# CORPORA AND CORPUS LINGUISTIC APPROACHES TO STUDY BUSINESS LANGUAGE

Nafruza Esanboyevna Azizova
Samarkand State Foreign Language Institute

**Abstract**: This article deals with the corpora and corpus linguistic approaches to study business language. The term corpus is understood more specifically as a compilation of naturally-occurring texts stored electronically and available for quantitative and qualitative analysis

**Keywords:** Corpus linguistics, corpora, Sketch Engine, Corporate Social Responsibility(CSR), British National Corpus (BNC), English for Specific Purposes (ESP), English for Academic Purposes (EAP).

Corpus linguistics is primarily concerned with studying language on the basis of large collections of real-life linguistic data normally known as corpora. The term corpus has been used in linguistics generally for some time to describe a sample of a language or language variety. However, in corpus linguistics, the term is understood more specifically as a compilation of naturally-occurring texts stored electronically and available for quantitative and qualitative analysis (McEnery and Hardie 2012) [2, 208]. The development of corpus linguistics has been largely fuelled by advances in computer technology and the availability of linguistic software that allows linguists to search through corpora rapidly and reliably. Insights derived in this way have significantly increased our understanding of language use by providing empirical evidence for the existence of regularities and patterns that are not immediately visible to the naked eye or simply defy linguists' intuition. As John Sinclair, the father of corpus linguistics, pointedly remarked: "The language looks rather different when you look at a lot of it at once" [6, 100].

Most work in corpus linguistics, particularly in the early stages, was concerned with the development and study of large reference corpora of national varieties including spoken and written registers. The British National Corpus (BNC), which contains 100 million words of English, is a good example of such large compilations, sometimes also referred to as mega-corpora [1, 131-149]. Currently, in the era of Big Data, we are in a phase of giga-corpora with compilations reaching billions of tokens such as the enTenTen12 corpus of English (4.65 billion tokens) available on Sketch Engine, a web-based corpus linguistic software (6. 103). Yet, while large reference corpora have proven invaluable in linguistic analyses, they are less suitable for investigating language use in specific professional domains.

This is largely due to the fact that they are compiled with a view to being a representative sample of a language, and their composition is carefully balanced to include texts and genres that are seen as important in a particular culture [6, 115]. Hence, text types or genres that are typical only of a specific domain are likely to be excluded from such compilations. Also, certain texts may not be included because they belong to the category of occluded genres that are, for example, difficult to access due to confidentiality or intellectual property concerns. Hence, researchers who are interested in studying language patterns in specific contexts are not much helped by large reference corpora.

It is for this reason that, since the mid-1990s, many scholars have started using the tools and techniques of corpus linguistics to build and interrogate smaller specialised corpora focusing on selected genres, registers and domains. The impetus for this kind of work, perhaps not surprisingly, came from the fields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP), where, given the growing importance of English in professional and academic communication globally, there was an urgent need to create teaching resources (Tribble 1997, 2002; Flowerdew 1998; Ghadessy, Henry, and Roseberry 2001) [ 4, 309]. It is within this realm that the first corpora of business language were created, leading to corpus-based research on language use in a variety of business genres, registers and domains.

Before discussing the main corpus resources of business language and the ways in which corpus-based research has enhanced our understanding of business communication, we outline here the key analytical tools and procedures commonly adopted in corpus research. These include: frequency, concordancing, collocation, keyword, cluster, and corpus annotation. To demonstrate the benefits and also limitations of these analytical tools, examples will be drawn from CORES, a one-million-word corpus of Corporate Social Responsibility (CSR) reports obtained from ten major oil companies and published over the last five years. CSR reports form part of corporate disclosure, and are the most public and visible documents offering insights into organizations actions and goals in relation to their role in society and their stakeholders.

Corpus-based research often begins with frequency counts. In the simplest terms, frequency can be defined as the number of times an item occurs in a corpus, where “item” can include a word, word form, part of speech (if a corpus was annotated accordingly), keyword or cluster [7, 23]. Frequency lists are useful tools in determining the focus of a given data set and pointing to features that are typical of a particular genre, register or domain. For example, Table 1 shows a frequency list of the top 12 words from CORES and two other corpora, the British National Corpus (BNC) and the British Academic Written English (BAWE) 2, which contains over 6

million words of academic assignments collected at British universities in various disciplines.

