FOREIGN LANGUAGES AND VIETNAMESE

PARALLEL CORPORA IN TRANSLATION AND USING SKETCHENGINE-ONLINE PLATFORM

LE CHI HIEU*

ABSTRACT: The role of a parallel corpus in translation studies has shifted from a research interest to an essential tool for professional translators. While comparable corpora are the standard for training machine translation, parallel corpora offer something authentic translation and native linguistic patterns. In the SLT-HaUI, we introduce step by step the parallel corpora in teaching and practicing for students of linguistics. It helps them to have knowledge of it and have electric tools to use it. The aim of this paper is to review some relevant factors related to the corpus linguistics and parallel corpus 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 comparable corpus, parallel etc... Then, we attempt to present tools for exploiting the comparable corpus such as Sketch Engine, one of the most powerful online platform for research linguistics.

KEY WORDS: Corpus; parallel corpus; corpus in translation; SketchEngine.
RECEIVED: April, 04th, 2026. EDITED: June, 27th, 2026

1. Introduction

The use of real textual examples in the study of translation is not new issue in applied linguistics. However, using parallel 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. In Vietnam, building many corpora for Vietnamese language is real demands and necessary. The fact that introducing parallel corpora in translation practice in SLT school-HaUI is necessary for learning and researching approach. In many courses, we try to introduce the online platform SketchEngine into translation teaching. It is very efficient for both students and professors.

2. What is reference corpus and parallel corpus

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 translation. Then, Baker (1999) starts to analyze similarities and differences between sets of translation into the same language by different translators. Baker's research is very suggestive in translation and becomes the inspiration for various projects of corpus's influence in translation. Later, 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) utilise the parallel corpora BNC and Gigaworld 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. In school of Language and Tourism (SLT-HaUI), the

**Dr; School of Languages and Tourism-HaUI; Email: hieulc@haui.edu.vn*

Corpus linguistics and Computational linguistics are official courses for students of 3th academic year. So we hope that this is the new stage of applying language in translation profession.

We look together two interfaces of concordance which we take from online platform SketchEngine. We try to use the word in English (power) and Vietnamese (sức mạnh) to analyse it in real context.

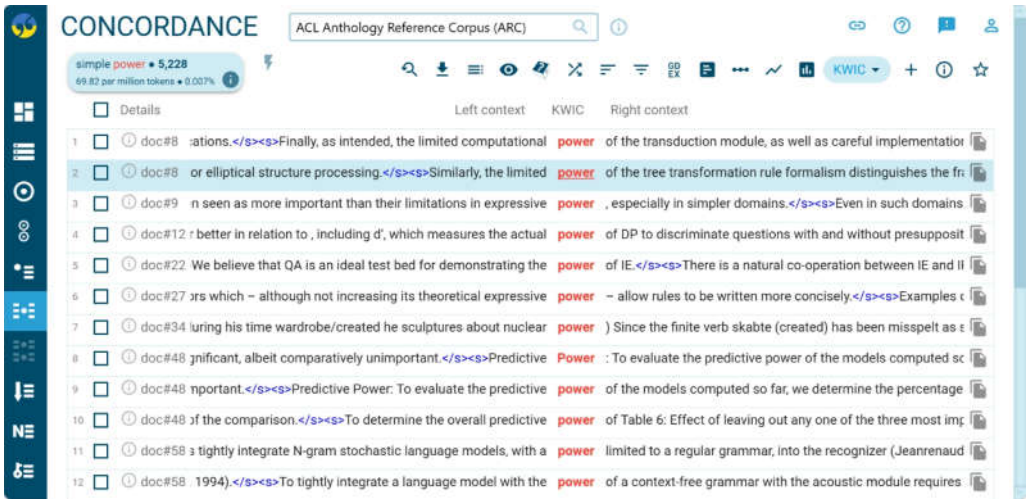

Illustration 1. *Example for concordance of word "power" in SketchEngine*

Illustration 2. *Example for concordance of word "sức mạnh" in SketchEngine*

In fact, in translation, a parallel corpus consists of sets of texts in the source and target languages that are assembled based on identical design criteria such as topic, genre, register, and time period but are written by native speakers independently of one another. For example, if you are translating a medical study on cardiology from English to Vietnamese, a parallel corpus would consist of original cardiology papers written by Vietnamese doctors for Vietnamese journals, rather than translations of English papers. In linguistics and data science, a parallel corpus is a collection of texts in two or more languages (or varieties of the same language) that are matched based on similar content, genre, and purpose which are translations of each other called "twin" texts. They cover the same ground but were written independently by native speakers in their respective languages.

