Hue University Journal of Science: Social Sciences and Humanities
ISSN 2588-1213
Vol. 134, No. 6B, 2025, p.p. 99–121, **DOI**: 10.26459/hueunijssh.v134i6B.7629

BUILDING THE ENGLISH-VIETNAMESE BILINGUAL CORPUS OF TOURISM AS AN ESP MATERIAL FOR STUDENTS FROM ENGLISH FACULTY, UNIVERSITY OF FOREIGN LANGUAGES AND INTERNATIONAL RELATIONS, HUE UNIVERSITY

Phan Thi Thanh Thảo

University of of Education, Hue University, 34 Le Loi St., Hue city, Vietnam

* Correspondence to **Phan Thi Thanh Thảo** < ptthao@hueuni.edu.vn>

(*Received: September 17, 2024; Accepted: January 07, 2025*)

Abstract: The current ESP resources on tourism are of necessity and diversity such as English textbooks, magazines, periodicals, newspapers, books, etc., which can be used in both electronic and printed formats on the topic of Tourism. However, there is a lack of bilingual English-Vietnamese corpus on Tourism serving the increasing demand for students due to its convenience, ease of use, and vocabulary lookup as a reference in learning ESP of tourism and translation study as well. This article presents the building of the Engish-Vietnamese bilingual corpus on tourism used by students at ESP Tourism classes at the University of Foreign Languages and International Relations, Hue University (HUFLIS). This study was conducted in two stages: 1/Building the corpus during 6 months (January-June 2023) and 2/ Practice of corpus with the participation of 350 students during the first semester of 2023-2024 school-year. Qualitative and quantitative approaches are applied in this study with research instruments including questionnaire, and face-to-face interviews. The research reveals the current realities of students' corpus use and their attitudes towards the benefits and challenges of using this corpus at ESP Tourism classes.

Keywords. Bilingual corpus, ESP material, ESP for tourism, students of English Faculty

1. Introduction

Building ESP resources has become an issue of great interest in higher education systems in Vietnam for many recent years, in particular with universities of foreign languages where learning and teaching ESP of various domains such as economics, tourism, medicine, construction, and so on has been taken into great consideration. Since the tourism industry is currently considered one of the three key economic sectors, receiving investment attention, constantly developing and making positive contributions to Vietnam's national economy, the training students in English for Tourism major plays a crucial role in providing human resources for the Central Highlands region in particular and the whole country in general. However, ESP has some characteristics depending on the specific needs of learners of different majors, for example, it will focus on the appropriate language for each activity in terms of grammar, vocabulary, skills, discourse, and genres.

Currently, the materials of Tourism learning at universities as well as on the market are very rich and diverse, such as English textbooks on Tourism, magazines, journals, newspapers, books, etc. The documents can be used in both electronic and printed forms, which is very convenient for learners. However, among these learning resources, there is still no bilingual English-Vietnamese corpus on Tourism that can be used flexibly in both electronic and paper forms. Meanwhile, the need of using documents like this corpus is increasing day by day for English students due to its convenience, ease of use in looking up vocabulary in translated documents in the tourism domain. In response to that need, building a bilingual English-Vietnamese corpus on Tourism is extremely vital for English students' use as a reference in their learning of ESP for Tourism.

A glance at previous studies on corpus-based language use in the second language (L2) classroom reveals both advantages and disadvantages in language learning. Some studies have been conducted to examine the effectiveness of using authentic and real-life examples. Corpora can be integrated in language classrooms to support students of various levels of English, especially ESP of different domains. In fact, a corpus could offer a learning technique to highlight how certain language forms, vocabulary items, and expressions of English are used naturally in real life. Nevertheless, the corpus use also requires the IT knowledge and skills which had better be trained for a short of time. Furthermore, the use of corpus examples which are still inconclusive and hard to understand in some cases may force the learners to select carefully and consider different types of examples in both the first language and the second language.

What is more, there has been little research on building a bilingual English-Vietnamese corpus on Tourism, which can be used as a reference for students to learn ESP of tourism at Vietnam's universities. Hence, this study focuses on the main steps of building a bilingual English-Vietnamese corpus on Tourism, shows the way how students of English Faculty use this corpus as an ESP refrerence in their learning English of tourism, as well as their attitudes towards the advantages and disadvantages of using this corpus. This study aims to answer the following research questions:

1. What are the needs of building the English-Vietnamese bilingual corpus?
2. What are the main steps of building the English-Vietnamese bilingual corpus?
3. What are the benefits and drawbacks of using this corpus?

2. Corpora

2.1. Definition

There exists a variety of ways to define a corpus in different domains. In the linguistics domain, corpus definitions are based on their origin, size and function. In fact, the word "corpus" deriving from Latin originally means the 'body', which contains a large body of texts which can be stored and processed in an electronic form (Oakes, 2012). According to many linguists, a corpus is defined as "a representative collection of pieces of a language that are selected according to explicit linguistic criteria and reflect natural chunks of this language to be used for a linguistic analysis" (McEnery et al., 2006; Sinclair, 2004). In other words, this definition implies the close relation between corpora and their linguistic features.

In sight of many linguistics and language studies, a corpus also refers to a large and structured set of texts (spoken or written) used for linguistic analysis and study. Actually, corpora can be manipulated as essential tools for being aware of as well as conducting an investigation into various aspects of language, including its structure, usage, and patterns. Moreover, considering the size of a corpus, Patsala and Michali (2020) recognizes that it forms a representative sample of language, while its machine-readable format allows annotation, and various types of analysis related to the criteria set and the tools used, for instance, part-of-speech, frequencies, key-word-in-context, etc. (Rayon, 2015). It is evident that the corpus has been compiled with a collection of texts ranged from small to big size, and from general to specialized fields.

According to Sinclair (2004), a corpus typically contains a vast number of words (e.g. millions of words) since it is acknowledged that the natural language's creativity has brought such immense variety of expressions appearing in the recurrent patterns which become the clues to the language's lexical structure. Take an example of the British National Corpus (BNC) including more than one hundred million of words, a corpus is collected surely with a huge amount of words from both spoken and written texts belonging to various domains. Furthermore, Kuble and Zinsmeister (2015) confirm that many linguistically analyzed and publicly available corpora have also been increased significantly which provide a rising amount of data for analyzing language and other applications on computational linguistic purposes.

Corpora provide many benefits in various applications in diferent domains. First, in the field of natural language processing, in machine learning technology, corpus is used to help computers "learn" how to process language automatically by labeling and recognizing languages. Second, corpus is widely used in machine translation, for example, based on bilingual corpus (manually translated), the computer will automatically give translation rules to be able to automatically translate other texts with similar styles, fields, and genres (Zanettin, 2014). Third, the corpus has many other applications related to Computational Linguistics such as speech recognition, speech synthesis, etc. In addition, in education and training, the corpus also serves the compilation of textbooks, dictionaries, exam questions, teaching support, especially subjects on translation and language practice (reading and writing skills).

2.2 Classification of corpora

According to Baker (1995, p.234), electronic corpora are divided into three types for translation research and language teaching, including multilingual corpora, comparative corpora, and parallel corpora. Multilingual corpora are defined as a collection of two or more monolingual corpora in different languages, constructed in the same organization or in different ways based on similar design criteria, i.e. this type of corpora includes texts in the mother tongue and does not contain any translated texts. Meanwhile, comparative corpora include two separate collections of texts in the same language: one corpus consists of original texts and the other includes the target language translation from a given source language. Finally, the parallel corpus includes the original texts, the source language, and their translated versions into the target language (Baker, 1995, p. 230).

Based on Baker's corpus classification, linguists have proposed refining it into two types: comparative and parallel corpora. In comparative corpora, documents are collected based on "textual similarity" (texts are collected based on similarity in topic, text type, communicative function, etc.), whereas in parallel corpora, texts are grouped together on the basis of "translation similarity" (i.e. one text can be considered as a translation of the other and vice versa). With this new perspective, comparative or parallel corpora can also be multilingual. According to Fernandes (2006), in translation research, a corpus can be divided into six types: subject (translation/language), domain (general/specialized), mode (written/spoken), types of relations between texts (comparative/parallel), historical (diachronic/synchronic), number of languages (monolingual/bilingual/multilingual), and orientation (unidirectional/bidirectional/multidirectional).

In terms of the number of languages, a corpus can be classified as monolingual (only one language), bilingual (two languages), or multilingual when there are more than three languages. In this study, our corpus has been constructed as an English-Vietnamese bilingual with the original text genre and with translations aligned sentence by sentence, in a unidirectional manner from English to Vietnamese.

