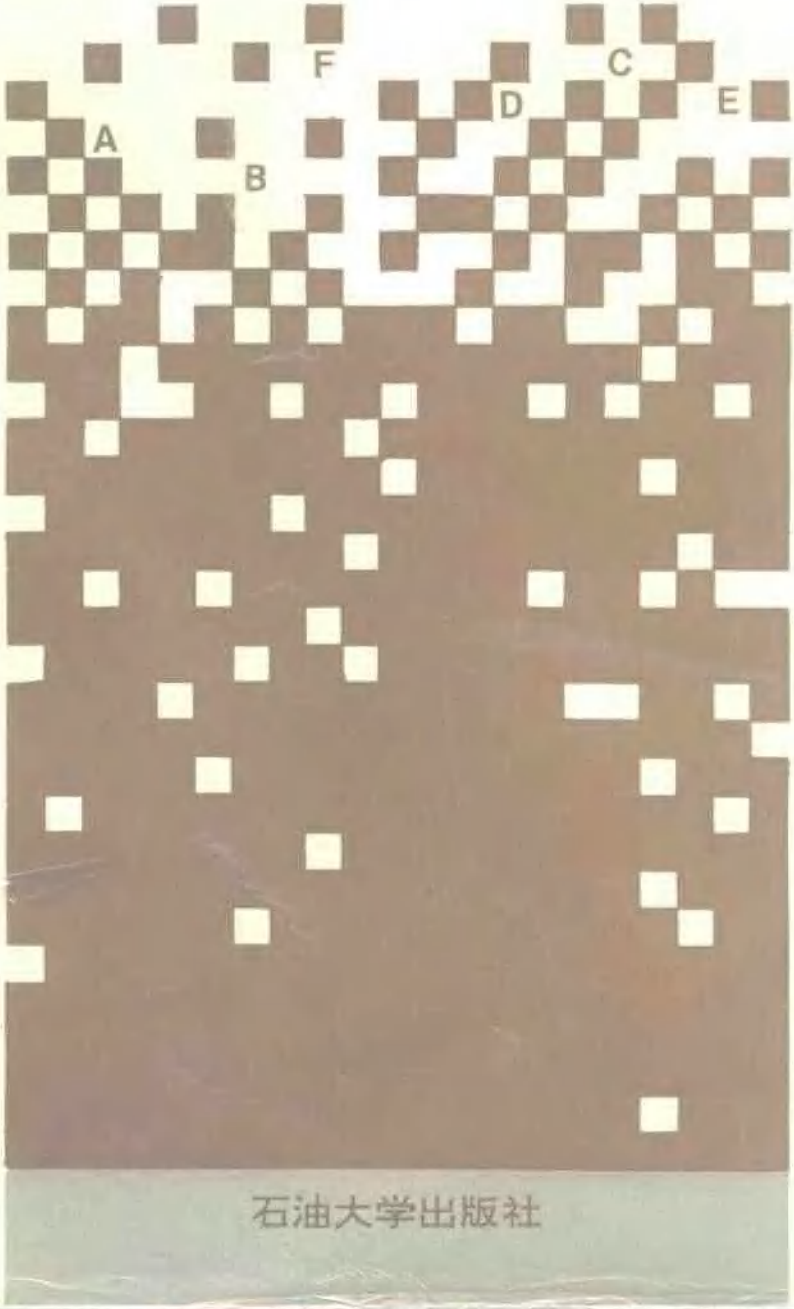
# 石油英语频率词典

# WORD FREQUENCY BOOK OF PETROLEUM ENGLISH

祝启波 编著




石油大学出版社

# Word Frequency Book of Petroleum English

# 石油英语频率词典

ZHU QIBO
祝启波著



石油大学出版社

# 内容简介

本书深入地阐述了广州石油英语语料库的建设方法、过程及结果。全书共三章。第一章介绍目前世界上英语语料库建设的概况及广州石油英语语料库建设的重要性。第二章讲解了石油英语语料库的建设方法及过程。第三章是本书的主干部分，介绍了计算机对该语料库进行分析后所得出的两个重要词表、一个标准频率指数分布表以及这些表中各参数的计算方法。

