged according to certain linguistic criteria (Scherer, 2006). The data of the corpus may consist of a text (written) and a spoken (speech) form that has been digitized and can be searched on a computer. The language of the corpus can be single (monolingual), bilingual, or multilingual. The corpus of a bilingual corpus can be arranged in a parallel form (Parallel Corpus) (Mcenery & Wilson, 2011). The corpus is also used as the basis for the quick and efficient preparation of various modern and historical dictionaries. The role of corpus linguistics is further confirmed when research on corpus shows that corpus can be used to build skills and check grammar in the process of foreign language teaching and translation. According to Chomsky, the object of corpus linguistics is language performance, as the linguist must seek to model language competence rather than performace. “The reasons for the revival of corpus linguistics should now be quite obvious. It is, in some ways, an attempt to redress the balance in linguistics between the use of naturally occurring data…Corpus linguistics is, and should be, a synthesis of introspective and observational procedures, relying on a mix of artificial and natural observation” (Mcenery, 2019, p.19). The corpus can reflect the linguistic phenomena that the language community uses in practice, and has high value in language learning and research. With a large, reasonably sampled corpus, most linguistic phenomena will be reflected more accurately, vividly, and up-to-date than in a dictionary (Dinh, 2018).

As global communication and collaboration become the norm, the role of translation in bridging linguistic and cultural gaps has grown in significance. Manual translation must be a time-consuming process. Translation memory becomes an important language techonology to make translator’s task faster, more efficient and easier. Translation quality and speed have been greatly improved with the emergence of neural machine translation or corpus-based translation. Vietnameseranks among the less widely spoken languages, thereby accounting for the comparatively limited corpus research dedicated to the Vietnamese language. Combining both automatic strategies and manual methods allows us to mitigate the problems and take advantage of the benefits. In this context, the development of customized bilingual corpora for specific languages is a critical effort, especially in the field of translation teaching and learning. This paper presents a comprehensive exploration of trends and advances in building Vietnamese bilingual corpora to support translation teaching and learning.

The research aims to answer the following questions:

1. What are the current concerns, trends, and research gaps that can guide future investigations in the teaching and learning of foreign language translation?

© 2026 Industrial University of Ho Chi Minh City

2. Which are Innovative and effective approaches to integrate Vietnamese computational linguistics into translation teaching and learning practices?

2 METHODS

This review focuses on relavant research in building VBC for translation teaching and learning within academic institutions. This study adopted a scoping study framework that followed the stages of the framework (Arksey & O'Malley, 2005). This framewok has the following stages:
(1) Identifying the research question,
(2) Identifying relevant studies,
(3) Study selection,
(4) Charting the data,
(5) Collating, summarising and reporting the results.

3 RESULTS

3.1. Identifying the research question

Our research have two main questions: (1) What are the concerns, trends, and gaps to lead future research for teaching and learning foreign language?; (2) Which are Innovative and effective methods for applying Vietnamese computational linguistics to teaching and learning practices?

3.2. Identifying relevant studies

Conducting research by utilizing databases to retrieve pertinent scientific articles, both in the broader context of scientific journals and with a specific focus on Computational Linguistics. These databases encompass a wide range of sources, both within and outside Vietnam, including Da Nang Journal of Science and Technology University, and the Science Journal of Hue University, ACL Anthology, China National Knowledge Infrastructure (CNKI), ResearchGate, Institute of Electrical and Electronics Engineers (IEEE), Multidisciplinary Digital Publishing Institute (MDPI).
The references are suitable for research requirements and meet the following criteria:
- The references are written in Chinese, English, and Vietnamese;
- The entire text is readily available for access;
- The researchs focus encompasses a bilingual corpus that integrates the Vietnamese language.

3.3. Study selection

Following the selection of appropriate articles for our review, we categorized them based on the following criteria: publication time, research object, written language, form and size of the corpus, the application of the corpus. A total of 28 articles were used for the review, during the period from 2003 to 2022, of which 24 articles were written in English, 3 articles were written in Vietnamese and 1 articles were written in Chinese.

3.4. Charting the data

In terms of structure, the article “CanVEC - the Canberra Vietnamese-English Code-switching Natural Speech Corpus” by Nguyen & Bryant (2020) builds a corpus consisting of speech corpus, while the rest of the articles build a corpus in the form of text.
The majority of the bilingual corpus has been created for English-Vietnamese, consisting of 4 out of 28 articles, while 1 article has been created for Vietnamese - English. Furthermore, 2 articles are dedicated to the development of the Vietnamese - French corpus, and 2 articles are devoted to creating the Chinese-Vietnamese corpus. In particular, the paper by Trieu & Le Nguyen (2017) built a corpus of four Asian languages (Indonesian, Filipino, Malay, and Vietnamese) - English. In addition, the research by Hoang & Phan (2014) studies the Vietnamese - Ede corpus, one of the less common language pairs. Besides, there is also 1 research paper on Korean - Vietnamese corpus (2019).

