# GENRE ANALYSIS IN BIOLOGICAL SCIENCE TEXTS

by

Grisel Ma. García Pérez

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS
in the Faculty
of
Education


SIMON FRASER UNIVERSITY
August, 1992

APPROVAL

Name: Grisel Maria Garcia-Perez

Degree: Master of Arts

Title of Thesis: Genre Analysis in Biological Science Texts

Examining Committee:

Chair: Mike Manley-Casimir

Gloria P. Sampson
Senior Supervisor

Judith Scott
Assistant Professor

Juan Sosa
Assistant Professor
Department of Spanish and
Latin American Studies
Simon Fraser University
External Examiner

Date Approved August 14, 1992

PARTIAL COPYRIGHT LICENSE

I hereby grant to Simon Fraser University the right to lend my thesis, project or extended essay (the title of which is shown below) to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users. I further agree that permission for multiple copying of this work for scholarly purposes may be granted by me or the Dean of Graduate Studies. It is understood that copying or publication of this work for financial gain shall not be allowed without my written permission.

Title of Thesis/Project/Extended Essay

GENRE ANALYSIS IN BIOLOGICAL SCIENCE TEXTS

Author: ____
(signature)

Grisel Maria Garcia-Perez
(name)

August 14, 1992
(date)

## Abstract

### Genre Analysis in Biological Science Texts

Defining the distinctive features of the variety of English used in scientific contexts has been a special tendency in the past few years Leech, (1983); Swales, (1985); Halliday, (1988); Johns, (1991); Reid, (1991). This is so because English has emerged as the predominant medium of scientific discussion and progress, hence theoretical and practical applications on the teaching of English have become a powerful need in all parts of the world.

Although these studies are grounded on a firm basis and offer a rationale for genre analysis, Douglas Biber (1988) in his book *Variation Across Speech and Writing* offers a model in which texts can be compared along dimensions of linguistic variation. This is the most sophisticated study on genre differences that has been published so far.

Few studies on genre analysis offer hard data on specific fields such as medicine, physics, biology and math. The present study presents the results of a search for the percentage of linguistic features specifically shared by texts in the field of biological science. The findings are compared to the general science corpora described in Biber's analyses.

The study includes 700-word text samples extracted from larger texts. All texts were automatically included in readable codes for the computer, words were automatically counted, and 10 of the

linguistic features (those which most frequently co-occurred in Biber's general science corpora) were automatically identified. After the computational analysis was done, inspection by the analyst of the computer results to check for errors was also mandatory as the program used, AnyTEXT™, did not always recognize all the linguistic features.

The frequency counts of linguistic features were normalized to a text length of 1000 words. By summing up the frequency of each of the linguistic features in the texts, I was able to average the factor score for each text across all texts in the biological science genre and compute a mean dimension score for the genre. I then used this mean dimension score to compare and to specify the relations among three sub-genres: biology, microbiology and biochemistry.

The findings of this analysis show that narrowing the corpus to a specific field provides an array of linguistic dimensions which do not necessarily coincide with the results of the general science corpora described in Biber's analysis.

This finding provides a firm foundation for curriculum developers who rely on these general science corpora when developing teaching materials to create more accurate and relevant teaching materials. The study should also continue to prove useful to the investigation of the co-occurrence of linguistic features in specific fields such as history, foodscience, pharmacy, etc.

*To my mother Clara*

Acknowledgements

Many people have supported and encouraged me in writing this thesis.

First, I want to thank both the University of Havana and Simon Fraser University for having had confidence in me to initiate the collaboration program between both institutions and for providing me with the opportunities to work and study.

I am indebted foremost to the many professors, classmates, students and friends whom I have come to know. Their patience, encouragement and support served as an enduring source of inspiration.

I also owe a professional debt of a very special kind to my supervisor, Gloria Sampson. Her listening to my ideas, careful reading of the drafts and thoughtful comments were invaluable.

A word of thanks is also given to all those who gave me strength and financial support to complete my degree. For this I would like to express my gratitude to Jorge García, Teresa Kirschner, Mike Manley-Casimir, Vern Loewen, Bruce Claymann, Pablo Dobud, Sharon Bailin, Stanley Shapiro, Bob Brown, and Colin Yerbury.

Thanks go, in particular, to my Chilean sisters Ester Erickson and Monica Lee. Their unfailing support, love and care helped me make time for writing and reminded me when it was time to stop and do other things.

Last but not least, I wish to give special thanks to the continued love, support and understanding of my mother Clara, my father José, my aunt Lydia and my sisters Maysa and Marlene. Although their contributions are less obvious, they are in many ways greater than any of the others.

Table of Contents


Approval (ii)

Abstract (iii)

Dedication (v)

Acknowledgements (vi)

List of table of Contents (viii)

1. Chapter I

Introduction 1

Summary 22

2. Chapter II

Methodology 23

Method 23

Text selection 24

Linguistic features 26

Frequency counts 29

Factors 33

Summary 37

3. Chapter III

Data collection 38

Summary 54

4. Chapter IV

Analysis of the data 55

Biochemistry sub-genre 68

Microbiology sub-genre 73

Biology sub-genre 74

Summary 75

5. Chapter V: Conclusions 76

References 84

Appendixes

Appendix A 91

Appendix B 93

List of Tables x

List of Figures xiii

List of Tables

Table 1.0 Descriptive statistics for the corpus of biological science texts as a whole 47

Table 1.1 Descriptive statistics for the corpus of the biochemistry sub-genre 48

Table 1.2 Descriptive statistics for the corpus of the microbiology sub-genre 49

Table 1.3 Descriptive statistics for the corpus of the biology sub-genre 50

Table 2.0 Description of the distribution of the linguistic features that had a co-occurrence of 9.6 and over within each sub-genre and across sub-genres compared to the general corpus 53

Table 2.1 Description of the distribution of the linguistic features that had a co-occurrence of less than 9.6 within each sub-genre and across sub-genres compared to the general corpus 53

Table 3.0 Mean frequencies for the Biological Science Corpora as a whole 57

Table 3.1 Mean frequencies for the Science Corpora presented in Biber's study 58

Table 3.2 Comparison between the mean frequencies in the Biological Science Corpora and the Science Corpora presented in Biber's study 59

Table 4.0 Distribution of linguistic features from those having the highest mean to those having the lowest mean in the biochemistry sub-genre compared to the general biological science corpora and to Biber's study 70

Table 4.1 Distribution of linguistic features from those having the highest mean to those having the lowest mean in the microbiology sub-genre compared to the general biological science corpora and to Biber's study 71

Table 4.2 Distribution of linguistic features from those having the highest mean to those having the lowest mean in the biology sub-genre compared to the general biological science corpora and to Biber's study 72

Table 5.0 Distribution of the mean frequencies of agentless passives across the study 77

Table 5.1 Distribution of the mean frequencies of by- passives across the study 78

Table 5.2 Distribution of the mean frequencies of third person pronouns across the study 78

Table 5.3 Distribution of the mean frequencies of pronoun *it* across the study 79

Table 5.4 Distribution of the mean frequencies of demonstrative pronouns across the study 79

Table 5.5 Distribution of the mean frequencies of 80

possibility modals across the study
Table 5.6 Distribution of the mean frequencies of predictive modals across the study 80
Table 5.7 Distribution of the mean frequencies of nominalizations across the study 81
Table 5.8 Distribution of the mean frequencies of perfect aspect across the study 81
Table 5.9 Distribution of the mean frequencies of conditionals across the study 82

List of Figures


Figure 1 39

Figure 2 40

Figure 3 41

Figure 4 42

Figure 5 43

Figure 6 44

# Chapter I
# Introduction

The tradition that language is viewed as a self-contained system or code, as a set of arbitrary rules and conventions, manipulated as a tool by speakers and writers is a way of thinking so deeply rooted that it continues to operate in much innovative literary theory and rhetoric (Rosenblatt, 1989). The influence of Ferdinand de Saussure is important in this tradition. His formulation of a two-element relationship between word and object (Saussure, 1916), has lent itself to the narrow conception of language as an autonomous system, independent of its instrumental role.

In contrast, Charles Sanders Peirce, the American founder of semiotics, formulated a three-element relationship between sign, object and interpretant. "A sign," Peirce wrote, "is in conjoint relation to the thing denoted and to the mind... The sign is related to its object only in consequence of a mental association, and depends on habit" (Peirce, 1931-35). Although he did not see mind as an entity, his model grounds language and the processes involved in speaking, listening, reading and writing firmly in the individual's relations with the world.

