EXCELLENCIA
INTERNATIONAL MULTI-DISCIPLINARY JOURNAL OF EDUCATION
Volume 02, Issue 12, 2024 ISSN (E): 2994-9521

The History of British National Corpus

Ataboyev Nozimjon Bobojon o'g'li [1]

[1] Doctor of Philology (DSc), Bukhara State University

Astanova Gulnora Maqsudovna [2]

[2] Master student of Bukhara State University

Annotation:

Lexical units of one language do not always have exact analogues in another language. Therefore, to solve this problem, the translator often uses translation transformations (conversions). Their skillful use ensures the adequacy of the translation: the translated text accurately reflects the content of the source text1.

Keywords: Text corpus, corpus linguistics, methods and mathematical statistics, machine translation, linguistic phenomena, stress and intonation

The characteristic “universal” usually refers to the main tool of science and implies a different set of properties for different fields of knowledge. In corpus linguistics, the characteristics “universal” and “representative” are synonymous. Thus, the main goal that linguists set in the 80s was to create a corpus that would represent a given language at a certain stage (or stages) of existence in all the variety of genres, styles, territorial and social variants, etc. We noted above that the BNC became the first corpus in this category. The word “national” in the title of the “British National Corpus”, intended to distinguish a variant of the language described by the corpus, is used to denote a representative corpus of any language and denotes a whole class of text corpora in this science.

Two important features of the national corpus should be highlighted.

First, it is characterised by representativeness, which implies a balanced composition of texts. This means that the corpus represents practically all types of written and spoken texts in a given language (fiction texts of different genres, journalistic, academic, scientific, business, colloquial, dialectal, etc.) and that all these texts are included in proportion to their share in the language of the corresponding period.

It should be borne in mind that a high degree of representativeness is achieved only with a significant volume of the corpus (tens and hundreds of millions of word uses).

Secondly, the corpus contains special, additional information about the properties of the texts included in it, which is called markup, or annotation. This property is inherent not only in national corpora and it allows to distinguish the corpus of texts from electronic libraries. However, we should not forget that libraries are created for reading, while text corpora are created for linguistic research.

The British National Corpus is a collection of written and spoken word samples taken from a wide variety of sources. The total number of words in the collection is approximately 100 million. The main purpose of the corpus is to represent the British version of English (both written and spoken) at the end of the 20th century. Official website is http://www.natcorp.ox.ac.uk/.

A monolingual corpus

The BNC is monolingual in that it only deals with English, specifically British (UK) English. The corpus only contains texts written by people who were born or adopted in the United Kingdom. In the case of spoken material, respondents who choose to record their conversations required to be fluent in English.

Other languages spoken in the UK, such as Scottish Gaelic and Welsh, have legitimate claims to be included in a British corpus, but the project partners' aim was to focus only on English.

A general corpus

It encompasses a wide range of genres and styles and is not restricted to any one topic, genre, or register. It offers instances of both spoken and written language in particular.

The BNC has a diverse range of samples of language in use, sufficient to support the claim that it is representative of modern British English. From Mills and Boon to Iris Murdoch, it contains fictional and non-fiction publications.

It includes biographies, scientific and academic expositions, essays written by high school and university students, electronic mail from a football supporters' mailing list and booklets about restaurants, patenting, tourism attractions, driving recommendations; package vacations, social work and plane-spotting publications; transcriptions of council meetings, business meetings, parliamentary debates, school classes, television news broadcasts and radio phone-ins, as well as daily conversations of a company director, a nurse, students, an aircraft engineer, a courier, and a machine operator, national and local newspapers from the Belfast Telegraph to the Daily Mirror.

A synchronic corpus

Synchronic research examines a language at a single point in time, without consideration for the historical events that led to its current state. The BNC reflects this trend: it deals with present British English, intending to present a snapshot of how English was used by British people from 1975 to 1994. It will gain diachronic importance with time, especially if other corpora of similar composition will be created.

A sample corpus

The majority of the books chosen for inclusion in the BNC are represented by samples rather than the entire text. The use of samples allows a wider range of texts to be represented: about three times as many books can be included if samples of 40-45,000 words are used instead of the average 120,000-word full work.

