

## FOREIGN LANGUAGES AND VIETNAMESE

**BUILDING CORPORA AND USING  
CORPUS-BASED METHODS IN TRANSLATION STUDIES****LE CHI HIEU\* - TRAN THI KIM HUE\*\* - TRAN THI THUY DUNG\*\*\***

**ABSTRACT:** The aim of this paper is to review some relevant factors related to the corpus linguistics such as using corpus-based method and its using in translation studies. Firstly, we describe some definitions of corpus, its developments in linguistics, especially in context of Vietnam (it is called corpus linguistics) and some kinds of corpora such as parallel corpus, comparable corpus, etc... Then, we attempt to present some tools for exploiting the corpus called concordance such as Antconc, Concordance, etc... They are useful in using the electric corpora. Finally, we outline the implementation of corpora in translation studies.

**KEY WORDS:** Corpus; parallel corpus; using corpus in translation.

**RECEIVED:** October, 23<sup>rd</sup>, 2025.

**EDITED:** November, 14<sup>th</sup>, 2025

**1. Introduction**

Corpus linguistics was born with John Sinclair and the Cobuild project at the University of Birmingham. He defined it as “a collection of naturally occurring language texts, chosen to characterize a state or variety of a language” (Sinclair, 1991: 171). In general, corpus is a collection of natural language (text, and/or transcription of speech) constructed with a specific purpose. Ideally, a corpus is a set of language production samples designed to be representative of a language through careful selection. Until 90's decade, with the development of ICT, corpus and corpus linguistics are clear outcomes of these technological advances.

Now, the linguists apply more and more the corpus in researching and studying of translations. It is useful to discover the term in real context, to help them creating the automatic translating tools and to support teachers and students comparing language A to language B through parallel and comparable corpus. Especially, corpus is so necessary in learning new words and find collocations or scientific terms in the original contexts.

**2. General review of using corpora**

The use of real textual examples in the study of translation is not new issue in applied linguistics. However, using corpora in translation has developed considerably in the last decades. The availability of computer and electric tools helps linguistics and students getting data quickly and easily. Since 90s and 2000s years, the universities in Europe and France trained corpus linguistics courses in their learning curriculums. For example, University Paris 7 develops some tools to help students using corpora in learning English in EILA department. Besides, colleagues in Malaysia start to build corpora serving students and teachers in learning and teaching ESL such as a corpus of English language of Malaysia School Students as EMAS Corpus created by the University Putra Malaysia. These two corpora were developed by the University of Malaya. The first is the MACLE corpus (Malaysia corpus of Learner English) based on student's essays. The second is the COMEL corpus (Corpus of Malay English) based on spoken corpus project. In Vietnam, building many corpora for Vietnamese language is real and necessary demand. But until now we don't have yet one corpus in national level. The fact that introducing learning corpora in translation practice in SLT school-HaUi is new trend of learning and researching approach.

\*Dr; School of Languages and Tourism-HaUi; Email: hieulc@hau.edu.vn

\*\*School of Languages and Tourism-HaUi; Email: tranthikimhue@hau.edu.vn

\*\*\*School of Languages and Tourism-HaUi; Email: dungttt@hau.edu.vn

The screenshot shows the British National Corpus (BNC) search interface. At the top, there are tabs for SEARCH, FREQUENCY, CONTEXT, and ACCOUNT. Below the tabs, there are search filters and a list of concordance results. The word 'president' is highlighted in green in the search results. The interface also includes a search bar, a list of search filters, and a table of search results.