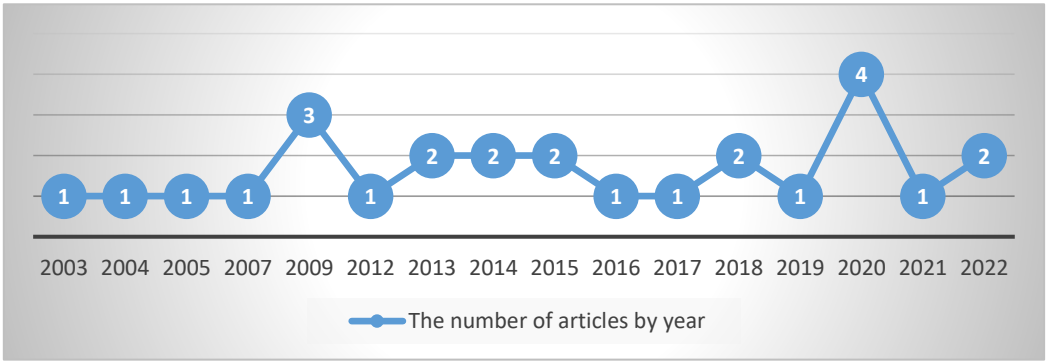


Figure 1: The number of academic articles on building bilingual corpus containing Vietnamese (2003 – 2022).

Based on the charts, it can be observed that the years 2009 and 2020 yielded the most fruitful results in this field of research. Although the research output in 2022 was lower compared to the previous two years, the level of attention and interest remains sustained.

Year	Author	Object	Size
2003	Dinh & Hoang, 2003	English - Vietnamese	more than 9,500 sentences
2005	Dinh, 2005	English - Vietnamese	more than 5,000,000 words
2009	Le, V. B. et al., 2009	Vietnamese - French	50,300 pairs of sentences
2012	Ngo & Winiwarter, 2012	English - Vietnamese	more than 800,000 pairs of sentences
2014	Hoang & Phan, 2014	Vietnamese - Ede	more than 22,000 sentences
2015	Le & Sadat, 2015	Vietnamese - French	20,000 pairs of sentences
2017	Trieu & Le Nguyen, 2017	Asian Languages	more than 1,100,000 pairs of sentences
2018	Nguyen et al., 2018	Vietnamese - English	more than 12,000 pairs of sentences
2018	王浩Wanghao, 2018	Chinese - Vietnamese	5,000 sentences
2019	Nguyen et al., 2019	Korean - Vietnamese	5,000,000 words

2020	Nguyen & Bryant, 2020	English - Vietnamese	10 hours of recording
2022	Tran et al., 2022	Chinese - Vietnamese	more than 400,000 pairs of sentences

Table 1: Reporting VBC important results (2003 – 2022)

The primary objective of the majority of corpora is to employ technological advancements in the creation of comprehensive bilingual and multilingual parallel language corpora, with the ultimate goal of enhancing machine translation.

3.5. Collating, summarising and reporting the results

Currently, the discourse concerning the utilization of VBC mainly centers on machine translation rather than language pedagogy or translation pedagogy. Most of the research aims to apply scientific technology to corpus construction in order to expand corpus capacity and enhance the effectiveness of machine translation.

The use of native speaker data arises was a useful choice for learners and teachers. Because Corpus can provide many examples of patterns of use, they have application to language classroom. Instructors can use it to improve their teaching methods with an upper-intermediate group, or introduce the application of corpora to students to help students prepare lessons. A teacher may advise students to search a corpus to find out what preposition most commonly folloew bargain as a verb. The advantage of looking up a lexico-grammatical query in a corpus is that it provides us with many examples of the search item in its context of use. However, a corpus will not tell us the meaning of the word or phrase, making it specifically suitable for an upper-intermediate group. Combining a dictionary and corpus can be valuable route in a pedagogical context. Nevertheless, corpora are expensive and difficult to put together. From an applied corpus linguist's perspective, it played a critical role but is not easily accessible to teachers and students (O'Keeffe et al., 2007).

Corpora have not only revolutionized linguistic research over the last few decades, but they have also had an impact on second language learning and teaching. More and more researchers and practitioners in the field of applied linguistics value what corpus linguistics has to offer language pedagogy. Nonetheless, corpora and corpus tools have yet to be widely adopted in pedagogical settings (Römer, 2011).

Farr (2008) pointed out that the use of corpora met some identified difficulties. The negative aspects of learning through the use of corpora are indicated below: The technical aspects are discouraging, with instances of corrupted data in certain corpora; The existence of numerous varieties makes it complex to apply the technology in the classroom, posing challenges for classroom application; The required time can be frustrating; Accessing and analyzing mega corpora is challenging; There is a lack of confidence in personal interpretations at times; Colloquial language use is confusing, particularly as spoken language tends to be messy. However, the continued integration of corpus-based instruction in the language content component of language teacher education programs should be encouraged.

