Introduction to Corpus Linguistics

Digital Text Corpora

Introduction

The history of digital text corpus generation and usage presents an exciting narrative. It shows how technology has brought about a resurgence in the discipline of linguistics, which otherwise turned its attention towards a direction of no return. We have briefly described the formation and content of some of the most widely known digital text corpora so far developed in English and some other languages.

The introduction of digital corpora dates back to the 1960s—with the advent of computer technology. It is an important milestone in the history of linguistic research and application. In the year 1961, two linguists at Brown University, USA, namely, Nelson Francis and Henry Kučera, first initiated an attempt to develop a text corpus of 1 million words from English texts written and used in America. This corpus is globally accepted as the first language corpus generated in digital form.

The generation of text corpora is not confined to a few widely privileged languages like English, French, German, or Spanish. Many lesser-known and unprivileged languages also are emerging with corpora of various types for various kinds of applications. This makes it possible to find corpora of various types in the most advanced as well as less advanced languages. In essence, digital text corpora are already developed in almost all languages, bar a few that are yet to have an opportunity to deploy computer technology facilities used by most others.

In India, for instance, the languages that have already developed text corpora of some kind or another in digital form or are in the process of developing these, include Assamese, Bangla, Bodo, Gujarati, Hindi, Indian English, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Sindhi, Tamil, Telugu and Urdu. Similar efforts have also been made for a few dialects spoken in the Indian subcontinent.

How are Corpora Formed?

A large number of text and speech corpora were developed in various parts of the world within six decades, either through personal initiatives or under the patronage of some academic or research institutes. The most striking and notable aspect of the recent phenomenon was that scholars started realizing the applicational importance of language corpora in language research and education, as well as in the development of tools and systems of language technology and communication. This has led scholars to convert present and past text materials into digital versions in the form of corpora so that these materials are made available to people for access and utilization. This has also resulted in the publication of a high percentage of research papers based on the analysis of old text materials now available in digital form.

A lot of activities are underway for assembling language corpora of various types for most of the languages of the world. Moreover, with the generation of bilingual and multilingual corpora in several languages, new horizons in bilingual and multilingual research and comparison are opening up. analyses of these corpora are making notable contributions to the development of new kinds of scholastics for the new generation of scholars.

In this report, we shall briefly describe the formation and content of some of the most widely known digital text corpora so far developed in English as well as some other languages; the patterns of their formation, the type of content included in them, and the way these corpora are used in various linguistic works. This will give some ideas about how a digital corpus should be developed in a language (including many Indian languages) following the guidelines and methods already adopted by most corpora developed so far.

The Brown Corpus

The first corpus generated in digital form, the *Brown Corpus of the Standard Sample of Present-Day American English (*henceforth*, Brown Corpus)* was developed in the year 1961 by Nelson Francis and Henry Kučera at Brown University, USA. The corpus was developed with a grant from the Cooperative Research Program of the U.S. Office of Education, USA. It consists of 1,014,312 words of running texts of edited English prose printed in the USA during the calendar year 1961. It contains written text samples relating to different subjects that were composed by native speakers of American English. Although all the text materials included in this corpus first appeared in printed form in 1961, some of the texts were undoubtedly written much earlier.

- The *Brown Corpus* is made from 500 text samples, with each sample having around 2,000 words. Each sample text is started at the beginning of a sentence, but not necessarily at a paragraph or some other larger division, and ends at the sentence ending with 2,000 words.
- The samples represent a wide range of styles and varieties of prose. Texts from verse are not included in the corpus on the ground that texts belonging to poetic composition present special linguistic issues different from those of the prose texts.
- The short verse passages quoted in prose samples are, however, kept in the corpus because their deletion from the context may have otherwise distorted the cohesion of the texts.
- Similarly, texts from dramas and plays have been excluded from it since these texts are considered to be an imaginative recreation of spoken discourse rather than being true representations of the written discourses of a language.
- Some text samples from fiction are included in the corpus, and texts that contain more than 50% dialogue are excluded due to the above reasons. The samples are mainly selected due to their representative feature rather than for any subjectively determining excellence.

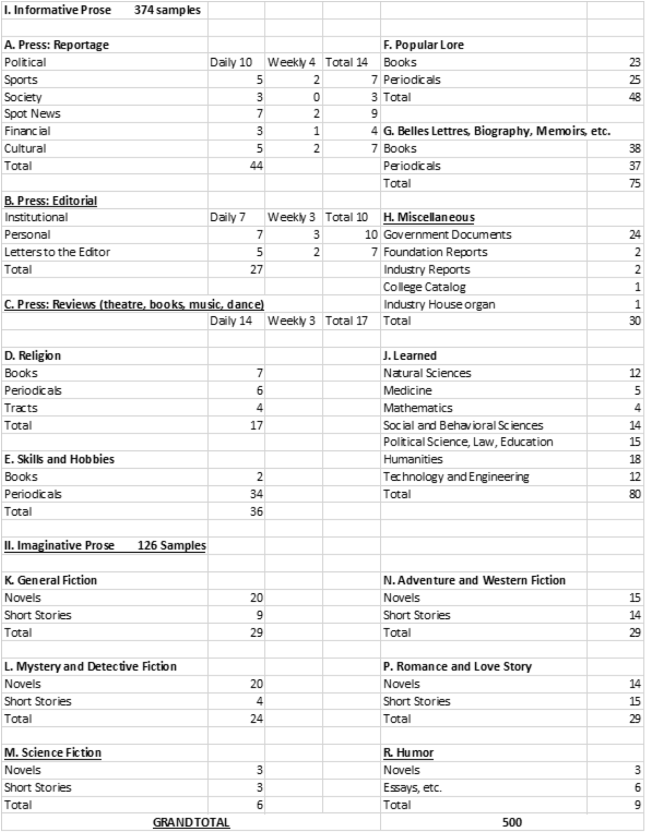
- Since the preparation and input of data is a major bottleneck in computer-based linguistic work, the intent is to make available a carefully chosen and prepared body of text materials of considerable size in a standard format.
- The corpus may prove to be a standard in setting patterns for the preparation and presentation of further bodies of data in English or other languages. The text selection procedure is controlled by three strategic parts:
    - Initial subjective classification of texts;
    - The number of samples of each category is used;
    - Random selection of samples within each category
- For most of the text categories, the library collection of Brown University, as well as that of the Providence Athenaeum, are treated as the main sources from which random selections of texts are made. However, for certain other text categories, it has been necessary to go beyond the scope of the collection.
- For samples from daily newspapers, the microfilm files of some American newspapers from the New York Public Library are also used. For some other categories that are ephemeral, text samples are collected arbitrarily. Periodical materials in the categories Skills and Hobbies and Popular Lore are chosen from the contents of a second-hand magazine store in New York City, NY, USA.
- The list of main text categories and their subdivisions was first drawn up at the conference held at Brown University in February 1961. All these figures were averaged to obtain the preliminary set of figures used for the formation of the corpus. Finer subdivision of text categories was based on proportional amounts of actual publication of the texts during the year 1961. The list of text categories with their principal subdivisions and the number of samples are provided in the following table (Table 1).