2.3 Corpus building and analysis

First, building a corpus has a profound impact on the success of meeting the users' needs for specific purposes. According to Ngula (2017, pp.210-211), designing a corpus first requires planning, deciding, and building a framework to guide the collection and processing of documents for that corpus. Before proceeding data collection, the corpus builder needs to answer the following questions:

1. What types of documents are involved?

2. What will be the size of the corpus?

3. Will the collection consist of cited documents or the entire corpus?

4. How many text samples are there?

Many researchers believe that building a corpus requires many criteria such as size, sampling, balance, representativeness, rights, and copyright. According to Biber (1993, p.243), representativeness is the extent to which a sample covers the full range of variation in a population. Balance is achieved if there is a full range of genres or text types in the corpus. The way in which each text excerpt or entire text is selected for inclusion in the corpus is called sampling. Baker (2010, p.96) thinks that "since a corpus must be representative of a particular language, linguistic variety, or topic, the texts within it must be carefully sampled and balanced to ensure that some texts do not skew the overall corpus. Copyright and rights are also considered when designing a corpus, however, there is still much debate among linguists as to whether these factors are applied when constructing corpora in practice.

With the advancements in corpus building, the cloud-based Smart CAT was used to build a bilingual English-Vietnamese corpus on Tourism. In fact, Smartcat is a cloud-based CAT tool with a connection of professional translators, language companies and business organizations" (Smartcat, 2021). According to Ariyaratne (2019), this software is one of the market leaders in the computer-assisted translation (CAT) tool field. Malenova (2019) believes that Smartcat brings together easy-to-understand functional elements and a user-friendly interface, making it an ideal tool for teaching translation technology from the ground up.

Second, corpus analysis tools have also been developed to meet the needs of processing and analyzing large corpora. Thanks to corpus analysis software, corpora can be analyzed quickly with good results. Baker (2010, p.102) says that a standalone corpus is not really useful to support language analysis, and corpora are often used in conjunction with software that can count, sort, and present language features. Hunston (2006, p.234) gives an outline of what corpus software does, including: i/ Searching the corpus for a given purpose; ii/ Counting the number of copies (or corresponding translations) of the target language in the corpus and calculating the relative frequencies; iii/ Displaying the copies (or corresponding translations) so that users can conduct further research on the corpus.

So far corpus analysis has been carried out with diverse tools which can perform many different functions. Many popular corpus analysis tools are in use today such as WordSmith, AntConc, ConcGram, and Sketch Engine, etc. These tools allow users to upload corpora for analysis. Each corpus analysis tool possesses a separate package of independent functions. According to Ngula (2017), although corpus analysis is often considered a methodology, it actually involves many independent analysis methods such as frequency lists, keyword lists, concordance, cluster/n-gram analysis, and collocates. In this study, we used the Sketch Engine corpus analysis software, one of the most powerful corpus analysis tools that allows users to build and develop large corpora with many outstanding functions. Currently, the software contains 500 ready-to-use corpora in over 90 languages, each containing about 50 billion words that can provide a reliable representative sample of the language (Sketch Engine, 2023).

2.4 Advantages and drawbacks of corpus use

Language teaching in the period of technology advancement has been changed significantly due to the great impacts of some linguistic domains, especially corpus linguistics (Alqadoumi,2013; Bennett, 2010; Boulton, 2012 and 2017). In fact, corpora use in teaching and learning a second language can offer various benefits since corpora studies may have some impacts on the language teachers' methodology. For example, teachers can create many activities in their lessons for students to explore the language features with computer programs such as the concordance which is used to identify words and their common collocations in a particular corpus (Sinclair, 2004; Carlota de Jesue, 2021). Here are some advantages of incorporating corpora in the second language teaching and learning as follows:

First, corpora can improve the programs and material design courses by providing authentic examples of language use which allows learners to encounter real-life language in context. In Alcantar and Jose's opinion, since corpora inform us the use of language elements in a way which we cannot do by ourselves, they can be used as an effective tool to illustrate the syllabus, and supply with most teaching materials, for example, the modals' use and their form presented in textbooks of English (Carlota de Jesue, 2021). With some corpora-based studies on modal verbs (e.g., *would, can, might*, etc.) conducted by researchers (Carlota de Jesue, 2021; Carter, 2003; McEnergy et al., 2006), they recognize that textbooks are not only used to teach full types of modal languages, but also offer some confusing explanations for the language they teach.

Second, corpora have certain influences on language awareness. Carter defines language awareness as the learners' development in their consciousness of the language forms and functions (Carter, 2003). Hence, when learners work with corpora, they can discover language themselves. Van Lier also gives an example of language awareness in teaching activities which shows the way the language is focused on the learners, not the teachers (Van Lier, 2001). For instance, when using a corpus, learners easily search for vocabulary with its concordances. Through observing and memorizing these concordances, learners could use the words or phrases more correctly and naturally. Besides, corpora analysis allows learners to identify common lexical and grammatical patterns, enabling them to learn language structures more efficiently. This data-driven approach helps learners understand how words and structures are used in inauthentic communication.

Third, using corpora also develops learners' and teachers' autonomy and independence. Actually, corpora use can promote students' self-study, since students are responsible for their own language learning, i.e. they can search for language features in various language resources. Students can use corpora to learn different structures and grammatical patterns and vocabulary as well as themselves without any support from their teachers. That is to say, their learning autonomy has been improved. Moreover, through using corpora, both learners and teachers are able to work independently with specific examples of 'real life' language, which makes them feel more confident about the language they use to present in classrooms. Furthermore, due to the increasing number of contributions in the field of corpus linguistics, more and more linguists use corpora in their language study and analysis. This idea penetrates researcher's mind in a series of language domains like cognitive linguistics, metaphor analysis, language learning, corpus stylistics, dictionary creation, translation (Carter, 2003).

Fourth, incorporating corpora into language learning also encourages the learners to develop their cultural aspect of language use. They can understand idiomatic expressions, cultural references, and the subtleties of communication that are often challenging to teach through traditional methods. Moreover, when learners work with authentic materials from the available or newly built corpora, they find out that learning a language is more interesting and beneficial since it can bring relevance to their studies and helps them see the practical applications of what they have learnt.

Finally, through corpora use, learners surely will be able to enhance their research skills. They can explore linguistic data, analyze patterns, fostering a more independent and critical thinking approach to language learning.

Nevertheless, there are some drawbacks of corpus use that are still challenging for those who have used it at the first time. First, the IT knowledge must be required so that the users might be keen on building, applying all the functions in their teaching/ learning a second language; hence, it takes a big amount of time to learn the way of using a corpus. Second, it might be costly to build a corpus since building a corpus needs the collaboration of a group of experts who obtain various expertise on many domains like corpus linguistics, machine learning, natural language processing, and so on. Besides, there exist many problems relating to personal computers connected the Internet when the corpus users operate the sketch engine containing certain corpora. Being unable to accessing the websites must be a considearbel obstacle for people in the process of using a corpus in some cases.

In brief, the advantages of corpus use in teaching and learning a second language often outweigh its disadvantages. That is why more and more people tend to construct and use more corpora in various specific domains.

2.5 Previous studies

In corpus linguistics, building a corpus has been conducted to meet the needs of language use in a specific field. Since the first corpus was born in 1961, many corpora have been built and developed such as: Brown University Corpus - containing about one million words and phrases used, marked by word morphology; Lancaster/Oslo-Bergen Corpus (LOB) of about one million words and phrases used with two sub-corpuses, Leeds-Lancaster Treebank and Lancaster Parsed Corpus marked by syntax; British National Corpus (BNC) - is the largest English corpus with 100 million words and phrases used; COCA (Corpus of Contemporary American English) is one of the largest corpora in the world with a total of more than 1 billion American English words stored from a variety of genres such as film scripts, spoken language, literary works, newspapers and academic texts.

Regarding the building of the English-Vietnamese bilingual corpus, Vietnamese researchers have carried out many studies for different purposes, which have contributed to the development of English-Vietnamese automatic translation and natural language processing domains. Some studies have been conducted in the building of English-Vietnamese corpora like the Master's thesis "Research on building a bilingual corpus for Vietnamese language processing" (Quynh, 2011) to study and build a corpus containing English-Vietnamese sentence pairs from various sources such as websites, dictionaries, books, newspapers, documents, etc. in various formats such as XML, TXT, DOC, etc. The researcher collected English-Vietnamese corpus documents and presented methods for using parallel English-Vietnamese corpora to create a language database for automatic translation, natural language processing and English learning. Another study on "Building a bilingual English-Vietnamese corpus for machine translation" (Ngo & Winiwarter, 2012) indicates the building of a parallel English-Vietnamese corpus to create a Vietnamese-English automatic translation system. This paper describes the techniques of data collection for the corpus, language tagging, bilingual annotation, and specially developed tools for manual annotation. The English-Vietnamese bilingual system was built with more than 800,000 sentence pairs and 10,000,000 English and Vietnamese words collected and aligned at the sentence level, and more than 45,000 sentence pairs in this corpus were aligned at the word level. The results of this study have played an important role in the field of English-Vietnamese automatic translation. The above studies have contributed to improving the methods and tools for constructing English-Vietnamese parallel corpora as well as improving the quality and efficiency of English-Vietnamese corpora, promoting the corpus-based approach in language research.

