BUILDING AND EXPLOITING VIETNAMESE ECONOMIC CORPUS

Nguyen Thi Thuy*, Tran Minh Chau*

Date received the article: 04/08/2022
Date received the review results: 03/02/2023
Date published the article: 28/02/2023

Abstract: *This article presents the steps of building a Vietnamese economic corpus, describes the corpus after its completion, and initially gives some suggestions for exploiting the corpus in teaching, learning, and researching. The methods used in this study include data collecting, processing, and analysis. The result of this study is a Vietnamese economic corpus consisting of 55,698 sentences, equivalent to 1,754,0659 words stored in the software. The research has a special practical value, providing a Vietnamese financial corpus which can be used in research, teaching, or studying in the current 4.0 technology context.*

Keywords*: corpus linguistics, corpus, and history of corpus linguistics*

I. Introduction

Corpus linguistics is a new branch of science that has made very important and even revolutionary contributions in the fields related to the study of language. The core foundation of corpus linguistics is corpus, built from collecting and storing a large volume of linguistic documents produced by users in real contexts. It is not merely a collection but also a system of tightly organized linguistic documents labelled with linguistic information such as word morphology, word form, semantics, syntax, and pragmatics. Therefore, each corpus is a huge repository of authentic linguistic resources that researchers can exploit to test hypotheses, build new laws and theories about language, and construct new theories. Exploited for other language-related purposes such as compiling dictionaries, teaching, studying, and translating.

Currently, most corpus repositories are in electronic form. With the support of rapidly developing computer technology, the corpus has helped users save a lot of time and effort in collecting and processing documents, especially with a huge volume of data. Modern corpus software systems are often set up with a number of functions such as statistics, frequency calculation, concordance, keywords, and collocations to meet a variety of research, teaching or learning purposes. Corpus linguistics shows great potential in applying information technology to language-

* National Economics University

related work and strongly promotes this trend in research communities.

Today, hundreds of corpora of many languages, most notably English. These corpora can be very large in size and organized into many different domains. In addition, researchers also build smaller corpora of a specific field, such as literature, economics, politics, law, etc., to serve specific needs related to those areas. In this article, we present the steps to build a Vietnamese corpus in the economic field and initially suggest exploiting linguistic information in the corpus for teaching Vietnamese as a foreign language, especially in the economic field.

II. Conceptual framework

2.1. Definition of corpus

The term *corpus* (plural form: corpora) stems from Latin, meaning body. The term *corpus* has been recorded since 1961 with the introduction of the first electronic corpus, the Brown corpus. However, the corpus was built and exploited a long time ago. The pre-Brown archives were mainly collected, stored, and processed manually. A corpus usually needs to ensure three criteria: authenticity, representativeness, and size. Among the most important criteria of the corpus, very few scientists mention electronic properties. Therefore, we define a corpus as follows: "A corpus is a large collection of spoken or (and) written samples used in practice, selected systematically, based on certain criteria, and collected manually or electronically, for language research and other related work".

2.2. Theories of corpus linguistics

The term *corpus linguistics* was first used by Aarts and Van den Heuvel in 1982, but according to Léon (Ramesh Krishnamurthy & Wolfgang Teubert, 2007), the term was not widely used until the 1990s with the rapid proliferation of publications and especially the introduction of the International Journal of Corpus Linguistics (IJCL).

Many researchers now believe that corpus linguistics is a new branch of science, and its formation is attached to the role of information technology in general and computer science in particular. Mc Enery (2012) defines corpus linguistics as "the study of linguistic data on a large scale – the machine-assisted analysis of a rich collection of speech or written transcripts". Author Dao Hong Thu also has a similar view on the role of computers in the formation of corpus linguistics. The author argues that "corpus linguistics (the term the author uses is equivalent to the term corpus in the article) is the intersection between language science and computer science, formed at the end of the 20th century on the basis of digital electronics, is the science of researching and building linguistic blocks, researching data processing methods and using data blocks" (Dao Hong Thu, 2007). Some other researchers do not include computers in the definition of corpus linguistics. For example, Sadinha (2004) argues that corpus linguistics "focuses on collecting and applying a corpus, or a carefully collected linguistic data set, to serve as a research resource, or language variants" (Carlos, 2019). Nguyen Thien Giap (2016) gives the following definition: "corpus linguistics is the study of language as expressed in samples of real texts".

III. Research methods

Used research methods include:

(1) Methods of data collection. Data collection is a method implemented

with many principles to achieve the criteria of a corpus, such as an authenticity, representativeness, and size-appropriateness. After the data collection step, the resulting product is an unprocessed corpus, known as raw corpus.