Sometimes we ask ourselves: “Why we have to use a the parallel corpus in translation?”. In fact, they show some advantages such as: Naturalness (Because they are translations, they show the most natural way to express an idea); Terminology discovery (If we are translating a new technical field (like AI or FinTech), a parallel corpus helps you find the "*organic*" terms native experts are currently using); Cross-cultural studies (They allow researchers to see how different cultures discuss the same topic differently).

Furthermore, one of the tool helping for using corpus and translation is online SketchEngine platform. This is a leading web-based corpus management and text analysis tool. It is used by lexicographers (or dictionary makers), translators, linguists, and researchers to study how language is actually used in massive collections of text. Essentially, while a search engine like Google helps you find information, Sketch Engine helps you find linguistic patterns. It stands out because of its specialized algorithms that process billions of words in seconds. Its core tools include:

- Word Sketch: The signature feature. It provides a one-page automatic summary of a word’s grammatical and collocational behavior. For example, searching "*coffee*" will instantly show you which adjectives (strong, hot, black), verbs (brew, drink, spill), and objects it most frequently interacts with.

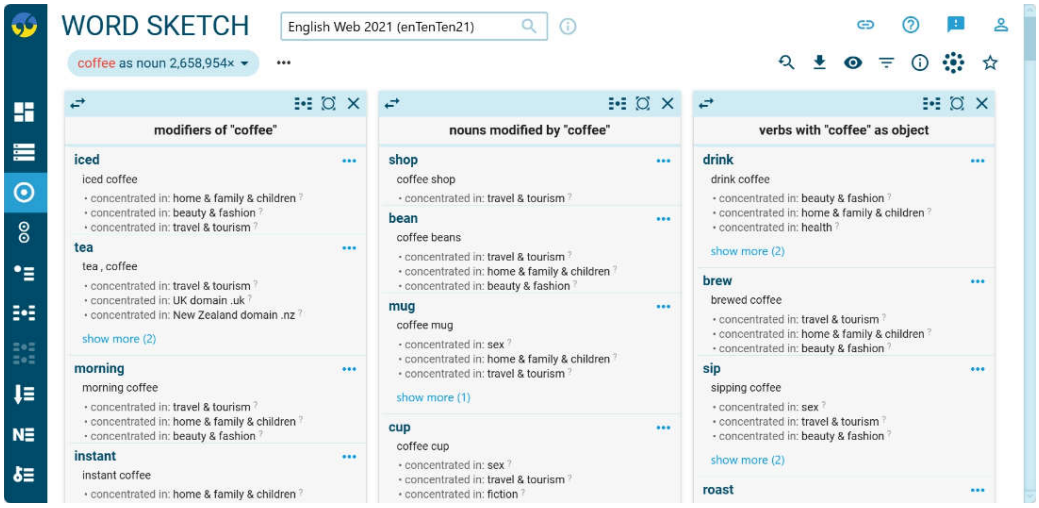

***Illustration 3**. Interface of Sketch Word of “coffee”*

- Concordance: This is a "Key Word In Context" (KWIC) search. It shows every instance of a word or phrase in a corpus, centered on the screen, so you can see the surrounding text.

- Thesaurus: Automatically identifies synonyms or semantically related words based on how they are used in similar contexts.

- N-grams: Generates lists of the most frequent two, three, or four-word combinations (for example: "*on behalf of, in term of,…*").

3–4-grams word

	Word	Count		Word	Count
1	one of the	303,265	11	to be a	100,285
2	as well as	291,495	12	a number of	98,337
3	part of the	151,966	13	is one of	90,869
4	I do n't	148,894	14	of the most	90,028
5	a lot of	145,674	15	out of the	87,383
6	the United States	144,719	16	the use of	82,268
7	be able to	132,872	17	you want to	81,888
8	in order to	128,631	18	there is a	80,923

***Ilustration 4**. N-gram from SketchEngine*

- Keywords & Terms: Compares two corpora to find which words are uniquely frequent in one (For example: finding medical terms in a doctor's journal compared to general English).

So who uses it and why? In fact, it serves many clients such as translators to find the most natural collocations in a target language. Lexicographers prefer to use it to write dictionary definitions based on real-world evidence rather than intuition such as Oxford, Cambridge, and Collins dictionary. Teachers and students use it to create vocabulary lists or check if a specific phrase is actually used by native speakers; NLP researchers to clean, tag (POS tagging), and prepare massive datasets for machine learning.