3. Research Methodology

3.1 Process of building the E-V bilingual corpus of Tourism

In this section, we introduce all the procedures of building a parallel corpus which can be used for students learning English as a second language as well as in translation studies. As Kennedy states that there are three stages of corpus compilation such as corpus design, text collection or capture and text encoding and markup (Kennedy, 1998). Adolphs also confirms three stages in the corpus compilation process including data collection, annotation and markup, and storage (Adolphs, 2008). There is an overlap between these two opinions which can be seen as an annotation step. However, in our study, we focus on three main stages presented in this building process as follows:

Stage 1: Preparation of the database for the corpus.

During the preparation, we need to decide the following issues:

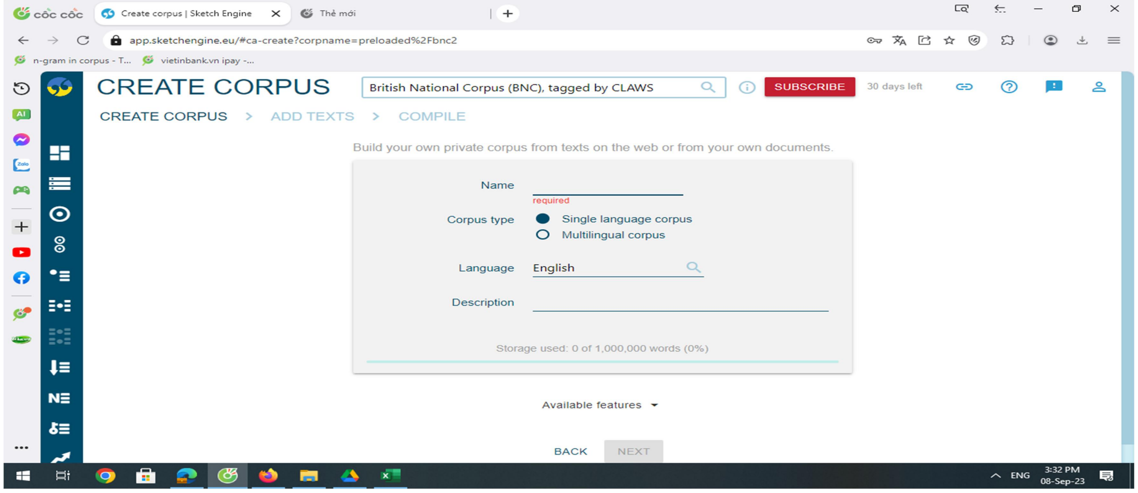
1. Purpose of corpus: This corpus is built to help Vietnamese students learn ESP (English as Specific Purposes) in the tourism domain
2. Kind of corpus: English-Vietnamese parallel corpus
3. Specific domain: Tourism
4. Size of corpus: about one million words
5. Resources: electronic articles, news, textbooks, journals, etc.
6. Features of corpus: This corpus is an electronic body of linguistic data (texts) extracted from the larger texts which have been translated from English (the source language) to Vietnamese (the target language). This parallel corpus can help the learners search for terms, phrases, collocations, and sentences that are aligned in both English and Vietnamese.
7. Format of data: Since The data prepared for the corpus are saved in the file formats like .xlsx, .xls or .txt, which can provide the builders or designers the alignments of the texts in both languages since most corpus investigation software will not read the kind of complex embedded formatting associated with common word processing packages like Microsoft Word or pdf.

Step 2: Work with Sketch Engine

Sketch Engine is known as a corpus manager and text analysis software developed by Lexical Computing CZ s.r.o. since 2003 (Sketch Engine, 2023). To support people to study languages, for example, the second language learners, researchers in corpus linguistics, and translators are able to search large text collections according to complex and linguistically motivated queries. Here are some major steps to deal with Sketch Engine.

First, you have to create an account and sign up to the Sketch Engine website: sketchengine.eu.

Second, you should choose the *New corpus* and select languages before *Creating Corpus* and then name your corpus (*see Figure 1*).



cốc cốc
Create corpus | Sketch Engine
Thẻ mới
app.sketchengine.eu/#ca-create?corpname=preloaded%2Fbnc2
n-gram in corpus - T...
vietinbank.vn ipay -...
British National Corpus (BNC), tagged by CLAWS
SUBSCRIBE
30 days left
ENG
3:32 PM
08-Sep-23

CREATE CORPUS

CREATE CORPUS > ADD TEXTS > COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name
required

Corpus type ● Single language corpus
○ Multilingual corpus

Language English

Description

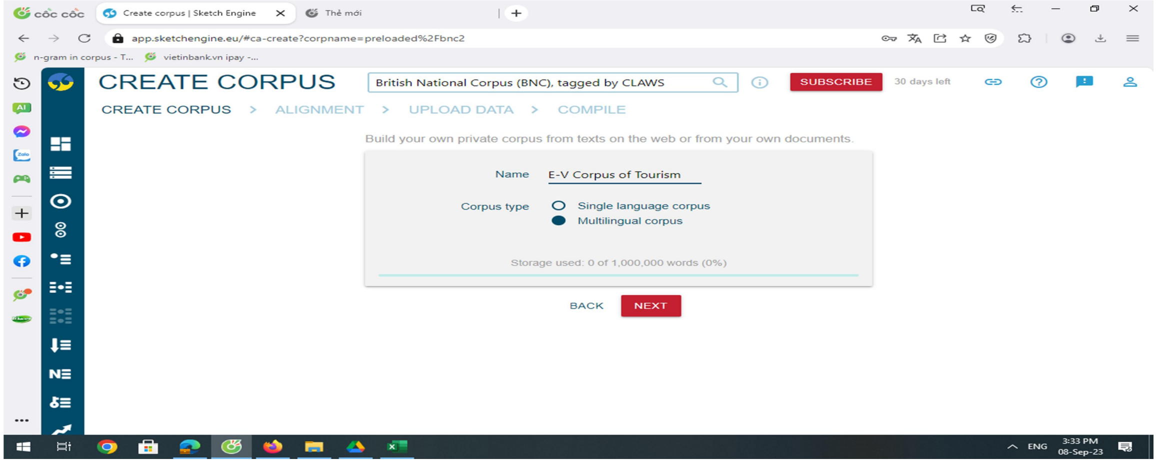
Storage used: 0 of 1,000,000 words (0%)

Available features

BACK NEXT

Fig. 1 Selecting languages and creating a corpus with Sketch Engine

If you want to create a monolingual corpus, you can select *Single language corpus*; otherwise, *Multilingual corpus* is your option if you would like to build a parallel corpus (*see Figure* 2).



cốc cốc
Create corpus | Sketch Engine
Thẻ mới
app.sketchengine.eu/#ca-create?corpname=preloaded%2Fbnc2
n-gram in corpus - T...
vietinbank.vn ipay -...
British National Corpus (BNC), tagged by CLAWS
SUBSCRIBE
30 days left
ENG
3:33 PM
08-Sep-23

CREATE CORPUS

CREATE CORPUS > ALIGNMENT > UPLOAD DATA > COMPILE

Build your own private corpus from texts on the web or from your own documents.

Name E-V Corpus of Tourism

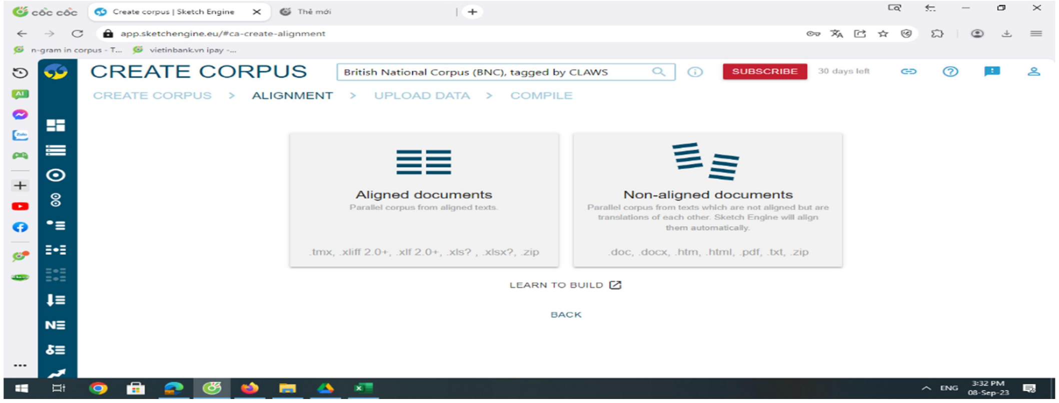
Corpus type ○ Single language corpus
● Multilingual corpus

Storage used: 0 of 1,000,000 words (0%)

BACK NEXT

Fig.2 Selecting Multilingual corpus if the corpus relates to many languages

Moreover, you will choose non-aligned documents if your documents can be aligned automatically (*see Figure 3*).