(2) Methods of processing corpus. The data processing method includes two phases: corpus preprocessing and language labelling. The corpus preprocessing includes steps such as cleaning the corpus, normalizing the corpus, and proofreading the corpus. Language labelling is related to linguistic knowledge, such as labelling word boundaries, sentence boundaries, word form, semantics, and grammar. This topic looks at labelling word boundaries, sentences, word types and entities. These tasks are performed semi-automatically, which means that the data will be processed and labelled by a computer. But after being processed automatically, the labeller will have to manually check for and correct the inconsistencies based on labelling principles.

(3) Analyzing method. After a Vietnamese economic corpus of economic terms is built, data extracted from the corpus will be analyzed. Then, suggestions and recommendations are made for Vietnamese language teaching and learning, a compilation of teaching materials, and a number of other related issues.

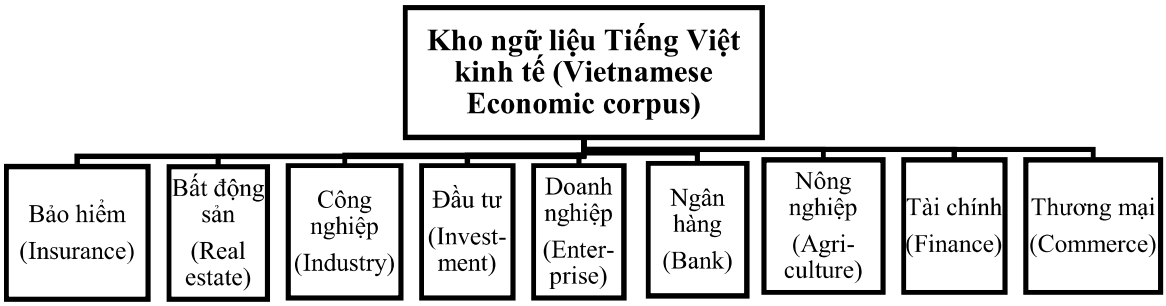
IV. Results and discussions.

4.1. A Vietnamese economic corpus

After being collected and processed, the Vietnamese economic corpus is stored in computer software built explicitly for the corpus. In general structure, the completed corpus consists of two basic components: corpus and software program.

a) The corpus.

Our corpus comprises nine fields divided into nine sub-groups, including 55.69 sentences and 1,754,0659 words.



The general model of the Vietnamese corpus in economic field

The basic figures of the corpus are presented in the following table

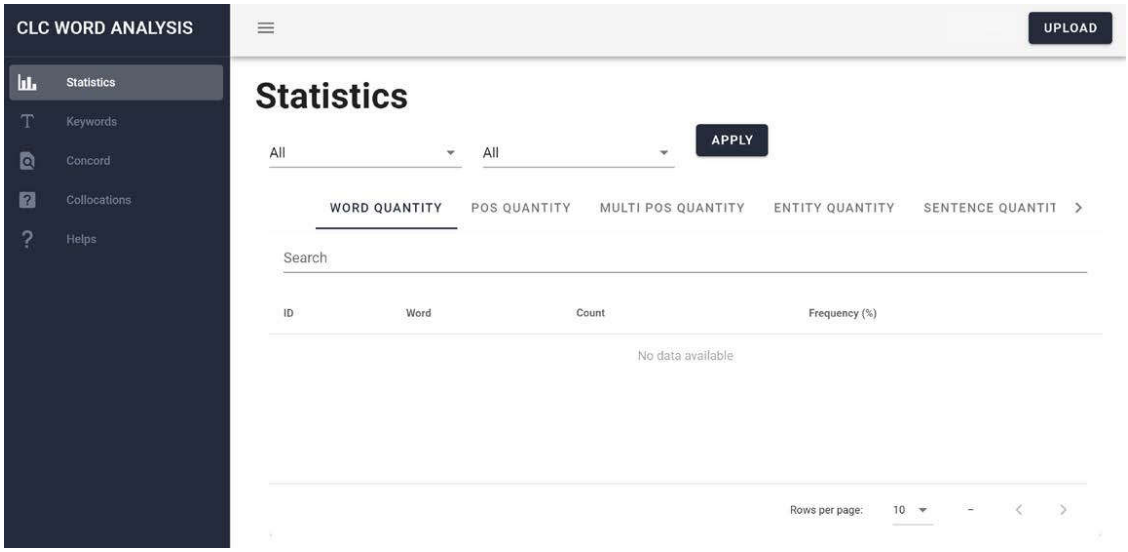
No.	Field	No. of document	No. of sentences	Word token†	Word type‡	Sentence length
1	Bảo hiểm (Insurance)	282	5.573	203.69315	7708	36.55
2	Bất động sản (Real state)	170	7.468	211.04568	9775	28.26
3	Công nghiệp (Industry)	237	5.568	182.0736	8331	32.70
4	Đầu tư (Investment)	177	4.792	152.48144	7837	31.82
5	Doanh nghiệp (Enterprise)	256	6.739	188.0181	11433	27.90

No.	Field	No. of document	No. of sentences	Word token†	Word type‡	Sentence length
6	Ngân hàng (Bank)	89	5.162	213.1906	7997	41.30
7	Nông nghiệp (Agriculture)	263	7.072	203.03712	10400	28.71
8	Tài chính (Finance)	374	7.348	209.71192	9048	28.54
9	Thương mại (Commerce)	276	5.976	190.81368	9116	31,93
	Total	**2.115**	**55.698**	**1.754.0659**	**36.775**	**31.49**

Statistic of the Vietnamese corpus in economic field

b) The corpus software

Below is the interface of the corpus software.



