COMPARISION BETWEEN LOB BROWN & KCIE

Kaushal Pallavi

***Abstract*:** The aim of this paper is to give general information about Brown Corpus, LOB and Kolhapur corpus for Indian English. The comparative study of Brown Corpus, LOB and Kolhapur corpus is also represented. Nowadays corpus is a basic need in this era of technology. It will play a major role in globalization of languages.

Keywords: Corpus, Brown Corpus, Lancaster-Oslo/Bergen corpus (LOB), The Kolhapur Corpus of Indian English (KCIE).

I. INTRODUCTION

Corpus building is an activity that takes time. Corpus is always designed for a particular purpose; the usefulness of a ready-made corpus must be judged with regard to the purpose to which a user intends to put it. There are thousands of corpora in the world, but most of them are created for specific research and are not publicly available. While abundant corpus resources for languages other than English are also available now, this research paper is only about the comparison of BROWN, LOB & KCIE corpora, which are grouped in terms of their primary sources so that readers will find it easier to choose a suitable corpus. It is used simply to give better primary uses of the relevant corpora. The comparison between three famous text corpora i.e., BROWN, LOB & KCIE is discussed below:

II. BROWN CORPUS:

The compilation of Brown Corpus of American English was completed in 1960s. It was the first electronic corpus. According to Kennedy, "in the face of massive indifference if not outright hostility from those who espoused the conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky"[i]

The Brown Corpus consists of a selection of texts published in the United States in 1961 and the first version of it was available on computer in 1964. Compared to later corpora, the Brown Corpus is relatively small, mainly because compiling a corpus with the technology available at that time was very laborious. According to Meyer, "It had to be keyed in by hand, a process requiring a tremendous amount of very tedious and time-consuming typing"[ii].

According to Kennedy, "The corpus was originally recorded on 100,000 punch-cards[iii] with "70 characters per line plus location information identifying the texts and line numbers". The cards were transferred to magnetic tape later and recently to CD-ROM.

Brown Corpus was the first computer-readable text corpora based on modern English prepared for linguistic research. It was developed by Henry Kuchera and Nelson Francis in the year 1961 at the Brown University, USA.[iv] Name of the corpus is also based on the name of Brown University. It was funded by the Cooperative Research Program of the U.S. Office of Education & Brown University.

STRUCTURE:

The basic structure of brown Corpus is divided into 500 samples which have more than 2000 words[v]. Each sample begins with a sentence but not necessarily of a paragraph or other larger division, and ends at the first sentence ending after 2000 words. The samples represent a wide range of styles and varieties of prose. Verse was not included because it presents special linguistic problems different from those of prose. Drama was also excluded as being the imaginative recreation of spoken discourse, rather than true written discourse. Fiction was included, but no samples were admitted which consisted of more than 50% dialogue. Samples were chosen for their representative quality rather than for any subjectively determined excellence. The use of the word standard in the title of the Corpus does not in any way mean that it is put forward as "standard English"; it merely expresses the hope that this corpus will be used for comparative studies where it is important to use the same body of data. Text published by the permanent resident of USA was included only.

The selection procedure was divided into two phases: one is an initial subjective classification and the other is decision as to how many samples of each category would be used, followed by a random selection of the actual samples within each category. Maximum texts of Brown Corpus were collected from the Library

of Brown University along with Providence Journal. The list of American newspapers which were the New York Public Library keeps with microfilms files was used. Some periodical materials in the categories Skills and Hobbies and Popular Lore were chosen from the contents of one of the largest second-hand magazine stores in New York City. Brown University, USA organized a conference in 1963. In that conference, the list of main text categories and sub-categories for brown corpus was prepared by delegates. In the conference, the delegates also discussed on each category, analyzed the samples of each category and also gave their views individually. The average of these statistics was to get for preliminary set. So that it could be used for corpus but later on some changes took place in it. The corpora consist of 1014312 evaluated words of English. The list of sub-division of Brown Corpus and number of its samples along with text categories are as follows:

Informative Prose Texts		
A	Press: Reportage (political, sports, society, spot new financial, cultural etc.)	44
B	Press: Editorial (institutional, personal letter to editor)	27
C	Press: Reviews (theatre, books, music, dance)	17
D	Religion (books, periodicals, tracts)	17
E	Skills & Hobbies (books, periodicals etc.)	36
F	Popular lore (books, periodical etc.)	48
G	Belles letters, biography, memories, periodicals	75
H	Miscellaneous (Govt. documents, foundation report, industry reports, college catalogue, industry, house organization)	30
J	Learned & Scientific writings (natural sciences medicine, mathematics, social & behavioral science, law, education, humanities and engineering etc.)	80
Imaginative Prose Texts		
K	General fiction (novels, short stories)	29
L	Mystery & detective fiction (novels, short stories)	24
M	Science fiction (novels, short stories)	6
N	Adventure & Western fiction (novels, short stories)	29
P	Romance & love story (novels, short stories)	29
R	Humour (novel, essays etc.)	9
Total		500