Psychologists' studies of children's acquisition of language support the Peircean triad (Werner and Kaplan, 1962; Vygotsky, 1962). Werner and Kaplan, in their work on symbol formation , stated

that the vocalization of a sign becomes a word when the sign and its object or referent are linked with the same "organismic state". Vygotsky pointed out that the sense of a word was "the sum of all the psychological events aroused in our consciousness by the word" (Vygotsky, 1962).

We know that language is a socially generated public system of communication. But it is sometimes forgotten that language is always internalized by an individual human being in a particular environment. Lexical concepts must be shared by speakers of a common language, yet there is room for considerable individual difference in the details of any concept. Bates (1979) explains the sense of words using the image of an iceberg; the tip of the iceberg representing the public aspect of meaning, and the submerged base representing private meaning.

The individual's share in a language is that set of features of the public system which has been internalized in the individual's experiences with words in the life situation. The residue of such relations in particular natural and social contexts constitutes a kind of linguistic-experimental reservoir. Human beings make sense of a new situation by applying, revising, or extending individual words selected from their personal linguistic-experimental reservoir.

Although we speak of individual words, we know that words do not function in isolation, but always in particular verbal, personal and social contexts. This idea has to be taken into consideration in Language Teaching (LT).

For effective LT, the teacher's own model of language should encompass all that the student knows language to be and take

account of the student's own linguistic experience in its richest potential. The language should be relevant to the students' experiences and to the linguistic demands of society, where they are surrounded not by grammars and dictionaries, but by text or language in use in a situation. Different studies show that situational language teaching is valid in teaching a first and a second language (Halliday, 1973; Savignon, 1983; Ellis, 1986; Long, 1990).

Although research on first language acquisition shows an array of information worthy of being studied, (Bloomfield, 1914; Saussure, 1916; Sapir, 1921; Vygotsky, 1939; Chomsky, 1951; Winograd, 1972-1983), emphasis in this work will be given to second language acquisition.

Second Language Acquisition (SLA) is a relatively new, interdisciplinary field of inquiry. Most empirical research on SLA have been conducted since 1960 by researchers drawing heavily upon theory, research findings, and research methods in a variety of fields (Ausubel, 1964; Rivers, 1973; Valdman, 1980; Sampson, 1990). These fields include education, anthropology, psychology, linguistics, applied linguistics, foreign languages and English Language Teaching (ELT), the field on which my thesis, from a general point of view, will be based.

Throughout its history, the general world of ELT has been divided into different groups: English as a Mother Tongue (EMT, English learnt as a first language), English as a Second Language (ESL, English taught to speakers of other languages in an English speaking environment), and English as a Foreign Language (EFL,

English taught to speakers of other languages, not in an English speaking environment).

ESL/EFL can be divided into two different groups: General English (GE, English taught in primary, secondary and tertiary levels), and English for Specific Purposes (ESP, English taught to prepare learners "for chosen communicative environments": Mohan, 1986). Within the field of ELT, special emphasis will be given in this paper to ESP.

Although several authors have defined the phrase ESP, English for Special or Specific Purposes, (Mohan, 1986; Phillips, 1981; Crandall, 1984) the following definition encompasses the elements widely considered essential to ESP:

> "An ESP course is purposeful and is aimed at the successful performance of occupational or educational roles. It is based on a rigorous analysis of student's needs and should be 'tailor-made'. Any ESP course may differ from another in its selection of skills, topics, situations and functions and also language. It is likely to be of limited duration. Students are more often adults but not necessarily so, and may be at any level of competence in the language: beginner, post-beginner, intermediate, etc. Students may take part in their ESP course before embarking on their occupational or educational role, or they may combine their study of English with performance of their role, or they may already be competent in their occupation or discipline but may desire to perform their role in English as well as in their first language".
>
> (Robinson, 1980)

Because of the fact that the content of the texts used in ESP courses have to do with the learner's needs in a specific area, there

is a tendency to use the phrases 'ESP course' and 'Content-area ESL course' nondistinctively.

Content-area ESL courses use content (e.g. mathematics, science, art) to provide the context for language instruction (Mohan, 1986; Rivers, 1990; Dow and Ryan, 1990). Ideally two instructional objectives _content-area knowledge or skill and increased language skill_ are achieved in each class. In these classes the students need to comprehend the material and the teacher's message, and teachers need to understand the students' messages well enough to provide feedback (Mohan, 1986, Rivers, 1990; Dow and Ryan, 1990). Language learning is stimulated by the rich context the subject matter provides, by the inherent interest and relevance of the content, and by the fact that the learners focus on messages and not on language form (Krashen, 1982).

English for specific or special purposes is that area of English language teaching which focuses on preparing learners for a chosen communicative environment. Its identity is given by the fact that the learner has a purpose, a purpose that is 'not restricted to linguistic competence alone but does involve the mastery of the language skills in which language forms an integral part' (Phillips, 1981).

Content-area ESL and ESP share several guiding principles:

a) the importance of context;

b) the importance of attending primarily to meaning and to language form; and

c) the importance of taking into consideration the needs of the learners.

However, in many ways the two types of instruction are very different:

a) Content-area ESL has a broader objective: improving overall English proficiency, in addition to teaching particular language skills required for understanding the content; whereas ESP courses are structured to promote efficient and effective acquisition of particular language and communicative skills.[1]

b) Content-area ESL aims to teach content. Even when the content material is carefully organized to support language learning as recommended by Mohan (1986), the organization considers both content and language, and logically, the content varies. ESP courses prepare the learners in very specific environments and this has been its main goal since its appearance.

The end of the Second World War in 1945 can be related to the birth of ESP. During that age there was an enormous and unprecedented expansion in scientific, technical and economic activity on an international scale. This expansion created a world dominated by science and technology which suddenly generated a demand for an international language. For various reasons, most notably the economic power of the United States in the post-war world, this role fell to English.

A whole mass of people wanted to learn English, not for pleasure or prestige of knowing the language, but because English was the key to international currencies of technology and commerce.

---

[1] See Krashen (1982, pp. 169-170) for a discussion of the difference between ESL through subject-matter teaching and ESP.

There appeared then a new generation of English learners who knew specifically why they were learning a language _business-persons who wanted to sell their products, doctors who needed to keep up with developments in their field, and a whole range of students whose course of study included textbooks and journals available only in English. They needed the language, and most important, they knew why they needed it.

The general consequences of this wanting to learn English exerted pressure on the language teaching profession to deliver the required goods, and along with this growing demand for English courses for specific needs, new ideas began to emerge in the study of language. These studies shifted attention away from defining the formal features of language usage to discovering the ways in which language is used in real communication (Widdowson, 1978). One finding of this research was that the language we speak and write varies considerably, and in a number of different ways, from one context to another.

In ELT this promoted the view that there are important differences between, for example, the English of science and the that of commerce. These ways of thinking pushed the development of English courses for specific groups of learners. The rationale was: if language varies from one situation to another, it should be possible to determine the features of specific situations and then make these features the basis of the learner's course.

The greatest expansion of research into the nature of particular varieties of English can be seen in the late 1960s and early 1970s, mainly describing written scientific and technical

English (Ewer and Latorre, 1969; Swales, 1971; Selinker and Trimble, 1975).

The concept of 'register' started to appear with the works of Halliday, McIntosh and Strevens in 1964;[2] Ewer and Latorre in 1969 and Swales in 1971. The main objectives in their works was to identify the grammatical and lexical features of registers in different fields of study, so teaching materials took these linguistic features as their syllabus. Then there was a tendency to have the students negotiate meanings according to the demands of the specific settings. This process of negotiating meaning according to the context of the situation is what linguists call 'register'.

A register is a way of encoding social processes linguistically (Halliday, 1988; Swales, 1990). The way in which these social processes are put into codes may take an oral form or a written form. These forms may be seen in a reading class, for example:

Before asking the students to read the text, teachers usually start a dialogue to draw out the students' background knowledge on the content of the text to be read (this process is called 'brainstorming'). Afterwards, the class engages in a silent reading activity, and this is followed by a discussion of what has been read. Generally speaking, these classes conclude with a written activity. Thus, the language used in this particular setting by these particular persons may be said to be 'the language reading register' of that English lesson.