Despite undeniable progress in the field of applied corpus linguistics, there is still significant room for advancement. A number of tasks can be developed to improve pedagogical corpus applications and help corpora contribute more to improving language learning and teaching (Römer, 2011). Since the mid-1980s, the role of corpora in language learning has received increasing attention in research. Chambers (2019) explain the gap between research and practice by using corpora and proposed a solution to bridge the research–practice gap. The true motive in this field of research is the aim to enable teachers and learners to benefit from corpus linguistics resources and approaches. The starting point in many studies is the needs of teachers and learners, but the corpus is usually dominant and lacks strategies to apply to translation teaching and learning. Researchers would also serve to inform the content of the supplementary resources which major publishers could provide to encourage language teachers to actively explore their corpus resources with their learners. In addition, the alignment with the CEFR (Common European Framework for Reference) used in the pedagogical enhancement of the corpora could also make an important

contribution to the integration of corpus data alongside other language resources in the teaching environment. Finally, the training of learners for corpus consultation and analysis would be an important consideration.

Both instructors and students have benefited from corpus-based translation education. The complexity of corpora can occasionally scare students. A corpus suitable for various courses also requires a lot of effort by the organization that researches and develops advanced concordancers (Abdinazarovna & Rustamovna, 2022). Ma et al. (2022) indicate that teachers not only developed the corpus-based language pedagogy primarily from a strong knowledge of corpus technology but also drew on solid pedagogical knowledge. VBC's development is an efficient way of meeting the needs of foreign language teaching and translation teaching that requires assistance from a variety of aspects.

Abdinazarovna & Rustamovna (2022) explore three approaches for educators to offer learners hands-on corpus activities. Initially, educators have the option to incorporate material obtained from corpus searches, enabling students to engage with the prepared content provided by the teacher. Additionally, teachers can utilize online corpora, with a particular emphasis on four user-friendly options (COCA, Time, MICASE, and MICUSP) within this section. Lastly, instructors can introduce established corpora or develop tailored corpora for their classes, such as those derived from readings or student papers, encouraging students to actively interact with the corpora. The authors point out experiencing language in this manner can assist learners in gaining a clearer understanding of usage patterns and the words that commonly appear alongside new vocabulary. In addition, Dinh D (2018) has indicated the benefits of using language corpora in translation work. In the process of learning translation, learners can explore collocates, analyze concordance lines, and summarize the meanings of the new words. Ma et al. (2022) identified two effective measures for promoting computer-assisted technologies in teaching and learning. First, provide educators with instruction and training in the theoretical aspects of corpus-based language pedagogy and research. Addressing technical issues related to corpus use and exploring strategies to motivate educators to adopt a seemingly complex new pedagogy - corpus-based language pedagogy.

The application of corpus to language teaching is a contemporary trend. The corpus based on the Vietnamese language is currently not large in scale, and its capacity and quantity are still severely limited. Therefore, the use of the Vietnamese corpus does not play its proper role in language teaching or translation teaching. Vietnamese corpora are currently only available in a few languages, the majority of which are Vietnamese and English corpora. Other language corpora are generally scarce. More vigor is required in this regard. Although many corpora have been created, they are only available to research organizations and are not recommended to the general public. As a result, teachers and students find it difficult to apply corpora to teaching and learning.

The innovative and effective application of Vietnamese computational linguistics to language teaching needs to draw on the research achievements of scholars. To begin, bridge the gap between research and practice; Second, to align with the CEFR used in the pedagogical enhancement of the corpus; Third, to develop a learning corpus for language teaching; Fourth, to improve corpus-based language pedagogy so as to facilitate the use of corpus.

4 CONCLUSION

Vietnamese is one of the less popular languages in the world, which is why research on Vietnamese corpora, especially bilingual corpora, is very limited. Nevertheless, it is encouraging to note that the size and quality of bilingual corpora containing Vietnamese have improved over time, although no studies have been found on the application of established corpora to translation teaching and learning. Therefore, future research focusing on the application of bilingual corpora is encouraged, particularly in the context of teaching and learning Vietnamese translation.

5 REFERENCES

[1] Abdinazarovna, S. G., & Rustamovna, U. R. (2022). *Using corpora in the language teaching and learning*. Web of Scientist: International Scientific Research Journal, 3(3), 500-508.

[2] Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. International journal of social research methodology, 8(1), 19-32.

[3] Chambers, A. (2019). *Towards the corpus revolution? Bridging the research–practice gap*. Language Teaching, 52(4), 460-475.

