BOWN CORPS AND FIRST GENERATION BRITISH CORPS: COMPARING THEM

Otajonov Sunnatillo Baxtiyorovich

University of Exact and Social Sciences

Annotation: *To learn and analyze the English language linguistically, one must thoroughly examine and analyze English corpora. Similar language corpora, such as Brown's, were familiar with all the features of language, including some of the innovations and scientific research in its study. Details regarding the construction of the first and second generations' buildings are given.*

Keywords: *language corpus, brown corpus, tags, Programming, word frequency, corpus versions.*

The challenge addressed by English linguistics is dealing with language corpora that were initially developed with flawless language acquisition in mind. However, the field of corpus linguistics has just recently gained popularity in science. It is widely acknowledged that two linguists, Henry Kucera and Nelson Francis, at Brown University, developed the first computerized corpus in contemporary times between 1961 and 1964. That example, they made statistical data available to the world when they released their seminal work "Computational analysis of Present-Day American English." Kucera and Francis put up a huge and diversified oeuvre for different computerized or programmed investigations that included parts of sociology, psychology, statistics, and languages. Houghton-Mifflin, a publisher in Boston, contacted Kucera shortly after the initial lexicostatistical analysis was published in order to obtain a database including one million words and three lines of citations for their upcoming American Heritage Dictionary. When it was released in 1969, this new lexicon was the first to be assembled using word frequency data gathered from corpus linguistics [1-2].

Brown's first collection only had words and a list of where each word should be placed. Sentence break tags were used for a while after that. The tagging tool by Greene and Rubin helped a lot with this, but it needed a lot of help.

A review of the books on the subject. The brown corpus was made up of about 80 sentence fragments. It used special indexes for compound sentences, abbreviations, foreign words, and other situations. It could be used as an example for many later corpora, like the Lancaster-Oslo-Bergen corpus. As Andrew McKee programmed and explained in English grammar books, corpus labeling made it possible to do much more complicated statistical analysis [3].

One interesting finding is that even for very large groups, putting words in descending order of how often they appear shows an exaggeration: the frequency of the nth most common word is about 1/N. That's why "the" makes up about 7% of Brown's corpus and "To" and "Of" make up more than 3%. About half of the 50,000 words are hapax legomena, which are words that only show up once in the corpus. The Zpha law is the name for this simple link between power and frequency that was written down by George Kingsley for a very wide range of events.

Although Brown was a pioneer in the area of corpus linguistics, the usual corpora that are used now (such as the Corpus of Modern American English, the British National Corpus, or the International Corpus of English) are significantly bigger in size and consist of roughly 100 million words.

To argue that Brown's corpus was crucial in the development of current corpus science is not an exaggeration. Not only does it serve as a model for New English, but it also acts as a model for all modern national businesses, and it is still often utilized as a data set in a variety of research. I have previously said that the whole language corpus, which is comprised of works that were published in the United States between the years 1961 and 1964, is approximately one million words in length and contains more than five hundred samples of the English language [4-6].

The approach of research... After seven years in usage, a new edition was published in 1971. This edition includes information about the text that was mostly published during that time period. Not only does the present edition contain information about later versions of the corpus, but it also includes information about the "marked" text that was finished at Brown University in 1979. This makes the current edition more comprehensive. Errors in the preparation of the original tape, which were corrected in the newly published copies, and subsequent typographical errors and anomalies in the main text, which were noted in the description of individual samples on pages 33-176 (other cases) were the two types of corrections that were made as a result of two full readings of the corpus.

Scholars are linguists who have labored to construct alternate versions of the Herald corpus. Some of the scholars who have contributed to this endeavor are M. Rubin, Barbara Green Levin, Sandra Pierce, Patricia Strauss, Stephen Ritz, Andrew Mackie, Jostein Haug, and Donald Sherman, among individuals. There were more than 160 copies of the corpus in circulation at the time that this article was written, and a current bibliography of published works that used or referenced the corpus comprises 57 items.

There are 1,014,312 words of text that are included in this standard corpus of modern American English. These words were taken from the edited English edition that was released in the United States in the year 1961. There is little question that part of the content was written earlier than 1961, despite the fact that it was completely published for the first time in 1961. Nevertheless, there is no content that has been provided that is known to be a reprint or a second edition of the text that was previously included [7].

In all, there are 500 examples, each consisting of more than two thousand words. Each sample starts at the beginning of a sentence, but it does not always begin at the beginning of a paragraph or any other bigger portion, and each sample, after 2,000 words, concludes at the conclusion of the first phrase. The examples consist of a diverse

assortment of different sorts and styles of writing. In contrast to prose, stanza is not included since several language issues prevent it from exist. (However, the brief poems that were included among the samples of prose have been preserved.) A good illustration of this would be the exclusion of drama because it is a fictitious replica of spoken language rather than genuine composed language. Samples that had more than fifty percent conversation were not accepted, although fiction was included in the collection. The selection of the specimens was not based on a subjectively defined level of perfection, but rather on the quality of their representativeness. It is not the case that the inclusion of the term "standard" in the title of a corpus in any way suggests that it is being marketed as "standard English." He only expects that the corpus will be utilized for comparison studies in which it is essential to use the same data. Due to the fact that data preparation and entering constituted a big challenge in computer operation, the objective was to supply content that had been meticulously picked and prepared, and that was of a substantial size, in a format that was standardized. Utilizing the corpus as a standard for developing a template for the purpose of producing and presenting extra material in either English or another language is possible [7-11].