本书力求概念清楚，各统计参数计算准确。几个重要表格中的数据及典型程序备份完整，可供上机查阅。该书适合于从事石油英语研究、石油英语教材编写，科技英语教学、应用语言学研究、人工智能、机器翻译工作的人员使用。也可供高校石油英语专业、计算机专业、应用语言学专业及各有关专业的师生参考。

**石油英语频率词典**

祝启波 著

*

石油大学出版社出版

山东省 东营市

各地新华书店发行

石油大学印刷厂印刷

*

开本 787×1092 1/16 19.5 印张 462 千字

1991 年 6 月第 1 版 1991 年 6 月第 1 次印刷

印数：1—1500 册

ISBN 7-5636-0184-8/$Z_3$ · 01

定价：15.00 元

# PREFACE

Gui Shichun

The build—up of corpora requires a Brobdingnagian effort; apart from pooling an enormous amount of human resources, nearly all of the well—known corpora such as the Brown corpus, the LOB corpus, the AHI (American Heritage Intermediate) corpus, and Birmingham English text corpora made extensive use of modern technology, mainly the mainframe and the optical character reader. So when Mr. Zhu approached me in 1985 as an MA candidate with the thesis proposal of setting up a specialized corpus of petroleum English on an IBM PC/XT, I thought the idea was admirable and audacious, and was quite willing to be his supervisor, because what emerged as final products of his project would be not only a corpus of petroleum English, but also a model of doing the complicated job of setting up a specialized corpus of a few hundreds of thousands of words on an inexpensive computer available to most people.

In the following year, Mr. Zhu embarked on his project of huge dimensions practically single—handed; collecting samples, typing materials on the keyboard, managing data—files, computing corpus statistics, and, abeve all, writing up programmes that could exploit the potentials of a personal computer for handling large quantities of data. And he reaped the reward to which he was undoubtedly entitled, Word Frequency Book of Petroleum English, the first of its kind in China, was testimony of his ingenuity and hard work, more eloquent than any words that I could find to express my appreciation.

The book is, of course, not addressed to the general readership, but I am sure, lexicographers, syllabus designers, text—book writers, and teachers of petroleum English will draw on invaluable information from these seemingly dry figures, so that their academic speculations will be more solidly data—based.

# ACKNOWLEDGEMENTS

First of all, my sincere thanks go to Professor Gui Shichun, my supervisor. He is responsible for nurturing my interest in computer—oriented research in linguistics and helping me to shape the idea of setting up GPEC ———— Guangzhou Petroleum English Corpus. Later, his shrewed comments and suggestions on the overall plan and strategies for various stages of the project saw my work through to its timely completion. Without his patient guidance, prompt encouragement and meticulous care, this study would never have come into being.

I owe a great debt of gratitude to Professor Yang Guanghua, President of Petroleum University. who showed great interest in my work when he was on an inspection tour to his Guangzhou base in 1990; to Professor Wu Mingfang, Dean of the Department of Foreign Languages of Petroleum University, who, with his profound knowledge in the field of corpus linguistics, carefully looked into the development of my corpus and recommended the publication of the book; to Professor Shi Mingxian, Director Wu Hanmin, Professor Lu Jimeng, Professor Shen Benshan and many other professors and experts of the Publication Foundation for Academic Works of Petroleum University, who verified the validity of my research and decided to fund the publication of this book during the Foundation's first annual meeting in 1990.

I acknowledge the valuable advice of Professor Wang Zongyan of Zhong Shan University, of Professor Wu Qianguang, Professor Li Xiaoju and Professor He Zhiran in the English Department of Guangzhou Foreign Language Institute; I acknowledge the kind reception offered by Professor Huang Renjie in the English Department of Jiao Tong University in Shanghai when I was visiting the Wang An computer room on the campus for corpus—building information; I acknowledge the generous aid of material by Professor Chen Yuming in the Department of Foreign Languages of Fu Dan University in Shanghai. Without their help, some of the computational and statistical analyses could not have been carried out.

I am under very particular obligations to the Educational Department of the Ministry of Petroleum Industry (MPI) and its foreign language training center in Guangzhou (GFLTC), for granting me the fund necessary to accomplish this investigation and for placing at my disposal a well—equipped computer room with an IBM Personal Computer and other accessories. Lie Guodong, Shu Zhi and Shen Yaonian, the three principal administrators in GFLTC jointly assumed many administrative duties to facilitate my work when I was in difficulties in 1987. Their unfailing enthusiasm and interest in the project encouraged me a lot. Furthermore the continuing dialogue I enjoyed with Director Shen Yaonian contributed a great deal to many aspects of this research. Also Director Pan Sheyin is quite instrumental in getting this book published.

