Brown Corpus

-Dishi Gupta

Introduction

The Brown Corpus was the first million-word electronic corpus of English. Compiled by *W.N. Francis and H. Kucera*, Brown University, the corpus consists of one million words of American English texts printed in 1961. The texts for the corpus were sampled from 15 different text categories to make the corpus a good standard reference. Today, this corpus is considered small, and slightly dated. However, the corpus being 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.

This Standard Corpus of Present-Day American English (the Brown Corpus) consists of *1,014,312 words* of running text of edited English prose printed in the United States during the calendar year 1961. Although all of the material first appeared in print in the year 1961, some of it was undoubtedly written earlier.

The Corpus is divided into 500 samples of 2000+ words each. Each sample begins at the beginning of a sentence but not necessarily of a paragraph or other larger division, and each end at the first sentence ends after 2000 words. The samples represent a wide range of styles and varieties of prose.

The selection procedure was in two phases: an initial subjective classification and a decision as to how many samples of each category would be used, followed by a random selection of the actual samples within each category.

The list of main categories with their principal subdivisions and the number of samples in each are given below:

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**	

Versions of The Corpus

The Corpus is available in six versions. All contain the same basic text, but they differ in typography and format.

1) ***Form A***. This is the original form of the Corpus, as it was prepared in 1963-64. The limitations of computer printing facilities at that time required that it use an elaborate coding procedure.
2) ***Form B***. This is the 'stripped' version, from which all punctuation symbols and codes except hyphens, apostrophes, and symbols for formulas and ellipses have been omitted.
3) ***Form C***. This is the '*tagged*' version, which makes use of a partially stripped text in which only proper name capitalization and those punctuation marks which are of grammatical significance have been retained.
4) ***Bergen Form I***. This version and the following were prepared at the Norwegian Computational Center for Humanistic Research at the University of Bergen under the direction of Dr. Jostein Hauge

5) ***Bergen Form II***. In this version, typographical information is somewhat reduced and a new longer line is used. This version is available on microfiche, together with a complete KWIC concordance.
6) ***Brown MARC Form***. This version was prepared at Stanford University. It is designed to be compatible with two commonly used research techniques that are appropriate for large textual corpora:
    i. searching for and retrieving full-sentence citations using single words or word + context as retrieval criteria;
    ii. generating KWIC-form concordances which can be organized according to varying arrangements of a keyword and its preceding or following verbal context.

Coding Procedure of Form A

The basic coding procedure followed in Form A is that devised for the U. S. Patent Office, described in the pamphlet *A Notation System for Transliterating Technical and Scientific Texts for Use in Data Processing Systems,* by Simon M. Newman, Rowena W. Swanson, and Kenneth Knowlton. This was the only complete coding system that could be found in 1963.

Text. The text normally occupies spaces 1-70 of each line, running continuously from line to line without regard to word endings.
Headings and Paragraph Divisions. The heading of the largest subdivision of the text that falls within the sample is called the 'mayor heading'. No indications of capitals, italics, and other graphic features are made within headings.
Special Types. A passage in italics is marked by the beginning symbol *= and the closing symbol *$. If the italicized passage is smaller than a word, these symbols are included in the word without spacing; thus, in*cred*ible would be coded IN*=CRED*$IBLE.
Abbreviations. The period marking an abbreviation is coded as **., to distinguish it from the ordinary sentence-ending period. Abbreviations not marked by a period are treated as *symbols*.
Symbols. Combinations of letters without a following period (except at the end of a sentence) and not constituting a genuine word are defined as symbols and are preceded by the code marker **J.
Formulas. Combinations of letters, numbers, and other symbols which also include operator symbols (such as +, =, exponents, and subscripts) are defined as formulas and are replaced by the code **F.
Numbers. Numbers that are not part of formulas are reproduced normally, including the decimal point, which is distinguished from the period by the fact that it is not followed by space.
Quotations. The Patent Office procedure of placing all punctuation marks after the close-quote symbol was followed. This leads to ambiguity, in that the distinction between an exclamation point or question mark inside a quotation and one outside it is lost.
Typographical Errors and Inconsistencies. No alterations have been made to the original text, even in the case of obvious typographical errors, misspellings, typographical inconsistencies, etc.

The Tagged Version

In the tagged version of the Corpus (Form C), each word is furnished with a brief tag that assigns it to a specific word class. There are 82 of these tags, which are of six kinds:

(a) major form classes (Parts of Speech);
(b) function words;
(c) certain important individual words: *not*, existential *there,* infinitival *to;*
(d) punctuation marks of syntactic significance;
(e) inflectional morphemes, notably noun plural and possessive;

The tagging of the Corpus has been a long and arduous process, extending over several years and involving quite a few different people.

