Linguistic Interpretation of English Literary Works

— Contextual Information in Francis Bacon's Essays

秦平新 编著

英国文学作品的语言学解读

——培根随笔语篇信息研究

Linguistic Interpretation of English Literary Works

中国矿业大学出版社

China University of Mining and Technology Press

Linguistic Interpretation of English Literary Works
—Contextual Information in Francis Bacon's Essays

英国文学作品的语言学解读
——培根随笔语篇信息研究

秦平新　编著

中国矿业大学出版社

内 容 提 要

《英国文学作品的语言学解读——培根随笔语篇信息研究》运用语料库语言学的最新研究成果，基于计算机软件的信息抽取功能对英国著名的哲学家、近代实验科学的创始人弗朗西斯·培根(Francis Bacon 1561—1626)随笔《人生论》进行语篇信息解读。主要内容包括：一、英文文本；二、语篇分析：文本总体统计特征，包括文本总的字数，句子数量，音节数，每个单词的平均字母数，音节数，每个句子的平均字数，文本的可读性统计数据，包括难词、长词、词汇密度等统计信息；三、词汇信息：主要包括词频统计，词汇分布特征，词长统计，高频词统计，主题词信息等内容；四、高频词的语料库对比分析等。基于计算机辅助的文学作品语篇信息解读目前只散见于零星的学术论文，而运用语言学的方法系统分析、解读某一完整的文学作品在国内尚属首例。

图书在版编目(CIP)数据

英国文学作品的语言学解读：培根随笔语篇信息研究 /
秦平新编著. —徐州：中国矿业大学出版社，2009.9
ISBN 978-7-5646-0422-6

Ⅰ.英… Ⅱ.秦… Ⅲ.计算机应用—随笔—文学语言—
文学研究—英国—近代 Ⅳ.I561.076-39

中国版本图书馆 CIP 数据核字(2009)第 139299 号

书　　名　英国文学作品的语言学解读——培根随笔语篇信息研究
编　　著　秦平新
责任编辑　周　红　潘俊成
责任校对　於娟娟
出版发行　中国矿业大学出版社(江苏省徐州市中国矿业大学内　邮编 221008)
营销热线　(0516)83885307　83884995
网　　址　http://www.cumtp.com　**E-mail**:cumtpvip@cumtp.com
排　　版　中国矿业大学出版社排版中心
印　　刷　徐州中矿大印发科技有限公司
经　　销　新华书店
开　　本　787×1092　1/16　**印张** 24.75　**字数** 615 千字
版次印次　2009 年9月第 1 版　2009 年9月第 1 次印刷
定　　价　36.00 元
(图书出现印装质量问题，本社负责调换)

前　言

我国语料库的建设始于20世纪80年代，当时的主要目标是汉语词汇统计研究。进入90年代以后，语料库方法在自然语言信息处理领域得到了广泛的应用，建立了各种类型的语料库，研究的内容涉及语料库建设中的各个方面。90年代末到新世纪初是语料库开发和应用的进一步发展时期，除了语言信息处理和语言工程领域以外，语料库方法在语言教学、词典编纂、现代汉语和汉语史研究等方面也得到了越来越广泛的应用。

本书语料库通常指为语言研究收集的、用电子形式保存的语言材料，由自然出现的书面语或口语的样本汇集而成，用来代表特定的语言或语言变体。经过科学选材和标注、具有适当规模的语料库能够反映和记录语言的实际使用情况。人们通过语料库观察和把握语言事实，分析和研究语言系统的规律。语料库已经成为语言学理论研究、应用研究和语言工程不可缺少的基础资源。

本书运用语料库语言学的最新研究成果，基于语料库的信息检索功能对英国著名的哲学家、近代实验科学的创始人弗兰西斯·培根(Francis Bacon，1561—1626)随笔《人生论》文本信息进行基于语料库的语言学解读。主要内容包括：一、英文文本；二、文本分析：文本总体统计特征，包括文本总的字数，句子数量，音节数，每个单词的平均字母数，音节数，每个句子的平均字数、文本的可读性统计数据，包括难词，长词，词汇密度等统计信息；三、词汇信息：词频统计，词汇分布特征，词长统计，高频词统计，主题词信息等内容。基于语料库的文学作品语篇信息解读目前只散见于零星的论文，而系统分析、解读某一完整作品在国内尚属首例。