The preparatory work of acquiring the samples was both laborious and time—consuming. It is with pleasure here that I acknowledge my indebtness to Mai Yongquan, head of the English language teaching section in GFLTC, to Peng Dezhi, my ex—student and now the senior interpreter in the TOTAL Petroleum Company, to Ou Yiang, chief librarian in the East China Petroleum Institute, for what was, for the most part, sheer drudgery in the course of the material collection.

My thanks also go to Li Wei, my computer teacher in Guangzhou Foreign Language Institute, who first

armed me with the necessary prerequisites for computer programming with some illustrative sample programmes , later encouraged me to do the programming on my own; to Wang Huihua, associate Professor in the computer section of GFLTC, Cai Shuping and Yiao Qun in the computer room of GFLTC, for the favors they did me in the early stage of the project; to Zhang Tianhang and Zhang Genglin, engineers from South China Computers Company, for their suggestions about my computing work and their efforts to improve the hardware enviornment of my computer.

Last but not least, my special thanks go to Professor Huang Haoshu, to Carol Maxwell, Su Weiqin, Du Wei, Yu Mengxin, Wu Shaoyu and many other (foreign) teachers and staff members of GFLTC for their understanding assistance which went much far beyond the call of their duty.

Zhu Qibo Guangzhou Training College, Petroleum University
February, 1991

## CONTENTS

# CHAPTER I
## INTRODUCTION

Corpus linguistics can be said to be in its infancy compared with other time—honoured traditions of the study of language like phonology and lexicology. This branch of linguistics is made possible only by the advent of the computer. It is now gaining wider and wider acceptance among linguists and language teachers because of the many contributions it makes to language teaching and linguistic studies. With the growing sophistication of hardware and software, corpus linguistics is now developing at a rapid speed.

The first standardized, edited computer corpus in the world is the Brown University Standard Corpus of Present—Day American English ( often referred to as "the BROWN corpus") (Yang Huizhong 1985: 94). It has been available for computational analysis for about 30 years. The BROWN corpus is a collection of texts of 1,014232 running words from 500 samples, each approximately 2,000 words long. The genre categories covered are as shown in table I—A.

TABLE I—A

THE TEXT CATEGORIES OF THE BROWN CORPUS

| CATEGORY | TEXTS IN THE CATEGORY |
|---|---|
| A. PRESS: REPORTAGE | 44 |
| B. PRESS: EDITORIAL | 27 |
| C. PRESS: REVIEWS | 17 |
| D. RELIGION | 17 |
| E. SKILLS AND HOBBIES | 36 |
| F. POPULAR LORE | 48 |
| G. BELLES LETTERS, BIOGRAPHY, MEMOIRS, ETC | 75 |
| H. MISCELLANEOUS | 30 |
| J. LEARNED | 80 |
| K. GENERAL FICTION | 29 |
| L. MYSTERY AND DETECTIVE FICTION | 24 |
| M. SCIENCE FICTION | 6 |
| N. ADVENTURE AND WESTERN FICTION | 29 |
| P. ROMANCE AND LOVE STORY | 29 |
| R. HUMOR | 9 |
| GRAND TOTAL | 500 |

(Francis 1979:196)

The corpus is aimed to provide a source of contemporary American English texts in computer-accessible form for further studies. Since its publication, it has served as a standard of comparison for studies of other texts ( e. g. British English and American English; Indian English and American English; adult language and child language) and a model for the construction of other corpora. Now it is being used at many computer centers all over the world (Francis 1979:193).

Another sizeable corpus of present—day English is the Brithish counterpart of the Brown corpus - the Lancater—Oslo/Bergen corpus ( referred to as "the LOB corpus "). This corpus originated in 1970, containing approximately 1 million running words. It was expected to serve as a British English data base parallel to the Brown corpus for multi—purpose comparisons and studies. The sampled materials were mostly British. The principles of sampling matched as closely as possible with the Brown corpus, departing only marginally in respect of the distribution of some categories (as shown in table I—B).

