CORPUS-BASED METHODS IN LINGUISTIC RESEARCH

*Nguyen Thi Hong Ha MA **

Abstract: *Corpus-based methods in linguistic research was initiated by Paul Baker in 2006 through the publication* 'Using Corpora in Discourse Analysis' *(London: Continuum). Since then, the method has been applied by many linguists in great variety of research, especially in making dictionaries, teaching languages, and comparing translations. With a number of advantages, the method enables researchers to quantify linguistic patterns, and to come to solid conclusions. Thanks to the technology development, user-friendly software has motivated the method to move rapidly, promising to bring many useful linguistic applications to our life.*

Key words: *corpus, method, research, linguistics, application.*

1. Introduction

As far as we know, quantitative research concentrate on how much or how many there is/are of a particular characteristic or item of a linguistic aspect. The strong point of this method is that it enables researchers to compare ralatively large numbers of things, people by applying a comparatively easy index. Quantative data use statistical methods for analysis, that is, particular methametics tools allow researchers to conduct on numeric data. *Corpus-based methods* are regarded as effective research methods in linguistics, and they *'should not be considered as only quantitative, but rather than an approach which can combine both qualitative and quantitative processes' (Paul Baker in Litosseliti, 2010: 93).*

This article mentions corpus-based methods in details, from basic concepts, their applications, to building and annotating corpora in linguistic research.

2. Theoretical Concepts of Corpus Linguistics

1.1. Definition

Corpus linguistics is a popular field of linguistics involving *'the analysis of very large collections of electronically stored texts aided by computer software'* (Paul Baker in Litosseliti, 2010: 93). The word *'corpus'* in Latin means *body*, so a corpus can be understood a 'body' of texts. According to Mc Enery and Wilson (1996:1), corpus linguistics features a *'methodology'* rather than a traditional branch of linguistics like semantics, grmammar, or phonetics.

Corpus linguistics with characteristics of empirical, inductive forms of analysis relies on instances of language in real-life use. The aims of corpus linguistics are to find out patterns, rules or explore trends about the ways people actually use their language in everyday life.

2.2. Advantages of Corpus Linguistics

There are several great advantages of corpus-based methods in linguistics as follow:

(i) Corpus-based methods enable researchers to restate or refect hypotheses about language use;

* Dean of English Faculty B,
Hanoi University of Business and Technology

(ii) Corpus-based methods allow researchers to bring out new questions and theories about language;

(iii) Corpus-based methods help researchers to quantify linguistic patterns, reaching more solid conclusions on language;

(iv) Corpus-based methods with large corpora lead researchers to obvious evidence of rare or unusual instances of language, and confirmation on very common phenomena of language.

2.3. Research questions of corpus-based methods

The overarching questions of corpus-based methods is 'how do people really use language', around which many research questions related to different fields in linguistics are raised. For example, in the field of language teaching it can be seen: *'Is the language used in textbooks actually reflected the language that people encounter in everyday life?'* (Mindt, 1996); or in study on language genres: '*Has writeen language become more informal over recent years?'* (Kennedy, 1998).

Moreover, the comparative trend also appears in research questions within studies on corpus-based linguistics, such as *'How does the use of linguistic feature X differ in usage between language varieties A and B in terms of frequency and/ or typical usage?'* or 'What associations are triggered by the use of linguistic item X, based on its typical uses?' Not only to discover similarities between the features of languages, but also do corpus-based methods help find out differences, in spite of small difference, or even no difference which is worth researching. In addition, corpus linguistic research approaches language patterns that people are unaware of, but they may still strongly influence users.

3. Types of corpora

Corpora in existence are divided into three main pairs (Baker, P. 2006 in Litosseliti, 2010), based on their characteristics and aims:

3.1. General corpora and specialized corpora

A general corpus aims to represent a particular language like the British National Corpus BNC or The Bank of English BoE., which is extremely large and takes a long time to build and annotate. More importantly, they are very useful resources for a wide range of research purposes. They play the role of *'benchmark'* about a typical language in comparison with a specilized corpus. Meanwile, *a specialized corpus* is much smaller and has more limited sets of texts in restrictions on time, genre or place/language variety. For instance, a specialized corpus of just newspapers published in October, 2019 in Vietnam. Speacilized corpora are generally easier to collect and for specific research questions.