---

[2] According to Halliday, the term 'register' was first used for 'text variety by [Thomas Bertram] Reid (1956); the concept was taken up and developed by Jean Ure (Ure and Ellis, 1972)', and by Halliday, McIntosh and Strevens (1964). See now Halliday (1988).

However, Sampson (1992) points out that 'sometimes a particular activity recurs regularly in the social context in which the linguistic register is found' and that 'if that particular activity becomes stereotypic or standardized, then the linguistic manifestation of that activity also tends to become stereotypic or standardized'. This activity is then called 'genre'.

*The New Webster's Dictionary and Thesaurus of the English Language* defines the word 'genre' as "a type or category specially of works of arts and literature"; whereas *The Collins Cobuild English Language Dictionary* defines it as "a type of literature, art, music, etc., which people consider to have the same style or subject". As can be seen, the words 'type' and 'category' are included in the definitions plus the word 'subject'. Arts, literature and music are social manifestations which at the same time are backed up by linguistic manifestations. One could very well understand 'types of linguistic manifestations'.

Swales (1990) in his descriptive powerful model of *Genre Analysis* considers the use of the term 'genre' in four fields:

a) folklore

b) literary studies

c) rhetoric

d) linguistics

In folklore studies one group of approaches examines genres as a classificatory category (Ben-Amos, 1976); for example, a story may be classified as myth, legend or tale.

Another group of approaches sees genres as *forms* , and these forms through tradition are taken as permanent; for example,

proverbs. Swales states that not all folklorists support the concept of 'permanence of form'; some are rather interested in the evolution of genres as an answer to the changing world.

The last group of approaches supports the functional aspect of narratives in the community.

The fact that folklorists see genres as functional and that some of them see genres as 'permanent forms' with an open door to change, this permanency depending on the evolution of the world shows, on one hand that they see genres as means to an end, and on the other hand that they understand genres as social processes.

In literary studies Swales says that 'literary critics and theorists have special reasons for de-emphasizing stability, since their scholarly activity is typically designed to show how the chosen author breaks the mould of convention and so establishes significance and originality.' This has to do with the stylistic devices writers use when writing, and how certain authors are characterized by a given style. He then refers to Todorov's way of seeing genre generation as the transformation of one or several old genres by inversion, by displacement and by combination, and points out how Todorov sees the influence of society on genre. The value of taking genre into consideration when analyzing a piece of literature falls on the fact that genre provides an interpretative and evaluative structure for a work of art, and that this structure gives the same importance to text, culture, history and society.

In rhetoric, two main points of view can be seen: the deductive and the inductive. Those on the deductive side construct a closed system of categories; for example, expressive, persuasive, literary,

and referential. The classification is based on the element in the communication process which receives the primary focus:[3]

a) if on the sender, expressive;

b) if on the receiver, persuasive;

c) if on the linguistic form, literary;

d) if on the realities of the world, referential.

Swales argues that although these classifications have considerable organizing value, the tendency for an 'early' categorization may lead to a failure to understand particular discourses in their own terms.

The other point of view, lead by the inductivists, take context more into account and give gender a more central place. Swales quotes Jamieson (1975) who stated the following:

> Three bodies of discourse may serve as evidence for the thesis that it is sometimes rhetorical genres and not rhetorical situations that are decisively formative. These bodies of discourse are the papal encyclical, the early state of the union addresses, and their congressional replies. I will argue that these discourses bear the chromosomal imprint of ancestral genres. Specifically, I propose to track essential elements of the contemporary papal encyclical to Roman imperial documents and the apostolic epistles, essential elements of the early state of the union addresses to the 'King's Speech' from the throne, and essential elements of the early congressional replies to the parliamentary replies to the king.
>
> (Jamieson, 1975)

With the previous statement Jamieson offers a way of studying discourse development over time. This offer provides a way

---

[3] See Kinneavy's (1971). *A Theory of Discourse: The Aims of Discourse*

to clarify certain social aspects of rhetoric that can be missed if analyzed differently. Miller (1984) shares this view, but goes further in her analysis because she extends the scope of genre analysis to types of discourse. She argues that a definition of genre must be centered not on the form of discourse, but on the action it is used to accomplish, and she gives special attention to how genres fit into the scales of human affairs.

Generally speaking, rhetoricians made an essential contribution to linguistics when providing a historical context to the study of genre, when proposing a way to organize genres, and when reinforcing the purposeful role of genre.

Anthropologists and some linguists view genre as 'speech events', restricted to activities or aspects of activities that are directly governed by rules or norms for the use of speech (Hymes, 1974). They also see them as 'types of communicative events', and are specifically interested in knowing which communications are generically typed in a specific community and which labels the community uses, because both aspects show elements of verbal behavior considered to be very important by the community (Saville-Troike, 1982).

Linguists have also considered genre in relation to 'register' (Gregory and Carroll, 1978; Frow, 1980; Martin, 1985; Couture, 1986). Register, according to Gregory and Carroll, is 'a contextual category correlating groupings of linguistic features with recurrent situational features'. Three variables have been given to this category: field (type of activity where the discourse operates), tenor (status and role relationships of the participants), and mode

(written or oral communication). These categories provide a conceptual basis for analysis. Frow does not see any distinction between genre and register; Martin, however, thinks that 'genres are realized through registers, and registers in turn are realized through language' . He gives the following definition of genre:

> Genres are how things get done, when language is used to accomplish them. They range from literary to far from literary forms: poems, narratives, expositions, lectures, seminars, recipes, manuals, appointment making, service encounters, news broadcasts and so on. The term genre is used here to embrace each of the linguistically realized activity types which comprise so much of our culture.
>
> (Martin, 1985)

Martin sees that genre influences the ways in which register variables can be combined in a society; so he points out that genre is a system underlying register. He points out that 'verbal strategies can be thought of in terms of states through which one moves in order to realize a genre'. This provides a good rationale for carrying out an analysis of discourse structure.

Couture says that 'unlike register, genre can only be realized in complete texts or texts that can be projected as complete, for a genre does more than specify kinds of codes in a group of related texts; it specifies conditions for beginning, continuing and ending a text'. For him, genres are whole structured texts and registers are generalizable stylistic choices.

Swales draws three conclusions from contributions of linguists . They view genres as:

a) types of goal-directed communicative events,

b) having schematic structure,

c) disassociated from registers or styles.

Taking the conclusions of the four approaches into account, Swales define genre as:

> "... a class of communicative events, the members of which share some set of communicative purpose. These purpose are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre... Exemplars of a genre exhibit various patterns of similarity in terms of structure, style, content, and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community..."
>
> (Swales, 1990)

He shares Martin's (1985) viewpoint that genres are goal oriented, social processes.

As can be seen, genres introduce a certain stability into a discourse community and are flexible enough to participate in social changes, so from this point of view they function as language itself. Because of this , they have become key points in some investigations carried out in ESP.

For example, Swales' analysis of genre (1981, 1984, 1990) has served as a reference for different studies on the teaching of ESP using a genre-based approach (Widdowson, 1983; Crookes, 1986; Marshal, 1991; Nwogu, 1991). The model Swales uses in his analysis is in line with schema-theoretic principles of information processing developed by Rumelhart (1981). Adopting strategies similar to those embodied in schema-theoretic models, Swales

(1981) posits a four 'move' schema for article introduction in ESP courses and specifically for scientific discussions. His study demonstrates not only an attempt to chunk texts into identifiable knowledge structures, but a concern with characterizing the linguistic features of each move and the means by which information in the move is signalled.

The moves he proposes are the following:

I. Establishing the field: Move one
    1. Showing centrality
        a) by interest
        b) by importance
        c) by topic-prominence
        d) by standard procedure
    2. Stating current knowledge
    3. Describing characteristics

II. Summarizing previous research: Move two
    1. Author orientations
        a) strong
        b) weak
    2. Subject orientations

III. Preparing for present Research: Move three
    1. Indicating a gap
    2. Question raising
    3. Extending a finding

IV. Indicating Present Research: Move four
    1. Giving the purpose
    2. Describing present research

a) by this/the present signals

b) by move 3 take-up

c) by switching to first person pronouns

Defining the distinctive features of the variety of English used in scientific contexts has been a special tendency in the past few years (Barber, 1962; Ewer and Latorre, 1969; Swales, 1984; Halliday, 1988; Reid, 1991). This is so because English has emerged as the predominant medium of scientific discussion and progress, hence theoretical and practical applications on the teaching of English have become a powerful need in all parts of the world.