- For most of the text categories, the library collection of Brown University, as well as that of the Providence Athenaeum, are treated as the main sources from which random selections of texts are made. However, for certain other text categories, it has been necessary to go beyond the scope of the collection.

- For samples from daily newspapers, the microfilm files of some American newspapers from the New York Public Library are also used. For some other categories that are ephemeral, text samples are collected arbitrarily. Periodical materials in the categories Skills and Hobbies and Popular Lore are chosen from the contents of a second-hand magazine store in New York City, NY, USA.

This particular corpus has been used over the years in several works of linguistics across the world. Starting from computational analysis of *American English (Kučera and Francis 1967)*, it has been used heavily to study the patterns of *use of punctuation in texts (Meyer 1986)*, *patterns of lexical collocation (Kjellmer 1994)*, *POS-tagging of English words (Belmore 1994)*, *digital access to texts (Jones 1987)*, dictionary compilation for *modern American English (Kjellmer 1994)*, use of *personal pronouns in texts (Nakamura 1989a, b)*, distribution of *grammatical tags in texts (Nakamura 1990)*, distribution of *vocabulary items across text types (Nakamura 1991)* and many other linguistic studies.

Today, this corpus is considered small, and slightly dated. However, the corpus is still in use, much of its usefulness lies in the fact that the Brown corpus layout has been copied by other corpus compilers. The LOB corpus (British English) and the Kolhapur Corpus (Indian English) are two examples of corpora made to match the Brown corpus. The availability of corpora that are so similar in structure is a valuable resource for researchers interested in comparing different language varieties, for example.



I. Informative Prose 374 samples					
A. Press: Reportage				**F. Popular Lore**	
Political	Daily 10	Weekly 4	Total 14	Books	23
Sports	5	2	7	Periodicals	25
Society	3	0	3	Total	48
Spot News	7	2	9		
Financial	3	1	4	**G. Belles Lettres, Biography, Memoirs, etc.**	
Cultural	5	2	7	Books	38
Total	44			Periodicals	37
				Total	75
B. Press: Editorial					
Institutional	Daily 7	Weekly 3	Total 10	**H. Miscellaneous**	
Personal	7	3	10	Government Documents	24
Letters to the Editor	5	2	7	Foundation Reports	2
Total	27			Industry Reports	2
				College Catalog	1
C. Press: Reviews (theatre, books, music, dance)				Industry House organ	1
	Daily 14	Weekly 3	Total 17	Total	30
D. Religion				**J. Learned**	
Books	7			Natural Sciences	12
Periodicals	6			Medicine	5
Tracts	4			Mathematics	4
Total	17			Social and Behavioral Sciences	14
				Political Science, Law, Education	15
E. Skills and Hobbies				Humanities	18
Books	2			Technology and Engineering	12
Periodicals	34			Total	80
Total	36				
II. Imaginative Prose 126 Samples					
K. General Fiction				**N. Adventure and Western Fiction**	
Novels	20			Novels	15
Short Stories	9			Short Stories	14
Total	29			Total	29
L. Mystery and Detective Fiction				**P. Romance and Love Story**	
Novels	20			Novels	14
Short Stories	4			Short Stories	15
Total	24			Total	29
M. Science Fiction				**R. Humor**	
Novels	3			Novels	3
Short Stories	3			Essays, etc.	6
Total	6			Total	9
GRAND TOTAL				**500**	