CREATE CORPUS
British National Corpus (BNC), tagged by CLAWS
SUBSCRIBE
30 days left
CREATE CORPUS > ALIGNMENT > UPLOAD DATA > COMPILE
Aligned documents
.tmx, .xliff 2.0+, .xlf 2.0+, .xls? , .xlsx?, .zip
Non-aligned documents
.doc, .docx, .htm, .html, .pdf, .txt, .zip
LEARN TO BUILD
BACK

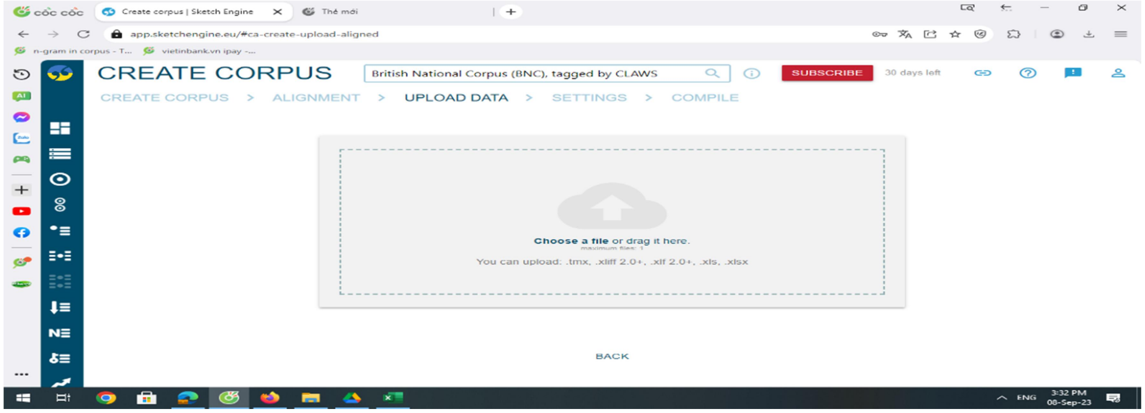
Parallel corpus from aligned texts.

Parallel corpus from texts which are not aligned but are translations of each other. Sketch Engine will align them automatically.

Fig. 3 Selecting aligned or non-aligned documents

Then you will upload your files remembering that your files should be accepted in the .doc, .docx, .html, .pdf formats.

-


CREATE CORPUS
British National Corpus (BNC), tagged by CLAWS
SUBSCRIBE
30 days left
CREATE CORPUS > ALIGNMENT > UPLOAD DATA > SETTINGS > COMPILE
BACK

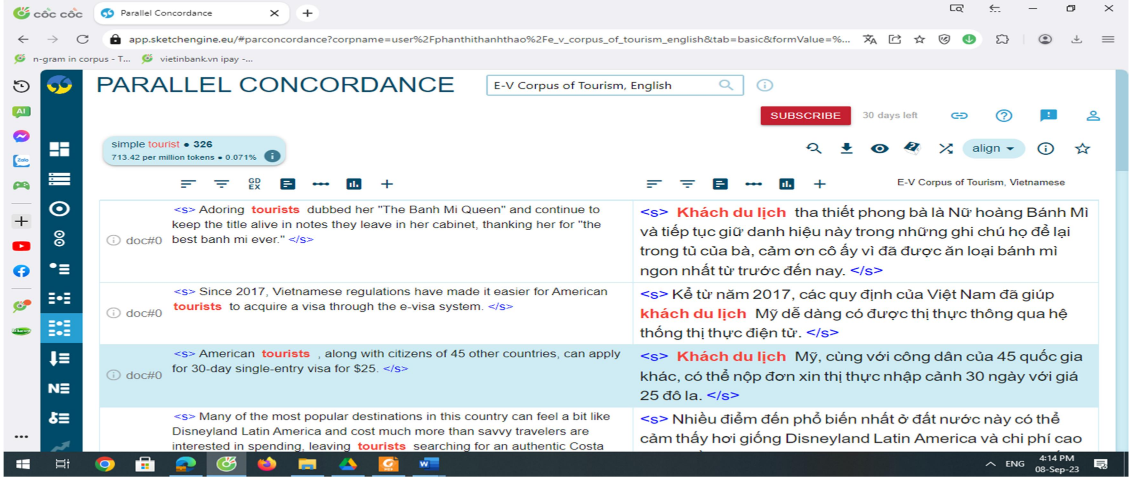
Choose a file or drag it here.
maximum files: 1
You can upload: .tmx, .xliff 2.0+, .xlf 2.0+, .xls, .xlsx

Fig. 4 Uploading data in Sketch Engine

Just after several minutes, the corpus will have been built in Sketch Engine.

<u>**Stage 3:**</u> Corpus analysis

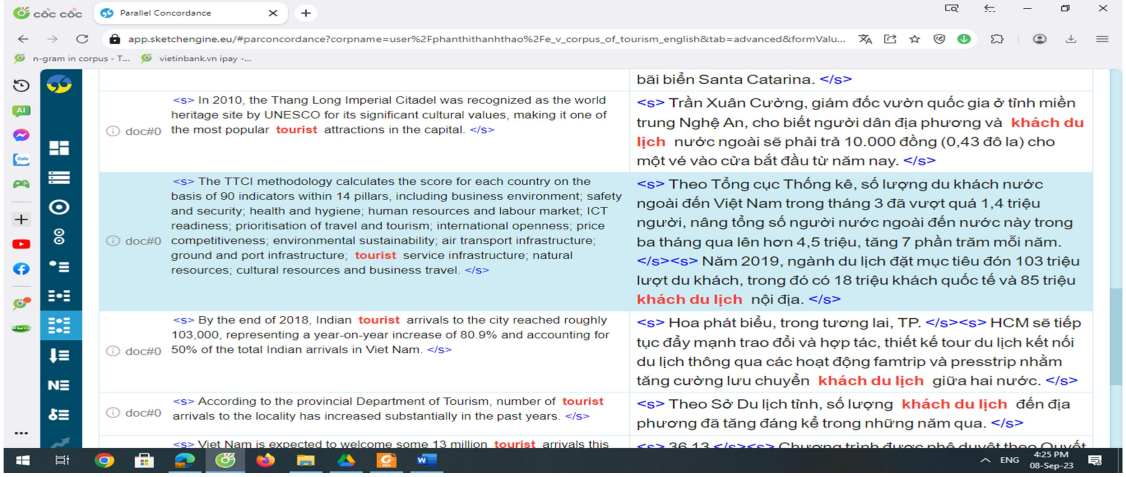
Dealing with corpus analysis, we should pay attention to the evaluation of the frequency of words' occurrences, which is considered one of the simplest, but best-known possible metrics. Take an example of Parallel Concordance in the English-Vietnamese parallel corpus that we have just designed, we may search for any words in English and their equivalents in Vietnamese in this corpus. For instance, we can find the plural noun "*tourists*" with its frequency of 326 times, 713.42 per million tokens accounting for 0.071%. Figure 5 below shows how we find out the concordances of terms, collocations, or sentences in both English and Vietnamese. Therefore, we can compare the linguistic features of this language pair.


cốc cốc
Parallel Concordance
app.sketchengine.eu/#parconcordance?corpname=user%2Fphanthithanhthao%2Fe_v_corpus_of_tourism_english&tab=basic&formValue=%...
n-gram in corpus - T...
vietinbank.vn ipay -...
PARALLEL CONCORDANCE
E-V Corpus of Tourism, English
SUBSCRIBE
30 days left
align
simple tourist • 326
713.42 per million tokens • 0.071%
GD EX
E-V Corpus of Tourism, Vietnamese
ENG
4:14 PM
08-Sep-23

doc#0	<s> Adoring **tourists** dubbed her "The Banh Mi Queen" and continue to keep the title alive in notes they leave in her cabinet, thanking her for "the best banh mi ever." </s>	<s> **Khách du lịch** tha thiết phong bà là Nữ hoàng Bánh Mì và tiếp tục giữ danh hiệu này trong những ghi chú họ để lại trong tủ của bà, cảm ơn cô ấy vì đã được ăn loại bánh mì ngon nhất từ trước đến nay. </s>
doc#0	<s> Since 2017, Vietnamese regulations have made it easier for American **tourists** to acquire a visa through the e-visa system. </s>	<s> Kể từ năm 2017, các quy định của Việt Nam đã giúp **khách du lịch** Mỹ dễ dàng có được thị thực thông qua hệ thống thị thực điện tử. </s>
doc#0	<s> American **tourists** , along with citizens of 45 other countries, can apply for 30-day single-entry visa for $25. </s>	<s> **Khách du lịch** Mỹ, cùng với công dân của 45 quốc gia khác, có thể nộp đơn xin thị thực nhập cảnh 30 ngày với giá 25 đô la. </s>
	<s> Many of the most popular destinations in this country can feel a bit like Disneyland Latin America and cost much more than savvy travelers are interested in spending, leaving **tourists** searching for an authentic Costa	<s> Nhiều điểm đến phổ biến nhất ở đất nước này có thể cảm thấy hơi giống Disneyland Latin America và chi phí cao

Fig. 5 Comparing the linguistic features of the word "*tourists*" in the parallel corpus

Similarly, if you want to find some phrases with the singular noun "*tourist*", there are only 72 phrases in our corpus with 157.56 per million tokens (0.016%).