3.2. Written corpora and spoken corpora

Written corpora contain computer-mediated texts such as e-mails, text messages or websites or mixture of all three while *spoken corpora* are usually smaller due to complexities surrounding, gathering and transcribing data. Written corpora with the access to the Internet are easier to build because most of the texts are already electronically coded. This type of corpora is expected to be increasingly popular when societies more frequently use electronic forms of communication.

3.3. Multilingual corpora and Parallel Corpora

The comparison of different languages has emerged as a growing area of corpus

linguistics. This area is useful in fields of language teaching, language testing, and translation. *A multilingual corpus* usually involves equal amounts of texts from a number of differents languages in the same genre. There is no need to translate directly from one language into another for such texts. Meanwhile, *a parallel corpus* is a special type of miltilingual corpus, in which texts are exact equivalents of each other. Parallel corpus are often carefully designed, sentence-aligned, that means tags are added to the corpus to identify which sentences are translations of each other. With the aid of the right software, researchers view translations of sentecnes side by side, and recognise the differences between translations and the original.

3.4. Learner Corpora

A learner corpus is produced by learners of that language, that is very useful for teachers to realize common errors at different stages of development, and indicate over- and underuses of vocabulary or grammar of learners in comparison with an equivalent corpus of native speakers. The Longman Learner Corpus and the International Corpus of Learner English both receive contributions from a great variety of learners all over the world, helping researchers find out the extent to which a student's first language is likely to influence the way they acquire English.

4. Applications of corpus-based methods in linguistics

4.1. Application in linguistic description

Corpus-based methods can aid researchers in making dictionary with real-life examples of words in use. Hunston (2002) researches the senses of the verb 'KNOW' shown in examples in three dictionaries, one of which did not use a corpus and the others used a corpus. These are the findings of the study (Hunston, 2002: 97):

- The 1987 Longman Dictionary of Contemporary English (without a corpus): 20 senses;
- The 1995 Longman Dictionary (with a corpus): over 40 senses;
- The 1995 COBUILD Dictionary (with a corpus): over 30 senses.

Obviously, corpus-base methods show their advantages of enhancing senses of meaning of words in dictionaries, reflecting the reality of diversely using language in everyday life.

4.2. Aplication in translation studies

Corpus-base methods, especially with the types of multilingual and parallel corpora show their strengths in comparative transaltion and interpreting studies. When conducting research on punctuation in Hans Christian Andersen's stories and in their translations into English, Malmkjer (1997) concludes that in translations, punctuation tends to be strenghtened, with commas aften being replaced with semicolons or full stops, and semicolons being transferred to full stops, too. This leads to long, complex sentences being divided into shorter and simpler clauses in translations to reduce the complexity of sentence structures in the original.

In another study, Mauranen (2000) reveals that translators usually make optional cohesive markers explicit in the translations, even though they are not available in the original text. This results in a tendency to spell things out rather than make them implicit.

4.3. Application in forensic linguistics

Coulthard (1993) carries out his study on witness statements used as evidence in

the trial of Derk Bentley (who was excuted in Britian in 1953 for his involvement in a policeman's death). He compares the frequencies of words in Bentley's statements with that in general written and spoken English, and that of other policemen and witnesses. He notes that Bentley had a higher frequency of using word 'then' than others. However, this word is a very typical feature of the police. This, together with other corpus-based eveidence, the researcher argues that Bentley aged 11 had not made his own statement, but it had been written for him.

4.4. Application in Critical Discourse Analysis CDA

In the area of CDA, Baker (2006) demonstrates how corpus-based methods can be used to express the '*incremental effect of discourse'*. He states that an association between two words, occuring many times in naturally occurring language is much better evidence for an underlying hegemonic discourse made explicit through the word combination than a single case. Furthermore, Mautner (2007) examines a corpus from a wide range of language sources to see how the elderly construct their discourse. The researcher recognizes that their discourse is expressed as ill-health victims who are in need of care more often than as empowered or independent citizens.