| ENTRY                  | TEXT                                                                                                                                                                          |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 D95 S_meeting        | a situation which is the shame of our Country, that's signed by the <b>President</b> of the British Pensioner's and Fred (---) the Secretary. That's issued by                |
| 2 F87 S_meeting        | ... they come from the records of Age Concern Scotland of which I am the <b>president</b> . Here are a couple who are doing very awkward cooking over an open fire            |
| 3 G3U S_meeting        | everyone we could think of. To the company, the union, even the <b>president</b> . (SP-G3UP5002) (reading) Our employers took notice of the letters. They learnt of the       |
| 4 HYE S_meeting        | the U K recession. In the U S A we've appointed John (---) <b>President</b> and so we've got the structure that we wanted, with Peter (---) as                                |
| 5 HYE S_meeting        | With a Managing Director in U K, Trevor (---) and a Managing Director, <b>President</b> in the U S A, John (---) and er, John (---) had a                                     |
| 6 CH1 W_newsp_tabloid  | cripple home-owners and industries far beyond his own borders. In Europe, only French <b>President</b> Francois Mitterrand gets anywhere near to the status of politi         |
| 7 CH1 W_newsp_tabloid  | 12 members of the public and a leading anti-Maastricht politician. Here was a back-to-the-wall <b>President</b> who BELIEVED in what he was saying and who shifted og         |
| 8 CH1 W_newsp_tabloid  | fence-sitting gutlessness of Honest John Major, our very own world leader and EC <b>President</b> who refused to take part in the French TV debate for fear of upsettir       |
| 9 CH1 W_newsp_tabloid  | has stirred up royal fever -- but not everyone is enjoying it. Like Russian <b>president</b> Boris Yeltsin who still stands accused of continuing the cover-up over the execu |
| 10 CH1 W_newsp_tabloid | lesbians.' -- Pat Robertson, US TV evangelist and campaigner for the re-election of <b>President</b> Bush in the Los Angeles Times, August 27. # FINGERS ON THE TRIGGGI       |
| 11 CH1 W_newsp_tabloid | the pound was suffering, the French gloated at our humiliation. Then France's <b>President</b> Mitterrand and Germany's Chancellor Kohl met to sort out their own ver         |
| 12 CH1 W_newsp_tabloid | sceptical EC country convulsed by troubles over Europe. And he must meet the French <b>President</b> , Francois Mitterrand, who, in alliance with German Chancellor H         |

*Illustration 1. Example for concordance of word “president” in BNC*

A parallel corpus is defined as corpus which contains source texts and their translations. It can be bilingual or multilingual corpus. They can be uni-directional (from English to French or from French to English, for example) or bi-directional (having both English source text with their French translations as well as French source texts with their English translations). Baker (1995) point out that a parallel corpus consists of texts originally written in language source alongside their translations into a language target. In fact, parallel corpus has already been compiled in several pair of language such as English-French, English-Italian, English-Norwegian, English-German, etc... (but not yet for Vietnamese language).

Parallel corpus can be used to provide information on translational behavior and to show the equivalence relationship between lexical items or structures in source and target languages or to study the phenomenon of translation (Schmied and al, 1996). In addition, it includes translator training, bilingual lexicography and machine translation. It supplies useful data for psycho linguists which investigate the differences between first and second language acquisition (Malmkjler, 1993).

Long time before, Sinclair (1992) pointed briefly out the corpus-based methods in translation “the new corpus resources are expected to have a profound effect on the translation of the future. ...In principle, the corpora can provide the information”. This is considered as the dawn toward corpus-based in translation.

Early, Baker (1999) starts to analyze similarities and differences between sets of translation into the same language by different translators. Baker’ research is very suggestive in translation and becomes the inspiration for various projects of corpus’s influence in translation. Then, Laviosa (2008) builds the English Comparable Corpus (ECC), created by putting together the Translational English Corpus (TEC) and texts extracted from the British National Corpus (BNC).

In China, Huang and Wan (Farahani M., 2023) utilize the comparable corpora BNC and Gigaword Chinese Corpus to explain the principles of translation. It suggests that the cultural differences in source and target languages can be used to underline the nature of translation.

In Vietnam, research of using corpora in translation is limited in Scientific journal and theoretical practice. We try to introduce the using corpora in Translation Practice discipline for students leaning Vietnamese language or from different foreign languages Departments. We hope that this is the new stage of applying language in Translation profession.

### 3. Methodology

#### *Using corpus-based method in translation studies:*

For a translator, corpus is one of the source of linguistic information when searching for translation equivalences or when seeking to produce a translational function. Firstly, most translators still rely on dictionaries to seek the lexical information. Bowker and Pearson (2002, 15) list five problems with dictionaries, especially in the context of LSP (language for specific purposes):

- Incompleteness: it takes a long time to compile a dictionary. So in many case it is not reflect the current state of knowledge or language.
- Size: normally, most dictionaries are still compiled for printed version.
- Lack of context's information.
- Lack of frequency information: the translators always want to know the frequency of lexical items.
- Difficult to find the relevant information: in printed dictionaries how to find the lexical items where it locates.