cốc cốc
Parallel Concordance
app.sketchengine.eu/#parconcordance?corpname=user%2Fphanthithanhthao%2Fe_v_corpus_of_tourism_english&tab=advanced&formValu...
n-gram in corpus - T...
vietinbank.vn ipay -...
ENG
4:25 PM
08-Sep-23

		bãi biển Santa Catarina. </s>
doc#0	<s> In 2010, the Thang Long Imperial Citadel was recognized as the world heritage site by UNESCO for its significant cultural values, making it one of the most popular **tourist** attractions in the capital. </s>	<s> Trần Xuân Cường, giám đốc vườn quốc gia ở tỉnh miền trung Nghệ An, cho biết người dân địa phương và **khách du lịch** nước ngoài sẽ phải trả 10.000 đồng (0,43 đô la) cho một vé vào cửa bắt đầu từ năm nay. </s>
doc#0	<s> The TTCI methodology calculates the score for each country on the basis of 90 indicators within 14 pillars, including business environment; safety and security; health and hygiene; human resources and labour market; ICT readiness; prioritisation of travel and tourism; international openness; price competitiveness; environmental sustainability; air transport infrastructure; ground and port infrastructure; **tourist** service infrastructure; natural resources; cultural resources and business travel. </s>	<s> Theo Tổng cục Thống kê, số lượng du khách nước ngoài đến Việt Nam trong tháng 3 đã vượt quá 1,4 triệu người, nâng tổng số người nước ngoài đến nước này trong ba tháng qua lên hơn 4,5 triệu, tăng 7 phần trăm mỗi năm. </s><s> Năm 2019, ngành du lịch đặt mục tiêu đón 103 triệu lượt du khách, trong đó có 18 triệu khách quốc tế và 85 triệu **khách du lịch** nội địa. </s>
doc#0	<s> By the end of 2018, Indian **tourist** arrivals to the city reached roughly 103,000, representing a year-on-year increase of 80.9% and accounting for 50% of the total Indian arrivals in Viet Nam. </s>	<s> Hoa phát biểu, trong tương lai, TP. </s><s> HCM sẽ tiếp tục đẩy mạnh trao đổi và hợp tác, thiết kế tour du lịch kết nối du lịch thông qua các hoạt động famtrip và presstrip nhằm tăng cường lưu chuyển **khách du lịch** giữa hai nước. </s>
doc#0	<s> According to the provincial Department of Tourism, number of **tourist** arrivals to the locality has increased substantially in the past years. </s>	<s> Theo Sở Du lịch tỉnh, số lượng **khách du lịch** đến địa phương đã tăng đáng kể trong những năm qua. </s>
	<s> Viet Nam is expected to welcome some 13 million **tourist** arrivals this	<s> 36,13 </s><s> Chương trình được phê duyệt theo Quyết

Fig. 6 Collocations with the singular noun "*tourist*" in the parallel corpus

3.2 Experiment with the Corpus

3.2.1 Participants and investigation procedures

The participants in this study are 350 students from English Faculty, University of Foreign Languages and International Relations, Hue University who are pursuing English courses on Tourism and interpreting & translating in the first semester of 2023-2024 academic year.

3.2.2 Research instruments for data collection and analysis

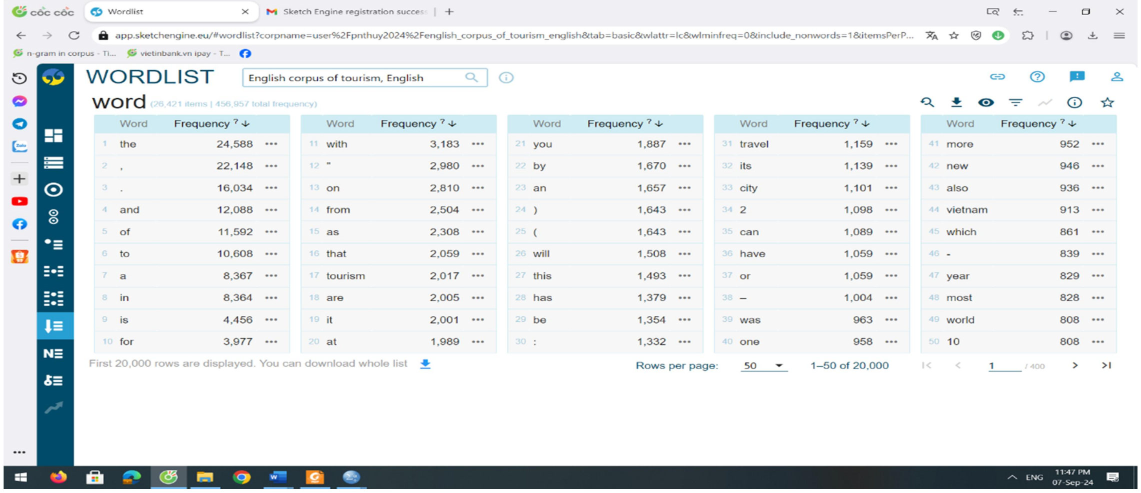
The data were collected by means of the following procedures: 1. Questionnaires (50 questions) were designed and administered to EFL students to measure their understanding of corpus linguistics and their attitudes towards the need of corpus use in their ESP learning; 2. Follow-up interviews with 20 students (10 in-depth questions) were used to ask for more details in order to clarify findings obtained through questionnaires. Those findings were further demonstrated on tables based on the 5- Likert scale: 1-Strongly disagree (M=1.00-1.80), 2-Disagree (M=1.81-2.60), 3-Neutral (M=2.61-3.40), 4-Agree (M=3.41-4.20), 5-Strongly agree (M=4.21-5.00), the collaboration of qualitative and quantitative approaches was employed.

4. Results and discussion

4.1 E-V bilingual corpus of Tourism built of one million words

The English-Vietnamese bilingual corpus of cultural tourism has been built with approximately 1 million words which can support students in learning ESP of tourism and translating/ interpreting. This corpus building has been carried out during 6 months in 2023. Here are some examples showing how to use the English – Vietnamese bilingual corpus of tourism.

To find out the frequency of words including lemmas, adjective, adverb, conjunction, noun, prepositions, pronoun, verb, numeral, tags, lemposes, pos, students can select “all starting with or ending with, or containing, matching regex, from this list”, the total number of frequency of many kinds of words will appear . Figure 7 shows the frequency of “the” article (24,588), “a” (8,367), “ and” conjunction (12,088), “of” preposition (11,592) , “to” (10,608), etc.


cốc cốc
Wordlist
Sketch Engine registration success
app.sketchengine.eu/#wordlist?corpname=user%2Fpnthuy2024%2Fenglish_corpus_of_tourism_english&tab=basic&wlattr=lc&wlminfreq=0&include_nonwords=1&itemsPerP...
n-gram in corpus - Ti...
vietinbank.vn ipay - T...
ENG
11:47 PM
07-Sep-24

WORDLIST

English corpus of tourism, English

word (26,421 items | 456,957 total frequency)

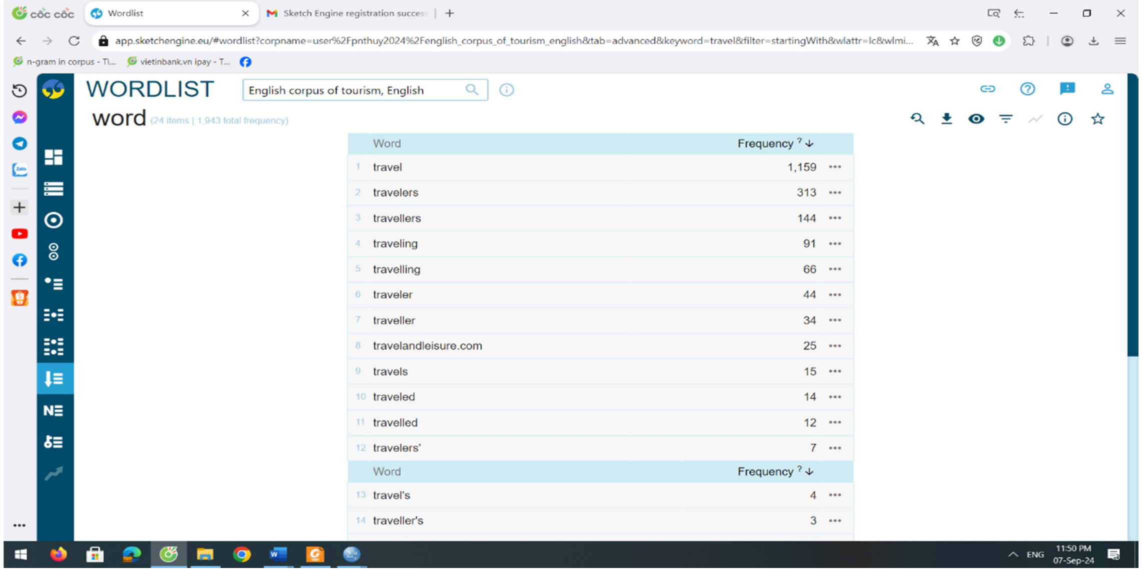
	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	the	24,588	11	with	3,183	21	you	1,887	31	travel	1,159	41	more	952
2	,	22,148	12	"	2,980	22	by	1,670	32	its	1,139	42	new	946
3	.	16,034	13	on	2,810	23	an	1,657	33	city	1,101	43	also	936
4	and	12,088	14	from	2,504	24	)	1,643	34	2	1,098	44	vietnam	913
5	of	11,592	15	as	2,308	25	(	1,643	35	can	1,089	45	which	861
6	to	10,608	16	that	2,059	26	will	1,508	36	have	1,059	46	-	839
7	a	8,367	17	tourism	2,017	27	this	1,493	37	or	1,059	47	year	829
8	in	8,364	18	are	2,005	28	has	1,379	38	–	1,004	48	most	828
9	is	4,456	19	it	2,001	29	be	1,354	39	was	963	49	world	808
10	for	3,977	20	at	1,989	30	:	1,332	40	one	958	50	10	808