**TABLE I—B**

**THE TEXT CATEGORIES OF THE LOB CORPUS**

| CATEGORY | TEXTS IN THE CATEGORY |
|---|---|
| A. PRESS: REPORTAGE | 44 |
| B. PRESS: EDITORIAL | 27 |
| C. PRESS: REVIEWS | 17 |
| D. RELIGION | 17 |
| E. SKILLS, TRADES, AND HOBBIES | 38 |
| F. POPULAR LORE | 44 |
| G. BELLES LETTERS, BIOGRAPHY, ESSAYS | 77 |
| H. MISCELLANEOUS | 30 |
| J. LEARNED AND SCIENTIFIC WRITINGS | 80 |
| K. GENERAL FICTION | 29 |
| L. MYSTERY AND DETECTIVE FICTION | 24 |
| M. SCIENCE FICTION | 6 |
| N. ADVENTURE AND WESTERN FICTION | 29 |
| P. ROMANCE AND LOVE STORY | 29 |
| R. HUMOR | 9 |
| GRAND TOTAL | 500 |

As a British equivalent of the BROWN corpus, LOB is equally successful in achieving its original aim.

Completed in 1971, the American Heritage Intermediate Corpus ( referred to as " the AHI corpus ") is one on an even larger scale. It comprises a total of 5,088,721 running words drawn from published materi-

als that are most likely to be encountered by school children of grade 1—9 (as shown in table I—C).

TABLE I—C

THE TEXT CATEGORIES OF THE AHI CORPUS

| CATEGORY | WORDS IN THE CATEGORY |
|---|---|
| READING | 1182971 |
| ENGLISH AND GRAMMAR | 283367 |
| COMPOSITION | 57776 |
| LITERATURE | 277907 |
| MATHEMATICS | 387619 |
| SOCIAL STUDIES | 503620 |
| SPELLING | 210157 |
| SCIENCE | 510570 |
| MUSIC | 209364 |
| ART | 47887 |
| HOME ECONOMICS | 83387 |
| SHOP | 65375 |
| LIBRARY FICTION | 303603 |
| LIBRARY NONFICTION | 374885 |
| LIBRARY REFERENCE | 271040 |
| MAGAZINE | 314643 |
| RELIGION | 4595 |
| GRAND TOTAL | 5088766 |

(Carroll 1971 : 3)

What is unique about this corpus is that sophisticated statistical techniques are used to estimate the probabilities of words' occurrences in a theoretical sample of infinite size. With its uniqueness and practicality, the AHI corpus has offered a different approach to the building of corpora.

As a corpus describing the English for Science and Technology (EST), the Jiao Tong University Corpus for EST ( often shortened as "the JDEST corpus" or "JDEST") in China deserves mentioning. This corpus is a million—word corpus designed "to meet the needs in the study of EST " (Yang Huizhong 1985: 94). The subject matters of the corpus cover 10 specialized fields as shown in table I—D.

TABLE I—D

THE TEXT CATEGORIES OF THE JDEST CORPUS

| CONTENT | WORDS |
|---|---|
| COMPUTERS | 100,383 |
| METALLURGY | 105,823 |
| MACHINE BUILDING | 104,025 |
| PHYSICS | 111,393 |
| ELECTRICAL ENGINEERING | 109,928 |
| CIVIL ENGINEERING | 98,826 |
| CHEMICAL ENGINEERING | 113,271 |
| NAVAL ARCHITECTURE | 114,168 |
| ATOMIC ENERGY | 107,131 |
| AIRCRAFT MANUFACTURING | 107,457 |
| TOTAL | 1,072,405 |

(Huang Renjie et al. 1989 : 404)

Incorporating words in some well known lists such as Thorndike's Teachers' Wordbook of 30,000 Words, M. West's General Service List of English Words into the frequency list of the corpus, a service list of about 6,000 words for EST learners has been obtained. The list has been used by syllabus designers for EST courses in China. More exploitation of the corpus is now under way.

Birmingham University's collection of English texts is a very influential corpus. Comprising over twenty million English words, it is claimed to be the largest corpus now in the world (Yang Huizhong 1985:94).

Although many large and medium size corpora have been or are being built in the world, none can be adopted for suitable use for the teaching and learning of petroleum English (PE). The average PE teacher's ignorance of the PE lexicon has resulted in waste in money and time. There is now an acute need for a comprehensive description and study of the PE lexicon. It is with this perspective that the present corpus ——— Guangzhou Petroleum English Corpus (GPEC) has been established and analysized.