Table 1 *Text samples in the Brown Corpus*

The LOB Corpus

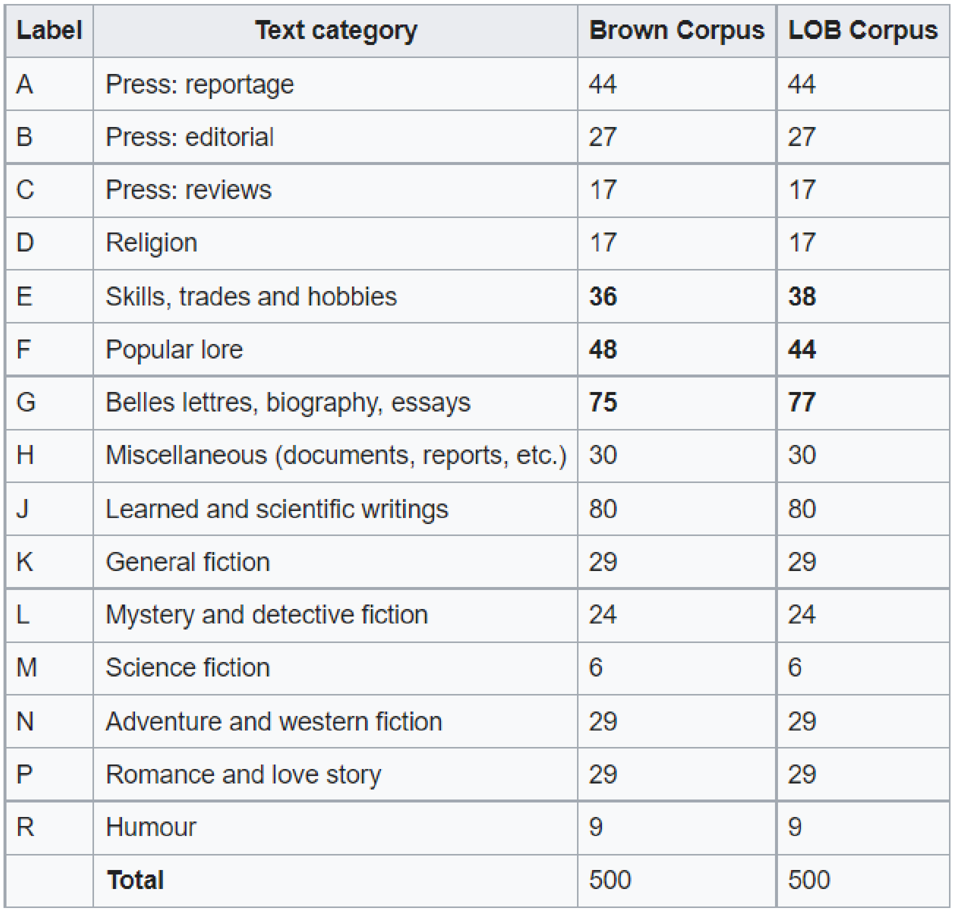
The *Lancaster–Oslo/ Bergen (LOB) Corpus* is an outcome of a mutual collaborative work carried out at the University of Lancaster, UK; the University of Oslo, Norway; and the Norwegian Computing Centre for the Humanities, Bergen, Norway.

It was supported by grants from the Longman Group and the British Academy, UK. In 1977, the project was moved to the Department of English, at the University of Oslo, Norway. Eventually, the project was completed in 1978 with adequate financial and technical support from the Norwegian Research Council for Science and the Humanities.

This corpus is the British counterpart of the Brown Corpus of American English, which contains texts printed in the same year so that comparisons between both varieties could be made. Similar to the Brown Corpus, it contains 500 text samples of nearly 2,000 words each, distributed over 15 text categories as stated in Table 2.

The generation of the LOB Corpus has made interlingual research and applications feasible, which was hardly possible before this corpus was made available.

- Both of these corpora, rather than concentrating on the limited texts used in specific works, aim at a general representation of text types for future research on a broad range of aspects of the respective language varieties.
- In addition, they facilitate a combined use of texts from two countries across domains to match the British English texts as closely as possible with the texts of American English.



Label	Text category	Brown Corpus	LOB Corpus
A	Press: reportage	44	44
B	Press: editorial	27	27
C	Press: reviews	17	17
D	Religion	17	17
E	Skills, trades and hobbies	**36**	**38**
F	Popular lore	**48**	**44**
G	Belles lettres, biography, essays	**75**	**77**
H	Miscellaneous (documents, reports, etc.)	30	30
J	Learned and scientific writings	80	80
K	General fiction	29	29
L	Mystery and detective fiction	24	24
M	Science fiction	6	6
N	Adventure and western fiction	29	29
P	Romance and love story	29	29
R	Humour	9	9
	Total	500	500

Table 2 *Composition of the Brown Corpus and the LOB Corpus*

- Similar to the Brown Corpus, the LOB Corpus contains 500 printed text samples of about 2,000 words each (about 1 million words in all). Although the year of publication and sampling principles are identical for both corpora, there are certain differences in the process of text selection.

- The text categories and subcategories of the Brown Corpus are analyzed in more detail and are matched with the corresponding categories and subcategories selected for the LOB Corpus. The list is given in Table 2, which also summarizes the composition of the LOB Corpus when compared with the Brown Corpus. The materials within the main text categories of the LOB Corpus are arranged to match that of the Brown Corpus as closely as possible.
- In collecting text samples from books, *The British National Bibliography Cumulated Subject Index 1960–1964* is used to ensure that books published in 1961 but not cataloged until 1962, are included in the sampling. This method of sampling strictly adheres to the subject divisions recorded in the Dewey Decimal Classification Scheme of *The British National Bibliography* so that subject subcategories of the corpus do not deviate from the standard matrix.
- The sampling of texts from periodicals as well as from newspapers is made based on *Willing's Press Guide (1961)*. Sampling for the periodicals is made by matching the corpus categories with the subject divisions of the class index of *Willing's Press Guide.*
- Where no suitable index heading is available from which sampling can be made, the entire class index is inspected. All suitable periodicals are enumerated and a simple random sampling process is carried out based on the numbering of samples. Periodicals sampled in this way are then excluded from any subsequent sampling of the index division under which they are listed.
- In contrast to the difficulties faced in sampling for the periodicals, the indexing of *Willing's Press Guide* makes the sampling for newspaper categories fairly easy. Index of *Daily Newspapers* is sampled for both provincial and national dailies. Since all the national dailies, except *The Guardian*, are published in London, UK, and since the index is subdivided by place of publication, a separate sampling of the national dailies is a comparatively simple matter.

- The sampling system of the governing documents is based on the Catalogue of Government Publications, 1961. The overall method in text sampling has been to randomly select titles from the bibliographical sources, and then to randomly sample particular items for the page at which to start the 2,000-word extract.

It is generally argued that the LOB Corpus is not at all a good representation of British English in a strict statistical sense. The issue of representativeness of the LOB Corpus arises from its deliberate attempt to include the relevant categories and subcategories of texts rather than from some blind statistical choices. The random sampling system simply ensures, within the stated guidelines, that the selection of individual texts is free of the conscious or unconscious influence of personal taste or preference.

The Australian Corpus of English

The *Australian National Corpus* is a national meta-collection that includes assorted examples of the Australian English text (published and unpublished), transcriptions, and audio and audio-visual materials from individual collections provided by collaborative institutions. Several retrospective corpora and content represented in the Australian National Corpus include linguistic occurrences that can be analyzed for both academic research and teaching purposes.