The software has several particular functions, including:

- Statistics function (Statistics). The statistical function is a fundamental function of the corpus. The functions of our software include word quantity statistics, POS quantity statistics, Mullti POS quantity statistics, entity quantity statistics, sentence statistics, and document quantity statistics.

- Extended full-text view function:

When a user wants to see the full text of a file, they just need to select the corresponding file, the full text of that file will be displayed on the software interface.

- Context index function (Concord). The context index function helps to list all occurrences of a word with its left and right context. This function applies to any word or a word with its type. When you want to see the whole source text of a particular line of context, click on that index line in the list, and the program will connect the user to the full text.

- Collocation function. The collocation function calculates the

† Word count (including word occurrence).

‡ Word count (not including word occurrence).

number of times words appear next to each other. Accordingly, if searching for co-occurrence of a word, that word will be called the center word, L1 is the word located to the left of the central word, L2 is the word located to the left of the L1 word, and L3 is the word located to the left of the L2 word, and so on. Similarly, we have R1, R2, R3… are the words located to the right from the center word. One center word statistically shows five other words on its left and five on its right.

4.2. Some applications of the Vietnamese corpus in economic field

The language in the corpus is authentic, so working with the corpus helps strengthen learners' and teachers' confidence in the knowledge they discover. In addition, with the advantage of its size and statistical functions, the corpus can provide information about both frequently occurring common and rarely used language cases in practice, thereby making suggestions for selecting appropriate teaching and learning content and methods. For example, based on frequency, a teacher can create a list of high-frequency vocabulary from the entire corpus or create lists for each specific field.

Below is a list of 20 economic terms we have taken from our corpus and sorted in descending order based on their occurrence and frequency.

No.	Term	Occurrence	Frequency[§]
1	Đầu tư (investment/to invest)	8145	2.33
2	Phát triển (development/to develop)	6439	2.43
3	Doanh nghiệp (enterprise)	5688	2.48
4	Thị trường (market)	5405	2.51
5	Dự án (project)	5185	2.52
6	Ngân hàng (bank)	4275	2.61
7	Sản xuất (produce)	4049	2.63
8	Ngành (field)	3916	2.65
9	Kinh tế (economic)	3654	2.68
10	Bảo hiểm xã hội (social insurance)	3614	2.68
11	Sản phẩm (product)	3427	2.70
12	Giá (price)	3426	2.70
13	Vốn (capital)	3189	2.70
14	Công ty (company)	3155	2.74
15	Bảo hiểm (insurance)	2977	2.77
16	Khách hàng (customer)	2803	2.79
17	Thương mại (commercer)	2635	2.82
18	Dịch vụ (service)	2600	2.82
19	Kinh doanh (business)	2431	2.85
20	Xuất khẩu (export)	2378	2.86

List of 20 economic terms in the corpus with high occurrence and frequency

[§] Calculated using the formula f = –lg (n/N), where n is the number of occurrences of a particular unit and N is the total number of units of the same type in the corpus. The smaller this number (minimum is 0), the more occurrences of the unit and vice versa.

With the bank sub-group, we make statistics by taking the first 100 words out of a total of nearly 8000 words in this sub-store, filtering out the wrong words and other common words, to finalise a list of terms related to the banking sector as follows.

No.	Term	Occurrence	Frequency
1	Ngân hàng (Bank)	2650	1.93
2	Phát triển (Development/Develop)	1055	2.30
3	Khách hàng (Customer)	920	2.36
4	Tín dụng (Credit)	908	2.37
5	Vay (loan/to loan)	904	2.37
6	Dịch vụ (service)	881	2.38
7	Hệ thống (system)	830	2.40
8	Vốn (sapital)	799	2.42
9	Nợ (debt/to owe)	735	2.46
10	Kinh tế (economic)	693	2.48
11	Thanh toán (payment/to pay)	620	2.53
12	Chính sách (policy)	581	2.56
13	Doanh nghiệp (enterprice)	542	2.59
14	Ngân hàng nhà nước (state bank)	530	2.60
15	Tài chính (finance)	479	2.64
16	Tiền (money)	474	2.65
17	Giao dịch (transaction/to exchange)	468	2.65
18	Ngân hàng thương mại (commercial bank)	366	2.76
19	Thẻ (card)	315	2.83
20	Tài sản (property)	308	2.84

List of 20 economic terms with high occurrence and frequency in the bank subgroup

We can also extract from the corpus lists of phrases as teaching materials based on the collocation function. For example, below is a list of 20 phrases with the centre word "insurance" and the actual word immediately following it.