通过语篇信息的解读对作品的文体特征，修辞特点，作者的写作风格，词汇使用的倾向性以及文本结构等会有一个相当全面的了解，有助于更好地理解作品的内容，写作特色等。无论是从欣赏经典文学的角度，还是用于阅读和写作教学，本书都不失为一本有益的颇具工具性价值的书籍，细心研读，相信读者一定会受益匪浅。

本书的主要读者对象是大学英语专业高年级学生和对运用语料库进行语篇分析具有一定了解且对外国文学感兴趣的读者。书中所介绍的内容是对文学作品的文本信息进行较为基础性的分析。读者通过本书的学习可以对运用语料库语言学的方法解读文本信息有一个基本的了解，为以后深入学习和研究奠定一定的基础。

笔者自2002年开始对语料库语言学发生兴趣，并一直关注和从事语料库语言学的相关研究，也一直尝试基于语料库的外语教学探索，包括基于语料库的词汇研究，翻译研究以及

语篇分析研究等。在教学过程中发现，运用语料库方法对文学作品进行文本分析相对于传统的教学方法而言具有独特的、不可替代的优势，即通过量化数据挖掘文本信息，这也是目前文学研究的一个趋势，这种研究方法科学，易于操作，并且结论可靠，属实证性研究，这也是语言学研究的基本要求和发展趋势。

在英国文学作品里，培根的随笔独树一帜，一直为广大读者所喜爱，据不完全统计，仅译成中文的版本就不下十余种。英语专业的学生在接触培根随笔的英语原文时总感到学习起来有很大的难度，要求其对文本信息进行分析就更是难上加难，常常不知道如何着手进行。鉴于此种现象，笔者就萌发了写作本书的想法，通过本书向学生介绍一些运用语料库方法进行文学文本分析的基本方法、思路，为今后进一步研究奠定良好的基础。本书在编写上采用由易到难，循序渐进的方法，先由文本的基本统计特征入手，逐步向纵深过渡，让读者由浅入深逐步了解文本分析的整个过程，便于读者遵循一定的方法学习和模仿，最后达到熟练运用的目的。

限于篇幅，本书中尚有大量的信息无法囊括进来，比如文本的词频索引（Concordance）和文本信息检索数据、词频统计信息以及关键词的扩展语境等，这些内容对全面理解文本信息都有很好的帮助，不能不说是个遗憾。此外，本书中的内容仅只是文本分析的基本统计数据，只有定量的分析而缺乏定性的分析，也是一个缺憾，待今后修订时进一步完善。

本书在写作过程中，采用了 Bartleby 的版本，并参考了中国友谊出版公司的版本。本书是河南城建学院规划教材，在撰写和出版过程中得到了河南城建学院教材建设委员会、教务处教材科、科研外事处、中国矿业大学出版社等单位的大力支持。借此机会对给予本书的出版提供帮助的单位和个人一并深表谢意。

编者

2009年5月

Table of Contents

Preface

BRIEF INTRODUCTION OF TEXT ANALYSIS

I Methodological background

The following is an attempt briefly to sketch a methodology for elementary text-analysis, with particular emphasis on how to approach a text one does not know well.

A. Kinds of text-analysis

Throughout "text-analysis" should be taken to mean "the analysis of text with the aid of algorithmic techniques".

An *algorithm* may be defined as a step-by-step procedure capable of being run on a computer—i. e. , an unambiguous and completely stated description of what the computer is to do. It can be expressed by a computer program but need not be; often the specifics of how an algorithm is implemented in a particular programming language would obscure the essentials. Text-analytic methods cover a spectrum between the completely algorithmic and the *exploratory*: in exploratory work we do not have a specific goal or procedure to follow but instead we look for leads. Most work mixes approaches from various points in the spectrum: we may make a word frequency list by algorithmic methods, but the results always need to be interpreted and investigated further, usually by much less algorithmic means.

Text-analysis may be divided into the following kinds, usually practiced at different places along the algorithmic-exploratory spectrum:

- **Concording** and related transformations of the textual data. These constitute the primary focus of attention here.
- **Content analysis.** This is a closely related if not overlapping kind, often included under the general rubric of "qualitative analysis", and used primarily in the social sciences. It is "a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding" (Stemler

2002). It often involves building and applying a "concept dictionary" or fixed vocabulary of terms on the basis of which words are extracted from the textual data for concording or statistical computation.

