Using The Century Of Prose Corpus

by Louis T. Milic [1]
Cleveland State University

The Century of Prose Corpus (COPC) is one of a number of compilations of texts that have been developed during the last three decades to facilitate a certain kind of linguistic analysis with computers. Unlike the Corpus Thomisticus (the whole of the works of Thomas Aquinas), for example, the kind of corpus I am talking about is a descendant of the Brown Corpus, devised in 1961 by Henry Kucera and Nelson Francis of Brown University. The Brown Corpus consists of a million words of edited American prose, all published during that one year and taken from a great variety of kinds and genres of printed materials, from humorous fiction to articles in learned journals. The devisers assumed that their corpus was large enough to represent nearly every type of linguistic unit that might be of interest to scholars. Although it was intended to be machine-readable, it has generated two large volumes of analysis and documentation in which alphabetic and rank-ordered word-lists provide a view of the American vocabulary at that period, among many other valuable pieces of information about the language, to say nothing of the other areas of knowledge that are served by this work.

It would not be inaccurate to compare the Brown Corpus and other corpora that have sprung up since to anthologies, such as those that serve as textbooks in courses in literature, history and other fields. The anthology is more than anything else a sample of the writing of a field or period, representative and typical of the totality of the population. Of course, it is not a statistical sample because of its preference for the best, the best-known, the most influential..., but it is a sample nonetheless. Someone who has read through an anthology has a grip on the writing and thinking, the preoccupations of a genre, time period, nation... Similarly, anyone who had been living on another planet during 1961 and on his return read through the million words of the Brown Corpus would have a pretty complete idea of what went on during that year. But of course that was not the intention of Francis and Kucera: their compilation was primarily a tool for research in language. Their Corpus gave rise to similar ones of Spoken English and of British English. But beyond that, it led others to create more specialized corpora. The COPC is one such.

The COPC is intended to represent a norm for the study of the English of Britain during the eighteenth century. Its actual delimitations are the years 1680-1780 and its dimensions are approximately 500,000 words. It is composed of two parts: A) the major authors:

Addison, Berkeley, Bolingbroke, Boswell, Burke, Chesterfield, Defoe, Dryden, Fielding, Gibbon, Goldsmith, Hume, Johnson, Locke, Adam Smith, Smollett, Steele, Swift, Temple, Walpole;

B) the 100 background writers.

Part A (the major authors) contains 15,000 words from each of the twenty most prominent authors in three selections of 5,000 words each drawn from various stages of each author's production. This part totals 300,000 words or 60% of the Corpus. Part B can best be visualized as a ten by ten matrix, in one dimension representing decades of years:

1=1680-1689 2=1690-1699 3=1700-1709 4=1710-1719

5=1720-1729 6=1730-1739 7=1740-1749 8=1750-1759

9=1760-1769 0=1770-1779.

and in the other ten different genres:

1 Biography (A)	6 History (G)
2 Periodicals (B)	7 Memoirs/Letters (H)
3 Education (D)	8 Polemics (K)
4 Essays (E)	9 Science (N)
5 Fiction (F)	10 Travel (Q)

It will be noticed that there is a blending here of genre and subject matter, which can be rationalized by the claim that subject matter dictates conventions that amount to genre.

There is a text of 2000 words in each cell. Consequently there are ten selections (20,000 words) for each decade (one from each genre) and ten for each genre (one from each decade), the whole consisting of one hundred selections of 2000 words each or 200,000 in all, 40% of COPC.

Each sentence of each text is identified by means of a header block. An excerpt of one of the Part B texts follows:

5N03(1728)0001/021-P1 Language is a set of words which any people have agreed upon, in order to communicate their thoughts to each other.
5N03(1728)0002/079-P0 The first principles of all languages, Buffier observes, may be reduced to expressions signifying first the subject spoken of; secondly the thing affirmed of it; thirdly the circumstances of the one and the other: but as each language has its particular ways of expressing each of these; languages are only to be looked on as an assemblage of expressions, which chance or caprice has established among a certain people; just as we look on the mode of dressing, etc.

The header block *5N03(1728)0001/021-P1* is analyzed thus:

5N03: identifier of text (decade 5, genre N, accession no. 03)

1728: date of publication

0001: sentence number 1 of selection

021: number of words in sentence

P1: sentence begins a paragraph.

The entire Corpus holds on three high-density 3 1/2" diskettes (or on tape) and may soon be available on CD. It can be used on mainframes or on 386-type personal computers. In its present form, it requires the user to have access to a program package (such as EYEBALL, ARRAS, Word Cruncher...) or to be able to program in a string-manipulation language, such as SNOBOL or one of its derivatives (e.g., SPITBOL). I have devised several programs with which I have analyzed the various texts for later statistical treatment. I shall mention two of these.

The LETTER program performs the following:

1. Counts the length of each sentence
2. Produces a sequential list of sentence lengths
3. Calculates and prints
   a. mean sentence-length in words
   b. standard deviation of the sentence-length
4. Displays a frequency distribution of the letters in the text, both raw scores and percentage
5. Displays the rank-order of the letters according to frequency, compared to the Brown Corpus and other corpora
6. Displays a frequency distribution of word sizes

7. Displays frequency distributions of word-initial and word-final letters for words greater than five letters in length
8. In a summary, provides the following:
   a. total words
   b. hyphenated words
   c. number of sentences
   d. number of interrogative sentences
   e. net number of letters in the text
   f. calculated vowel-consonant ratio
   g. mean sentence length
      1. in letters
      2. in words (by a method different from 1)
   h. mean word-length in letters.