Although these studies are grounded on a firm basis and offer a rationale for genre analysis, Douglas Biber (1988) in his book *Variation Across Speech and Writing* offers a model in which texts can be compared along dimensions of linguistic variation. This is the most sophisticated study on genre differences that has been published so far. By computing factor scores, that is, by summing up the frequency of each of the linguistic features in a factor for each text, he was able to average the factor score for each text across all texts in a genre and compute a mean dimension score for the genre. He then used these mean dimension scores to compare and to specify the relations among genres.

The utility of genre analysis in the teaching of ESL has been demonstrated by all the studies presented in the literature review herein. However, few of them offer hard data on specific fields such as medicine, physics, biology and math. Because Biber's study provides a foundation for cross-linguistic research, the present study will search for the linguistic features which are shared

specifically by texts in the field of biological science and compare the findings to the general science corpora that have been described in Biber's analyses.

The findings of this analysis of concepts will help curriculum developers who rely on these general science corpora when developing teaching materials to create more accurate and relevant teaching materials.[4]

The terms 'dimension', 'genre' and 'text' are key words in this study. Getting clearer about the meaning given to these three concepts in the present work is the step to be taken now.

Biber (1988) compares texts along 'dimensions' of linguistic variations. He states that researchers have found out that texts are related along particular situational or functional parameters. e.g. formal/informal, interactive/non-interactive, literary/colloquial, restricted/elaborated. These parameters can be considered as

---

[4] In 1987, a group of ESP professors from the University of Havana, started planning for organizational change of the existing curriculum for first-year ESP-university students. The existing program at the time did not meet the needs of the students. The majority of the reports sent by the different faculties to the department expressed lack of students' motivation in learning English, lack of continuity between pre-university and university studies, and very little use of the reading books which were supposed to be used by the teachers in the classroom.

Knowing that in the development of a curriculum it is essential to be sure of the educational purposes which the school should seek to attain (in our specific case the most important one being reading comprehension), on the basis of the interests of the students, and their needs, the instructional objectives were set out.

After stating the objectives, some professors analyzed the materials which the students were using. The primary sources of language input for the learners were a reading book and a workbook. The existing materials had to be adapted because the majority of the readings included in the textbook (Training in Effective Reading I) were not related to the students' specialties; therefore, they were not interesting for the learners. A group of professors prepared sets of Reading Selections specially for the program following the opinions of experts from different faculties as to which articles to include in the book. Different sets were issued for each one of the faculties, so as to group authentic 'texts' related to the students' interests and specialties. The corpus of text which will be analyzed in this study will be drawn from one of these sets: *Reading Selections for Biological Science Students.*

dimensions because 'they define continuums of variations rather than discrete poles.

According to Biber, dimensions may also contribute to comparing texts in terms of their linguistic characterization: nominal/verbal, having a complex/simple structure, and the like.

In his work he uses frequency counts of particular linguistic features as a means to give exact quantitative characterization of a text; however these counts do not identify linguistic dimensions. Linguistic dimensions are characterized on the basis of a consistent co-occurrence pattern among features; that is, the consistent co-occurrence of a cluster of features in texts define a linguistic dimension.

The approach used by Biber in his study completely differs from previous studies. Other studies began with a situational or functional distinction and afterwards identified linguistic features associated with that distinction; Biber identifies the clusters of features in terms of shared function, but without necessarily representing a linguistic dimension. Biber uses quantitative techniques to identify the groups of features and then interprets them in functional terms. The linguistic rather than the functional dimension is given priority.

He bases this approach on the idea that "if certain features consistently co-occur, then it is reasonable to look for an underlying functional influence that encourages their use". For example, let us consider some frequency counts in two texts (conversational and scientific), and let us analyze the frequencies for two linguistic features: passive constructions and nominalizations.

a) Conversational text (the type that would appear in an ESL/EFL textbook):

A: "Hi Barb, What are you up to?"

B: "I'm taking these books back to the library."

A: "Will you be free after that?"

B: "I'm afraid not. I have to drop my sister off at the airport at noon."

A: "What will you do after you drop her off?

B: "I have a meeting with my supervisor at 3 o'clock."

A. "You're really busy today!.Would you like to get together after the meeting?"

B. "I'd love to, but by that time, I'll probably be dead tired!"

A. "That's fine with me. Let's make it another day. See you tomorrow!"

B: "See you then, Allan. Take care!"

A: "You too."

b) Scientific text:

Until recently it was considered that the main problem for populations in developing countries was undernutrition due to inadequate food supplies. Frequent infections were seen as a result of microbial exposure in individuals who were immunodeficient secondary to undernutrition. This concept has now changed, following a number of studies showing that undernutrition most often occurs as a result of frequent infections. It is known that infections lead to loss of appetite and to fever, which is energy consuming. Moreover, effects on the intestinal mucosa during acute gastroenteritis may disturb food absorbtion. These effects result from activities of the host defense.

(Source: Reading Selections for Foodscience Students, pp.15)

Both texts used (100 words each) are quite different in relation to these linguistic features. The scientific text has 3

passives and 9 nominalizations, whereas the conversation has neither passives nor nominalizations. If these two texts were representative of their kind, one could easily say that passives and nominalizations tend to co-occur and thus, belong to the same linguistic dimension.

Now, if in the same texts, we were to analyze the frequency for two different linguistic features, let's say pronouns (1st and 2nd) and contractions, we would see that the conversational text has 15 pronouns and 7 contractions while these features are not present at all in the scientific text.

We can then say that the passive-nominalization dimension and the pronoun-contraction dimension are part of the same dimension because they pattern in a consistent way. The relation may be as follows between conversational and scientific texts:

When there are many passives, there are many nominalizations; when there are many passives and nominalizations, there are few pronouns and contractions and vice versa. When there are many pronouns and contractions, there are few passives and nominalizations. That is, the marked features in one text presuppose the absence of the unmarked feature in the other text.

Once the linguistic co-occurrence patterns are identified, the resulting dimensions can be interpreted in functional terms. The approach moves from stating WHAT features co-occur to explaining the WHY of their co-occurrence.

After identifying and interpreting the linguistic dimensions, they can be used to specify 'textual relations'. Textual relations are

defined by a simultaneous comparison of the texts with respect to all dimensions.

There are several issues related to the notion of texts used in corpus studies of variation; for example, text length. Text samples must be long enough to represent the linguistic characteristics of the full text, but not so long as to add unnecessarily to the work required to compile and use a corpus. Texts in this study will be identified as 'continuous segments of naturally occurring discourse' (Biber and Finegan, 1991).

'Text types' and 'genre' have been words used nondistinctively in studies of textual variation (Biber, 1988). A comprehensive response to this problem of nondistinctive use may be the fact that genres are said to be 'text categories readily distinguished by mature speakers of English' (Biber and Finegan, 1991). In Biber's view then, the text categories used in his corpora are genres. Text types, on the contrary, have a strictly linguistic basis; they are sets or grouping of texts so that the texts within each group are linguistically similar, while the groups themselves are characterized by being linguistically different.

A complete linguistic description of a genre should include, according to Biber (1988), a characterization of the central tendency of the typical text and a characterization of the range of variation. Genres have a wide range of variation because:

a) they include distinguishable subgenres. For example, scientific texts may include articles from biology, physics, math, etc. These subgenres are often significantly different in their linguistic characterization (Biber, 1988); and

b) they show considerable differences in the extent to which they have found a norm, for example, fiction and essays had a wide range of variation in English during the 18th and early 19th centuries reflecting the different purposes and audiences for these genres (Biber, 1988).

In these cases the wide range of variation does not invalidate genre category. The ranges reflect various functional and developmental characteristics of that category.

Summary

Different ideas have been raised in this first chapter:

1. General background on language acquisition
    a) first language
    b) second language
2. General World of ELT
    a) EMT
    b) ESL/EFL
        - GE
        - ESP
3. Historical background on the birth of ESP
4. Some insights into studies in varieties of English; definitions of:
    a) 'registers' and 'genres'
    b) 'dimensions'
    c) 'text' and 'text types'
5. The basic rationale for this thesis.