List of Tags

Tag	Description	Examples	Tag	Description	Examples	Tag	Description	Examples	Tag	Description	Examples
.	sentence closer	. ; ? !	JJS	semantically superlative adjective	*chief, top*	HVN	*had* (past participle)		VBD	verb, past tense	
(	left parenthesis		JJT	morphologically superlative adjective	*biggest*	HVZ	*has*		VBG	verb, present participle/gerund	
)	right parenthesis		MD	modal auxiliary	*can, should, will*	IN	preposition		VBN	verb, past participle	
*	negation words	*not, n't*	NC	cited word (hyphenated after regular tag)		JJ	adjective		VBZ	verb, 3rd. singular present	
--	dash		NN	singular or mass noun		JJR	comparative adjective		WDT	*wh-* determiner	*what, which*
,	comma		NN$	possessive singular noun		WPS	nominative *wh-* pronoun	*who, which, that*	WP$	possessive *wh-* pronoun	*whose*
:	colon		NNS	plural noun		WQL	*wh-* qualifier	*how*	WPO	objective *wh-* pronoun	*whom, which, that*
ABL	pre-qualifier	*quite, rather*	NNS$	possessive plural noun		BEZ	*is*		PPL	singular reflexive/intensive personal pronoun	*myself*
ABN	pre-quantifier	*half, all*	NP	proper noun or part of name phrase		CC	coordinating conjunction	*and, or*	PPLS	plural reflexive/intensive personal pronoun	*ourselves*
ABX	pre-quantifier	*both*	NP$	possessive proper noun		CD	cardinal numeral	*one, two, 2, etc.*	PPO	objective personal pronoun	*me, him, it, them*
AP	post-determiner	*many, several, next*	NPS	plural proper noun		CS	subordinating conjunction	*if, although*	PP$	3rd. singular nominative pronoun	*he, she, it, one*
AT	article	*a, the, no*	NPS$	possessive plural proper noun		DO	*do*		PPSS	other nominative personal pronoun	*I, we, they, you*
BE	*be*		NR	adverbial noun	*home, today, west*	BED	*were*		NRS	plural adverbial noun	
BER	*are, art*		PP$$	second (nominal) possessive pronoun	*mine, ours*	BEDZ	*was*		OD	ordinal numeral	*first, 2nd*
DOD	*did*		QL	qualifier	*very, fairly*	BEG	*being*		PN	nominal pronoun	*everybody, nothing*
DOZ	*does*		QLP	post-qualifier	*enough, indeed*	BEM	*am*		PN$	possessive nominal pronoun	
DT	singular determiner	*this, that*	RB	adverb		BEN	*been*		PP$	possessive personal pronoun	*my, our*
DTI	singular or plural determiner/quantifier	*some, any*	RBR	comparative adverb		DTX	determiner/double conjunction	*either*	RN	nominal adverb	*here then, indoors*
DTS	plural determiner	*these, those*	RBT	superlative adverb		EX	existential *there*		RP	adverb/particle	*about, off, up*
HVD	*had* (past tense)		UH	interjection, exclamation		FW	foreign word (hyphenated before regular tag)		TL	word occurring in title (hyphenated after	
HVG	*having*		VB	verb, base form		HL	word occurring in headline (hyphenated after regular tag)			regular tag)	
WRB	*wh-* adverb	*how, where, when*	HV	*have*		TO	infinitive marker *to*				

The Noun Phrase*.* The model for this consists of a head preceded by a determiner sector and a modifier sector. The center of the determiner sector is the determiner itself, of which three basic kinds are recognized: articles, *a/an , the,* tagged *AT*.
The tagging system makes provision for three kinds: *adjectives, participles, and nominals*. The problem lies in the fact that by compounding (open, hyphened, or closed), suffixation, or simple adjunction.

The Verbal Phrase*.* Verbs in the base form, regardless of syntactic function, are tagged *VB*. Modal auxiliaries, regardless of tense, are all tagged *MD*.

Pronouns*.* Personal pronouns have tags beginning with *PP*, followed by one or more letters indicating case, concord, and sometimes number.

Adverbials*.* The general tag for adverbs is *RB*, with *RBR* and *RBT* for inflectional comparatives and superlatives.

Connectives*.* Coordinating conjunctions (*and, or,* etc.) are tagged *CC* and subordinators *(since, because, if*) *CS*. Prepositions are tagged *IN*. The word to is tagged *TO* when used as the infinitive marker.

Miscellaneous Items*.* The existential subject *there is* tagged *EX* and thus distinguished from the homonymous adverb. Exclamations of various sorts, which have no syntactic function, are tagged *UH*; they occur mostly in the dialogue of the fictional samples.

Capitalized Words, Titles, and Proper Nouns. The conventions of (non-sentence-initial) capitalization in English are complex and to a considerable degree variable, unlike most other aspects of the writing system. The aim has been to identify capitalized uses as much as possible but also to associate capitalized words with their lower-case alternatives.