The Australian Corpus of English (ACE) was compiled to match Australian data from 1986 with the American (Brown Corpus) and British (LOB) corpora of written English from the 1960s. It includes 500 samples of published texts taken from 15 different categories of nonfiction and fiction, including newspapers, reportage, editorials, and reviews; magazines and journals: popular, academic; government, and corporate documents; fiction monographs and short stories (both popular and literary).

Due to the variation in time scale, differences are obvious not only concerning the geographical region but also concerning time among the Brown Corpus, the LOB Corpus, and the ACE. However, this is of considerable interest to researchers in terms of showing the direction of influence in the latter part of the century.

One of the prime objectives of the ACE was to find a balance of genres represented in the Brown Corpus and the LOB Corpus, as well as to create a more or less equivalent set of 2,000-word samples for each of the text categories.

- Within each text category, the sampling procedures are usually strategic rather than random, because the corpus needs to match the subgenres and subject areas wherever possible with the categories of its model corpora.
- In some categories (e.g., fiction) the corpus requirements are such that the designers have to sample almost every Australian monograph published in that year, and thus, the representation in the ACE is almost the total.
- The corpus designers give preference to those that are held in multiple libraries in several states wherever possible and therefore probably have a greater readership and impact on the public. Among serials—both the popular and scholarly—the selection is usually dictated by the subject matter to ensure a spread of interests and disciplines, like the broad ranges captured by their predecessors.

The Freiburg–LOB Corpus

The *Freiburg–LOB Corpus of British English (FLOB)* was the result of the effort from Christian Mair in 1991 when he took the initiative to compile a corpus that would match the Brown Corpus and the LOB Corpus. This corpus has aimed to represent the language of the early 1990s in contrast with the language of 1961 that was included in the first two corpora, this makes a notable difference in the FLOB Corpus concerning the other two renowned and widely used language corpora.

When the project truly started in 1991, necessary funding was provided by the German Research Foundation to speed up the process of data collection and compilation. The comparative analysis of the corpora was to provide linguists with a suitable empirical basis to study the language change in progress over the years.

- The compilation of the FLOB Corpus eventually enabled linguists to test the following aspects of modern English:
    - An opportunity for scholars to verify some of the current hypotheses on linguistic change in present-day English;
    - A systematic comparison of the frequency of use of lexical items, particularly of closed class items, gives a huge scope to detect changes in English not noticed previously;
    - Deal systematically with one of the major methodological issues in the study of ongoing linguistic change across the varieties of English;
    - To study the interdependence of the two regional varieties (British vs. American) at the synchronic level;
    - Generates a scope for studying variations in style and treatment at the diachronic level with the same variety (British English in 1961 vs. British English in 1991);
    - The databases of present-day British and American English are supplied to draw comparisons with Indian, Australian, and New Zealand English.
- The sampling principle used to compile newspaper texts for the Brown Corpus and the LOB Corpus was random. Wherever possible, the same magazines and periodicals used in the LOB Corpus were also used for the FLOB Corpus.
- In the sampling of monographs, great care was taken to select books of equal topics rather than randomly selecting titles from bibliographical sources. This was because the aim was to achieve a kind of close comparability with the LOB Corpus.

- Instead of using a complex coding system applied to the texts of the LOB Corpus, the FLOB Corpus uses a highly simplified version of the Standard Generalized Markup Language (SGML) coding system drawn up for coding the International Corpus of English (ICE).
- To ensure that the FLOB Corpus is 'readable' as far as possible, the markup symbols are kept to a minimum. For instance, the use of double codes in the text is carefully avoided so that the corpus users are not confused in terms of the identity or functions of the texts. For instance, the use of double codes in the text is carefully avoided so that the corpus users are not confused in terms of the identity or functions of the texts.

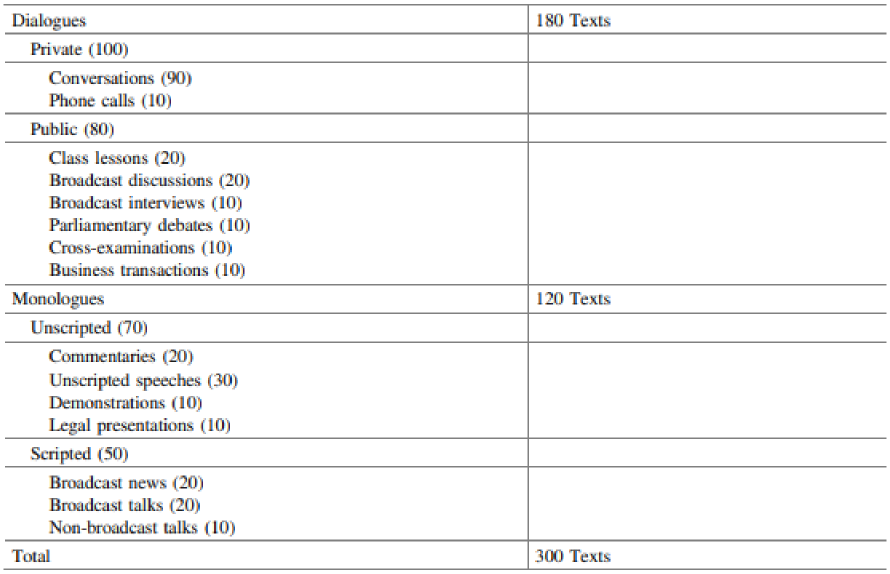
The International Corpus of English

The *International Corpus of English (ICE)* is a set of corpora representing varieties of English from around the world. Over twenty countries or groups of countries where English is the first language or an official second language are included. The father of the project, *Sidney Greenbaum*, insisted on the primacy of the spoken word, following Randolph Quirk and Jan Svartvik's collaboration on the original London-Lund Corpus (LLC). This emphasis on word-for-word transcription marks out ICE from many other corpora, including those containing, e.g., parliamentary or legal paraphrases.