Then Kubler N. (2003) cites the connection of corpora and translation that can be seen from two perspectives: Learning using corpora to translate: using corpora as the tools to find linguistic information in the translation process; Learning to translate using corpora: studying the process of translating using corpora (Beeby, 2009). Baker and al. (1993) publishes her paper “corpus linguistic and translation studies: implication and applications” in honor of John Sinclair and predicts “the availability of large corpora of both original and translated text ... will enable translation scholar to uncover the nature of translated texts”. Two years later in the “corpora in translation studies: an overview and some suggestions for future research”, that idea was further developed by suggesting the research of designing parallel, comparable, multilingual corpora. It was the dawn of corpus-based translation.

In fact, using corpus in translation becomes the new subfield of research concerning the practice and training of translator which links strongly with disciplines such as contrastive analysis, language for special purposes (LSP), terminology, lexicography and computational linguistics. In addition, Gavioli and Zanetti (1997) carried out a study where translator trainees designed, compiled and analysed a comparable English-Italian corpus of medical research articles. Hatim (1999) praised corpus-based translation “...is truly new wave of research...”. Furthermore, Baker (1995, 232) used the term “multilingual corpus” to refer “set of two or more monolingual corpora in different languages, ...”.

So multilingual corpus consists of texts that have not been translated. It can be used for contrastive linguistic works and bilingual lexicography. However, Baker consider the usefulness of multilingual corpora in translation studies, claiming that “a natural way ... to find out how to say something naturally in language A and language B”. In fact, corpus provides interesting technics for statistical items and even collocational patterns in vast quantities of texts (Clear, 1993). And such techniques are being extended to bilingual corpora. Therefore, corpora and the software for processing texts provide translators the powerful tools for studying the very nature of translation.

Baker (1995) proposes using comparable corpus in translation studies. It consists of a collection of translated texts (from one or more languages) into English. She outlines the using of comparable corpora in translation “translated texts tend to be more explicit, unambiguous and grammatical conventional”. In fact, comparable corpus technics allow much more powerful about translation. For example, translated texts display a high type-token ratio, low lexical density and low sentence length (Scott, 1996).

In context of this research, we apply the corpus-based method for students in School of Language and Tourism (SLT) and in University of Foreign Trade of Hanoi (UFTH) with about 80 students. Fortunately, in SLT, we officially teach Corpus Linguistics course for student from 2025. So using corpus is quite simple for them. Otherwise, Student from UFTH are very new for this method in learning the foreign languages and applying in translation studies. So they have some difficulties in using it.

### *Building a corpus:*

One of the principal issues in corpus-based method is to create up a corpus. So we have to answer the questions below: How big the corpus is? How many texts should be included? What genres and text types should be represented in the corpus? In fact, since the corpus is used in linguistic research, the question “how building up a corpus?” is always discussed. For serving to build up a corpus, we concentrate to documents in domain of restaurants and hotels. We collect about 100 documents in both English and Vietnamese to build up 2 corpora, one by English and another by Vietnamese. We save these corpora under the format of .txt and coding of UTF-8. Then we named these under the name Vncorpus and Engcorpus. Each corpus has about 2-3000 words. By these corpora, we can find the words most used in the field of restaurants and hotels which serve for our research in both two languages: English and Vietnamese.

## 4. Results and Discussion

### 4.1. Results: Corpora in Translation Profession

In fact, most translators have to use the computer in their profession because the translations are usually produced in format of electronic texts. Thus, the translation works are made in computer-assisted translation (CAT) environment. Actually, translation memory software contributes to change the translation profession because it involves specific skills and translational knowledge. It helps the translators saving time, ensure the accuracy and productivity of texts. So the translation memory database becomes a sort of parallel corpus in which a new translated sentence is matched to the phrases of source texts (it means the source texts are similar to the translated one). In those cases, most translational text materials can be practiced by one scanning, and the translators have only to translate the revised sentences. Translation memory is able to generate parallel concordances from the word or collocation in the source texts.

Finally, translation profession demands the translators the skills and practice in corpus linguistics. This is indispensable to the translator's competences. They can build up a monolingual or bilingual corpora and use some corpus analysis concordance to find out the information of the words, collocations, terms in both source and target languages and mostly the parallel corpora to solve the questions in translational works. The translators can create up a small corpus and process the documents by using search engine implemented by ad hoc programs and online services such as Antconc and others tools, etc...