- **Statistical analysis.** This involves counting particular features of the textual data and then applying one or more mathematical transformations. The simplest type produces frequency lists of word-forms, usually arranged from the most to the least frequent. We will pay some attention to such lists here. More powerful and complex types of statistical analysis are used for example in stylometry and authorship studies; see Buyyows 1992, Holmes 1998.

B. Application to unseen or poorly known texts

There are two reasons why one might legitimately be using text-analytic techniques on a text one does not know well. First, corpora of use in the humanities are approaching and some are already past the point at which a human being could read through their contents in a lifetime—especially given when that person might begin his or her reading; furthermore, some of these are not intended for normal reading, such as the non-literary collections meant for historical or linguistic purposes. Second, and more importantly, text-analysis is fundamentally different from manual methods and so reveals aspects of even well-known texts that one is likely not to have considered before. To the degree to which these texts are made new by the change in perspective, understanding will be aided by text-analytic techniques.

The first reason, that corpora tend to be too large, can be put in more positive terms: a good command of these techniques will make it practical for the ignorant but intelligent person to profit from materials outside his or her own field. Thus interdisciplinary research tends to be fostered.

II Prior knowledge

We assume, then, application of the first kind of analysis, concording, with some use of frequency lists, to unseen texts.

Nevertheless the place to begin is with whatever you know about the given body of text (known as the *corpus*). It is unlikely that you will know absolutely nothing at all about it, but in any case read around in it *briefly*, picking up what you can. Consider:

A. **Genre.** What kind of a text do you have? Novelistic, poetic, bureaucratic, legal? Was it originally written, or was it delivered orally? What are the formal features you would expect such a text to have, which can you spot when you look at it?

B. **Rhetoric and vocabulary.** Genre will tend to define a particular way of speaking or writing and to shape the vocabulary, including how frequently particular words appear.

C. **Social or psychological circumstances.** Familiarity with the social circumstances surrounding the creation of the text *may* be relevant; so also the known or suspected psychology of the author or speaker.

D. **Historical circumstances.** The more you know about the historical circumstances under which the text was produced the better. Awareness of the historically (and to a certain extent, regionally) defined vocabulary will give you hints as to where you might begin in a search for interesting terms.

E. **Nature of the artefact.** The physical object from which the text has been taken, usually a printed book, may be relevant. Stephen's book, *Monday Night Class*, gives several indications of its time and subculture of origin; likewise, the photographs included in it and on the back cover reinforce the historical fact of the seriousness with which his words were taken. These, again, give reason to press forward with the analysis, and the clearly religious character of the assemblies to which he spoke direct you to the corresponding language.

In other words, the seemingly disembodied electronic text has several contexts essential to a full understanding of it. The more of that understanding you can have the better, though because the focus here is on technique, the point is not to dwell on acquiring knowledge of the contexts, only to get what you can quickly.

III Steps in the analysis

The methodology outlined here is like a fishing expedition: you go at the text with a quiet, open mind, having little or no idea what you are going to catch. If you are after something in particular, then of course it is a different kind of activity. Even in a focused enquiry, however, software allows you to ask certain kinds of questions so easily and get answers back so quickly that curiosity is given a much freer reign; you can afford to play, ask even apparently improbable questions, and so raise the chances that you will be surprised by an important result you had little reason to expect. Thus a certain amount of fishing is recommended even for the focused questioner.

A. **High-frequency words.** A quite crude but useful technique is to look through a list of the most frequent word-forms for anything that is unusual or particularly characteristic of the text in question. Frequency of word-forms is only roughly related to what a text says, but it is related, and so is useful to work with.

Two examples spring to mind from both the Simpson and Stephen corpora: the verb "know" and the first-person singular pronoun "I". (Note, in the comparison study outlined in Corpus analysis of meaning, how so little information says so much about both, how it draws a contrastive parallel between the two men.)