Table 1

The 12 most frequent words in BNC, BAWE and CORES

| BNC | | | BAWE | | | CORES | | |
|---|---|---|---|---|---|---|---|---|
| Word | Raw Freq. | Norm. Freq. | Word | Raw Freq. | Norm. Freq. | Word | Raw Freq. | Norm. Freq. |
| THE | 6,055,105 | 630.4 | THE | 491,471 | 705.3 | THE | 42,832 | 491.1 |
| OF | 3,049,564 | 317.5 | OF | 270,136 | 387.7 | OF | 30,462 | 349.3 |
| AND | 2,624,341 | 273.2 | AND | 207,623 | 298.0 | AND | 27,974 | 320.7 |
| TO | 2,599,505 | 270.6 | TO | 188,666 | 270.8 | IN | 21,284 | 244.0 |
| A | 2,181,592 | 227.1 | IN | 137,911 | 197.9 | TO | 18,749 | 215.0 |
| IN | 1,946,021 | 202.6 | A | 125,736 | 180.4 | A | 9,513 | 109.1 |
| THAT | 1,052,259 | 109.6 | I S | 110,721 | 158.9 | FOR | 8,813 | 101.0 |
| IS | 974,293 | 101.4 | THAT | 78,781 | 113.1 | OUR | 6,838 | 78.4 |
| IT | 922,687 | 96.1 | AS | 62,128 | 89.2 | IS | 6,488 | 74.4 |
| FOR | 880,848 | 91.7 | BE | 58,053 | 83.3 | WE | 5,652 | 64.8 |

Considering these lists, we can see that the most frequent items in all three corpora are function words, which is not surprising given that in most languages function or grammatical words are used with high frequency. Despite the similarities, these three short lists also reveal a number of differences, for example, in the use of personal pronouns. Whereas I is a very frequent item in general English (it occurs 76.3 times per ten thousand words), in CORES it is very infrequent with only three occurrences per ten thousand words. By contrast, the plural pronoun we figures among the top ten most frequent words in CORES, just behind a related pronominal reference our. Finally, personal pronouns are completely absent from the BAWE top ten and occur much lower on the frequency list, pointing to the impersonal style of academic writing. The use of pronouns can signal personal meanings and identity relations and thus hint at specific aspects of a genre or discourse type that are considered worth investigating further.

The very frequent use of “we” in CSR reporting places the genre firmly within business discourse, of which the use of this self-reference seems to be a typical

feature [7, 94]. However, depending on purpose, we might convey different pragmatic and discursive meanings and these could be further investigated by conducting concordance analysis. Concordances are lists with lines that display all occurrences of a search term, often referred to as a node. The node is normally positioned in the middle with a few words to the left and to the right. Such a display is known as a KWIC, short for Key Word in Context. Table 2 provides an example of a concordance for the pronoun 'we' from CORES. In contrast to frequency lists, which display items in isolation, a concordance analysis enables the researcher to discover lexico-grammatical patterns that offer clues to the uses and meanings of the search term in context.

Table 2

A random sample of 10 concordance lines for 'we' from CORES

| | | |
|---|---|---|
| energy mix. In looking at these pathways | we | seek to identify forms of energy that can |
| providing energy security. In transport, | we | believe that making car engines much more |
| costs and technology. In power and heat, | we | believe an effective pathway would create |
| production as well as meeting new demand. | We | believe that the oil industry will be required |
| improvements in SAGD technology, and | we | will incorporate technologies and operating |
| technologies, and it is these skills that | we | will bring to our projects. Examples where |
| to use carbon capture and storage (CCS)? | We | recognize that CCS could be a longer- term |
| to Trinidad and Angola to Azerbaijan. | We | will be following the same principles |

The above concordance shows that, in the context of CSR discourse, we functions as the corporate we. As such, it is used to represent an organisation to the outside world as a unified whole, despite internal hierarchies and sometimes conflicting interests or views. Examination of the verbs to the right of the pronoun reveals that most of them belong to the category of mental verbs, such as believe or recognise, which describe human experiences, perceptions and attitudes and do not necessarily involve volition or action [4, 314]. Action verbs do occur to the right as well, for example, incorporate or produce, but mostly in the infinitive form following the marker of future tense "will" or another mental verb such as aim or strive. This short concordance already points to some interesting discursive features of CSR reporting. A personified corporate voice expresses a unified commitment to the improvement of social and environmental concerns, not in terms of concrete evidence or immediate action but as a broad intention to "do good" in future [5, 2011].