1. bảo hiểm xã hội (social insurance)	11. bảo hiểm đầu tư (investment insurance)
2. bảo hiểm y tế (health Insurance)	12. bảo hiểm thực hiện (insurance to perform)
3. bảo hiểm bắt buộc (compulsory insurance)	13. bảo hiểm chấm dứt (insurance ends)
4. bảo hiểm tự nguyện (voluntary insurance)	14. bảo hiểm hưu trí (Pension Insurance)
5. bảo hiểm sức khỏe (health Insurance)	15. bảo hiểm tài sản (property insurance)
6. bảo hiểm xe cơ giới (motor vehicle insurance)	16. bảo hiểm thai sản (maternity insurance)
7. bảo hiểm tiền gửi (deposit insurance)	17. bảo hiểm nhân thọ (life insurance)
8. bảo hiểm việt nam (Vietnam insurance)	18. bảo hiểm phi nhân thọ (Non-life insurance)
9. bảo hiểm PIV (PIV insurance)	19.bảo hiểm FWD (FWD insurance)
10. bảo hiểm Bưu điện (postal insurance)	20. bảo hiểm thất nghiệp (Unemployment Insurance)

List of 20 collocations of 'insurance"

Besides exploiting and building a list of typical vocabulary and phrases to teach economic Vietnamese, the Vietnamese economic corpus can also be applied in the following fields:

- Studying language: providing materials to study economic press discourse such as title setting, article structure, or context, etc., research on the use of metaphor in the field of economic language, research on the characteristics of word combinations in economic press documents, research on the aspects of using names of organizations, places, titles, labels, and brands, etc.

- Dictionary compilation: With the Vietnamese economic corpus, we also see the potential of building a dictionary specifically for learners of economic Vietnamese. This dictionary is very useful for learners because it can localize the vocabulary that often appears in economic documents, helping learners focus on the economic field they are interested in and save time looking up. The compilation is also based on the frequency of words through statistical results. The examples to illustrate the entries are extracted from the context index function of the software itself, which is an advantage when compiling a dictionary of economic terms.

V. Conclusion

In many countries, corpus linguistics has made significant progress. Therefore, their research and other related jobs, such as teaching, dictionary compiling, or translation, are massive. Corpus linguistics has only been known in Vietnam for about two decades, so corpus research still needs to be improved. The article initially studies the construction of a specific corpus of economic Vietnamese, including nine sub-groups with an actual size of 55,698 sentences, equivalent to 1,754,0659 words. The corpus is entered, stored, and exploited by separate software. The software has some basic functions such as statistics of words, sentences, types of words, text, context index (Concord), and collocations. We also initially exploit vocabulary lists and word combinations as the first and typical units easily exploited for teaching economic Vietnamese from the corpus. Besides those applications, we also predict the potential of the corpus being exploited in a number of fields, such as language research or dictionary compilation. We hope that in Vietnam, corpus linguistics will be promoted to develop more strongly, contributing more to the teaching and research of languages in general and of Vietnamese in particular.

References:

In Vietnamese

[1]. Điền, Đ. (2018). *Ngôn ngữ học khối liệu.* Thành phố Hồ Chí Minh: NXB Đại học quốc gia Thành phố HCM.

[2]. Giáp, N. T. (2016). *Từ điển khái niệm Ngôn ngữ học.* Hà Nội: NXB Đại học Quốc gia, Hà Nội.

[3]. Thu, Đ. H. (2007). *Ngôn ngữ học khối liệu (Corpus).* Hà Nội: Số 7, Tạp chí Ngôn ngữ và đời sống.

In English

[4]. Carlos Assunção, Carla Araújo. (2019). *Entries on the history of corpus linguistic.* Sao Paulo.

[5]. Ramesh Krishnamurthy, Wolfgang Teubert (2007). *Introduction to Corpus Linguistics: Critical Concepts in Linguistics* https://www.researchgate.net/ publication/282649636_Introduction_to_Corpus_Linguistics_Critical_Concepts_in_Linguistics_6_volumes

[6]. Stefanowitsch, A. (2020). *Corpus linguistics A guide to the methodology.* Berlin: Language Science Press.

[7]. Tony McEnery, A. H. (2012). *Corpus Linguistics Method, Theory and Practice.* New York: Cambridge University Press.

Author address: National Economics University

Email: thuyngth@neu.edu.vn