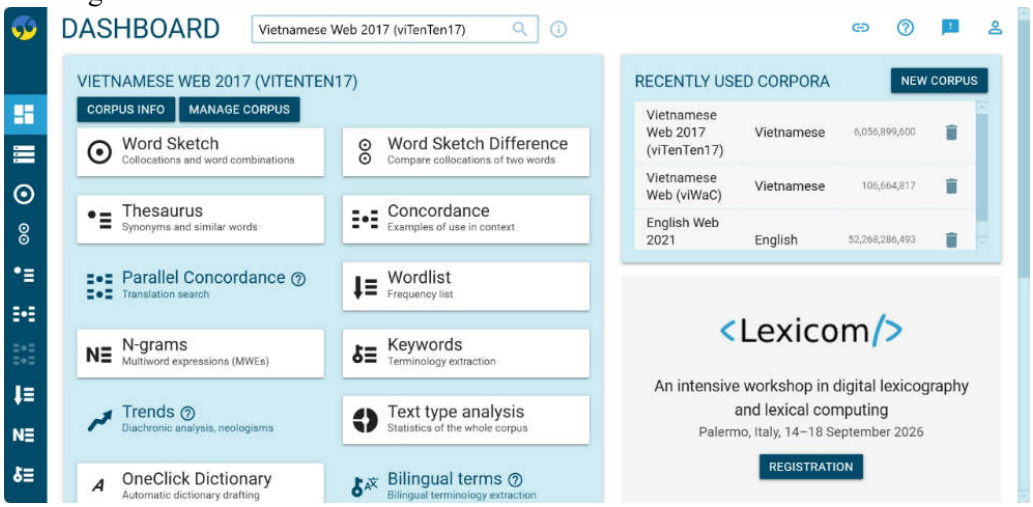

***Illustration 5**. Interface of online platform SketchEngine*

Actually, Sketch Engine supports over 100 languages, including robust support for Vietnamese (Vietnamese web and Vietnamese web 2017 in total of more than 7.000.000 words). It handles specific challenges like Vietnamese word segmentation effectively, making it a favorite for linguists working with Southeast Asian languages. It is a subscription-based service, though many universities provide institutional access. They also offer a free trial and a Sketch Engine for Language Learning

version which is a simplified, free interface for students. In which, a parallel corpus isn't just a statistic folder; it is a dynamic toolset used to find equivalent ways of speaking across different languages or varieties without relying on direct translations.

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 (1997) praised corpus-based translation "…*is truly new wave of research*…". Furthermore, Baker (1993, 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 (1993) 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, 1997).

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. In SLT, the students of Linguistic can have access to the online platform and to serve it for learning of foreign language and translation. This help them to have knowledge in both technology and linguistics. After some courses in the class, we receive a lot of optimistic attitudes about using and practicing these tools for learning and researching.

4. Why does the translators use parallel corpora?

Translators often face the trap-the tendency for translated text to sound slightly stiff or unnaturally influenced by the source language's syntax. In fact, the parallel corpora solve this through:

- Natural Collocations: They reveal how native speakers naturally group words. For exemple: A dictionary might give you five synonyms for "*large*," but a comparable corpus shows that in a real context, native speakers specifically use "huge".

- Genre Conventions: Different cultures have different moves in writing. An English business "*Letter of Intent*" may be more direct than its Vietnamese equivalent. A comparable corpus allows the translator to intend the cultural expectations of the target audience.

- Terminology Discovery: For emerging fields (such as AI), by looking at native texts in the target language, translators can find the most current terms being used by experts in that country.

By this table, we can compare the parallel and comparable corpus:

	Parallel Corpus	**Comparable Corpus**
Source	Original + Translation	Original A + Original B
Goal	Finding equivalences	Finding native authenticity
Utility	Great for "what was said"	Great for "how they would say it"

***Illustration 6**. Comparison of Corpus types*

5. Building a Learner's corpus and Parallel 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 to build up a corpus?" is always discussed. For serving to buil 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 200.000-300.000 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. Then, we put our corpora in QR code for coding it, one for English, one for Vietnamese corpus. To read the corpora, we can read in the link: https://qrco.de/bglZiq and scan the QR codes below:

The tool we use to explore the comparable corpora is Sketch Engine. This is most powerful comparable feature. Instead of searching for pre-existing translations, we can create our own corpora from the live web.

- Way of working: You provide technical terms in both English and Vietnamese.