## Chapter II
## Methodology

Method

So far, researchers have investigated linguistic textual variations using either a microscopic or a macroscopic analysis or a combination of the two (Schiffrin, 1981; Besnier, 1983; Biber, 1988).

Microscopic analysis identifies the linguistic features and genre distinctions to be included in a macro analysis, and provides a functional analysis of the features, so as to be able to interpret the textual dimensions in functional terms.

Macroscopic analyses pinpoint the underlying textual dimensions in a set of texts, enable the general description of a general account of linguistic variations among texts, and provide a framework for the discussions of the similarities and differences among texts and genres.

Both micro and macro analysis will be used here:

a) A macroscopic outlook to analyze the co-occurrent patterns among ten linguistic features in 24 texts, identifying two textual dimensions; and

b) A microscopic analysis to identify the features and to interpret the dimensions in functional terms.

Biber identifies six textual dimensions in his study:

a) Involved versus Informational Production,

b) Narrative versus Non-Narrative Concerns,

c) Explicit versus Situation-Dependent Reference,

d) Overt Expression of Persuation,

e) Abstract versus Non-Abstract Information, and

f) On-line Informational Elaboration.

From these six textual dimensions, relevant salient loadings were reported in the analysis of the general science corpora in 'Explicit versus situation dependent' and 'Abstract versus Non-abstract information'. These two textual dimensions will be key points for the comparisons which will be established between the results of the analysis between the biological science corpora and the scientific texts included in Biber's study.

Text selection

There are several issues relating to the character of texts used in corpus studies of variation; one is the source. This study is based on a corpus of 24 texts from the field of biology selected from the book *Reading Selections for Biological Science Students*. This book is used by ESP professors at the University of Havana to conduct reading classes in English to second year university students from the Faculty of Biology.

All texts in the book were published between 1986 and 1989 (See Appendix A). The professors in charge of editing the articles included in the book (Figueredo, 1991, et al.) took into consideration the fact that the biological sciences are divided into three main branches in that faculty: Microbiology, Biology and Biochemistry. So they had to ensure balance among the branches when selecting the

articles representing each branch in the book. This subdivision in biological science leads to a subdivision in 'genres'. In this study these subdivisions will be named 'subgenres'. Subgenres, then are sub-classes of genres which may be different from one another. The study will also compare the relations of linguistic features among the three subgenres: biology, microbiology and biochemistry.

Another issue relating to the character of texts is text length. As it was previously mentioned, texts should be long enough to represent reliably the linguistic characteristics of the full text, but not so long as to add unnecessary information not to be used in the analysis. Few empirical investigations of variation within texts and optimal text sample length propose the analyses of the distribution of linguistic features across 1000-word texts samples extracted from larger texts (Biber 1990).

This study will include 700-word texts samples[5] (See Appendix B) extracted from larger texts, inasmuch as previous studies have indicated that such shorter extracts do reliably represent at least certain linguistic characteristic of a text (Biber,1988). To analyze all these texts without the aid of a computer would require several years, but the use of a computerized corpus in this study will enable automatic inclusion of the texts in readable codes for the computer with the use of the scanner, automatic counting of words, and automatic identification of

---

[5] Because balance had to be ensure among the three subgenres when selecting the texts from the book *Reading for Biological Science Students,* there were some articles which did not have 700 words. The distribution of words per article is as follows: Fifteen articles have 700 words and 9 have between 407 and 689 words. As all the counts were normalized to a text length of 1000, the difference between text length does not constitute a problem.

linguistic features in a collection of texts. The automatic identification of linguistic features will be done with the use of AnyText™, a Hypercard® based program that allows one to do fast word searches on any text-only files. After the computational analysis is done, inspection by the analyist of the computer results to check for errors is also mandatory.

Linguistic features

For the purpose of this study, Biber's research was surveyed to identify the relevant features characteristic of the scientific genre. Among the 67 linguistic features for all genres that Biber identifies, I selected 10 of those that co-occurred the most in scientific writing and grouped them into six major grammatical categories:

a) Passives
    1. agentless
    2. by-passives

b) Pronouns
    1. third person pronouns
    2. it
    3. demonstrative

c) Modals
    1. possibility
    2. predictive

d) Nominalizations (*-tion, -ment, -ness, -ity* including the plural forms)

e) Perfect tense

d) Conditionals

Although these features have been organized in grammatical categories, they may be said to be functional markers in the texts. As Biber (1988) says, 'features from the same grammatical category can have different functions, and features from different grammatical categories can have a shared function'. This provides a rationale for determining the underlying functional dimensions.

In Chapter I reference has been made to some of the functions related to certain linguistic features; however, a detailed description of the features has not been given so far. This is important first, to let readers determine exactly which forms were counted as instances of each feature; and second, to take into account the exact instances Biber took into account in his study:

1. Agentless passives

   a) BE + (Adv) + any past participial verb

   *Example: It was also found that...*

   b) BE + noun or pronoun + any past participial verb

   *Example: Were you told that...?* (questions)

2. By-passives

   a) BE + (Adv) + any past participial verb + by

   *Example: The samples were analyzed by the chemist.*

   b) BE + noun or pronoun + any past participial verb + by

   *Example: Were the samples analyzed by the chemist?*

3. Third person pronouns

*she, he, they, her, him, them, his their, himself, herself, themselves* (plus contracted forms).

*Example: They posit that...*

Third person personal pronouns mark relatively inexact reference to persons outside of the immediate interaction. Biber (1986) finds that they co-occur frequently with perfect aspect forms.

4. Pronoun *it*

*It* is the most generalized pronoun, since it can refer to animate beings or to abstract concepts. It can be substituted for nouns, phrases, or whole clauses.

*Example: It is a substance that can be used in ...*

5. Demonstrative pronoun

a) that/this/these/those (where that is not a relative clause)

b) that's

*Example: This demonstrates the validity of...*

6. Possibility modals

*can, may, might, could* (+ contractions)

*Example: Propolis may be used in...*

7. Predictive modals

*will, would, shall* (+ contractions)

*Example: They will face far more barriers ...*

8. Nominalizations

All words ending in *-tion, -ment, -ness, or -ity* (plus plural forms.

*Example: The full implications of infection...*

Nominalizations have been used in many register studies. Biber (1986) finds that they tend to co-occur with passive constructions.

9. Perfect aspect

a) HAVE + (adverb) + any past participial verb

*Example: No infection has occurred...*

b) HAVE + noun or pronoun + any past participial verb

*Example: Have they been infected by the virus?*

(These include contracted forms of HAVE).

Perfect aspect forms mark actions in past time with 'current relevance' (Quirk et al., 1985). They have been associated with certain types of academic writing (Biber, 1986).

10. Conditional subordinators

*if, unless*

*Example: It will become cancerous unless the bowel is removed.*

Frequency counts

The frequency counts of linguistic features will be normalized to a text length of 1000 words[6]. Normalizing text length is mandatory for any comparison of frequency counts across texts because, even though a text length may not be very relevant in relation to another, the fact that the amount of words differs, may lead to an inaccurate assessment of the frequency distribution in texts. For example, let us compare two texts different in length:

[6] Biber also normalized his corpora to a text length of 1000 words. In order to compare the results of this study to those of Biber's, the texts have to be normalized to the same amount of words.

Text I

"...One of the most important questions in the area of food and nutrition relates to requirements..."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 13)

Text II

"...In 1854 Pasteur began his brilliant studies on yeast and fermentation showing fermentation to be a vital entity, with yeast being the living entity responsible. His observations were correct; his interpretation of the data required adjustment only to acomodate the idea that fermentation could be carried out by extracts properly made from living yeast cells..."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 41)

The raw frequency count of nominalizations shows that Text II has more nominalizations (8) than Text I (3); but the amount of nominalization in Text II is based on a count of the number of 55 words of text, that is more than three times as many words as Text I just having 16 words. So there are many more opportunities for nominalizations to occur in the 55 word-length text than in the other one. Raw counts will never represent comparable frequencies of occurrence unless the texts are the same length. By computing how many features would occur if the text were 100 words, for example, the frequencies can be compared directly:

*Text I*

*3 (nom.) ÷ 16 (length of text) x 100 = 18.7 (nom.)*

*Text II*

*8 (nom.) ÷ 55 (length of text) x 100 = 14.5 (nom.)*

As observed, by normalizing the counts so that they represent the frequencies per 100 words, we can conclude that Text I uses more nominalizations than Text II.