First 20,000 rows are displayed. You can download whole list

Rows per page: 50 1–50 of 20,000 1 / 400

Fig. 7 Frequency of other words in the English corpus of tourism

If students would like to search the frequency of a specific word, for example, the word " travel" " travelers", " travelling", they can find their frequencies 1,159; 313, and 91 respectively as shown in Figure 8 below:



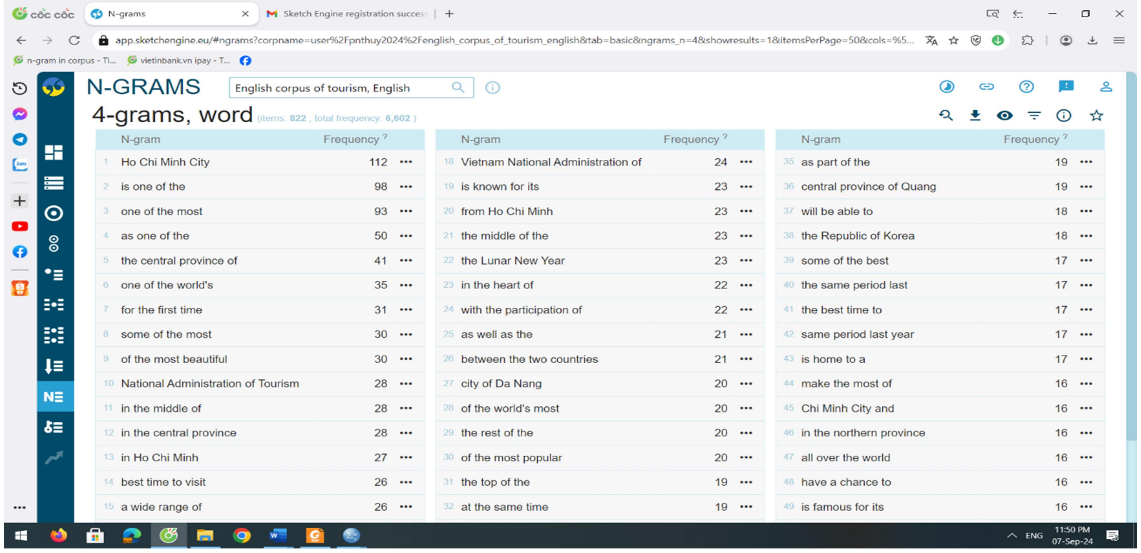
WORDLIST English corpus of tourism, English

word (24 items | 1,943 total frequency)

	Word	Frequency
1	travel	1,159
2	travelers	313
3	travellers	144
4	traveling	91
5	travelling	66
6	traveler	44
7	traveller	34
8	travelandleisure.com	25
9	travels	15
10	traveled	14
11	travelled	12
12	travelers'	7
13	travel's	4
14	traveller's	3

Fig. 8 Frequency of "*travel*" in the English corpus of tourism

Furthermore, students are able to select the function N-grams to search for the frequency of phrases containing 2,3,4,5 or N words in the corpus. Figure 9 presents 4-grams of the corpus with their frequencies, e.g. *Ho Chi Minh City* (112); *is one of* (98), *one of the most* (93), *as one of the* (50), *the central province of* (41), etc.



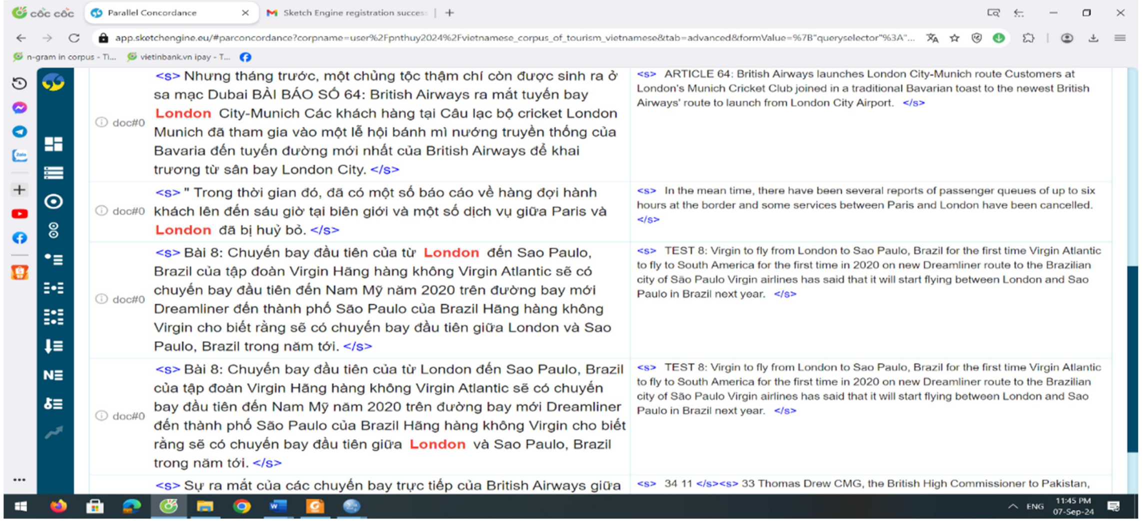
N-GRAMS English corpus of tourism, English

4-grams, word (items: 822 , total frequency: 6,602)

	N-gram	Frequency
1	Ho Chi Minh City	112
2	is one of the	98
3	one of the most	93
4	as one of the	50
5	the central province of	41
6	one of the world's	35
7	for the first time	31
8	some of the most	30
9	of the most beautiful	30
10	National Administration of Tourism	28
11	in the middle of	28
12	in the central province	28
13	in Ho Chi Minh	27
14	best time to visit	26
15	a wide range of	26
18	Vietnam National Administration of	24
19	is known for its	23
20	from Ho Chi Minh	23
21	the middle of the	23
22	the Lunar New Year	23
23	in the heart of	22
24	with the participation of	22
25	as well as the	21
26	between the two countries	21
27	city of Da Nang	20
28	of the world's most	20
29	the rest of the	20
30	of the most popular	20
31	the top of the	19
32	at the same time	19
35	as part of the	19
36	central province of Quang	19
37	will be able to	18
38	the Republic of Korea	18
39	some of the best	17
40	the same period last	17
41	the best time to	17
42	same period last year	17
43	is home to a	17
44	make the most of	16
45	Chi Minh City and	16
46	in the northern province	16
47	all over the world	16
48	have a chance to	16
49	is famous for its	16

Fig. 9 Frequency of 4-grams in the English corpus of tourism

In addition, students who are interested in translation domain can select the function of Parallel Concordance to search for some alignments including sentences, paragraphs, titles , headings, etc. in both languages English and Vietnamese. Figure 10 indicates the English texts and their translated Vietnamese versions after the corpus user selects the Paralelll concordance function.


cốc cốc
Parallel Concordance
Sketch Engine registration succes
app.sketchengine.eu/#parconcordance?corpname=user%2Fpnthuy2024%2Fvietnamese_corpus_of_tourism_vietnamese&tab=advanced&formValue=%7B"queryselector"%3A"...
n-gram in corpus - Ti...
vietinbank.vn ipay - T...
ENG
11:45 PM
07-Sep-24