4.5. Application in language teaching

Corpora can also help language teaching be more effective. Mindt (1996) studies a corpus of spoken English and realized that native speakers use the model verb *'will'* most frequently for referring to future time. Yet, in German textbooks used for teaching English, 'will' was introduced to students in the middle of the second year when they had already learnt other modal verbs, which were less frequent in use in the corpus. Such studies are very useful for writing textbooks and designing teaching syllabus in language teaching.

4.6. Application in stylistics

Corpus methods of analysis have been used in stylistics in order to enhance systematicity and decrease subjectivity. For example, Malhberg (2009) examines stylistics in writing literary works by Charles Dickens and states that the writer often mentions the ways characters use household objects as a way of drawing readers' attention to the characters' emotions. Indicating a number of examples in the works related to objects like a watering-pot or a knife and fork, the researcher through her corpus-based analysis concludes that in Dickens' novels these objects are consistently used to emphasize characters' emotional states.

5. Corpus-based research tools and building corpora

Difference across space/genres (variation) and over time (change) is the most commonly applied in corpus- based research. According to Baker, among them frequency data are 'indicator of markedness' (Baker 2010: p.125); wordlists are the most basic 'points of entry' (Baker 2010: p.133) to analyse; keywords are 'somewhat more sophisticated' (Baker 2010: p.134) means of research and concordances with associated information are involved in collocates and clusters.

The most frequently used data include:

(i) Corpus, text and sentence (average word length);

(ii) Standardized type/ token ration (STTR) and standard deviation (SD);

(iii) Significance (p-value).

Theoretically, any text and collection of texts are considered to be a corpus and the

analysis can be conducted on the corpus of very short texts. According to McEnery and Wilson (1996), a corpus usually contains a sample *'maximally representative of the variety under examination'* is *'of a finte size'*, exists in *'machine readable'* and *'constitutes a standard reference for the language variety which it represents'* (McEnery and Wilson, 1996: 22-23). It is clear that a corpus must be large enough to show some feature about frequencies of linguistic phenomena, helping researchers to discover the common as well as unusual things in language.

Baker in Litosseliti (2010: p.95) suggests three criteria for identifying the size of a corpus: aspects of language, type of language, and pratical reasons. Kennedy (1998: 68) states that *'a corpus of 100,000 words will usually big enough for the sudy of prosody'*, and an analysis of verb-form morphology will need half a million words. Meanwhile, according to Biber (1993) confirms that a million words will be enough for a study on grammar. Also, the more various the language is, the larger the corpus requires. The British National Corpus involving a very wide range of written and spoken language genres and as a standard reference for British English, has the size of 100 million words while a corpus of weather forecast with restricted language just needs a much more smaller size. The practical conditions affecting the building a corpus could be the avalability of texts, the amount of money and time spent on the study and permission from authors.

The key theoretical concepts in corpus linguistic include sampling, balcance and representativeness. A corpus must be representative of a particular language variety, so the texts need choosing carefully to make sure that the corpus must be as a whole. Perhaps, it is not necessary to take the whole text, but parts of it if they are too long novels. Equal-sized samples from texts also need to be considered carefully to assure the balance. For instance, different parts (beginnings, middles, ends) of texts are equally sampled for the corpus. If texts are quite short and from one writer, the whole texts might be included rather than different sections.

Corpora are usually applied with the assistance of analysis software doing counting, sorting and presenting language characteristics. The current software such as *WordSmith Tools, Xaira, Wmatrix, and AntConc* can be employed in conjuction with a range of corpora. First, each text within a corpus is usually saved on a separate file containing a 'header' which has information about its author, date of publication, genre, etc. It allows researchers to concentrate on specific types of texts or to compare different types. Second, a corpus is annotated or tagged with further information for more complex calculations carried out on them. Sometimes, standard generalized mark-up language (SGML) where tags are presented in codes (elements) inside matching angle brackets (< >) is applied. Finally, certain features of the language variety are represented with other codes (entities) starting with an ampersand character (&) and ending in a semi-colon.

Hand-checking is often required although tagging can be automatically done by computer as tagging software is not always 100% accurate. Computer programs usually work best on texts having grammatically predictable sentences and relatively familiar words. Apart from checking, only human beings can interpret

the results of calculations from computer, no computer software can cover this work.