We look together one screenshot of English-French concordance that we pick it up from Tradoo.it (tradoo.it.com), web of comparable corpus. This is the case of translation of the word “thanks to” from the English to the French. We see that it is the corpus of 200 million of words for each language. So it is very strong for translation aids and terminology searching.

| Anglais                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Français                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>EDMONTON, ALBERTA, May 23, 2009—Two Edmonton-based organizations will develop programs to help Aboriginal people prepare for employment, <b>thanks to</b> support from the Government of Canada.<br/> <a href="http://news.gc.ca/web/article-eng.do?nid=453299">http://news.gc.ca/web/article-eng.do?nid=453299</a><br/>           Source : Centre des nouvelles du Canada [ Voir le bilette ]</p>                                                                                                                               | <p>EDMONTON (ALBERTA), le 23 mai 2009 — Deux organismes d'Edmonton élaboreront des programmes pour aider les Autochtones à se préparer à l'emploi <b>grâce à</b> l'appui du gouvernement du Canada.<br/> <a href="http://nouvelles.gc.ca/web/article-fra.do?nid=453299">http://nouvelles.gc.ca/web/article-fra.do?nid=453299</a><br/>           Source : Centre des nouvelles du Canada [ Voir le bilette ]</p>                                                                                                                    |
| <p>SHELBURNE, Nova Scotia, June 11, 2008 - The heritage and history of the Black Loyalists will be celebrated in Shelburne this summer, <b>thanks to</b> an investment by the Government of Canada.<br/> <a href="http://news.gc.ca/web/article-eng.do?nid=404609">http://news.gc.ca/web/article-eng.do?nid=404609</a><br/>           Source : Centre des nouvelles du Canada [ Voir le bilette ]</p>                                                                                                                               | <p>SHELBURNE (Nouvelle-Écosse), le 11 juin 2008 - Le patrimoine et l'histoire des Loyalistes noirs seront célébrés à Shelburne cet été, <b>grâce à</b> un investissement du gouvernement du Canada.<br/> <a href="http://nouvelles.gc.ca/web/article-fra.do?nid=404609">http://nouvelles.gc.ca/web/article-fra.do?nid=404609</a><br/>           Source : Centre des nouvelles du Canada [ Voir le bilette ]</p>                                                                                                                    |
| <p>Do not these figures speak clearly to the fact that, year in and year out, the bulk of Canada's debt is being paid by the unemployed, <b>thanks to</b> the cynicism of the Minister of Finance?<br/> <a href="http://www.parl.gc.ca/HousePublications/.../Publication.aspx?Language=E&amp;Mode=1&amp;Parl=37&amp;Ses=1&amp;DocId=1385306">http://www.parl.gc.ca/HousePublications/.../Publication.aspx?Language=E&amp;Mode=1&amp;Parl=37&amp;Ses=1&amp;DocId=1385306</a><br/>           Source : HANSARD [ Voir le bilette ]</p> | <p>Est-ce que ces chiffres ne sont pas la preuve éloquent que la dette du Canada est, bon an mal an, payée en grande partie par les chômeurs, <b>à cause</b> du cynisme du ministre des Finances?<br/> <a href="http://www.parl.gc.ca/HousePublications/.../Publication.aspx?Language=F&amp;Mode=1&amp;Parl=37&amp;Ses=1&amp;DocId=1385306">http://www.parl.gc.ca/HousePublications/.../Publication.aspx?Language=F&amp;Mode=1&amp;Parl=37&amp;Ses=1&amp;DocId=1385306</a><br/>           Source : HANSARD [ Voir le bilette ]</p> |

*Illustration 2. Example of English-French comparable corpus (tradoo.it.com)*

Otherwise, we look another example for the translation of word “grâce à” in French to English “thanks to” in the comparable corpus:

| Français                                                                                                                                                                                                                                                                                                                                                                                           | Anglais                                                                                                                                                                                                                                                                                                                                                                                      |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Nous allons maintenant encore plus loin, comme nous l'avons appris hier <b>grâce à</b> un communiqué conjoint des leaders de l'opposition et du caucus libéral, les sénateurs Cowan et Carignan.<br>Source : Sénat du Canada [ <a href="#">Voir le bilette</a> ]                                                                                                                                   | Now, this Parliament is going even further, as we saw yesterday, <b>with</b> a joint news release from the Leaders of the Opposition and the Liberal Caucus, Senator Cowan and Senator Carignan.<br>Source : Sénat du Canada [ <a href="#">Voir le bilette</a> ]                                                                                                                             |
| Nous allons maintenant encore plus loin, comme nous l'avons appris hier <b>grâce à</b> un communiqué conjoint des leaders de l'opposition et du caucus libéral, les sénateurs Cowan et Carignan.<br>Source : Sénat du Canada [ <a href="#">Voir le bilette</a> ]                                                                                                                                   | Now, this Parliament is going even further, as we saw yesterday, <b>with</b> a joint news release from the Leaders of the Opposition and the Liberal Caucus, Senator Cowan and Senator Carignan.<br>Source : Sénat du Canada [ <a href="#">Voir le bilette</a> ]                                                                                                                             |
| C'est <b>grâce à</b> Dieu que cette opération s'est déroulée pendant le dernier mois de la mission; nous avions une expérience opérationnelle de cinq mois, qui nous a permis de nous en sortir.<br><a href="http://www.noscommunes.ca/DocumentViewer/_/tmoignages">http://www.noscommunes.ca/DocumentViewer/_/tmoignages</a><br>Source : Chambre des communes [ <a href="#">Voir le bilette</a> ] | It is only through the <b>grace of</b> God that this operation happened in the last month of the tour, when we had five months of seasoning and getting all of our operational experience ready.<br><a href="http://www.ourcommons.ca/DocumentViewer/_/evidence">http://www.ourcommons.ca/DocumentViewer/_/evidence</a><br>Source : Chambre des communes [ <a href="#">Voir le bilette</a> ] |

*Illustration 3. Example of French-English in comparable corpus (tradoo.it.com)*

In observing the two examples, we find that there are some differences in using of word “thanks to” in translation. In some case, “thanks to” is translated to “à cause de” (because of) instead of “grâce à” as we intend. Otherwise, in the corpus, we find that “grâce à” are translated to “with” and “grace of” instead of “thanks to”. So we see that there are some gaps in translation of both two languages. And the mission of translators is to find the suitable words or lexical structures.

#### 4.2. Discussion

Corpus linguistics is not same as obtaining data through the use of computer. This is the study and analysis of data obtained from a corpus. The main task of linguistics is not to find the data but to analyze it in the original context. Computer are so useful tools used in this process. So how to help students using well this tools is not easy for teachers in ESL because it demands not only knowledge in linguistics but also in linguistic computer. The second question is if electric corpora is research tools or research method. So It depend to academic aims of applied linguists.

Otherwise, teaching-oriented corpora are so useful in teaching LSP. In fact, linguists often choose small corpora based on specific needs such as Hongkong financial services corpus, Hongkong engineering corpus, ... So LSP corpora is so necessary for student and teachers in the context of Hanoi University of Industry where many faculties such as Electric, Automobile Technology, Chemistry,... need specialized English teaching as ESL. Perhaps, the applying is not easy and demand the time and cooperation between them.

In the system of Foreign Language Schools, students of last year in interpreting and translating branch base mostly in experience to learn the verbal terms, verbal collocations, ... For them, Google is useful tool. But we notice that it is search engine not corpus sample. If they want to discover the word in real context, they need use corpus and concordance. In fact, they can create up themselves one small specialized corpus of 300.000-500.000 words for their learning objectives. Otherwise, students can use some open big corpus such as BNC, Lancaster Corpus, Le Monde Corpus, ...

Corpora have a significant influence in machine assisted translation as well as in translation research and practice. They provide the translators the technology, tools and data which can be used to aid in translation process. So the creation of reliable corpora for translation studies is indispensable. Texts for corpus, either in format of electronic or printed one have to be standard. Then they can be enriched by lemmatization and annotated grammatically. In another way, parallel corpora must be aligned.

In fact, this is the first stage of corpus-based translation although it consists in theoretical studies, testing in corpus building and analyzing of linguistic features. Currently, the scope of research has

diversified. The translators want to make some investigations in translated texts, both in corpus - assisted translation and how the translation contributes to language development. However, the increasing of machine translated memory will make it possible to compare to human translation as Tymoczko (1998) said that the development of corpora and corpus-based methods “represent a long – term investment for the field of Translation Studies” and the vision of corpus–based translation as well as “intellectual endeavors unimpeded by time or space facilitated by intercommunication across the globe”.

### 5. Conclusion

Corpora is useful and suitable for vocabulary study. In fact, its contribution to study of vocabulary is remarkable. They allow students to analyze the meaning, context in which the words typically occur. This help them understand more and more the realistic picture of how the language and its vocabulary work. So the using of corpus in translation studies is useful to be researched seriously, totally and applied step by step in the universities of foreign language as Baker (1993) said “There is now an urgent need to explore the potential for using large computerized corpora in translation studies”.

