УДК 811.111’33

T. A. Vetskur
Volyn Institute for Economics and Management, Lutsk
E. K. Koliada
Lesya Ukrainka Eastern European National University, Lutsk

CORPORA IN LINGUISTICS: THEORETICAL FRAMEWORK AND APPLICATION

Стаття присвячена корпусним методам дослідження мовного матеріалу, сфері їх застосування, основним перевагам та недолікам. У праці представлено визначення та опис основних понять корпусної лінгвістики, а також короткий огляд історії розвитку цієї лінгвістичної течії.

Ключові слова: *дані, корпус, методологія, дискурс, колокація, лексикографія.*

В статье рассматриваются корпусные методы изучения языкового материала, сфера их применения, основные преимущества и недостатки. В работе содержится определение и описание основных понятий корпусной лингвистики, а также краткий обзор истории развития этого лингвистического направления.

Ключевые слова: *данные, корпус, методология, дискурс, коллокация, лексикография.*

The paper deals with corpus studies, their methodology, sphere of application and main strengths and weaknesses. The main notions of corpus linguistics are defined and described. The paper also presents a brief outline of the corpus linguistics history and development.

Key words: *data, corpus, methodology, discourse, collocation, lexicography.*

Nowadays corpus methods are considered to be practically a universal tool for analyzing linguistic phenomena in various domains such as discourse analysis, cognitive linguistics, sociolinguistics, psycholinguistics etc. **The problem of corpus studies** was investigated in great detail by D. Biber [2], G. Leech [5], T. McEnery [6], T. Virtanen [7] and A. Wilson [6]. Still, there remain some areas which require close and careful consideration.

The present paper aims at revealing the core points of corpus studies and delimiting the spheres of corpora successful application. The above objective presupposes the **following tasks**:

1) to provide definitions of the main notions used by corpus linguistics;

2) to give some examples of its application when processing linguistic data;

3) to indicate corpus linguistics strengths and weaknesses with proper validation and justification.

The origins of corpus studies can be traced back to the mid-18th century when Samuel Johnson used a corpus of texts to gather authentic uses of words, which he then included as examples in his dictionary of English. The focus was made on the regulatory function of the study, not on its representative character. However, present-day corpus studies make a sharp contrast.

Geoffrey Leech believes that two pioneers of modern Corpus Linguistics are Randolph Quirk and Nelson Francis, and the key dates are 1959 (when Quirk started his Survey of English Usage) and 1962 (when Francis, aided by Henry Kucera, started to collect Brown Corpus) [5, p. 155]. These two scholars both hit on the idea of collecting a large body of texts (and transcriptions) wide-ranging enough to represent, to a reasonable extent, the contemporary English language. In this, they must have been considerably influenced by the American structuralist school of the 1940s and 1950s, which placed fundamental emphasis on the need for a corpus of any language to be investigated.

Further on, Geoffrey Leech suggests that there are two defining goals that made Quirk and Francis founding fathers of modern Corpus Linguistics:

1) someone giving an account of a language should aim at what Quirk called "total accountability", i.e. all relevant data obtainable should be taken into account, not just the examples that the investigator finds useful or congenial;

2) a corpus, complied in the spirit of offering total accountability, should be made available as a resource for the world of scholarship at large [5, p. 156].

Among a great number of corpus definitions we consider the following to be the most relevant and trustworthy. T McEnery and A. Wilson state that **a corpus** in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration [6, p. 32].

Spanish scholars stress that corpus linguistics using the computer as a tool is currently regarded as a methodology involving an empirical approach to language which started to gain importance in the 1970s [3, p. 11]. A corpus can be used as a source of data in syntax, lexis, pragmatics, sociolinguistics, psycholinguistics, and virtually every branch of language study [5, p. 157]. Corpora reduce the burden of evidence that is often placed on intuitions to show how particular grammatical and lexical choices are regularly made [4, p. 89].

The term corpus when used in the context of modern linguistics tends most frequently to have more specific connotations which may be considered under four main headings:

- Sampling and representativeness;
- Finite size;
- Machine-readable form;
- A standard reference.