[4] Dien, D., & Kiem, H. (2003). *POS-tagger for English-Vietnamese bilingual corpus*. In Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond (pp. 88-95).

[5] Dinh, D. (2005). *Building an annotated English-Vietnamese parallel corpus*. MONKHMER STUDIES, 35, 21.

[6] Dinh, D. (2018). *Ngôn ngữ học ngữ liệu*. Nhà xuất bản Đại học Quốc gia TP.HCM.

[7] Farr, F. (2008). *Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users*. Language Awareness, 17(1), 25-43.

[8] Hoang, T. M. L. & Phan, H. K. (2014). *Giải pháp xây dựng kho ngữ liệu đa ngữ Việt-ÊĐê gán nhãn theo ngữ cảnh (Solution to build a Vietnamese-Ede multilingual corpus with contextual labels)*. Tạp Chí Khoa học Và Công nghệ - Đại học Đà Nẵng.

[9] Le, N. T., & Sadat, F. (2015, June). *Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross*-Linguistic *Projection*. In Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Démonstrations (pp. 12-13).

[10] Luong, A. V., Nguyen, D., & Dinh, D. (2020). *Building a corpus for Vietnamese text readability assessment in the literature domain*. Universal Journal of Educational Research, 8(10), 4996-5004.

[11] Ma, Q., Yuan, R., Cheung, L. M. E., & Yang, J. (2022). *Teacher paths for developing corpus-based language pedagogy: a* case *study*. Computer Assisted Language Learning, 1-32.

[12] Mcenery, T. and Wilson, A. (2011). *Corpus Linguistics*. Cambridge University Press.

[13] Mcenery, Tony. (2019). *Corpus linguistics*. Edinburgh University Press.

[14] Nguyen, L., & Bryant, C. (2020, May). *CanVEC-the Canberra Vietnamese-English Code-switching Natural Speech* Corpus. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 4121-4129).

[15] Ngo, Q. H., Dien, D., & Winiwarter, W. (2014, August). *Building English-Vietnamese named entity corpus with aligned bilingual news articles*. In Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing (pp. 85-93).

[16] Nguyen, P. T., Vu, X. L., Nguyen, T. M. H., & Le, H. P. (2009, August). *Building a large syntactically-annotated corpus of Vietnamese*. In The Third Linguistic Annotation Workshop-The LAW III (p. 6p).

[17] Nguyen, Q. P., Vo, A. D., Shin, J. C., Tran, P., & Ock, C. Y. (2019). *Korean-vietnamese neural machine translation system with korean morphological analysis and word sense disambiguation*. IEEE Access, 7, 32602-32616.

[18] Nguyen, T. H., Nguyen, T. M. H. & Nguyen M. H. (2018). *Xây dựng kho ngữ liệu du lịch song ngữ Việt–Anh gióng hàng mức câu cho* dịch *máy (Building a Vietnamese-English bilingual tourism corpus with sentence level alignment for machine translation)*. Tạp chí Thông tin và Truyền thông (pp. 9-16).

[19] O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

[20] Thao, P. T., Tri, T. Q., Kawazoe, A., Dinh, D., & Collier, N. (2007). *Construction of Vietnamese corpora for named entity recognition*. In Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (pp. 719-724).

[21] Ngo, Q. H., Winiwarter, W., & Wloka, B. (2013, October). *EVB Corpus-a multi-layer English-Vietnamese bilingual* corpus *for studying tasks in comparative linguistics*. In Proceedings of the 11th Workshop on Asian Language Resources (pp. 1-9).

[22] Römer, U. (2011). *Corpus research applications in second language teaching*. Annual review of applied linguistics, 31, 205-225.

[23] Scherer, C. (2006). *Korpuslinguistik. (Corpus linguistics)*. Universitätsverlag Winter.

[24] Thu, Y. K., Pa, W. P., Utiyama, M., Finch, A., & Sumita, E. (2016, May). *Introducing the Asian language treebank (ALT)*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 1574-1578).

[25] Tran, P., Nguyen, T., Vu, D. H., Tran, H. A., & Vo, B. (2022). *A method of Chinese-Vietnamese bilingual corpus* construction *for machine translation*. IEEE Access, 10, 78928-78938.

[26] Trieu, H. L., & Le Nguyen, M. (2017). *A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages*. In Proceedings of Machine Translation Summit XVI: Research Track (pp. 268-281).

[27] Van Huynh, T., *Van* Nguyen, K., & Nguyen, N. L. T. (2022, October). *ViNLI: a Vietnamese corpus for studies on open-domain natural language inference*. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3858-3872).

[28] Vo, D. N. & *Dinh*, D. (2020). *Các tiêu chí ngôn ngữ trong việc xây dựng kho ngữ liệu Tiếng Việt (Language criteria in building Vietnamese corpus)*. Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Nha Trang.

Received on August 20 – 2026
Revised on October 08 – 2026