---

*\*This research is funded by Hanoi University of Industry under Project, ...-2025-RD/HD-DHCN/Nghiên cứu này được tài trợ bởi Trường Đại học Công nghiệp Hà Nội, nhiệm vụ KH&CN mã số ...-2025\_RD/HD-DHCN.*

## REFERENCES

### English

1. Adam K (2014), *Corpora in English language teaching*, British Council.
2. Baker, M (1993), *Corpora in Translation Studies: An Overview and Suggestions for Future Research*, Scientific Research Publishing.
3. John Benjamins (2009), Beeby and colleagues. *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*,
4. Bowker L., Pearson J (2002), *Working with Specialized Language*, Routledge.
5. Claudio F., Federico Z (2015), *New directions in corpus-based translation studies*, Tạp chí Khoa học ngôn ngữ.
6. Farahani M (2023), *New Perspectives on Corpus Translations Studies*, *New voices in translations studies* vol.28.
7. Gena B (2010), *Using corpora in the language learning Classroom: Corpus linguistics for Teachers*, Michigan.
8. Hatim B (1997), *Communication across cultures: Translation theory and contrastive text linguistics*, University of Exeter Press.
9. Malmkjaer K (1993), *Who can make nice a better word than pretty? Collocation, translation, and psycholinguistics*, pp.213-232, Jhon Benjamins publishing company.
10. Laviosa S (2008), *Description in the translation classroom: universals as a case in point*, pp. 119-132, Jhon Benjamins.
11. Laura Gavioli (1997), Federico Zanettin. *Comparable corpora and translation: a pedagogic perspective*, first international conference on *Corpus Use and Learning to Translate*, Bertinoro.
12. Natalie Kübler (2014), *Mettre en œuvre la linguistique de corpus à l'université, Vers une compétence utile pour l'enseignement/apprentissage des langues?*, *Les cahiers de l'Acedle*.
13. Natalie Kübler (2011), *Working with different corpora in translation teaching*, *New trends in Corpora and Language Learning*, London.

14. Natalie Kübler (2003), Corpora and LSP translation, In Zanettin, F., S. Bernardini & D. Stewart (eds), Corpora in Translator Education, Manchester: St Jerome.
15. Schimied J (1996), Explicitness as a universal feature of translation, Hildergard.
16. Sinclair J (1991), Corpus, Concordance, Collocation, Oxford University Press.
17. Scott M (1997), PC analysis of keywords - And key keywords. System, pp.233-245.
18. Stubbs. M (1996), Text and Corpus Analysis, Oxford Blackwell Publisher Ltd.
19. Tymoczko M (1998), Computerized Corpora and the Future of Translation Studies, Meta, pp.652-660.

#### Vietnamese

20. Đinh Điền (2019), *Ngôn ngữ học ngữ liệu*, Nxb Đại học quốc gia TP Hồ Chí Minh.
21. Đào Hồng Thu (2009), *Ngôn ngữ học khối liệu và những vấn đề liên quan* (quyển I), Nxb Khoa học xã hội.
22. Đào Hồng Thu (2010), “Ngôn ngữ học khối liệu và phương pháp tiếp cận hiện nay”, Tạp chí *Từ điển học và Bách khoa thư*, số 3. 2010.
23. Hoàng Văn Vân (2005), *Nghiên cứu dịch thuật*, Nxb Khoa học xã hội.

#### **Xây dựng khối ngữ liệu và sử dụng phương pháp nghiên cứu dựa trên khối ngữ liệu vào nghiên cứu dịch thuật**

**Tóm tắt:** Mục đích của bài báo đề cập đến các vấn đề liên quan đến ngôn ngữ học khối liệu (corpus) như xây dựng khối ngữ liệu chuyên ngành; áp dụng phương pháp nghiên cứu dựa trên khối ngữ liệu và cách sử dụng chúng trong dịch thuật văn bản. Tác giả mô tả các khái niệm về corpus, quá trình phát triển của nó đối với ngôn ngữ học và giới thiệu các loại hình corpus như khối liệu song song, khối liệu dịch, v.v. Sau đó, giới thiệu một số công cụ để khai thác khối liệu ví dụ như Antconc, Concordance,... Đây là các công cụ rất hữu ích để khai thác được các corpus điện tử. Nội dung lý thuyết ứng dụng khối liệu trong dịch thuật được phân tích ở phần cuối của bài viết.

**Từ khóa:** Ngôn ngữ học khối liệu; ứng dụng trong dịch thuật; khối liệu song song; khối liệu so sánh.