The procedure is broken down into two stages: first, there is an initial subjective categorization and choice regarding the number of samples to use in each category; second, there is a random selection of the actual samples that are included in each category. Both the Providence Athenaeum Repository and the Brown University Library were considered to be a universe in which a random selection was made for a number of different categories. On the other hand, this meant breaking out of those two divisions for some categories. For the daily press, a list of American newspapers was utilized (with the addition of the Providence Journal), including the New York Public Library, which was in possession of microfilm files. In essence, certain material categories needed random selections to be made. For example, publications that fell under the "skills and hobbies" and "social sciences" categories were chosen from the inventory of one of the most extensive used magazine stores in New York City.

During a symposium that took place at Brown University in February of 1963, a list of the primary categories and the subgroups that fall under them was prepared. Additionally, participants at the conference voiced their thoughts on the amount of samples that were included in each category in an independent manner. The first set of numbers that were utilized was obtained by taking the average of these figures. In the aftermath of the contests, a number of adjustments were made in response to the knowledge obtained from the experience. Taking into account the relative amount of the actual output in 1961, the optimal division was determined.

Following the establishment of these categories, subcategories, and the total number of samples, the selection of the actual samples was carried out by means of a variety of random techniques, the most prominent of which was the use of a table of random numbers that was applied to the entire list of publications that were already in existence in the pertinent field. An further selection is made using a table of random numbers to determine the page from which the sample starts. Every sample starts with the very first full sentence that is found on the page that has been chosen. A number of elements, including titles and captions, footnotes, tables, and picture captions, have been done away with. It was estimated that there were around two thousand words, and the sample was finished during the subsequent sentence break. A word was defined in the final encoding (which will be explained further below) as any string of letters that included spaces on both sides, with the exception of the start-of-paragraph encoding signs. This was done within the context of these computational reasons. With the exception of some abbreviations, a sentence starts with a capital letter and ends with a symbol (.! or?), which is then followed by a space and another capital letter. There is a possibility that there is no space after the final character of the phrase in certain instances; the computer was used to make calculations that were higher in accuracy.

The copyright holder has granted permission for all copyrighted materials. The copyright permission information is contained in the distinct sample listing on pages 33-176.

The case is available in six different versions. Although they all contain the same fundamental text, they differ in terms of typography and format.

1. Form A. This is the original body shape as it was manufactured in 1963-64. At the time, the limitations of computer printing tools required him to use the detailed coding procedure described in Section 3 below.

2. Form B. This was the "drawn" version, from which all punctuation and codes were removed except hyphens, apostrophes, and formula symbols and ellipses. It is especially useful for those interested in individual words, and is also used to construct frequency tables in Coachman and Francis, A Computational Analysis of Modern American English (Providence: Broo's Univ. Press, 1967).

3. Form C. This is a "marked" version that uses partially deleted text, retaining only the initials of the corresponding name and grammatically significant punctuation marks. Each individual word (token) in this version has 81 lists of grammar tags, each of which defines a specific class of words.

4. Bergen Form I. This version and the following humanistic studies were prepared at the Norwegian Computing Center (EDB-center NAVF for humanistisk forskning) at the University of Bergen under the direction of Dr. Jostein Hauge. Both have upper- and lower-case letters, simple punctuation and a minimum of special codes. This version stores typographic data and uses the same division as the original, except that words at the end of a line are never divided.

5. Bergen Form II. This version reduces the typography slightly and uses a new longer line. This version taken from EDB-senter (Harald Haarfagresgt. 31, University of Bergen, N-5007 Bergen, Norway) is available in mikrofix along with full MOSC compliance.

6. Brown MARC form. This version was made at Stanford University. It is designed to fit two commonly used research methods suitable for large text corpora:

1. Search and retrieve full quotations from sentences using one word or word + context as search criteria;

2. Generate KOIC-form matches that can be organized according to different keyword orders and preceding or following verbal context.

Simply put, Brown's corpus and other first-generation corpora have a significant impact on the advancement of computational linguistics and natural language processing. They have been instrumental in providing researchers with a vast amount of data to analyze, leading to the development of numerous essential concepts and techniques that continue to be widely utilized in the field [11].

REFERENCES

1. Francis, W. Nelson & Henry Kucera. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press. 1967.

2. Biber, D. Nutq Va Yozish Bo‘Yicha O‘Zgarishlar . Kembrij: Kembrij Universiteti Nashriyoti, 1988 Yil.

3. Biber, D. Ko'p O'lchovli Yondashuvlar. In: Lüdeling, A.; Kytö, M. (Tahr.). Korpus Lingvistikasi - Xalqaro Qo'llanma. Berlin / Nyu-York: Valter De Gruyter, 2009 Yil.

4. Biber, D.; Conrad, S. Registr, Janr Va Uslub. Kembrij; Nyu-York: Kembrij Universiteti Nashriyoti, 2009. (Tilshunoslik Bo'yicha Kembrij Darsliklari).

5. Winthrop Nelson Francis and Henry Kučera. Frequency Analysis of English Usage: Lexicon and Grammar, Houghton Mifflin. 1983.

6. Cameron, L. Ta'lim Nutqidagi Metafora . London: Continuum, 2003.

7. McEnery, Tony & Wilson, Andrew. Corpus Linguistics. Edinburgh:Edinburgh University Press. 2001.

8. McEnery, Tony, Yukio Tono & Xiao, Richard. Corpus based Language Studies: An Advanced Resourse Book.London:Routledge. 2006.

9. Mukherjee,Joybrato. Anglistische Korpuslinguistik. Eine Einfuhrung. Berlin:Erich Schmidt. 2009.

10. Scherer, Carmen. Korpuslingvistik. Eine Einfuhrung. Heidelberg: Winter.2006.

11. Biber, Douglas et al. Corpus Linguistics: Investigating Language Structure and Use. Cambridge:CUP. 1998.