- Each corpus contains one million words in 500 texts of 2000 words, following the sampling methodology used for the Brown Corpus. Unlike Brown or the LOB Corpus, however, the majority of texts are derived from spoken data.
- With only one million words per corpus, ICE corpora are considered very small by modern standards. ICE corpora contain 600,000 words of orthographically transcribed spoken English.

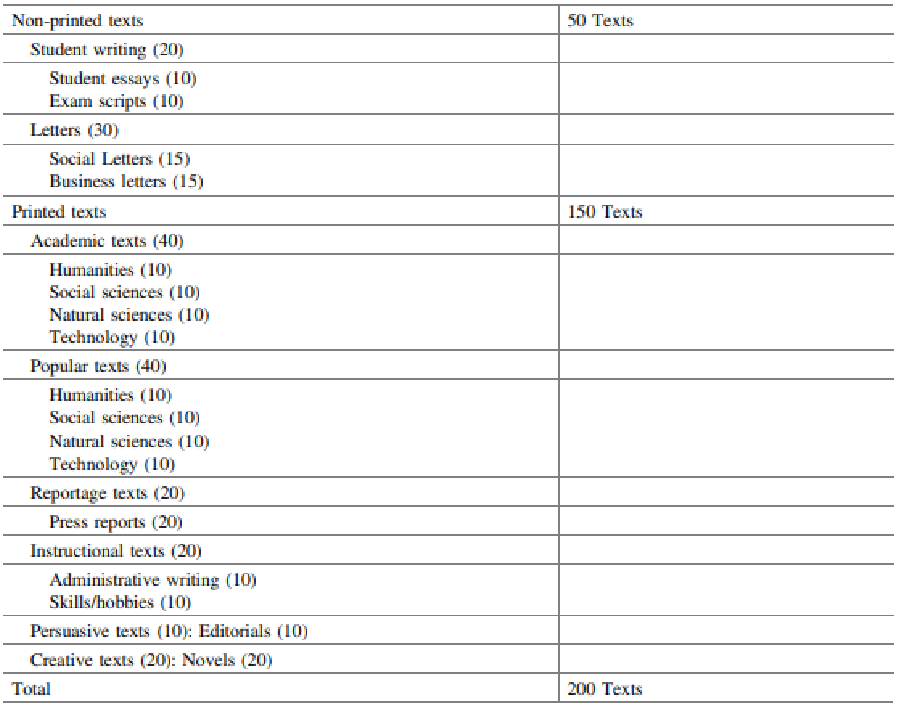
- The corpora consist entirely of data from 1990 or later. The subjects from which the data was collected are all adults who were educated in English and were either born or moved at an early age, to the country to which their data is attributed. There are speech and text samples from both men and women of many age groups.
- The British Component of ICE, ICE-GB, is fully parsed with a detailed Quirk et al. phrase structure grammar, and the analyses have been thoroughly checked and completed. This analysis includes a part-of-speech tagging and parsing of the entire corpus.
- To ensure compatibility between the individual corpora in ICE, each team is following a common corpus design, as well as a common scheme for grammatical annotation. Many corpora are currently available for download on the ICE official webpage, though some require a license.

All the parts of ICE are complete and released for public access. These are comprised of a one-million-word database of contemporary English used in each country. The corpora are fully parsed and they carry a large number of syntactic trees. With data retrieval software tagged with it, the ICE becomes a good resource for the study of English used all over the world. The final composite form of the ICE provides authentic materials for any kind of comparative research. The spoken and the written text categories of ICE are given in Tables 3 and 4 where the numbers in brackets indicate the number of texts containing 2,000 words in each category.



Dialogues	180 Texts
Private (100)	
Conversations (90) Phone calls (10)	
Public (80)	
Class lessons (20) Broadcast discussions (20) Broadcast interviews (10) Parliamentary debates (10) Cross-examinations (10) Business transactions (10)	
Monologues	120 Texts
Unscripted (70)	
Commentaries (20) Unscripted speeches (30) Demonstrations (10) Legal presentations (10)	
Scripted (50)	
Broadcast news (20) Broadcast talks (20) Non-broadcast talks (10)	
Total	300 Texts

Table 3 *Categories of spoken text samples in ICE*



Non-printed texts	50 Texts
Student writing (20)	
Student essays (10) Exam scripts (10)	
Letters (30)	
Social Letters (15) Business letters (15)	
Printed texts	150 Texts
Academic texts (40)	
Humanities (10) Social sciences (10) Natural sciences (10) Technology (10)	
Popular texts (40)	
Humanities (10) Social sciences (10) Natural sciences (10) Technology (10)	
Reportage texts (20)	
Press reports (20)	
Instructional texts (20)	
Administrative writing (10) Skills/hobbies (10)	
Persuasive texts (10): Editorials (10)	
Creative texts (20): Novels (20)	
Total	200 Texts

Table 4 *Categories of written text samples in ICE*

British National Corpus

The *British National Corpus (BNC)* is a large corpus of modern English texts. It contains hundreds of millions of words of representative samples from both written and spoken English texts from a wide variety of text types and disciplines. It is scientifically designed to represent much wider cross-sections of the current British English, reflected both in spoken and written practices of native British speakers.

The corpus-building activities were carried out and managed by a consortium built from direct collaboration between industry and academics. The project was led by the Oxford University Press with the participation of the majority of dictionary publishers in England. The project was funded by the Science and Engineering Council and the Department of Trade and Industry, under the programme of the Joint Framework for Information Technology. Additional support was provided by the British Library and the British Academy.