Tools to work with the Brown Corpus

A complete set of tools is available to work with the Brown corpus online (without registration) to generate:

- word sketch– English collocations categorized by grammatical relations
- thesaurus– synonyms and similar words for every word
- keywords– terminology extraction of one-word and multi-word units
- word lists – lists of English nouns, verbs, adjectives, etc., organized by frequency
- n-grams– frequency list of multi-word units
- concordance – examples in context

Availability

Access to the corpus is freely available for research.

How to access the raw documents from the Brown corpus using NLTK?

ID	File	Genre	Description
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*
F25	cf25	lore	Boroff: *Jewish Teenage Culture*
G22	cg22	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	ch15	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	cj19	learned	Mosteller: *Probability with Statistical Applications*
K04	ck04	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	cl13	mystery	Hitchens: *Footsteps in the Night*
M01	cm01	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	cn15	adventure	Field: *Rattlesnake Ridge*
P12	cp12	romance	Callaghan: *A Passion in Rome*
R06	cr06	humor	Thurber: *The Future, If Any, of Comedy*

Example Document for Each Section of the Brown Corpus

The above table gives an example of each genre of the Brown Corpus (for a complete list, see List of Samples).

As just mentioned, a text corpus is a large body of text. Many corpora are designed to contain a careful balance of material in one or more genres.
We can access the corpus as a list of words, or a list of sentences (where each sentence is itself just a list of words) by importing `brown` package from the `nltk` package. We can optionally specify particular categories or files to read:

```
>>> from nltk.corpus import brown
>>> brown.categories()
['adventure', 'belles_lettres', 'editorial', 'fiction',
'government', 'hobbies',
'humor', 'learned', 'lore', 'mystery', 'news', 'religion',
'reviews', 'romance',
'science_fiction']
>>> brown.words(categories='news')
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
>>> brown.words(fileids=['cg22'])
['Does', 'our', 'society', 'have', 'a', 'runaway', ',', ...]
>>> brown.sents(categories=['news', 'editorial', 'reviews'])
[['The', 'Fulton', 'County'...], ['The', 'jury', 'further'...],
...]
```

The Brown Corpus is a convenient resource for studying systematic differences between genres, a kind of linguistic inquiry known as **stylistics**. Let's compare genres in their usage of modal verbs. The first step is to produce the counts for a particular genre. Remember to `import nltk` before doing the following:

```
>>> from nltk.corpus import brown
>>> news_text = brown.words(categories='news')
>>> fdist = nltk.FreqDist(w.lower() for w in news_text)
>>> modals = ['can', 'could', 'may', 'might', 'must', 'will']
>>> for m in modals:
...     print(m + ':', fdist[m], end=' ')
...
can: 94 could: 87 may: 93 might: 38 must: 53 will: 389
```

Note: We need to include end=' ' in order for the print function to put its output on a single line.

Next, we need to obtain counts for each genre of interest. We'll use NLTK's support for conditional frequency distributions. Observe that the most frequent modal in the news genre is *will*, while the most frequent modal in the romance genre is *could*.

```
>>> cfd = nltk.ConditionalFreqDist(
...           (genre, word)
...           for genre in brown.categories()
...           for word in brown.words(categories=genre))
>>> genres = ['news', 'religion', 'hobbies', 'science_fiction',
'romance', 'humor']
>>> modals = ['can', 'could', 'may', 'might', 'must', 'will']
>>> cfd.tabulate(conditions=genres, samples=modals)
```

```
                  can could  may might must will
            news   93   86   66   38   50  389
        religion   82   59   78   12   54   71
         hobbies  268   58  131   22   83  264
science_fiction   16   49    4   12    8   16
         romance   74  193   11   51   45   43
           humor   16   30    8    8    9   13
```

The actual brown corpus data is **packaged as raw text files**. And you can find their IDs with:

```
len(brown.fileids()) # 500 sources, each file is a source.
print(brown.fileids()[:100]) # Prints first 100 sources.
```

You can access the raw files with:

```
print(brown.raw('cb01').strip()[:1000]) # First 1000 characters.
```

You will see that each word comes with a slash and a label and unlike normal text, we see that punctuations are separated from the word that comes before it, e.g.,

> The/at General/jj-tl Assembly/nn-tl ,/, which/wdt adjourns/vbz today/nr ,/, has/hvz performed/vbn in/in an/at atmosphere/nn of/in crisis/nn and/cc struggle/nn from/in the/at day/nn it/pps convened/vbd ./.

References:

- Brown Corpus Manual
- Brown Corpus
- Accessing Text Corpora and Lexical Resources
- Basic NLP with NLTK
- NLTK Sample Usage for Corpus