There are of course severe limitations on what you can do with a frequency list, especially if you are interested in *words* (dictionary headwords, such as "know" or "I") rather than word-forms (such as "knows" or "knew", or "me" or "we"), and much more if you are focused on ideas (such as cognition or the self) rather than words. If the former, then you need to find all the inflected forms of the word and combine their frequencies. If the latter, you need to find all the relevant synonyms and combine the frequencies of all the inflected forms; even then, since ideas are only tangentially related to words, the result would be incomplete. Very often, however, the raw frequency list will prove useful enough.

B. Collocations. A somewhat more sophisticated tool for relating word-forms to meaning generates information on what words tend to be found together, either contiguously, such as "I didn't know that", or within a specified proximity or *span*, e. g. "black" within 5 words of "bag". The idea here is that repeated collocations are more reliable indicators of meaning that repetitions of single word-forms. See Sinclair 1991 (chapter 8) for a full discussion.

The program *Monoconc* and others will generate lists of collocations ordered by frequency so that you can identify recurring phrases and associations of words quickly. Note that if you wish to study collocations over a wider span than the program permits, you can do this by following these steps:

(1) Set the concordance "window" (the number of characters shown on either side of the target word) to a sufficiently large number;

(2) Run a concordance of the word for which you wish to study the collocates;

(3) Save the concordance as a text-file;

(4) Use that file as input to the program, generate from it a frequency listing.

This listing will thus give you the frequencies of the collocates of your target word.

A government document, for example, will tend to have quite high frequencies of standard phrases; for a literary work, even two occurrences of a phrase may be highly significant. The *Monoconc*-style listing, of collocations within a span, is of course less bound to literal repetition—it will include together, for example, instances of the collocation of "don't" and "know" in the phrases "I don't know" and "I don't even know".

C. Concording. The essential idea behind the concordance, especially the KWIC, is to direct your attention to the immediate linguistic environment of the specified word. Hence when you find a potentially interesting word, often the next step is to run a concordance on it, then look down the concordance listing to see what patterns you can spot. With *Monoconc* generating collocation statistics will often immediately follow.

A KWIC is made considerably more useful by the ability to sort an on-screen listing according to the words to the left and right of the target words; *Monoconc* offers such ability, and the same can be done with other concordance software. Such sorting tends to bring out the patterns, since repetitions are grouped together.

Since current KWIC software deals only with word-forms rather than words, you willoften also need to concord the inflected forms. In English many of these can be caught by use of the appropriate wildcards, but not all. An example is *go*, *went* and *gone*; another is *I*, *me*, *my*, *mine*, *we*, *us*, *our(s)*, all forms of the first-person personal pronoun.

Synonyms, of course, are entirely your task to identify, but doing so is made considerably easier than it might be by the tendency in many writers and speakers to emphasise an idea by using a number of synonyms together or nearby each other. Thus the text can itself help you to build a reasonable list for further concording. Compiling such a list is a recursive activity—in the beginning, a new synonym will tend to turn up others; when the law of diminishing returns asserts itself, it is time to stop. The result we may call a "fixed vocabulary", to which can be added the contiguous collocations you have identified. All together these represent a translation of an idea, as it were, into data.

A fixed vocabulary can then be used to turn up passages in the text for study—as is commonly done in "content analysis". If you know the text well, then a very interesting further question to ask is, when does this vocabulary not identify passages in which the targeted idea clearly or arguably occurs? Why does it not? Some very interesting findings can result from pursuit of this question.

(Source: Text Analysis Info)

Chapter 1

OF TRUTH

What is truth? said jesting Pilate, and would not stay for an answer. Certainly there be, that delight in giddiness, and count it a bondage to fix a belief; affecting free-will in thinking, as well as in acting. And though the sects of philosophers of that kind be gone, yet there remain certain discoursing wits, which are of the same veins, though there be not so much blood in them, as was in those of the ancients.

But it is not only the difficulty and labor, which men take in finding out of truth, nor again, that when it is found, it imposeth upon men's thoughts, that doth bring lies in favor; but a natural though corrupt love, of the lie itself. One of the later school of the Grecians, examineth the matter, and is at a stand, to think what should be in it, that men should love lies; where neither they make for pleasure, as with poets, nor for advantage, as with the merchant; but for the lie's sake. But I cannot tell; this same truth, is a naked, and open day-light, that doth not show the masks, and mummeries, and triumphs, of the world, half so stately and daintily as candle-lights.