GPEC is a computer—assembled corpus composed of 710 sample texts of 500—600 words each ——— 411,612 running words in all. The sampled materials represent exclusively PE texts, covering a period of about 30 years ( mostly 1975—1986). The categories included are:

TABLE I—E

THE TEXT CATEGORIES OF GPEC

| CONTENT | WORDS | TEXTS |
|---|---|---|
| A1. PETROLEUM GEOLOGY AND PROSPECTING | 87,269 | 145 |
| B1. PETROLEUM REFINERY AND PETROCHEMISTRY | 116,826 | 199 |
| C1. OIL DRILLING | 81,632 | 142 |
| D1. PETROLEUM PIPELINE | 42,635 | 74 |
| E1. OFFSHORE OIL EXPLORATION | 83,250 | 150 |
| TOTAL | 411,612 | 710 |

The project which led to this corpus was conducted by the present author in 1986—1987, under the guidance of ProfessorGui Shichun, with the fund provided by GFLTC of MPI.

## CHAPTER II

## THE DEVELOPMENT OF GPEC

The main principle for the sampling of GPEC is that the samples to be drawn should strike a right balance between homogeneity and heterogeneity. On the one hand the size of the corpus should be sufficiently large and the materials should have something to do with the petroleum industry. On the other hand, subfields should be first defined so as to maximize the diversity. After a comprehensive study of the sampling schemes for different corpora, and some careful considerations of the nature of GPEC and some other factors, the first decision was made on the optimum size for GPEC ———— 400,000 running words. This number was thought to be fairly adequate to represent a single specialized field compared with the number of running words they alloted to a branch of science in corpora like BROWN, LOB, and JDEST. What is more, this would be just the amount of text that could be fed into the machine, proofread, corrected in about 7 months by a single individual and fully processed within the modest funds allowed. The second decision was made to ensure the diversity character: the lexical universe of PE was roughly divided into 5 areas with different proportions:

| | |
|---|---|
| A1. PETROLEUM GEOLOGY AND PROSPECTING | 80,000 |
| B1. PETROLEUM REFINERY AND PETROCHEMISTRY | 120,000 |
| C1. OIL DRILLING | 80,000 |
| D1. PETROLEUM PIPELINE | 40,000 |
| E1. OFFSHORE OIL EXPLORATION | 80,000 |

The proportions were weighed chiefly by constant consultations with veteran PE teachers, librarians, specialists and engineers in the institutions of petroleum industry ; by investigating the distribution of PE books in the main petroleum libraries; by referring to the Classification System for Books in Chinese Libraries; and also by appealing to the author's own experience of working as a PE teacher for many years. Considering the flexibility for generating citation context, the choice for words per text is 500—600.

The progress of securing satisfactory samples of PE was retarded by the relative scarcity of holdings of the PE books in GFLTC and in the Institute of Foreign Languages in Guangzhou. Worse still, even within the confines of several well—known libraries in Guangzhou only materials for one category ( category B1) could be collected. To tackle the problem the present author went to the East China Petroleum Institute in Shandong for materials of categories A1; D1 and E1; to South China Sea Oil Company in Zhan Jiang for materials of category C1. To minimize the effects of bias on the author's part, sample books were taken at approximately uniform intervals within each category, while pages were randomly chosen excluding pages devoted fully to lists, tables, indexes, graphs and other odd symbols.

Once the sample page was decided it was photocopied and registered with its catalogue number, control number and some other information.

Since we had no devices like optical character scanner, the traditional inputing technique of keyboarding had to be resorted to. Feeding all the 410,000 words into the machine by a non—professional typist or oper-

ator is really a daunting task. It took the author an average of 5 painful months for this tedious process. Samples were first read, counted and marked off before they were keyed in. Practical considerations led to some minor modifications of the sampled pages. For example a sampled page that contains less than 500 words were usually supplemented by the next comparable page ( in subject matter, the type of writing, and, if possible, the same calender year of publication). In order to incorporate as much as possible the graphic details of the original into the corpus, letters in formulas were fed into the computer as they stand, punctuation marks like asterics, slashes, parantheses were not ommitted though they were not counted as words.