- The Result: Sketch Engine crawls the web and builds two separate, independent corpora. One contains original English texts on the topic, and the other contains original Vietnamese texts..

This is the primary tool for analyzing comparable data. Unlike a parallel concordance (which shows side-by-side translations), a Bilingual Word Sketch compares the behavior of words. Firstly, we select your English corpus and your Vietnamese corpus. You search for the word "*bank*" in English and "*ngân hàng*" in Vietnamese. Then, we have the results: it shows a side-by-side comparison of collocations. For example:

English: "*bank*" + "*account*" …

1	various	exas was a republic " .</s><s>It permits the state to take over	bank	accounts , stocks and other personal property of persons mis
2	various	The bill , which Daniel said he drafted personally , would force	banks	, insurance firms , pipeline companies and other corporations
3	various	, sounded the opposition keynote when he said it would force	banks	to violate their contractual obligations with depositors and un
4	various	obligations with depositors and undermine the confidence of	bank	customers .</s><s>`` If you destroy confidence in banks , you
5	various	ice of bank customers .</s><s>`` If you destroy confidence in	banks	, you do something to the economy " , he said .</s><s>`` You t
6	various	mount `` to making a gift out of the taxpayers<g/>' pockets to	banks	, insurance and pipeline companies " .</s><s>His contention
7	various	l presentation was made to Mrs. Geraldine Thompson of Red	Bank	, who is stepping down after 35 years on the committee .</s>
8	various	etter public image for welfare " .</s><s>The wife of convicted	bank	robber Lawrence G. Huntley was arrested in Phoenix , Ariz. , la
9	various	nd last April for robbery of the Hillsdale branch of Multnomah	Bank	, also was charged with the store holdup .</s><s>Secret Gran
10	various	00 m.p.h. -- fast even for a modern jet plane touchdown -- and	banked	into the airfield .</s><s>Moments later the V-<g/>1 skimmed
11	various	50 a share) .</s><s>There was even a cable in French from a	bank	in Switzerland that had somehow learned about the Dallas stc
12	various	" , it said translated .</s><s>But E.G.T. could not let the Swiss	bank	have even 10 shares .</s><s>After it allotted shares to 41 unc

Vietnamese: "*tài khoản*" + "*ngân hàng*"

Left Context	Hit	Right Context
: hỗ trợ bởi Đối tác sự kiện chính là	Ngân hàng	United Overseas (UOB). Hình ảnh sự
hai trương tiêu biểu của năm, được	Ngân hàng	United Overseas (UOB) tài trợ: Khácl
: hỗ trợ bởi Đối tác sự kiện chính là	Ngân hàng	United Overseas (UOB). Hình ảnh sự
hai trương tiêu biểu của năm, được	Ngân hàng	United Overseas (UOB) tài trợ: Khácl
: hỗ trợ bởi Đối tác sự kiện chính là	Ngân hàng	United Overseas (UOB). Hình ảnh sự
hai trương tiêu biểu của năm, được	Ngân hàng	United Overseas (UOB) tài trợ: Khácl
ıi nên ưu tiên đặt cọc qua tài khoản	ngân hàng	chính chủ của khách sạn hoặc doanl
anh toán chỉ từ 1,5 tỷ để nhận nhà,	ngân hàng	hỗ trợ vay với lãi suất 0% trong 12
ng 5 sao Elisa ở TP HCM tiếp tục bị	ngân hàng	đem bán đấu giá (NLĐO) – Đây là lầ

We see that even if the texts aren't translations, the statistical patterns reveal that these words serve the same linguistic function in their respective cultures. Using a parallel corpus in Sketch Engine is one of the most effective ways to find natural, native-sounding terminology and phrasing. While a parallel corpus shows you how things have been translated, a comparable corpus shows you how people actually write in their own language.

Here is how you can use Sketch Engine to build and analyse a parallel corpus. Firstly, we build the own parallel corpus. If Sketch Engine doesn't already have the specific corpus we need, you can build a sibling pair using the WebBootCaT tool in NLP. This is the steps to do:

- Select "WebBootCaT" (Create corpus from the web).
- Choose the Source Language (e.g., English) and enter 5-10 "seed words" (technical terms related to your topic). Sketch Engine will crawl the web and build an English corpus.
- Repeat the process for your Target Language (Vietnamese) using the translated versions of those seed words.
- The Result: You now have two independent, topic-specific corpora that are comparable because they share the same subject matter.