- The BNC represents a wide range of modern British English. The written part covers 90% of the total text samples and includes extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, and many other kinds of written text.
- The spoken part comprises only 10% of the text, including a large amount of unscripted and informal conversation, and recorded speech selected from different age groups, regions, and social classes in a demographically balanced way.
- In addition, it contains speech samples collected from different interactional contexts such as formal business talks, government meetings, radio shows, individual telephonic conversations, and so on.
- To date, the BNC contains 500 million words comprising more than 5,000 text types. Of these, more than 1,000 text samples are transcribed from spoken conversations and monologues. Each text is segmented into separate orthographic ‘sentence units’, within which words are assigned a word class or as part-of-speech. The text

segmentation, as well as word classification, is carried out automatically by the Constituent Likelihood Automatic Word Tagging System (CLAWS).

- The text classification scheme used for the BNC distinguishes 65 parts of speech described in the documentation files that accompany the corpus. The corpus is encoded according to the guidelines of the Text Encoding Initiative (TEI) using the ISO standard SGML system to represent the output from CLAWS and a variety of other structural properties of the texts, e.g., headings, paragraphs, and word lists.
- The wide diversity of text types as well as systematic schemes of text annotation makes the BNC a highly useful resource for research in morphology, lexicography, lexicology, semantics, artificial intelligence, language technology, natural language processing, literary study, speech synthesis, speech recognition, culture studies, and many other domains of linguistics and related disciplines.

Croatian National Corpus

The *Croatian National Corpus (CNC)* was made from a large language database comprising nearly 30 million words of contemporary text samples of the Croatian language. The Analysis of this corpus produces a large list of non-lemmatized tokens; these are used to compile lexical resources including the Croatian National Dictionary and Croatian language teaching materials.

- The corpus was augmented with new data from representative Croatian texts carefully taken from older and contemporary texts. After compilation, the entire corpus was processed to generate a lexicon of 1 million Croatian words. Both the corpus and lexical database are used to collect information on the areas relating to:
    - Croatistics (study of the history of the Croatian language, its orthographic problems, process of inflection and derivation, and development of Croatian terminology, etc.);

    - Lexicography and lexicology (Croatian and Croatian–foreign language dictionaries, concordance, thesauri, dictionaries relating to terminology, neologisms, etc.);
    - Information science (indexing and searching of text data, natural language processing and interpretation, natural language generation, computational systems, techniques, and tools for Croatian language texts, etc.)
- The corpus is treated as having the status of fundamental research in humanistic sciences as well as a strategic resource for the Croatian language. The project that generated the Croatian corpus also set out to encompass the following areas:
    - To compile and process a multimillion-word corpus of old and contemporary Croatian language;
    - To provide a collection of selected text samples and a compilation of dictionaries of older Croatian authors.
    - The translation and conversion of great literary works of civilization (e.g., Bible, Talmud, Koran) into an electronic form that will be included in the Croatian corpus;
    - To supplement the dictionary of Croatian orthography with results obtained from analysis of written texts;
    - To conduct diachronic and synchronic investigations based on the small sub-texts included in the corpus;
    - To process and analyze the corpus for the identification and retrieval of instances of a neologism for Croatian terminology;
    - To publish linguistic results derived from the analysis of the corpus in digital form for global use;
    - To further analyze the corpus to serve lexicological and lexicographic projects of the Croatian language.

English–Norwegian Parallel Corpus

The English–Norwegian Parallel Corpus (ENPC), which was developed in 1994, consists of original English texts and their Norwegian translations. It was intended to be used as a general research tool, available for all types of applied and theoretical linguistic research. It is an outcome of a cooperative project between the Institute for British and American Studies at the University of Oslo.

- The corpus comprises four main parts:
    - English original text samples;
    - The translations of English original text samples into Norwegian;
    - Norwegian original text samples;
    - The translations of Norwegian original text samples into English
- Because of the unique composition of this corpus, text samples are used in various ways, as shown below:
    - To initiate translation studies (from English to Norwegian and vice versa);
    - To make comparative studies based on comparable original texts in the two languages;
    - To carry out comparative studies based on a particular text and its translation;
    - To conduct comparative studies between the original and translated texts.
- The ENPC consists of text segments of 10,000–15,000 words taken from the beginning of each text source (approximately 40–50 pages). The initial planning for 80 pairs of text samples was further extended throughout the project.
- Hence, at present, the corpus comprises 100 pairs of text samples, containing a total of 2.6 million words. It contains 60% imaginative prose texts (e.g., children's books, detective novels, and general literature) and 40% informative prose texts (e.g., popular science, government publications, legal texts, and tourist information).
- All the text samples are encoded in the SGML following the recommendations of TEI. Each sentence of one language has a pointer to the location of the sentence in another language.

- This corpus project cooperates with similar projects carried out in Sweden and Finland concerning English original texts and their translations in these two neighboring languages.
- The text samples are selected from several libraries located in Norway and England. In cases where relevant text segments are not available in electronic form, these are copied and converted by the OCR system. The division of running texts into sentences is carried out automatically based on punctuation marks and capital letters, which often act as sentence-terminal markers.
- In the scanning process, some text codes, such as the ends of the paragraph and bold text, are automatically incorporated as important text markers, while other codes are added manually.
- The database is managed and controlled with the SGML parser before it passes through indexing. The English texts are tagged at the sentence and part-of-speech levels using the Helsinki Tagger, while the Norwegian texts are encoded in a similar fashion using the Norwegian Tagger.

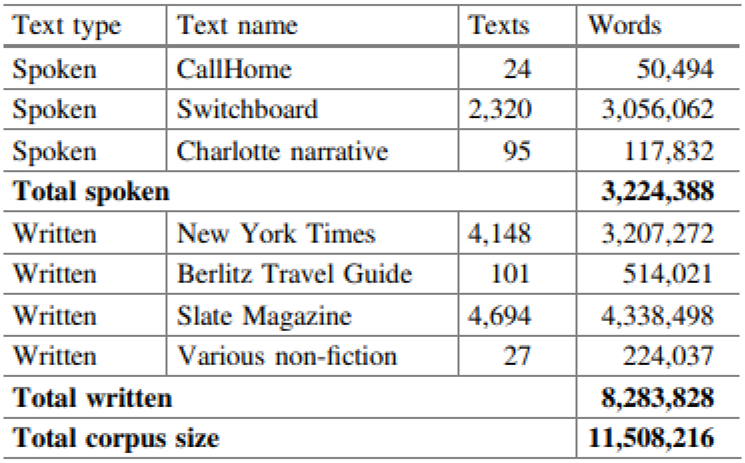
Besides tagging the parallel corpus, the software has also been developed for processing the parallel texts. This software for aligning sentences in texts uses a simple bilingual word list containing approximately 900 lines of words. Each line contains several words or word stems in one or the other languages. In addition, checking is carried out in terms of numbers as well as on possible proper nouns and words that exhibit identical spellings in the two languages.