It was Chomsky's criticism of early corpora that they would always be skewed: in other words, some utterances would be excluded because they are rare, other much more common utterances might be excluded simply by chance, and chance might also act so that some rare utterances were actually included in the corpus. Although modern computer technology means that nowadays much larger

corpora can be collected than these Chomsky was thinking about when he made these criticisms, his criticism about the potential skewedness of a corpus is an important and valid one which must be taken seriously. However, this does not mean abandoning the corpus analysis enterprise. Rather, consideration of Chomsky's criticism should be directed towards the establishment of ways in which a much less biased and more generally representative corpus may be constructed.

As well as sampling, the term "corpus" also tends to imply a body of a finite size, for example 1,000,000 words. This is not, however, universally so. At Birmingham University, for example, John Sinclair's COBUILD team have been engaged in the construction and analysis of a collection of texts known as a monitor corpus. A monitor corpus, which Sinclair's team often prefer to call simply "a collection of texts" rather than a "corpus", is an open-ended entity. Texts are constantly being added to it, so that it gets bigger and bigger as more samples are added. Monitor corpora are primarily of importance in lexicographic work, which is the main interest of the COBUILD group. They enable lexicographers to trawl a stream of new texts looking for the occurrence of new words or for changing meanings of old words. Their main advantages are:

- the age of texts, which is not static and means that very new texts can be included, unlike the synchronic "snapshot" provided by finite corpora;
- their scope, in that a larger and much broader sample of the language can be covered.

Today, corpus-based lexicographic investigations address six major types of research questions. These are:

1. What are the meanings associated with a particular word?
2. What is the frequency of a word relative to other related words?
3. What non-linguistic association patterns does a particular word have (e. g., to registers, historical periods, or dialects)?
4. What words commonly co-occur with a particular word, and what is the distribution of these "collocational" sequences across registers?

5. How are senses and uses of a word distributed?

6. How are seemingly synonymous words used and distributed in different ways [2, p. 23–24]?

Given the growing size of corpora, corpus analysis is increasingly quantitative, tending to focus on those features which are readily identifiable and countable for a computer. This generally means word-forms and their co-occurrences with other word-forms, n-grams (i.e. recurrent multi-word sequences) etc. These frequency counts can then be compared with those obtained for other words and co-occurrences, or with those for different corpora or sub-corpora, in order to identify statistically significant differences. We fully agree with the statement of G. Aston that a major strength of corpus analysis is that such counts can highlight patterns which may have eluded intuition [1, p. 7].

Nevertheless, this kind of investigation has its shortcomings. The main disadvantage of corpora is that they are not such a reliable source of quantitative (as opposed to qualitative) data about a language as they are constantly changing in size and are less rigorously sampled than finite corpora. With the exception of the monitor corpus observed, though, it should be noted that it is more often the case that a corpus has a finite number of words contained in it. At the beginning of a corpus-building project, the research plan will set out in detail how the language variety is to be sampled, and how many samples of how many words are to be collected so that a pre-defined grand total is arrived at.

With the Lancaster-Oslo/Bergen (LOB) corpus and the Brown corpus the grand total was 1,000,000 running words of text; with the British National Corpus (BNC) it was 100,000,000 running words. Unlike the monitor corpus, therefore, when such a corpus reaches the grand total of words, collection stops and the corpus is not thereafter increased in size. One exception to this is the London-Lund corpus, which was augmented in the mid-1970s by Sidney Greenbaum to cover a wider variety of genres.

The definition of a corpus we use clearly states the importance of machine-readable feature. Though, for many years, the term “corpus” could be used only in

reference to printed text. But now things have changed, so that this is perhaps the exception rather than the rule. One example of a corpus which is available in printed form is a Corpus of English Conversation (Starvik and Quirk 1980). This corpus represents the "original" London-Lund corpus (i.e. minus the additional examples of more formal speech added by Sidney Greenbaum in the 1970s). Although these texts are also available in machine-readable form within the London-Lund corpus, this work is notable as it is one of the very few corpora available in book format. The appearance of corpora in book form is likely to remain very rare, though the Spoken English Corpus has recently appeared in this format.

There is also a limited amount of other corpus data (excluding context-free frequency lists and so on, prepared from corpora) which is available in other media. A complete key-word-in-context concordance of the LOB corpus is available on microfiche and, with spoken corpora, copies of the actual recordings are sometimes available for, amongst other things, instrumental phonetic analysis: this is the case with the Lancaster/IBM Spoken English Corpus, but not with the London-Lund corpus.