The frequency of occurrence of the linguistic features to be anlayzed in the study will be given in five different values.

a) the mean frequency,

b) the maximum frequency,

c) the minimum frequency,

d) the range (difference between maximum and minimum frequencies), and

e) the standard deviation (a measure of the spread of the distribution).

Let us analyze the frequency of occurrence of the passive voice in the following example:

A: "... Today it can be said, with a certain degree of certainty that, at least in industrialized countries, some 50% of all poultry carcasses may be infected with salmonella."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 55)

B: "... It is known that, when virus on a food or food-contact surface is inactivated by a strong oxidizing agent such as chlorine, the yet-to-be-defined reaction to which the virus is subjected is driven by chemical energy."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 65)

C: "... The aflatoxins are a group of secondary fungal metabolites that are potent animal toxins and carcinogens and have been epidemiologically implicated as environmental carcinogens in man."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 81)

The three texts differ both in the number of words they have and in the occurrences of the linguistic feature to be analyzed: *passive voice.* Once the raw counts are normalized (let us suppose that the texts are 50 words each), the mean frequency, the minimum frequency, the maximum frequency, the range and the standard deviation can be determined.

The procedure is the following:

a) counting the words

*Text A: 28*

*Text B: 37*

*Text C: 26*

b) counting the occurrences of passives

*Text A: 2*

*Text B: 4*

*Text C: 1*

c) normalizing the counts

*Text A: 3.5*

*Text B: 5.4*

*Text C: 1.9*

d) calculating the mean frequency

*3.5 + 5.4 + 1.9 = 10.8 ÷ 3 = 3.6*

e) looking for the minimum frequency

*1.9*

f) looking for the maximum frequency

*5.4*

e) calculating the range

*5.4-1.9 = 3.5*

f) calculating the standard deviation

$3.5^2 + 5.4^2 + 1.9^2 = 12.25 + 29.16 + 3.61 = 45.02 \div 3 = 15.0$

*square values of normalized frequencies*

$3.5^2 = 12.25$

*square value of mean frequency*

*15.0 -12.25 = 2.75*

$\sqrt{2.75} = 1.6$

The standard deviation is a measure of the spread of frequency values of a feature. This standardization of values reflects the magnitude of a frequency with respect to the range of possible variation.

Factors

Factors represent an area of high shared variance in the data, a grouping of linguistic features that co-occur with a high frequency. Factors are defined by correlations among the frequency counts of linguistic features; that is, when several linguistic features are highly correlated, then a factor is defined. For example, let us consider the correlations between passives and that-clauses in the following texts:

A: "... Today it can be said, with a certain degree of certainty that, at least in industrialized countries, some 50% of all poultry carcasses may be infected with salmonella."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 55)

B: "... It is known that, when virus on a food or food-contact surface is inactivated by a strong oxidizing agent such as chlorine, the yet-to-be-defined reaction to which the virus is subjected is driven by chemical energy."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 65)

C: "... The aflatoxins are a group of secondary fungal metabolites that are potent animal toxins and carcinogens and have been epidemiologically implicated as environmental carcinogens in man."

(Source: *Reading Selections for Food Science Students,* University of Havana, 1991; pp. 81)

Text A has 28 words, Text B 37 and Text C 26. After a raw count of the linguistic features to be analyzed, the next step is normalizing the counts. Let us suppose that the texts are 50 words each. The following procedure follows:

| *text* | *length of text* | *raw counts* | | *frequency counts* | |
|---|---|---|---|---|---|
| | | *passives* | *that-clauses* | *passives* | *that clauses* |
| *A* | *28* | *2* | *1* | *3.5* | *1.7* |
| *B* | *37* | *4* | *1* | *5.4* | *1.3* |
| *C* | *26* | *1* | *1* | *1.9* | *1.9* |

As these features frequently co-occur, a factor can be defined; that is the passive/that-clause factor; and in this way the rest of the features are reduced and other factors are defined. Once the factors are defined they have to be analyzed.

The first step in a factor analysis is choosing a method for extracting the factors. In linguistics, the use of factor analysis is

generally exploratory. Although there are several options of factor analysis available, the study will include the most widely used known as 'common factor analysis' (Biber, 1988).

Common factor analysis extracts the minimum amount of shared linguistics features. So the first factor extracts the maximum amounts of shared linguistic features; that is the first factor would correspond to the largest group of co-occurrence in the data (passive-nominalizations, for example); the second would then extract the maximum amounts of shared linguistic features from the tokens left over after the first factor has been analyzed, and so on.

When interpreting factors, underlying functional dimensions seem to explain the co-occurrence patterns among features identified by the factors. The rule is:

*if a cluster of features co-occurs frequently in texts, the features will reveal a common function in those texts .*

However, interpretation of factors are tentative until they are confirmed in further research.

In the same way that the frequency of passives in a text may be called the passive score of that text, factor scores are computed for each text to characterize the text with respect to each factor. A factor score or dimension is computed by summing, for each text, the number of occurrences of the features having salient loadings on that factor. For instance, let us analyze the salient loadings of each of the following texts:

a) Conversational text

A: "Hi Barb, What are you up to?"

B: "I'm taking these books back to the library."

A: "Will you be free after that?"

B: "I'm afraid not. I have to drop my sister off at the airport at noon."

A: "What will you do after you drop her off?

B: "I have a meeting with my supervisor at 3 o'clock."

A. "You're really busy today!.Would you like to get together after the meeting?"

B. "I'd love to, but by that time, I'll probably be dead tired!"

A. "That's fine with me. Let's make it another day. See you tomorrow!"

B: "See you then, Allan. Take care!"

A: "You too."

b) Scientific text:

Until recently it was considered that the main problem for populations in developing countries was undernutrition due to inadequate food supplies. Frequent infections were seen as a result of microbial exposure in individuals who were immunodeficient secondary to undernutrition. This concept has now changed, following a number of studies showing that undernutrition most often occurs as a result of frequent infections. It is known that infections lead to loss of appetite and to fever, which is energy consuming. Moreover, effects on the intestinal mucosa during acute gastroenteritis may disturb food absorbtion. These effects result from activities of the host defense.

(Source: Reading Selections for Foodscience Students, pp.15)

The salient loadings in the conversational text are the use of contractions and third personal pronouns, and the salient loadings in the scientific text are the use of passives and nominalizations. The factor scores for each of these text would be the following:

*Conversational*

*15 pronouns + 7 contractions = 22 (factor score)*

*Scientific*

*9 nominalizations + 3 passives = 12 (factor score)*

Summary

In this second chapter, an overall view of the methodology to be followed in the study has been given:

I. Methodological aspects

1. Method

a) Microscopic

b) Macroscopic

2. Rationale for text selection

a) Source

d) Text length

3. Computational program used (AnyText™)

3. Linguistic features

II. Statistical aspects

1. Frequency counts

2. Factors as a basis to determine linguistic dimensions.

## Chapter III

## Data Collection

As it has been mentioned above, 24 texts extracted from larger texts included in the book *Readings for Biological Science Students* were included in the analysis. These 24 articles (See Appendix A) have been generally selected by the professors in our department to carry out intensive analysis of the text in class, the others are used for extensive reading. The texts were automatically converted to readable codes for the computer with the use of a scanner, and words in each text were automatically counted. Afterwards, each text was converted to a text-only file[7] as the HyperCard® based program used, AnyText™, operates with these codes.

AnyText™[8] was designed especially to work with the Greek, Hebrew/Aramaic, and English Biblical texts, but may be used with any text-only file.

Among the uses that AnyText™ provides, the following were included in our study:

a) Wild card string searches through the indexed word lists,

b) Area for collecting and taking notes, and

---

[7] Virtually all word-processing programs provide a SAVE AS option that allows you to do this.

[8] AnyText™ is a HyperCard-based Macintosh application created by Philip B. Payne, David Stringham and Michael Saia and produced by the Linguist's Sofware Inc., P.O. Box 580, Edmonds, WA 98020-0580, U.S.A.

c) Ability to create text files containing complete concordances and word lists or the partial word list and concordances resulting from proximity searches.