doc#0	<s> Nhưng tháng trước, một chủng tộc thậm chí còn được sinh ra ở sa mạc Dubai BÀI BÁO SỐ 64: British Airways ra mắt tuyến bay **London** City-Munich Các khách hàng tại Câu lạc bộ cricket London Munich đã tham gia vào một lễ hội bánh mì nướng truyền thống của Bavaria đến tuyến đường mới nhất của British Airways để khai trương từ sân bay London City. </s>	<s> ARTICLE 64: British Airways launches London City-Munich route Customers at London's Munich Cricket Club joined in a traditional Bavarian toast to the newest British Airways' route to launch from London City Airport. </s>
doc#0	<s> " Trong thời gian đó, đã có một số báo cáo về hàng đợi hành khách lên đến sáu giờ tại biên giới và một số dịch vụ giữa Paris và **London** đã bị huỷ bỏ. </s>	<s> In the mean time, there have been several reports of passenger queues of up to six hours at the border and some services between Paris and London have been cancelled. </s>
doc#0	<s> Bài 8: Chuyến bay đầu tiên của từ **London** đến Sao Paulo, Brazil của tập đoàn Virgin Hãng hàng không Virgin Atlantic sẽ có chuyến bay đầu tiên đến Nam Mỹ năm 2020 trên đường bay mới Dreamliner đến thành phố São Paulo của Brazil Hãng hàng không Virgin cho biết rằng sẽ có chuyến bay đầu tiên giữa London và Sao Paulo, Brazil trong năm tới. </s>	<s> TEST 8: Virgin to fly from London to Sao Paulo, Brazil for the first time Virgin Atlantic to fly to South America for the first time in 2020 on new Dreamliner route to the Brazilian city of São Paulo Virgin airlines has said that it will start flying between London and Sao Paulo in Brazil next year. </s>
doc#0	<s> Bài 8: Chuyến bay đầu tiên của từ London đến Sao Paulo, Brazil của tập đoàn Virgin Hãng hàng không Virgin Atlantic sẽ có chuyến bay đầu tiên đến Nam Mỹ năm 2020 trên đường bay mới Dreamliner đến thành phố São Paulo của Brazil Hãng hàng không Virgin cho biết rằng sẽ có chuyến bay đầu tiên giữa **London** và Sao Paulo, Brazil trong năm tới. </s>	<s> TEST 8: Virgin to fly from London to Sao Paulo, Brazil for the first time Virgin Atlantic to fly to South America for the first time in 2020 on new Dreamliner route to the Brazilian city of São Paulo Virgin airlines has said that it will start flying between London and Sao Paulo in Brazil next year. </s>
	<s> Sự ra mắt của các chuyến bay trực tiếp của British Airways giữa	<s> 34 11 </s><s> 33 Thomas Drew CMG, the British High Commissioner to Pakistan,

Fig. 10 An example of Parallel concordance from the E-V bilingual corpus of tourism

During the corpus building process, we have also created an English-Vietnamese glossary of tourism including 3,250 words. This glossary provides the users the terms or phrases relating to the Tourism topic that has been appreaciated helpful and effective for students to learn English for tourism.

4.2 Students' practice of corpus use

To further investigate students' opinions on the advantages and disadvantages of using bilingual English-Vietnamese corpus of tourism in learning ESP of tourism, we have experimented this corpus for 350 students from English Faculty, HUFLIS during the first semester of the academic year 2023-2024. The study results have been achieved from 339 respondents as follows:

4.2.1 Advantages

Firstly, in learning vocabulary, especially specialized terms in Tourism, the use of bilingual English-Vietnamese of Tourism has brought many significant advantages. With an average of M= 3.86, it can be seen that a large number of students completely agreed that this corpus use has helped them easiliy search for specialized terms in both English and Vietnamese without teachers' guidance. Moreover, the technique of searching for terms and vocabulary is quite simple. Surprisingly, the same proportion (M=3.86) of students agreed and completely agreed that using corpus could help them build a glossary of terms with an ease. Finally, according to our survey, the majority of students are aware of the benefits of using the bilingual corpus in learning specialized vocabulary in Tourism as well as the second language faster and more effectively (M=3.98). This is a typical and useful feature of corpus use for the second language vocabulary learners.

Second, in the translation industry, using this corpus has made a considerable contribution to learn translating texts on the same topic of Tourism from English to Vietnamese. With a volume of nearly one million words, this corpus has helped students translate documents better when using the structures of two languages: English and Vietnamese. M=3.82 shows that a high percentage of students agreed and completely agreed with this advantage of corpus use. Moreover, using this bilingual corpus of Tourism has saved time for students while they memorized specialized vocabulary in both languages, which supported them respond quickly during their translation work. With M=3.95, it can be affirmed once again that corpus use has aided students in their translation skills thanks to the consolidation and development of Tourism knowledge in both languages.

Third, for technology skills, corpus use helps students better master the idioms or phrases of both languages via modern electronic devices like computers, smartphones or tablets. A high percentage (M=3.89) of students agreed and strongly agreed with this opinion. Especially in the interview, 17/20 students affirmed that using corpus completely helps them gradually improve their IT knowledge and skills, as one student (ST8) shared " I've never known what Sketch Engine software was like before, but after using the bilingual English-Vietnamese corpus of Tourism, I have a deeper knowledge of this software, and feel very excited to approach and apply it in language research and translation". Sharing the same thought as ST8, ST10 said: "Using this bilingual English-Vietnamese corpus is really useful because it helps me learn a lot of IT knowledge, for example, via Sketch Engine, I learned how to analyze the characteristics, functions and structure of a language, how to build a corpus, and compare it with other available corpora".

As shown in Table 1, the standard deviation from .602 to .794 (<1) is a very small number indicating that there is no significant difference in the results, which affirms the reliability of the survey results of all the benefits that the use of this corpus has brought to students when they participate in ESP for Tourism courses at the Faculty of English, HUFLIS.

Table 1. Advantages of the corpus use as an ESP material

No	Advantages	Total	Mean	Standard deviation
1	Using this bilingual corpora on Tourism helps me easily search for specialized terms in both languages without the teacher's guidance.	N= 339	3.86	.612
2	The technique of searching for vocabulary becomes easy when I use the bilingual corpus.	N= 339	3.96	.602
3	Using this bilingual corpus helps me translate texts better in looking up the structures of the two languages.	N= 339	3.82	.633
4	Using this bilingual corpus helps me to compile statistics of terms to build glossaries more easily.	N= 339	3.86	.794
5	Using this bilingual corpus helps me easily access idioms or phrases with electronic devices such as computers, smartphones, or tablets.	N= 339	3.89	.646
6	Using this bilingual corpus on Tourism Culture helps me save time in memorizing specialized vocabulary in both languages	N= 339	3.95	.637
7	My knowledge of the Tourism Culture has been improved after using this bilingual corpus.	N= 339	3.94	.615
8	My IT knowledge has been improved when I regularly use the bilingual corpus.	N= 339	3.93	.700
9	My vocabulary of Tourism Culture of both languages has been increased significantly after this corpus use.	N= 339	3.94	.634
10	Using the corpus helps me learn Tourism Culture vocabulary as well as language knowledge faster and more effectively.	N= 339	3.98	.672

4.2.2 *Disadvantages*

In addition to the advantages gained from the students' use of the English-Vietnamese bilingual corpus of Tourism, there still exist some specific shortcomings as shown in Table 2 below:

Table 4. **Challenges of using the E-V bilingual corpus as an ESP material**

No	Disadvantages	Total	Mean	Standard deviation
11	I have difficulty downloading software applications when using the bilingual corpus on the Tourism.	N= 339	3.48	.672
12	Using the bilingual corpus is difficult because I do not have good knowledge of information technology	N= 339	3.60	.893
13	I sometimes encounter technical problems when using the bilingual corpus, so it takes a lot of time to search for information.	N= 339	3.61	.751
14	I have difficulty using the corpus to compile vocabulary or terminology to create a dictionary or a glossary of specialized terms.	N= 339	3.59	.780
15	I need to equip myself with learning aids such as computers, smartphones or tablets to be able to use the bilingual corpus, so I have financial difficulties.	N= 339	3.58	.818
16	I have difficulty downloading software applications when using the bilingual corpus on the Tourism.	N= 339	3.36	.826
17	Using the bilingual corpus is difficult because I do not have good knowledge of information technology	N= 339	3.56	.720
18	I sometimes encounter technical problems when using the bilingual corpus, so it takes a lot of time to search for information.	N= 339	3.68	.766
19	I have difficulty using the corpus to compile vocabulary or terminology to create a dictionary or a glossary of specialized terms.	N= 339	3.58	.689
20	I need to equip myself with learning aids such as computers, smartphones or tablets to be able to use the bilingual corpus, so I have financial difficulties.	N= 339	3.63	.728

In fact, there are many different opinions about the limitations of using the English-Vietnamese bilingual corpus of students participating in ESP courses of Tourism. Among 10 disadvantages mentioned above, the challenges relating to information technology skills and the ability of IT application such as downloading software to use this corpus via Sketch Engine, more than half of the students (M=3.48) agreed and strongly agreed that they had many difficulties in installing Sketch Engine and creating an account to use this corpus. Furthermore, many students (M=3.36) found it very difficult to use this corpus since they did not obtain good IT knowledge and skill.

Furthermore, a high proportion of students in the survey (M=3.61) occasionally encountered technical problems when using this corpus; hence, they had to spend a lot of time searching for information such as vocabulary, terms and phrases in both languages: source language and target language. In addition, only a few students did not have difficulty using a bilingual corpus to compile vocabulary or terms to build dictionaries or glossaries. Therefore, it is necessary to equip students with IT knowledge so that they can apply it well in learning ESP, this result is similar to Li's (2012) view on the importance of information technology knowledge in learning ESP.