Table-1 Text Samples in the Brown Corpus (1961)[vi]

III. THE LANCASTER-OSLO/BERGEN CORPUS:

The Lancaster-Oslo/Bergen corpus is a result of a mutual collaboration of the University of Lancaster, University of Oslo and the Norwegian Computing Centre for the Humanities, Bergen. It is also known as LOB corpus. The LOB project is compiled during the period of 1970- 1976[vii] at the Department of Linguistics and Modern English Language, Lancaster University, Bergen, UK under the direction of Geoffrey Leech. It is funded by the Longman Group and the British Academy, UK. In 1977, the LOB project is shifted to Oslo, Norway. Under the supervision of Stieg Johansson, the project was completed in 1978 at the Department of English, University of Oslo, with the financial & technical help of Norwegian Research Council for Science and the Humanities (NAVF)[viii].

STRUCTURE:

Simultaneously to the Brown Corpus, the LOB corpus is also contained one million words of Present-Day samples of British English. It also collected 500 text samples, each have more than 2000 words. The text categories of LOB are almost similar to the Brown corpus. The corpus is divided into the fifteen text categories which are mentioned below:

1. Press: Reportage (44 texts)
2. Press: Editorial (27 texts)
3. Press: Reviews (17 texts)
4. Religion (17 texts)
5. Skills, Trades and Hobbies (38 texts)
6. Popular Lore (44 texts)
7. Belles Letters, Biography, Essays (77 texts)
8. Miscellaneous writings (30 texts)
9. Learned and Scientific writings (80 texts)
10. General Fiction (29 texts)
11. Mystery and Detective Fiction (29 texts)
12. Science Fiction (6 texts)
13. Adventure and Western Fiction (29 texts)
14. Romance and Love Story (29 texts)
15. Humour (9 texts)

The first text sample of LOB corpus printed and published in the year 1961. The main aim of LOB corpus is to fix the assembling text of British English equivalent to the American English. The LOB corpus has made interlingual research feasible which were hardly possible before. Although, LOB corpus has many similarities with Brown corpus but there is difference in selection and collection of text pattern. The technique of text sampling of LOB corpus is quite interesting. The text samples are collected in more detail as compare to Brown Corpus. There are mainly three sources from where the text samples are collected i.e.

a. **Newspapers**: The collection of newspaper category is quite easy. It is randomly collected from the newspaper but the daily newspapers are sampled for both provincial and national dailies. The samples of daily newspapers, weekly provincial newspaper, and Sunday newspaper are collected separately.
b. **Periodicals**: The sampling of periodicals is made on the basis of Willing's Press Guide (1961) by matching the corpus categories with the subject division of a class index. Only specific articles were selected from Willing's periodicals for further random sampling.
c. **Governmental Documents**: The sampling of Governmental Documents is based on the catalogue of Governmental Publications. It is also selected from bibliographical sources randomly. Texts published by the Britishers are included in the LOB Corpus. Texts published by the non-Britishers are excluded.

In spite all, the composition of the categories and sub-categories of LOB corpus compare with the Brown corpus are summaries as given below:

Composition of Brown & LOB Corpus[ix]		
Text Category	**Text in each corpus**	
	Brown	**LOB**
Press: Reportage	44	44
Press: Editorial	27	27
Press: Reviews	17	17
Religion	17	17
Skills, Trades & Hobbies	36	38
Popular lore	48	44
Belles letters, biography, Essays	75	77
Miscellaneous	30	30
Learned & Scientific writings	80	80
General fiction	29	29
Mystery & Detective fiction	24	24
Science fiction	6	6
Adventure & Western fiction	29	29
Romance & love story	29	29
Humour	9	9
Total	500	500

Table-2

IV. THE KOLHAPUR CORPUS OF INDIAN ENGLISH

In 1978, Sh. S.V. Shastri was exploring under the leadership of Professor G.N. Leech. He started the project in 1980 with a substantial financial aid from the Shivaji University, U.G.C. and various other sources including personal funds. Shastri used the same format to make the Brown Corpus and LOB Corpus to build the Kolhapur Corpus[x]. They selected 15 different domains for KCIE.