The best results are obtained by taking into consideration that word order in English and Norwegian sentences are approximately identical. The lexical lists obtained from the automatic alignment of words are added to the anchor word list of particular texts.

American National Corpus

The *American National Corpus (ANC)* was designed and developed to gather a representative language database of modern American English comparable to the BNC. The analysis of the BNC demonstrated that due to several differences in language used in the two separate countries, the BNC could not be used as a useful referential resource for studying the American English variety. This led to the aim to generate a corpus of modern American English that could be used to significantly contribute to linguistic research, as well as to provide a rich national resource for education at all levels

- Since the corpus is in the process of generation, only the first part comprising a database of 10 million words has been released so far for public access (recently it has released 15 million words).
- Text samples included in the first part are those that were first received from the proposed contributors. Therefore, the corpus is not balanced in the true sense of the term. Moreover, manual verification and validation of XML tagging and part-of-speech annotation of text samples have not yet been done.
- The header files are minimal, although they contain fairly complete information concerning *domain, sub-domain, subject, audience,* and *mediums* of text samples. The list (Table 5) given below summarizes the contents of the first part of the ANC.



Text type	Text name	Texts	Words
Spoken	CallHome	24	50,494
Spoken	Switchboard	2,320	3,056,062
Spoken	Charlotte narrative	95	117,832
Total spoken			**3,224,388**
Written	New York Times	4,148	3,207,272
Written	Berlitz Travel Guide	101	514,021
Written	Slate Magazine	4,694	4,338,498
Written	Various non-fiction	27	224,037
Total written			**8,283,828**
Total corpus size			**11,508,216**

Table 5 *Components and total words of the first part of the ANC*

- The speech part of the ANC contains components of 'CallHome' that include transcripts and documentation files of 24 unscripted telephonic conversations made between native speakers of American English. The transcripts, a subset of the full 'CallHome' corpus obtained from the *LDC,* cover a continuous 10-minute segment of each call, comprising 50,494 words.
- The 'Switchboard' part of the ANC includes transcriptions of the original LDC 'Switchboard' corpus. It contains approximately 2,320 spontaneous conversations averaging six minutes in length and comprising about 3 million words of text spoken by over 500 speakers of both sexes from every major dialect of American English.
- The Charlotte Narrative and Conversation Collection contains 95 narratives, conversations, and some interviews that represent the residents of Mecklenburg County, North Carolina, as well as the surrounding communities.

The sub-corpus of the ANC, which is supplied by the Oxford University Press, contains approximately 250,000 words from non-fiction texts. The text samples were mostly obtained from some notable works authored by American writers. Selected texts mostly relate to the American Constitution, the textile industry, child development and child care, general biology, architecture, and so on.

Some Small-Sized Text Corpora

Besides some of the major language corpora discussed above, there are thousands of small corpora of written texts made in almost all the languages of the world. Since it is not possible to provide detailed information about the form and composition of all corpora, we provide below some short descriptions of some corpora, which are often referred to in linguistics and language technology.

The *Canterbury Tales Project* carried by Cambridge University Press contains a complete corpus of writings of Chaucer. The *Penn Treebank* of the University of Pennsylvania contains articles from Wall Street Journal along with classical, historical, and religious English texts. The *Tycho Brahe Corpus of Portuguese* contains nearly 5 million words of historical Portuguese texts. The *Institute for Dutch Lexicology* has already developed several large corpora in written Dutch that are intended for use in various academic and research purposes.

Bank of Swedish

Bank of Swedish (also called *Språkbanken*) was developed, on a national basis, at Göteborg University, Sweden. At present, it is available for general access in machine-readable form with a large set of linguistic data arranged in a systematic order.

- The corpus comprises nearly 40 million running words collected from fiction, legal documents, reports of the proceedings of the Swedish Parliament, and daily newspapers. It contains not only words but also graphemes, morphemes, idioms, phrases, and sentences of various forms and structures, both in normal and concordance forms.
- The collected texts, as well as the processed materials, are directly used to build up the Swedish Word Bank, as well as to supply necessary data, advice, and information to researchers working in the area of language processing and computational linguistics for Swedish.

- To date, the corpus has delivered a huge lexical database for developing systems for automatic spell-checking in Swedish to several Swedish and American word-processing companies.

Corpus del Español

The *Corpus del Español* contains 100 million words collected from various Spanish texts from the 1200s to 1900s. The corpus was created at the Illinois State University, USA, with funding from the National Endowment for the Humanities, USA.

- Initially, the corpus contained about 45 million words collected from various texts written in Spanish. At later stages, it was enlarged to form a 'full text' of 100 million words divided equally among the texts of literature, spoken texts, fiction, newspaper, and academic texts spanning the following chronological scale:
    - 20 million words from texts between the years 1200 and 1400;
    - 40 million words from texts between the years 1400 and 1700;
    - 40 million words from texts between the years 1700 and 1900;
    - 20 million words from texts between the years 1900 and 2000.
- Each word collected in the corpus is supplied with information regarding the frequency of its use in each century as well as its register variation in modern Spanish texts. The corpus is linked with several databases containing information about parts of speech and the lemma of Spanish words. The linked-up databases also contain annotations regarding synonyms and the etymology of the words.
- The unique aspect of the corpus lies in its use of several relational databases that contain annotation schemes for close interactions among different links for text and word processing.
- The unique network system underlying the corpus processing software makes the corpus a powerful resource enriched with various operational and active search engines.