Truth may perhaps come to the price of a pearl, that showeth best by day; but it will not rise to the price of a diamond, or carbuncle, that showeth best in varied lights.

A mixture of a lie doth ever add pleasure. Doth any man doubt, that if there were taken out of men's minds, vain opinions, flattering hopes, false valuations, imaginations as one would, and the like, but it would leave the minds, of a number of men, poor shrunken things, full of melancholy and indisposition, and unpleasing to themselves?

One of the fathers, in great severity, called poesy vinum doemonum, because it filleth the imagination; and yet, it is but with the shadow of a lie. But it is not the lie that passeth through the mind, but the lie that sinketh in, and settleth in it, that doth the hurt; such as we spake of before.

But, howsoever these things are thus in men's depraved judgments, and affections, yet truth, which only doth judge itself, teacheth that the inquiry of truth, which is the love-making, or wooing of it, the knowledge of truth, which is the presence of it, and the belief of truth, which is the enjoying of it, is the sovereign good of human nature.

The first creature of God, in the works of the days, was the light of the sense; the last, was the light of reason; and his sabbath work ever since, is the illumination of his Spirit. First he breathed light, upon the face of the matter or chaos; then he breathed light, into the face of man; and still he breatheth and inspireth light, into the face of his chosen.

The poet, that beautified the sect, that was otherwise inferior to the rest, saith yet excellently well: It is a pleasure, to stand upon the shore, and to see ships tossed upon the sea; a pleasure, to stand in the window of a castle, and to see a battle, and the adventures thereof below: but no pleasure is comparable to the standing upon the vantage ground of truth (a hill not to be commanded, and where the air is always clear and serene), and to see the errors, and wanderings, and mists, and tempests, in the vale below; so always that this prospect be with pity, and not with swelling, or pride. Certainly, it is heaven upon earth, to have a man's mind move in charity, rest in providence, and turn upon the poles of truth.

To pass from theological, and philosophical truth, to the truth of civil business; it will be acknowledged, even by those that practise it not, that clear, and round dealing, is the honor of man's nature; and that mixture of falsehoods, is like alloy in coin of gold and silver, which may make the metal work the better, but it embaseth it. For these winding, and crooked courses, are the goings of the serpent; which goeth basely upon the belly, and not upon the feet.

There is no vice, that doth so cover a man with shame, as to be found false and perfidious. And therefore Montaigne saith prettily, when he inquired the reason, why the word of the lie should be such a disgrace, and such an odious charge? Saith he, If it be well weighed, to say that a man lieth, is as much to say, as that he is brave towards God, and a coward towards men. For a lie faces God, and shrinks from man.

Surely the wickedness of falsehood, and breach of faith, cannot possibly be so highly expressed, as in that it shall be the last peal, to call the judgments of God upon the generations of men; it being foretold, that when Christ cometh, he shall not find faith upon the earth.

Text Analysis

1 Summary Statistics

1.1 General

	Overall	Sampled
Characters (all)	4,534	664
Characters (words only)	3,518	518
Words	841	124
Different Words	340	81
Sentences	24	5
Syllables	1,094	160

1.2 Averages

	Overall	Sampled
Characters per Word	4.18	4.18
Syllables per Word	1.30	1.29
Words per Sentence	35.04	24.80

1.3 Readability

	Overall	Sampled	Calculated Grading
Hard Words	50	7	
Long Words	124	18	
Lexical Density	40.43 %	65.32 %	
Gunning Fog Index	16.39	12.18	Hard
Coleman-Liau Grade	15.95	8.80	8th Grade
Flesch-Kincaid Grade Level	13.43	9.31	9th Grade (4 years)
Flesch Reading Ease	61.22	72.50	Fairly Easy: 6th Grade
ARI (Automated Readability Index)	21.48	10.65	10th Grade
SMOG(Simple Measure of Gobbledygook)	10.91	9.48	9 Years (Some high school)
LIX (Laesbarhedsindex)	49.79	39.32	Standard