When working with AnyText™, the analyst feels very comfortable because of the speed and accuracy of the searches. Once the AnyText™ folder is double-clicked, a screen appears with different bottons which provide the analysist with different functions in order to work with the files (See Figure 1). The first operation carried out with each of the 24 files was to build the index. After clicking the **Build Index** botton, another screen appears in which one sets up certain resources that are needed for sorting and indexing (See Figure 2). In all cases the selections given by the program were accepted.

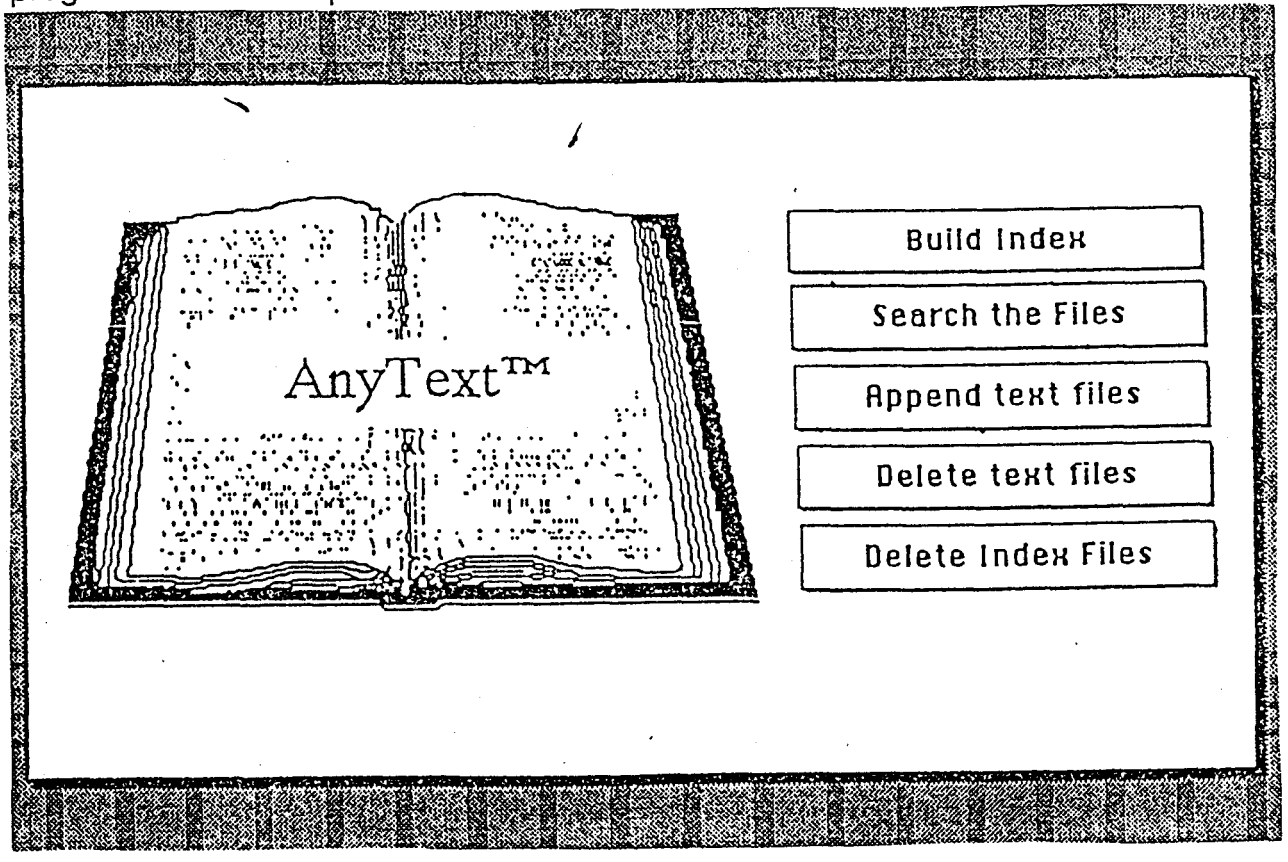


**Figure 1**

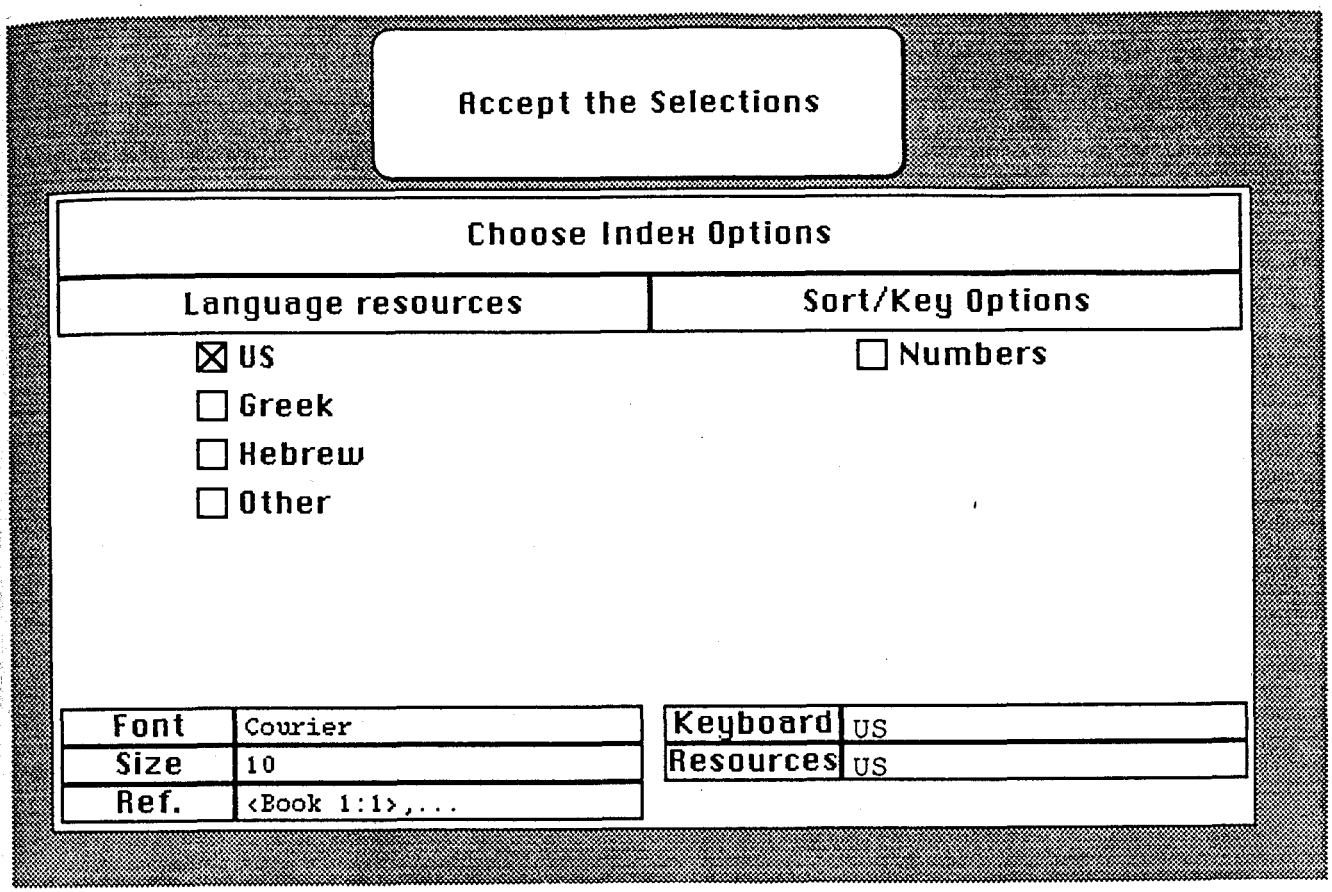


Figure 2

After clicking the **Accept the Selections** button, one is allowed to select a text file. Then one opens the file and the program gives messages indicating the progress and displays a 'beach ball' indicating that the computer is busy. This did not take a long time in this study, as the largest texts were just 700 words long. Then the **Search Screen** (or card) appears and the work starts.