Another benefit of employing samples is that publishers and authors do not have to worry about piracy: many believed that making entire texts in the corpus available to the public would allow for

unlawful re-use, but utilizing much smaller samples instead allayed most of the concerns. Samples were gathered from the beginning, middle and end of each book. Newspapers, magazines and journals with various authors were often included in their entirety. There was no need for sampling in the spoken corpus and full transcriptions were used throughout.

Creation of the British National Corpus

The BNC was created via the collaborative efforts of a broad number of people, including organizations and individuals. It was created between 1991 and 1994 by a consortium which comprised three commercial partners – Oxford University Press (OUP), Longman Group UK Ltd and Chambers Harrap – and two academic ones – Oxford University Computing Services (OUCS) and the Unit for Computer Research in the English Language (UCREL) at Lancaster University.

Creation was divided into two stages: planning (design stage) and execution (creation stage), which are further discussed below.

Planning/Design stage

The project began with a detailed planning step in which the design concepts for the corpus were developed. These were utilized to create a set of selection criteria that were then used to find appropriate texts for inclusion in the corpus. A vast number of classification elements were found for the texts in the corpus, in addition to the selection criteria for the written and spoken components.

Selection Criteria: Written texts

Texts were chosen for inclusion in the corpus based on three separate criteria: domain, period/time and medium. The following are the target percentage for each of these characteristics.

Domain

The domain of a text denotes the type of writing contained within it.

- 75 percent of the written materials were to be informative writings, with nearly equal amounts from applied sciences, arts, belief & idea, trade & finance, leisure, natural & pure science, social science, and international affairs.
- Twenty-five percent of the written materials were to be imaginative, i.e. literary and artistic works.

Medium

A text's medium identifies the type of publication in which it appears. The categorization system employed is extremely wide.

- Books were to account for 60% of all written writings.
- Periodicals were to account for 25% of the total (newspapers etc.)
- Other types of miscellaneous published content should account for 5 to 10% of the total (brochures, advertising leaflets, etc.)
- Unpublished written content, such as personal letters and diaries, essays and memos, etc., should account for 5 to 10% of the total.
- A tiny percentage (less than 5%) should come from information written to be spoken texts (for example, political speeches, play texts, broadcast scripts, etc.)

Time

The date of publication of a work is referred to as the time criterion.

- The BNC, as a synchronic corpus, should contain texts from the same time period. The goal was for no text to date back further than 1975.
- Because of their continuous popularity and consequent effect on the language, this condition was eased only for imaginative works, a few of which date back to 1964.

Classification features: Written texts

A huge number of categorization factors were identified for the texts in the corpus in addition to the selection criteria. The objective was to ensure that there was an appropriate level of diversity within each criterion, although no fixed proportions were set for these elements. The following are some of the classification criteria:

- Size of the sample (number of words) and its scope (start and end points)
- Author's name, age, gender, location of origin and domicile
- Topic or subject of the text
- Age and gender of the target audience
- Level of writing: the more literary or technical a book is, the "higher" its level (a subjective measure of reading difficulty).

Designing the Spoken Component

The 10-million-word spoken corpus is divided into two sections: a demographic section that contains transcriptions of spontaneous natural conversations conducted by members of the public and a context-governed section that contains transcriptions of recordings made at specific types of meetings and events.

When available, information on the participants was recorded, including age, gender, accent, and occupation.

References:

1. Baker P., Hardie A., McEnery T. Glossary of Corpus Linguistics. –Edinburg: Edinburgh University Press, 2006. 192 p
2. Cruden, A. *A Complete Concordance to the Holy Scriptures of Old and New Testament*. - London: Fleming H. Revell Company,1737. 756 p.
3. Davies M. Corpora: an introduction // The Cambridge handbook of Corpus Linguistics / ed. by D. Biber, R. Reppen. Cambridge University Press, 2015. P. 11–31.
4. Flowerdew L. The argument for using English specialized corpora to understand academic and professional language // Discourse in professions: perspectives from Corpus Linguistics / ed. by U. Connor, T. Upton. 2004. P. 11–33
5. Kennedy G. An Introduction to Corpus linguistics. –London and New York: Addison Wesley Longman limited, 1998. 315 p.