6. Conclusion

In summary, corpus-based methods are regared as potential to produce interesting findings about language, but as with many other methods, it is researchers' task to provide explantion for the findings. Corpus-based methods cannot cover all fields in linguistics, so it could be combined with other methods to maximize strengths and avoid limitations. Although it contains its weak points such as being time and money-consuming to build a corpus, a continuing need to update balanced reference corpora, researcher's ability to use computer fluently, identifying certain type of language patterns, it is worth applying to linguistic studies where comparisons must be done. The advantage of the corpus-based methods lies in employing fast and accurate techniques to discover patterns that human reserarchers would not recognize. Corpus analysis uses large amount of natural data, so a high degree of reliability and validity of linguistic research is usually achieved.

References

1. Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

2. Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh Sociolinguistics. Edinburgh: Edinburgh University Press.

3. Biber, D. (1993). 'Representativess in corpus design' *Literary and linguistics Computing* 8,4: 243-57.

4. Coulthard, M. (1993). 'On beginning the study of forensic texts: corpus concordance collocation'. In M.Hoey (ed.). *Data, Description, Discourse*. London: Harper Collins.

5. Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

6. Kennedy, G. (1998), *An Introduction to Corpus Linguistics*. London: Longman.

7. Litosseliti, L. (2010). *Research Methods in Linguistics*. Continuum.

8. Malhberg, M. (2009). 'Corpus stylistics and the Pickwickian watering-pot' in Baker (ed.). *Contemporary Approaches to Corpus Linguistics*. London: Continuum.

9. Malmkjer, K. (1997). 'Punctuation in Hans Christian Andersen's stories and in their translations into English', in F.Poyatos (ed.). *Nonverbal Communication and Translation. New Perspectives and Challenges in Literature, Interpretation and the Media*. Amsterdam and Philadelphia: Benjamins.

10. Mauranen, A. (2000). 'Strange strings in translated language: a study on corpora', in M. Olohan (ed.). *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*. Manchester: St. Jerome Publishing.

11. Mautner, G. (2007). 'Mining large corpora for social information: the case of elderly', *Language in Society*, 36 (1), 51-72.

12. Mindt, D. (1996). 'English corpus linguistics and the foreign language teaching syllabus', in J. Thomas and M. Shorty (eds), *Using Corpora for Language Research*. London: Longman, 232-247.

13. Mc Enery and Wilson (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

PHƯƠNG PHÁP DỰA VÀO KHỐI LIỆU TRONG NGHIÊN CỨU NGÔN NGỮ

ThS. Nguyễn Thị Hồng Hà *

Phương pháp dựa vào khối liệu trong nghiên cứu ngôn ngữ được học giả Paul Baker khởi xướng vào năm 2006 qua ấn bản 'Sử dụng Khối liệu trong Phân tích Diễn ngôn' (Nxb London: Comtinuum). Kể từ đó, phương pháp này được nhiều nhà ngôn ngữ học áp dụng trong các nghiên cứu của mình, đặc biệt trong việc biên soạn từ điển, dạy ngoại ngữ, và so sánh dịch thuật.

Phương pháp này có nhiều lợi thế, giúp nhà nghiên cứu định lượng hóa các mẫu ngôn ngữ, từ đó đi đến những kết luận đầy thuyết phục. Ứng dụng của phương pháp này trong nghiên cứu ngôn ngữ rất phong phú: từ việc mô tả ngôn ngữ trong làm từ điển, đến hỗ trợ dạy tiếng, ngôn ngữ pháp lý, phong cách học, đến nghiên cứu so sánh trong dịch thuật,…

Đây là một phương pháp mới trong nghiên cứu ngôn ngữ ở Việt Nam, nên tác giả muốn giới thiệu nó đến bạn đọc. Cùng với sự phát triển của công nghệ, các phần mềm hỗ trợ tính toán hiện đại đã làm cho phương pháp này ngày càng phát triển mạnh, mang lại nhiều ứng dụng thiết thực của ngôn ngữ cho đời sống đương đại.

Từ khóa: Phương pháp, nghiên cứu, khối liệu, ngôn ngữ học, ứng dụng.

Ngày nhận bài: 15/10/2019

* Chủ nhiệm Khoa Tiếng Anh B,
Trường ĐH KD&CN Hà Nội