Secondly, we use the Bilingual Word Sketch. This is the tool for comparable analysis. It allows you to see how a word behaves in Language A versus Language B. Then we select your English corpus and Vietnamese corpus. Search for a term like "*investment*" in English and "*đầu tư*" in Vietnamese. In next step, we compare in Sketch Engine. It will show you two columns of collocations (word partners). For example:

English: We can see "*foreign investment*," "*capital investment*," "*return on investment*.".

Vietnamese: We can see "*đầu tư nước ngoài*," "*vốn đầu tư*," "*lợi nhuận đầu tư*.".

At last, we will see that we aren't just looking at a dictionary but we are seeing the statistical frequency of how these terms are actually paired by native speakers. In extracting Keywords and Terms, if we have a parallel corpus, we can use the Keywords and Term Extraction tools. So have you ever made a question: “Why it's useful?”. In fact, this helps you build a bilingual glossary for a translation

6. Discussion and future prospects

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, …

Comparable and parallel 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 this article, 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 Tymoczo (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*". Beside, using online platform like SketEngine is most useful for translators and linguistic researchers. Using it is the way to shorten the time of reading, researching and translating.

7. Conclusion

Corpora is useful and suitable for vocabulary study. In fact, its contribution to study of vocabulary is remarkable. They allow students to analyse 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*". Then, to go further, parallel corpora are indispensable. With the development of technology, translators can build up themselves the corpora or use the online platform such as SketchEngine. They overpass the limit of classical translation and have access to the new methods of this. It is useful for both research and practical of translation.

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. Beeby and colleagues (2009), *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*, John Benjamins.
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. Hatim B (1997), *Communication across cultures: Translation theory and contrastive text linguistics*, University of Exeter Press.
8. Malmkjaer K (2008), *Who can make nice a better word than pretty? Collocation, translation, and psycholinguistics*, pp213-232, Jhon Benjamins publishing company. 1993.
9. Laviosa S. *Description in the translation classroom: universals as a case in point, pp. 119-132*, Jhon Benjamins.
10. Laura Gavioli, Federico Zanettin (1997), *Comparable corpora and translation: a pedagogic perspective*, first international conference on Corpus Use and Learning to Translate, Bertinoro.
11. 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*.
12. Natalie Kübler (2003), *Corpora and LSP translation*, In Zanettin, F., S. Bernardini & D. Stewart (eds), Corpora in Translator Education, Manchester: St Jerome.
13. Sinclair J (1992), *Corpus, Concordance, Collocation*, Oxford University Press.
14. Scott M (1997), *PC analysis of keywords - And key keywords*. System, p. 233-245.
15. Stubbs. M (1996), *Text and Corpus Analysis*, Oxford Blackwell Publisher Ltd.

Vietnamese

16. Đinh Điền (2019), *Ngôn ngữ học ngữ liệu*, Nxb Đại học quốc gia TP Hồ Chí Minh.
17. Hoàng Văn Vân (2005), *Nghiên cứu dịch thuật*, Nxb Khoa học xã hội.
18. Nguyễn Ngọc Vũ (2024), *Ngôn ngữ học tính toán: từ lý thuyết đến thực tiễn*, Nxb Đại học Sư phạm tp. Hồ Chí Minh.

Khối ngữ liệu song song trong dịch thuật và sử dụng nền tảng trực tuyến SketchEngine

Tóm tắt: Khối ngữ liệu song song là bước phát triển tiếp theo của Khối ngữ liệu người học (Learner's corpus). Nó là công cụ hữu hiệu cho các nhà ngôn ngữ học cần những thống kê ngôn ngữ hay cho các dịch giả trong dịch thuật. Tại Trường Ngoại ngữ-Du lịch, Đại học Công nghiệp Hà Nội, chúng tôi dần đưa các công cụ khối ngữ liệu song song vào giảng dạy cho sinh viên, giúp sinh viên có thể sử dụng các công cụ điện tử trong học tập và nghiên cứu. Sử dụng các công cụ online cũng đang là xu hướng của các dịch giả chuyên nghiệp. SketchEngine là một trong số các công cụ như vậy. Đây là một trong những công cụ online lớn nhất hiện nay, tích hợp hàng trăm khối ngữ liệu lớn (trong đó có tiếng Việt). Trong khuôn khổ bài báo, chúng tôi giới thiệu SketchEngine, một nền tảng trực tuyến, và cách sử dụng chúng trong nghiên cứu dịch thuật. Sau đó, chúng tôi xây dựng khối ngữ liệu song song và cách khai thác trong nền tảng đó.

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