Lexicology and lexicography are not the only domains where corpus approach can be successfully used. It can also yield significant results in specific dimensions of discourse singled out by T. Virtanen [7, p. 54-55]. Starting from (1) a "structural" dimension, present in much work on textuality, linguists can proceed to (2) a "content-based" dimension, typically opted for in rhetorically-oriented studies. The "cognitive" dimension (3) is omnipresent in studies of text and discourse, and it can thus be specifically foregrounded where expedient. The "interactional" dimension (4), originating in studies of spontaneous speech, cuts across much of the current discussion of discourse phenomena, highlighting the dynamism of discourse practices in both speech and writing. And the "socio-cultural" dimension (5), too, demands consideration of the reflexivity of text and discourse.

In the fifth dimension the focus is on situational and socio-cultural contexts in which people jointly engage and re-engage in social action through discourse, and in performances through which discourse takes shape; the concern is with ways of (co)-constructing such contexts and adapting to them, and of maintaining or altering them through discourse.

It is obvious that these five dimensions of discourse are not all equally accessible to users of corpus-linguistic methods. In view of the discussion of context in such investigations, corpus-linguistic approaches can be expected to focus predominantly on the structural aspects of discourse and the various content-based phenomena apparent in text and talk. In contrast, the interactional and socio-cultural dimensions of discourse lend themselves less well to corpus studies because what is examined here is the dynamism of discourse as social action.

Taking into consideration all of the above, we may summarize that corpus linguistics is broadly referred to as any linguistic framework which uses computer corpora as data and associated method of enquiry. It is highly relevant in lexicology, lexicography and some dimensions of discourse studies bent on quantitative analysis. The discovery of distributional patterns is the domain of corpus linguistics par excellence. Discourse linguists have benefited from corpus-linguistic methods to study variation across texts and discourses, including variation across time in historical linguistics. The usual text classifications include text/discourse types, genres, registers, styles and modes, while fictionality can also constitute a dividing line between text categories.

At the same time, corpus methods cannot be used in the cases when the context is of particular importance since corpora are the outcome of the processes of decontextualization and recontextualization of discourse. The corpus data are not the “original” or ”authentic” pieces of writing that they represent, nor they are studied in a communicative situation matching this of their writers or the expected readership.

References

1. Aston G. Applied Corpus Linguistics and the Learning Experience / G. Aston // Perspectives on Corpus Linguistics / ed. by V. Viana, S. Zyngier and G. Barnbrook. – Amsterdam : John Benjamins Publishing Company, 2011. – P. 1–16.
2. Biber D. Corpus Linguistics : Investigating Language Structure and Use / D. Biber, S. Conrad, R. Reppen. – Cambridge : Cambridge University Press, 2004. – 311 p.
3. Hornero A. M. Foreword / A. M. Hornero, M. J. Luzon, S. Murillo // Corpus Linguistics : Application for the Study of English. – Bern : Peter Lang, 2008. – P. 11–22.
4. Hyland K. A Convincing Argument : Corpus Analysis and Academic Persuasion / K. Hyland // Discourse in the Professions : Perspectives from Corpus Linguistics / ed. by U. Connor and T. A. Upton. – Amsterdam : John Benjamins Publishing Company, 2004. – P. 87–114.
5. Leech G. Principles and Applications of Corpus Linguistics / G. Leech // Perspectives on Corpus Linguistics / ed. by V. Viana, S. Zyngier and G. Barnbrook. – Amsterdam : John Benjamins Publishing Company, 2011. – P. 155–170.
6. McEnery T. Corpus Linguistics / T. McEnery, A. Wilson /2nd ed. – Edinburgh: Edinburgh University Press, 2004. – 247 p.
7. Virtanen T. Discourse Linguistics Meets Corpus Linguistics : Theoretical and Methodological Issues in the Troubled Relationship / T. Virtaten // Corpus Linguistics : Refinements and Reassessments / ed. by A. Renouf and A. Kehoe. – Amsterdam : Editions Rodopi B.V., 2009. – P. 49–66.