All data processing was performed at GFLTC of MPI. The computer used is an IBM PC XT. The source programming languages are IBM BASIC and dBASE III. These two languages were chosen out of the considerations of computer time, storage space and the efficiency for particular functions of the languages. It was also the author's hope that , if successful, the model of processing such a bulky data set with the readily available software and hardware would lend itself to the building of similar medium—sized ESP corpora.

During the early stage, a pilot study was conducted of the first 87,269 tokens keyed in so as to try out the programs, to test the validity of the calculation methods and to see whether the general arrangements were feasible. When these were found to be satisfactory, the project was given the go—ahead. Later when all the data sets were available, some more sophisticated programming languages such as Pascal, C, Prolog and Foxbase were used to corroborate the accuracy of the results.

## CHAPTER III
## INTRODUCING THE LISTS AND TABLES

Among other things, the computational analysis showed that in GPEC (411,612 running words) there are 24506 different word types among which 11259 word types are HAPAX LEGOMENA ( types appearing only once in the corpus). Lists and tables produced include an alphabetical list for the whole corpus ( 7678 words); a rank list for the whole corpus ( 6645 words); five rank lists for the five categories ( a total of approximately 12930 words); a PE and GE common word list ( 3287 words); a reversed alphabetical list of the PE and GE common word list ( 3287 words); a PE+ word list ( 2812 words); a GE+ word list ( 2503 words); a subtechnical word list ( 2143 words ); five technical word lists ( a total of 2706 words); and an SFI distribution table. However, for lack of space we can only focus on the Rank List, the Alphabetical List and the SFI Distribution Table here.

1. THE RANK LIST

The rank list is arranged in descending order of SFI and U. All the words included have an SFI of more than 47.5 in GPEC. It was felt that low SFI words can be assumed to be idosyncratic to particular text samples and thus not useful enough to justify the expansion of more than 100 pages. The following parameters are included in this list:

VOC —— vocabulary item, the word or letter.

F —— frequency, the actual occurrence(s) of a word.

D —— dispersion, a measure of the evenness or unevenness of a word's distribution over the categories. It is computed according to a measure of relative entrophy based on information theory statistics. The formula is:

$$D = [\log P + (-\text{sum}(pi * \log(pi))/P]/\log(n)$$

(Carroll 1971:xl)

If equal occurrences of a word are found in the 5 categories, D takes the value of something toward 1 (e.g. 0.9998 for "the"); if all occurrences of a word are found in a single category, D takes the value of 0 (e.g. "arsenic", f=11, D=0).

U —— index of usage, the estimated value for word usage measured by frequency—per—million. It is computed by the following formula:

$$U = (1000000/N) * [F * D + (1-D) * fmin]$$

(Carroll 1971:xl)

A more frequent word can have a lower U value than a less frequent word provided the dispersion (D) of the former is less than that of the latter.

SFI —— standard frequency index, an estimate of the probability of a word's occurrence in a theoretical sample of infinite size. The value is directly derived from U according to the formula:

$$SFI = 10(\log U + 4)$$

(Carroll 1971:xl)

If a word has an SFI of 90, one would expect to find it once in every ten words. If a word has an SFI of 80 one would expect to find it once in every 100 words. Table III—A is an interpretation of SFI:

TABLE III—A

INTERPRETATION OF THE STANDARD FREQUENCY INDEX

| SFI | PROBA OF WORD'S OCCURRENCE IN A THEORETICAL INDEFINITELY LARGE SAMPLE | | EXAMPLE OF A WORD WITH DESIGNATED SFI | |
|---|---|---|---|---|
| 90 | 1 IN EVERY | 10 WORDS | the | 88.9 |
| 80 | 1 IN EVERY | 100 WORDS | be | 80.1 |
| 70 | 1 IN EVERY | 1,000 WORDS | equipment | 69.3 |
| 60 | 1 IN EVERY | 10,000 WORDS | showed | 60.1 |
| 50 | 1 IN EVERY | 100,000 WORDS | hypothesis | 50 |
| 40 | 1 IN EVERY | 1,000,000 WORDS | adaptations | 39.9 |
| 30 | 1 IN EVERY | 10,000,000 WORDS | | |
| 20 | 1 IN EVERY | 100,000,000 WORDS | | |
| 10 | 1 IN EVERY | 1,000,000,000 WORDS | | |