Regarding the weakness in combining language knowledge and information technology skills, many students have difficulty in analyzing concordance, or collocations (phrases) in the corpus (means are 3.56 and 3.69, respectively). In a more in-depth interview of 20 students on their difficulty in using this corpus, student ST11 said “I hardly know how to use Sketch Engine software to open the corpus, even using an Excel file for the original corpus is quite difficult for me, I don’t know how to search for phrases (collocations) or concordance like my friends did.”. Student ST15 declared that her language knowledge as well as IT skills were problematic, i.e. that she was still a bit weak in processing the documents when using a software, since she mainly only knew how to use Word or PowerPoint”. What is more, many students (M=3.63) said that they sometimes had challenges finding translated terms of tourism laws in this corpus as they did not understand the search techniques, or their language knowledge was limited.

In summary, Table 2 with the mean score from 3.36 to 3.68 and the standard deviation from .672 to .826 (<1) shows that the study results are quite accurate and reliable, indicating that the use of English-Vietnamese bilingual tourism corpus by students still has some disadvantages apart from the advantages mentioned in section 4.2.1. According to many students, the combination of language knowledge and IT skills can help them easily use this document in learning ESP since this corpus has provided a very useful and highly practical document.

Those problems with which students have dealt during the bilingual E-V corpus use could be solved by giving some instructions online or face-to-face before their practice. With the lecturer’s guidance, students might have found it much easier to use this corpus as they could learn how to download the software apps or create a glossary, and solve the technical problems themselves. In addition, it is suggested to organize a workshop or seminar for students to introduce the corpus linguistics applications in their ESP courses at the beginning of the school-year which will be able to bring many opportunities for students to explore the benefits of corpora use in their ESP learning and practice.

4.2.3 Suggestions of building and using the E-V bilingual corpus of Tourism

As ananyzed benefits and drawbacks of the E-V bilingual corpus of tourism in the previous sections, apparently it is essential for students of English Faculty to build and use this corpus in their learning ESP of tourism courses. Via the survey, we also figure out that students had some suggestions of building and using an English-Vietnamese bilingual corpus of Tourism as shown in Table 3.

Table 3. Students' suggestions of buidling and using an E-V bilingual corpus on Tourism as an ESP learning source

No	Questions	**Total**	**Mea n**	**Standard deviation**
21	It is suggested that this corpus should be used as an ESP learning resource because it is a very useful for language learners.	N= 339	4.01	.810
22	Since bilingual corpora are very useful in learning vocabulary and specialized terminology, students majoring in English should approach the method of its use as soon as possible.	N= 339	4.02	.554
23	Students need to learn about building bilingual corpora so that they can build them themselves as soon as possible.	N= 339	4.05	.456
24	Students could suggest their teachers regularly use the corpus to help them improve their language and IT knowledge	N= 339	4.05	.547
25	Students can learn by themselves to build many bilingual corpora of different topics	N= 339	4.05	.621

Interestingly, there are similar rates of students suggesting that this corpus should be used as an ESP learning materials since it is very useful for language learners (M= 4.01) and the similar proportion of students commented that they should approach the method of this corpus use as soon as possible (M= 4.02). Therefore, they need to learn the way how to build the bilingual corpora so that they can build them by themselves for language study purposes. A high number of students (M= 4.05) agreed and strongly agreed with this suggestion. In addition, the same ratio of students making suggestions that their lecturers should regularly use the corpus to help them improve both their language and IT knowledge, and students desire to build many bilingual corpora of different topics except for Tourism (M= 4.05).

In brief, Table 2 with the mean score from 4.01 to 4.05 and the standard deviation from .456 to .810 (<1) shows that the findings are completely believable and exact, which are noticeable to build the bilingual corpora of many topics for learning ESP courses.

5. Conclusion and implications

In the trend of promoting the research on the construction and effective application of authentic materials in foreign language teaching, this article presents the building of an English-Vietnamese bilingual corpus of Tourism having a size of nearly one million words and its use by 350 students from the English Faculty, University of Foreign Languages and International Relations, Hue University at ESP for Tourism classes. The study also finds out students' practice of using this corpus in their ESP courses, especially in their language research and translation learning as well. Moreover, this article has shown the advantages of using this bilingual corpus such as improving knowledge and methods of memorizing vocabulary on Tourism, at the same time developing translation skills and information technology skills for students, although there are still limitations such as students spend a lot of time using this corpus because of difficulties in downloading the Sketch Engine software and using it. Therefore, the study has made recommendations to promote strengths and limit weaknesses to improve the quality of English language learning in general and ESP for tourism in particular. It can be said that this is an initial study for applying the results of corpus linguistics to the field of applied linguistics, so more in-depth research on this issue is needed in the future.

References

1. Adolphs, S. (2008). *Corpus and Context.* John Benjamins Publishing Company
2. Alqadoumi, O. M. (2013). Using corpus linguistics as a tool for reform in English language teaching and learning: the case of public schools in Arab countries. In *IEEE computer society: proceedings of a conference*, University of Bahrain, Bahrain, 246-252
3. Ariyaratne, M. (2019). The impact of translation technologies on the translation process to give a quality output: A study with special reference the government translators in Sri

Lanka. *International Journal of Research and Innovation in Social Science (IJRISS)*, *III*(XII), 134–150.

4. Baker, M. (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2), 223–243.
5. Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (Ed.), *Research methods in linguistics*, London: Continuum pp. 93–113.
6. Baker, P., & McEnery, T. (2015). *Corpora and discourse studies: Integrating discourse and corpora*. Palgrave Macmillan.
7. Bennett, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press. **DOI:** https://doi.org/10.3998/mpub.371534
8. Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.
9. Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, *50*(4): 483-506. DOI 10.1017/S0261444817000167
10. Boulton, A. (2012). Computer Corpora in language learning: DST approaches to research. Melanges Crapel. Centre de recherches et d'applications pédagogiques en langues, 33, 79-91.
11. Carlota de Jesus, A. D. and Carrillo, J. L. Q. (2021). Advantages of using corpora to teach English. ISBN: 978-607-8356-17-1. https://cenedic.ucol.mx/fieel/pdf/1.pdf
12. Carter, R. (2003). Language awareness. *ELT Journal*, *57*(1), 64-65.
13. Fernandes, L. (2006). Corpora in Translation Studies: revisiting Baker's tipology. *Fragmentos*, *30*, 87–95. https://doi.org/10.5007/fragmentos.v30i0.8217
14. Hunston, S. (2006). *Corpus linguistics*. Birmingham: University of Birmingham.
15. Kennedy, G (1998). An introduction to corpus linguistics. London: Longman
16. Kubler, S., Zinsmeister, H. (2015). *Corpus Linguistics and Linguistically Annotated Corpora*, Bloomsbury-London-New York.
17. Li, X. (2012). Information Technology Application in ESP Education. 2012 International Conference on Future Computer Supported Education. *IERI Procedia 2*, 771-774.
18. Malenova, E. (2019). Cloud technologies in a translation cassroom. *Trans-Kom*, *12*(1), 76–89.
19. McEnery, T., Xiao, R., Tono, Y.(2006). *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge Publisher.

20. McEnery, T., Hardie, A. (2014). Corpus Linguistics: Method, Theory, and Practice. Cambridge University Press.

21. Ngo, Q. H., & Winiwarter, W. (2012). Building an English-Vietnamese Bilingual Corpus for Machine Translation. *2012 International Conference on Asian Language Processing*, 157–160. https://doi.org/10.1109/IALP.2012.30

22. Ngula, R. (2017). Corpus methods in language studies. In: Kuupole, D. (ed.), *Perspectives on Conducting and Reporting Research in the Humanities,* Cape Coast: University of Cape Coast Press, 205-223.

23. Oakes, M.P. (2012). Corpus linguistics and language variation. In: Baker. P. (ed.). *Contemporary corpus linguistics*. London: Continuum, 159-183.

24. Patsala, P., Michali, M. (2020). Sharpening Students' Critical Literacy Skills Through Corpus-Based Instruction: Addressing the Issue of Language Sexism. In: *Handbook of Research on Cultivating Literacy in Diverse and Multilingual Classroom*. DOI: 10.4018/978-1-7998-2722-1.ch012

25. Quynh, H. . (2011). *A study on building bilingual corpus to serve Vietnamese language processing* [Da Nang]. https://doi.org/60.48.01.

26. Rayon, P. (2015). Computational tools and methods for corpus compilation and analysis. In: Biber, D., Reppen, R. (eds.) *The Cambridge Handbook of English Corpus Linguistics*, Cambridge University Press, 32-49.

27. Sinclair, J. (2004). Corpus and texts: Basic principles. In: *Developing linguistic corpora: a guide to good practice.*

28. Sketch Engine. (2023). https://en.wikipedia.org/wiki/Sketch_Engine.

29. Smartcat. (n.d.). Retrieved December 27, 2021, from https://en.wikipedia.org/wiki/Smartcat#cite_note-:1-1

30. Van Lier, L. (2001). Language awareness. In: Carter, R., Nunan, D. (eds.) *The Cambridge Guide to teaching English to Speakers of other languages*. Cambridge: Cambridge University Press, 160-165.

31. Zanettin, F. (2014). Corpora in Translation Retrieved from https://www.researchgate.net/publication/265968828