The INDEXER program does the following:

1. Alphabetical word-index with raw frequencies
2. Rank-ordered word-index of the 100 most frequent lexemes, with raw frequencies and percentages
3. Counts
   a. tokens
   b. types
   c. hapax legomena
4. Calculates
   a. type-token ratio
   b. hapax-token ratio
   c. hapax-type ratio.

And of course, the programs may be applied not only to individual texts, but to groups, to decades, genres, Parts and to the whole corpus.

As can easily be noticed, these two programs alone acting on each of the texts in COPC generate a very substantial amount of data which can be analyzed or treated in a number of ways. To illustrate one possibility out of many, I shall follow Newton's principle about the relation of data and hypotheses:

> For the best and safest method of philosophizing seems to be, first diligently to investigate the properties of things and establish them by experiment, and then to seek hypotheses to explain them. For hypotheses ought to be fitted merely to explain the properties of things and not attempt to predetermine them...

Inspection of the data - that is, the texts themselves and the output of the programs - had led me to observe that writings in the same genre showed a consistent use of certain variables. In order to examine this possibility, I organized the data of Part B into ten variables for each of the hundred texts in it and analyzed this by means of the SPSS statistical data analysis package. Although the number of variables is of course unlimited, I chose ten more or less at random. These variables consist of two sets: "standard" and arbitrary. The five standard variables (often found in the literature):

1. mean sentence-length (MSL)
2. mean word-length (MWL)
3. number of types (TYP)
4. number of hapax (HAP)
5. percentage sum of five most frequent function words (FW)

The five arbitrary variables are:

6. frequency sum of the letters "t," "i," and "o" (LET).

7. frequency of the letter "s" in final position (SFIN).
8. frequency of the letter "d" in final position (DFIN).
9. sum of the two most frequent function words (TOP2).
10. number of nouns in ranks 1-54 of each selection (NN).

The Correlation procedure in SPSS produces Pearson correlation coefficients for each pair of variables, when these have been arrayed in an appropriate form, as follows:

Text	MSL	MWL	LET	SFIN	DFIN	TYP	HAP	FW	TOP2	NN
6Q45	29.82	4.48	24.17	22.20	12.64	695	455	21.64	13.85	1
8Q16	40.86	4.65	23.12	17.79	15.06	762	534	19.04	10.14	7
0B90	38.63	4.49	24.18	15.56	18.63	846	615	19.64	9.90	3
0N74	62.47	4.85	24.63	28.49	10.53	737	510	21.65	12.25	10
4D83	45.52	4.70	24.70	23.54	15.32	637	385	18.97	10.23	6
8K78	34.02	4.55	25.80	17.46	13.73	683	440	19.12	10.59	6
6K82	46.88	4.54	25.25	20.89	9.93	578	339	20.30	9.68	8
9B73	38.32	4.70	24.78	24.11	11.17	740	495	20.88	11.57	7
9K98	34.00	4.64	24.40	25.08	14.98	627	395	19.79	11.62	9
2N09	34.50	4.62	23.55	27.04	12.07	669	442	21.90	13.10	7

The Pearson correlations are as follows:

Positive	.01		.001	
MSL-FIN	.24		MWL-SFIN	.46
MSL-FW	.29		MWL-TYP	.31
MWL-HAP	.28		MWL-FW	.55
SFIN-TYP	.29		MWL-TOP2	.59
SFIN-HAP	.28		MWL-NN	.46
SFIN-FW	.24		FW-NN	.43
SFIN-TOP2	.29		TOP2-NN	.46
Negative				
TYP-NN	-.28		LET-TYP	-.31
HAP-NN	-.30		LET-HAP	-.29

As can be easily seen, a good number of these are quite significant, some at the one percent, some at the next level, mostly positive, although a few are negative. Of course, some of the correlations are significant but meaningless, as they represent merely functional relationships, e.g., types and hapax, function words and "top two". But others suggest something factual and possibly important about the relationship of genre to the quantitative fabric of texts. To look into the possibilities of this relationship, we must go deeper and discover which genres select which variables. By subjecting the data to analysis of variance (ANOVA), we find the pattern in the following matrix:

	1	2	3	4	5	6	7	8	9	10
VariableA	B	D	E	F	G	H	K	N	Q	
MSL	-	+	-		+		+	+		-
MWL					-	+	-		+	
SFIN	-				-	+	-		+	
DFIN	+		-	-	+			-	-	+
LET	-				-			+		-
TYP	+	+	-	+	-	-		-	-	+
HAP	+	+	-	+		-		-	-	+
FW	+	-		-	+			+		
TOP2		+	-		-	+			+	
NN	+	-	+		-	+	-		+	-
Total	7	6	7	[3]	9	7	[4]	5	8	6

It is plain that certain genres select more significant variables than do others. Numbers 4 and 7 (essays and memoirs/ letters) seem less distinct than the others. Numbers 5 and 9, on the other hand (fiction and science), are much more distinctive.

Following Newton's recommendation, therefore, we are free next either to devise hypotheses about these relationships or try new experiments to deepen our understanding. A possible explanation might be that the conventions of fiction and science writing are much more strict than those of essays, memoirs or letters, and that this strictness manifests itself at the quantitative microlinguistic level. Another might be that the term "genre" is not as rigorous or as easily defined as is generally believed. At any rate, to feel confident about such hypotheses would require further analysis of factors by means of regression or other advanced statistical techniques.

This simple illustration is only intended to reveal a small fraction of the immense possibilities for study and research that are latent in a carefully constructed corpus of substantial size and extent.

[1] Presented at the IASSIST 92 Conference held in Madison, Wisconsin, U.S.A. May 26 - 29, 1992.