KCIE was drawn out from the publications in the year 1978[xi]. After the negotiation with the writers of early corpora, they took the decision and made it sure that the resemblance will not suffer much as result. On the other side,

it is felt that the worth of the KCIE is extremely intensified in general and in particular as an origin for the interpretation of Indian English. KCIE was initially planned to make maximum comparisons with Brown and LOB Corpus, but due to some rational and practical considerations significant distinctions have been made. In fact, KCIE Corpus failed to match the Brown Corpus and LOB Corps. Where samples were collected from the published papers in 1961 in Brown and LOB, there were collected samples published in KCIE in 1978. KCIE Corpus is full of a unique list of Indian words and periods. It was freed from the British shadow. Thus, the Kolhapur Corpus succeeded in performing in the Indian. The success of Kolhapur started the creation of new generation of corpora in India.

STRUCTURE

The KCIE is prototypical intended to be a representative corpus of illustrative sample texts printed and published in of 1978. A stratified sampling process was done for the adoption of texts. The composition of texts in the KCIE as compared to those in the other two corpora is given in table.

Text Categories		**No. of texts in each category**		
		BROWN	**LOB**	**KCIE**
A	Press: reportage	44	44	44
B	Press: editorial	27	27	27
C	Press: reviews	17	17	17
D	Religion	17	17	17
E	Skills, Trades and Hobbies	36	38	38
F	Popular lore	48	44	44
G	Belles Letters	75	77	70
H	Miscellaneous (Govt. Documents, foundation reports, industry reports, College catalogue, industry house organ).	30	30	37
J	Learned and scientific writings	80	80	80
K	General fiction	29	29	58
L	Mystery and detective fiction	24	24	24
M	Science fiction	6	6	2
N	Adventure (Western fiction)	29	29	15
P	Romance and love story	29	29	18
R	Humour	9	9	9
Total		500	500	500

Table 3: The Basic Composition of BROWN, LOB and KCIE Corpora.

V. COMPARISON BETWEEN BROWN, LOB AND KOLHAPUR CORPUS FOR INDIAN ENGLISH

The comparison of Brown, LOB and KCIE corpora can be divided on their categories. According to table no.1 one thing is common between the corpora that there are total 15 number of categories and the sum of their texts are 500. But there is the difference between distribution of materials in some categories so, here the category wise comparison of Brown, LOB and KCIE corpora are given below:

1. **Category A-C:** The KCIE is closer to the LOB in terms of category A-C means national and regional newspapers. This is done deliberately because, as in the case of the British or LOB corpus, there is a clear difference between national and regional newspapers in terms of both the distribution and circulation figures of Indian newspapers. The ratio of lessons taken from national and regional papers is 62% to 38% in KCIE corpus and 60% to 40% in LOB corpus which are divided in the table given below:

No.	**Name of newspaper (National Newspapers)**	**Number of Texts**		
		Daily	**Sunday/Weekly**	**Total**
1	The Hindu, Madras	9	2	11
2	Economic Times, New Delhi	1	2	3
3	The Statesman, Calcutta	5	1	6
4	The Hindustan Times, New Delhi	4	4	8
5	The Times of India, Bombay	11	8	19
6	The Indian Express (various editions)	3	3	6

	33	20	53

Table-4 Details of number of texts drawn from different National newspapers for KCIE.

No.	Name of newspaper (Regional Newspapers)	Number of Texts		
		Daily	Sunday/Weekly	Total
1	Business Standard, Calcutta	3	-	3
2	Deccan Herald Bangalore	2	1	3
3	The Tribune, Chandigarh	3	1	4
4	National Herald, Lucknew	1	1.5	2.5
5	Searchlight, Patna	1	1	2
6	The Assam Tribune, Gauhati	2	-	2
7	Amrit Bazar Patrika, Calcutta	1	-	1
8	Deccan Chronicle, Secunderabad	3	-	3
9	The Western Times, Ahmedabad	1	-	1
10	Madhya Pradesh Chronicle, Bhopal	2	-	2
11	Nagpur Times, Nagpur	2	0.5	2.5
12	Navhind Times, Panaji	1	0.5	1.5
13	Northern India Patrika, Allahabad	-	1.5	1.5
14	Poona Herald, Poona	3	-	3
15	Blitz Weekly, Bombay	3	-	3
	Sub-totals	28	7	35
	Grand Totals	61	27	88

Table-5 Details of number of texts drawn from different regional newspapers for LOB Corpus.

The subcategories of the texts and their distribution in the newspapers according to the newspapers on dailies, Sundays and Weeklies are more similar according to the given table, except that the two subcategories society and culture had to be merged into one. Such a sharp in the reports of the newspapers of the Indian Press which are included in KCIE. Another difference is KCIE is that there are fewer personal editorials in the KCIE.

2. **Category D**: In terms of subcategories, religious writers are categorized neither in Brown nor in the LOB corpus, but the makers of the LOB corpus examined the brown texts and sampled the texts, from 'Education to Popular'. Has decided to include a variety of texts. Committed writing '. The same procedure was followed in the selection of lessons for the KCIE except that the sub-sections 'tracts' were not displayed. The distribution of texts in books and magazines is almost maintained (see Table 7).