2 Word Analysis

2.1 Word Frequency Profile based on the Whole Vocabulary

Word Frequency	Number of Words	Cumulative Vocabulary	Cumulative Word Count	Percentage Vocabulary	Percentage Word Count
1	242	242	242	70.967,74	28.437,13
2	53	295	348	86.510,26	40.893,07
3	14	309	390	90.615,84	45.828,44
4	4	313	406	91.788,86	47.708,58
5	4	317	426	92.961,88	50.058,75
6	3	320	444	93.841,64	52.173,91
7	3	323	465	94.721,41	54.641,60
8	3	326	489	95.601,17	57.461,81
11	2	328	511	96.187,68	60.047,00
12	3	331	547	97.067,45	64.277,32
13	1	332	560	97.360,70	65.804,94
20	2	334	600	97.947,21	70.505,29
21	1	335	621	98.240,47	72.972,97
22	1	336	643	98.533,72	75.558,17

continued

Word Frequency	Number of Words	Cumulative Vocabulary	Cumulative Word Count	Percentage Vocabulary	Percentage Word Count
23	1	337	666	98.826,98	78.260,87
24	1	338	690	99.120,23	81.081,08
42	1	339	732	99.413,49	86.016,45
48	1	340	780	99.706,74	91.656,87
71	1	341	851	100.000,00	100.000,00

2.2 Analysis based on the Whole Vocabulary

Total vocabulary=341 types

Project wordcount=851 tokens

Types/tokens=0.400,705,05

Types/sqrt(tokens)=11.689,329,45

Yule's k=181.192,790,40

2.3 Word Distribution

Length	Count	Proportion
1 letter words	23	2.7%
2 letter words	181	21.3%
3 letter words	197	23.1%
4 letter words	137	16.1%
5 letter words	118	13.9%
6 letter words	56	6.6%
7 letter words	35	4.1%
8 letter words	38	4.5%
9 letter words	23	2.7%
10 letter words	15	1.8%
11 letter words	7	0.8%
12 letter words	4	0.5%
13 letter words	3	0.4%

2.4 Key Words

Keywords are the words in this text that are far more frequent, proportionally, than they are in a general reference corpus (here, the Brown corpus, whose 1 million words comprise 500 texts of 2,000 words on a broad range of topics—see Brown freqs).

The number accompanying each word represents the number of times more frequent the word is in this text than it is in the Brown corpus. For example, the first item in the output **891.75 saith** is calculated on the basic that **saith** has **4** natural occurrences in the Brown's 1 million words, but **3** occurrences in your 841-word text. These 3 occurrences are proportionally a lot more than the 4 occurrences in the Brown. Taken as a proportion of 1,000,000 words, these 3 occurrences represent 3/841×1,000,000 = **3,567** virtual occurrences. These 3, 567 occurrences are 891. 75 times more numerous than the 4 occurrences in Brown. The keyword list below contains all the words in this text that are at least **10 times** more numerous in this text than in the Brown reference corpus (the "keyness factor"). The greater the keyness factor, the more "key" a word is likely to be to the input text.

(1) 891.75 saith	(9) 42.46 minds	(17) 16.63 certainly
(2) 264.22 breathed	(10) 37.75 belief	(18) 16.40 below
(3) 125.67 truth	(11) 37.16 towards	(19) 14.86 rest
(4) 95.89 pleasure	(12) 26.48 upon	(20) 14.33 earth
(5) 84.93 judgments	(13) 24.10 stand	(21) 12.26 nature
(6) 82.00 false	(14) 22.22 price	(22) 10.86 clear
(7) 79.27 mixture	(15) 21.42 faith	(23) 10.21 love
(8) 54.05 lies	(16) 18.02 light	

At keyness cut-off of 10, *there are* 23 *keywords from a total of* 841 *words, for a keyword ratio of* 0.027.

2.5 KWIC (Key Word in Context)

OF TRUTH/What is	truth	? said jesting
OF TRUTH/What is	truth	? said jesting Pilate, and
men take in finding out of	truth	, nor again, that when it
But I cannot tell; this same	truth	, is a naked, and open day
daintily as candle-lights.	Truth	may perhaps come to the price
and affections, yet	truth	, which only doth judge itself
teacheth that the inquiry of	truth	, which is the love-making
of it, the knowledge of	truth	, which is the presence of
of it, and the belief of	truth	, which is the enjoying of
upon the vantage ground of	truth	(a hill not to be commanded
and turn upon the poles of	truth	To pass from theological
and philosophical	truth	, to the truth of civil business
philosophical truth, to the	truth	of civil business; it will