The first electronic compilation that was used for corpus-based research on business language is the Business English Corpus (BEC) created by Mike Nelson [3, 61]. This corpus consists of just over 1 million tokens with almost an equal proportion of spoken and written registers. It includes a variety of business genres, broadly divided into texts used for doing business and texts that talk about business [3, 69]. The former are texts produced by businesses for internal and external

purposes including annual reports, minutes and negotiations, whereas the latter are samples from business journalism and business education. The full composition of the corpus can be seen on the author's webpage at: http://users.utu.fi/micnel/BEC/bec2/corpus-makeup.htm.

There are also some samples of business language available as part of larger English corpora. For example, the written part of the BNC includes over seven million words of texts from the field of commerce and finance, while the spoken part contains a 1.3-million-word sample of spoken business communication covering sales demonstrations and business meetings. Also, the International Corpus of English (ICE) includes a 20,000 word sample of spoken business transactions.

One of the pioneering corpus studies on business language was conducted by Nelson [3, 55]. His main aim was to identify lexical features of Business English (BE) and establish how they differ from those of General English (GE). Nelson's research, which was based on the Business English Corpus (BEC) described above, reveals that BE shares most of the lexical features with GE, especially when measured by raw frequencies. At the same time, BE contains a number of unusually frequent words – keywords that constitute BE as a distinctive lexical variety. These include items such as "company", "market" and "customer". Moreover, Nelson's research shows that BE includes word combinations that often display prosodies unique to this variety. Interestingly, most of them seem to have positive associations focusing on success, dynamism and action. For example, the term manager shows a tendency to be linked with positive items such as excellent and forthright. On the basis of his research, Nelson suggests that business English teaching materials should focus not only on single lexical items but also on unique word combinations and prosodies. This might ensure that students are exposed to language which is more likely to accurately reflect the actual business world.

As much of the discussion above demonstrates, the use of corpus tools and methods has significantly enhanced our understanding of business language. It is true that the number of studies using corpus approaches to business communication is, if compared with other domains, relatively small. Yet research in this area has examined a wide range of linguistic phenomena, over a variety of genres, registers and contexts. The use of quantitative techniques (e.g., frequency, collocation, keywords or clusters) to study large amounts of business language has provided significant, empirically founded insights into the frequent and distinctive features of the variety, be these lexico-grammatical or pragmatic. It has also pinpointed the characteristics that business language shares with other varieties, thus bringing into focus its multidimensionality.

The purpose of corpora and corpus linguistic approaches to studying business language. The term corpus is understood more specifically as a compilation of

naturally-occurring texts stored electronically and available for quantitative and qualitative analysis (McEnery and Hardie 2012). We discussed the main corpus resources of business language and the ways in which corpus-based research has enhanced our understanding of business communication, we outline here the key analytical tools and procedures commonly adopted in corpus research. These include: frequency, concordancing, collocation, keyword, cluster, and corpus annotation.

Corpus studies on business style of the Uzbek language have not been conducted so far, because we are lack of a basis of large collections of real-life linguistic data normally known as corpora. However, the project "Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities" co-funded by the Erasmus+ Programme of the European Union which lasts from 2017 to 2020 seeks to pursue the objectives of improving the level of competences and skills of the involved higher education institutions by developing new and innovative education programmes in Uzbekistan. Due to this project we hope Uzbek National Corpus will be created in the near future, then we will be able to conduct corpus studies on business style of the Uzbek language.

## References

1. Tribble, Christopher. 2002. Corpora and corpus analysis: New windows on academic writing. In John Flowerdew (ed.), Academic discourse, 131 – 149. London: Longman.

2. McEnery, Tony & Andrew Hardie. 2012. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.

3. Nelson, Mike. 2000. A corpus-based study of business English and business English teaching materials. Ph.D. dissertation, University of Manchester.

4. Handbook of Business Communication by Gerlinde Mautner and Franz Rainer (Eds.) 2017 Walter de Gruyter Inc., Boston/Berlin – p. 713

5. Biber, Douglas & Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In Hilde Hasselgard & Signe Oksefjell (eds.), Out of corpora: Studies in honor of Stig Johansson, 181 –189. Amsterdam: Rodopi. (cf. Lischinsky 2011).

6. Sinclair, John. 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

7. Handford, Michael. 2010. The language of business meetings. Cambridge: Cambridge University Press. (Handford 2010; Lischinsky 2011).