		BROWN	LOB	KCIE
Category D. Religion				
	Books	7	9	8
	Periodicals	6	7	9
	Tracts	4	1	-
Category E Skills, Trades & Hobbies				
	Books	2	5	9
	Periodicals	34	33	29
Category F Popular Lore				
	Books	23	16	10
	Periodicals	25	28	34
Category G. Belles Letters etc				
	Books	38	41	29
	Periodicals	37	36	41
Category H Miscellaneous				
	Government Document	24	24	32
	Foundation Reports	2	2	-
	Industry Reports	2	2	2
	University Catalogue	1	1	1
	Ind. House Organ	1	1	1

Category J Learned				
	Natural Sciences	12	12	12
	Medicine	5	5	5
	Mathematics	4	4	4
	Social Science	14	14	14
	Political Science, Law, Education	15	15	15
	Humanities	18	18	18
	Technology & Engineering	12	12	12

Table 6- Categories D-J: The BROWN, LOB and KCIE Corpora compared

3. **Categories E-J**: The subcategories of texts in the KCIE are almost identical to the LOB corpus. However, it was not always possible to match individual texts in the three categories E, F and G as source type, book or regular format for content. While the books that are most popular in the case of Category E are: In the case of categories F and G (see Table 7) there is some under-representation. As mentioned earlier, we have deliberately changed the weight between category G and H. G is subtracted by seven lessons and H is multiplied by the same number. This was done to illustrate the Indian situation, in which, first of all, government documents divide themselves between the Central and State Governments. Documents and most of them are much more than any other printed material in English except press material. This is reflected in the greater representation of government documents in the KCIE. And the foundation report has not been submitted (see Table 7). The three texts correspond very closely to the three corporations. This is the only category that has one-to-one correspondence with sub-categories (see Table 7).
4. **Categories K-R:** This part of the collection, which is fictional, does not fit the maximum. As mentioned earlier the required number of lessons was not available and its possible consequences were discussed with the experts in the field and it was felt that there would be a slight loss of comparability. Subcategories i.e., K, L, M, N, and P all represent fiction, general fiction (K), mystery and detective (L), science fiction (M), adventure and western fiction (N), and romance and love story. (P). The classification is based on the theme / treatment and is often bound to overlap, and the corpus compile is interested in 'style'. In the sampling process, it is quite possible to run wide numbers in the form of text for a selected part of the work, especially in the case of novels. In view of all this, first of all, the subcategories were negatively defined, i.e., L, M, N, or P, was considered. Category K And especially selected texts from novels were examined and placed in such defined categories. The fact remains that the number of lessons corresponds only to the L and R categories. In the case of K, we have twice the number 58 instead of 29; Science fiction only 6 versus 2; Adventurous 29 instead of just 15; And the romance and love story are only 18 instead of 29. Again, mystery and espionage are mostly mysteries around the West for spies and death, murder, etc., but for KCIE, it includes other kinds of secrets in the sense of '. Mysterious or miraculous. Similarly, in the case of adventure and western fiction, there is nothing in India to match 'western fiction'. So, the sub-category is made up entirely of 'adventure'. Now, it is important to mention that if the imaginary prose part of the KCIE is not comparable to the brown or LOB on its face, then it is actually designed to represent KCIE.

VI. CONCLUSION

The above-mentioned comparison of BROWN, LOB & KCIE text corpora are some most famous corpora of the world. With almost the history of six decades of corpora (parallel and multilingual corpora) in various world-famous languages are created and the new horizons are open every day for the research, comparison and tool development. It becomes easier for any linguist or researcher to use those all kinds of corpora for buildup language specific and research specific purpose corpora. The availability of all the major and minor corpora across the world serve the needs of linguists, linguistics and people.

References

[i] http://www.helsinki.Fi/varieng/CoRD/corpora/BROWN/background.html#kennedey_1988
[ii] http://www.helsinki.Fi/varieng/CoRD/corpora/BROWN/background.html#meyer_2002
[iii] http://en.wikipedia.org/wiki/punchcard
[iv] https://en.wikipedia.org/wiki/Brown_Corpus#History
[v] https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
[vi] Niladari Shekhar dash. Language corpora. 2009. Mittal publication, new Delhi. P. 37
[vii] Niladari Shekhar dash. Language corpora. 2009. Mittal publication, new Delhi. P.38
[viii] Niladari Shekhar dash. Language corpora. 2009. Mittal publication, new Delhi. P.38
[ix] Niladari Shekhar dash. Language corpora. 2009. Mittal publication, new Delhi. P.39
[x] http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM
[xi] http://korpus.uib.no/icame/manuals/KOLHAPUR/INDEX.HTM