The Spanish Syntax Research Group at the University of Santiago de Compostela has developed another corpus of 1.5 million words of modern Spanish texts along with a syntactic database of 160,000 analyzed clauses. This group is also in the process of developing a corpus of medieval and classical Spanish texts to be added to their existing corpus.

COSMAS Corpus

The *COSMAS Corpus* of the modern German language stores more than 20 million words from running texts of various types. Due to restrictions imposed by the copyright policy of publishers, the corpus is available to common people in a restricted and limited version (nearly 11 million running words).

- The corpus is supported by several data search engines and text analysis tools that are capable of addressing the various needs of the corpus users. Information obtained from the corpus is used for the analysis of modern German from various perspectives and for designing language resources like dictionaries and teaching materials.

SUZANNE Corpus

The *SUZANNE Corpus* was developed with a sponsor from the Economic and Social Research Council, UK as a part of developing a comprehensive and fully explicit annotation scheme for the grammatical structures of English texts.

- The corpus, a by-product of the scheme of annotation, contains annotations of nearly 130,000 words collected from written American English of a subset of the Brown Corpus. After the completion of annotation in 1992, the corpus was globally released for open access and utilization in any kind of linguistic research and investigation.
- Because of its detailed and reliable annotation, the corpus is utilized in a great deal of gratifying research and development work in English grammar across the world.

Child Language Data Exchange System

The *Child Language Data Exchange System (CHILDES Database)* systematically includes a lot of transcribed texts collected from writings of children and adults who are learning English either as a first or second language.

- The corpus is made up of text samples of several small corpora mentioned below:
  - English corpora (obtained from texts written by normal English-speaking subjects);
  - English lexical databases with morphological tags (words are disambiguated by the part-of-speech tagging program);
  - Bilingual corpora (obtained from writings of bilingual and second language learning subjects);
  - Clinical corpora (obtained mainly from texts related to clinical subjects);
  - Frog story corpora (where narratives are elicited by using Mercer Mayer's Frog Story picture book);
  - Narrative corpora (where narratives are elicited with other pictures and stories);
  - German language corpora (obtained from the texts composed by normally developing children who are learning various Germanic and Nordic languages);
  - Romance language corpora (acquired from texts written by normal subjects learning a Romance language);
  - Other language corpora (obtained from texts composed by normal children learning various other languages).
- Several research centers investigating the nature and problems of child language acquisition use this corpus quite extensively to identify error patterns in language acquisition and generation in first and second-language learning as well as for providing necessary remedies for the removal of errors.

COMPARA Corpus

The bidirectional *COMPARA Corpus* contains a large collection of Portuguese–English and English–Portuguese texts, as well as their translations. The corpus designers have made no prior decision concerning the kinds of source texts and translations to be included in it. Although initial efforts have been made for assembling a corpus of published fiction, it has included other genres as well.

- So far, it is successful in terms of including extracts from contemporary and non-contemporary fiction composed by native authors and translated by people from Angola, Portugal, Mozambique, Brazil, South Africa, the UK, and the USA. To date, nearly 62 different Portuguese–English text pairs are included while new text pairs are being added regularly. However, the text pairs are yet to be processed to become machine-readable and searchable.
- This corpus was designed with specific goals. Researchers use this corpus to study translations and to compare and contrast countless different features of English and Portuguese.
- Professional translators and students of translation find it useful to discover how different words and expressions were translated in the past, and translation teachers use it in the classroom to tackle specific difficulties faced in Portuguese–English translation tasks.
- Portuguese learners of English and English learners of Portuguese also use it to study how similar meanings are expressed in the two languages, while language teachers can use it to create teaching materials for their students.

French–Norwegian Parallel Corpus

The *French–Norwegian Parallel Corpus* is developed at Bergen University, Norway with several Norwegian original texts available in French translation.

- In addition to the original Norwegian and French texts, translations of these texts are included in the corpus. The number of texts in the corpus is 30 pairs, approximately.
- During the last part of the project, several translations of English originals to German, Dutch, and Portuguese were added to the original corpus. Thus, some of the texts have been made available in several European languages (i.e., French, Norwegian, German, Dutch, English, and Portuguese).
- The corpus is used as a highly reliable resource for multilingual studies. To extend it to other European languages, people working on the project have started compiling translations of several imaginative and informative texts from 10 different languages used in Europe into French texts.

Estonian Corpus of Written Texts

The *Estonian Corpus of Written Texts* contains a large collection of samples from national newspapers (17.5%), official documents (1.2%), general essays and bibliographies (9.0%), hobbies (7.5%), fiction and stories (25.0%), encyclopedia (2.0%), propaganda (6.0%), popular science (15.0%), religion (0.8%), and natural science and engineering (16.0%). The corpus is available both in untagged and tagged versions for works related to mainstream linguistics, language teaching, dictionary compilation, grammar writing, and primer preparation. In addition, it is available for research and development works in language technology and computational linguistics.

Conclusion

The above corpora showcase the kind of formation and content of some of the widely known text corpora developed in English and other languages after the use of the computer in corpus generation.

We have referred to some big and small corpora available today with a focus on their formation, content, and utilization. All these activities show that with the presently available computer technology, corpus-building work is no more the capital-intensive enterprise that it was a few decades ago. Therefore, it is no longer a prized privilege entertained by a few well-funded institutions or organizations.

The discussion presented here clearly shows that a lot of activities are underway for assembling language corpora of various types for most of the languages of the world. Moreover, with the generation of bilingual and multilingual corpora in several languages, new horizons in bilingual and multilingual research and comparison are opening up.

The new intellectual inputs and corpus-designing principles remain as important as they were when the corpus-building enterprises started more than half a century ago. But the basic practicalities of assembling large corpora have become far less daunting now. It is now feasible for an individual to generate a corpus for language-specific and research-specific purposes. Therefore, along with large, general-purpose corpora, we find thousands of small, narrowly focused corpora developed across languages to serve the needs of linguistics and people.