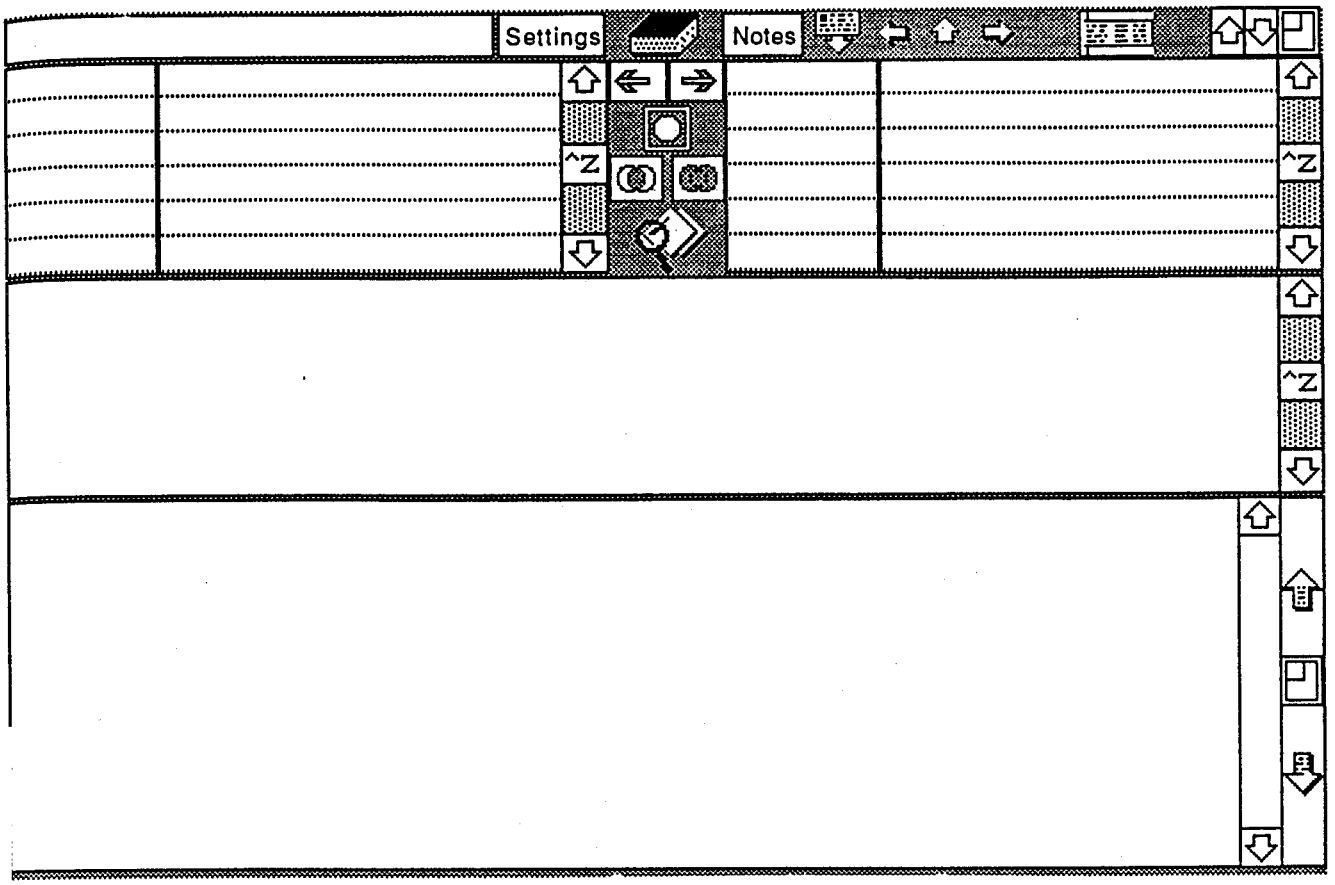


**Figure 3**

Once one clicks on the **Closed Book** icon in the middle of the screen at the top (see Figure 3) , the **Choose Index Options** window appears, one clicks on the **Accept the Selections** botton, and a **File Selection** window appears from which one selects a file by clicking on **Open.** If one opens the *Medicine-434 file* [9], for example, a list of all the unique words in the original text file will appear in two **Index fields.** The lists are in alphabetical order with **Index1** showing the first word in the list ("a") and **Index2** the last word ("xanthus") (See Figure 4).

---

[9] All the texts were given a name. If the texts were less than 700 words, the 'surname' of the text represented the number of words in the text.

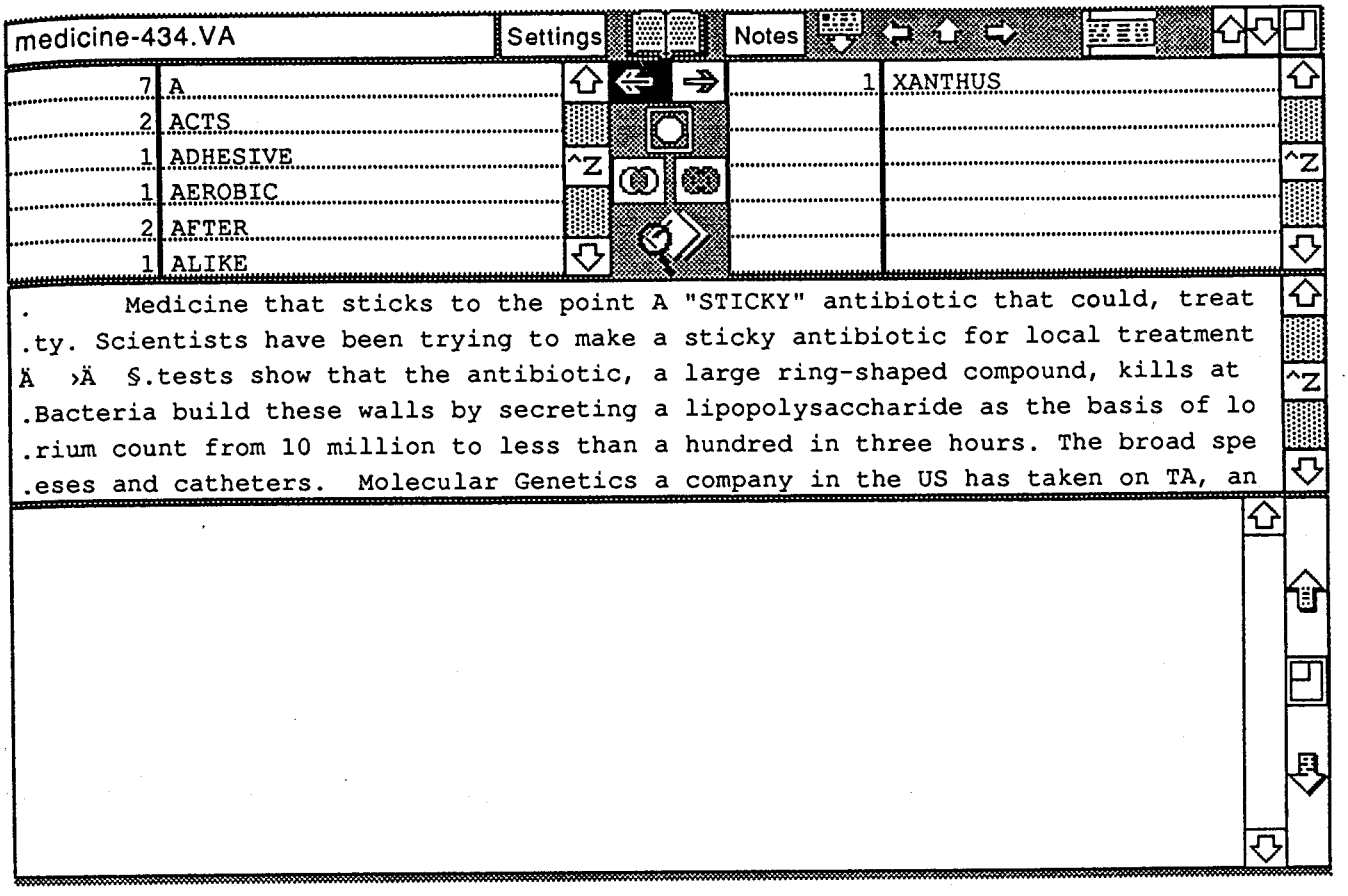


**Figure 4**

Between the two fields there are several buttons. The top arrow buttons show the field which is active. There is a scroll bar at the right side of each index fields that operates like a normal Macintosh™ scroll bar. With this scroll bar, words could be searched in just a few seconds. For example, if one is working with the *Medicine-434 file* and one wants to look for the occurrences of the

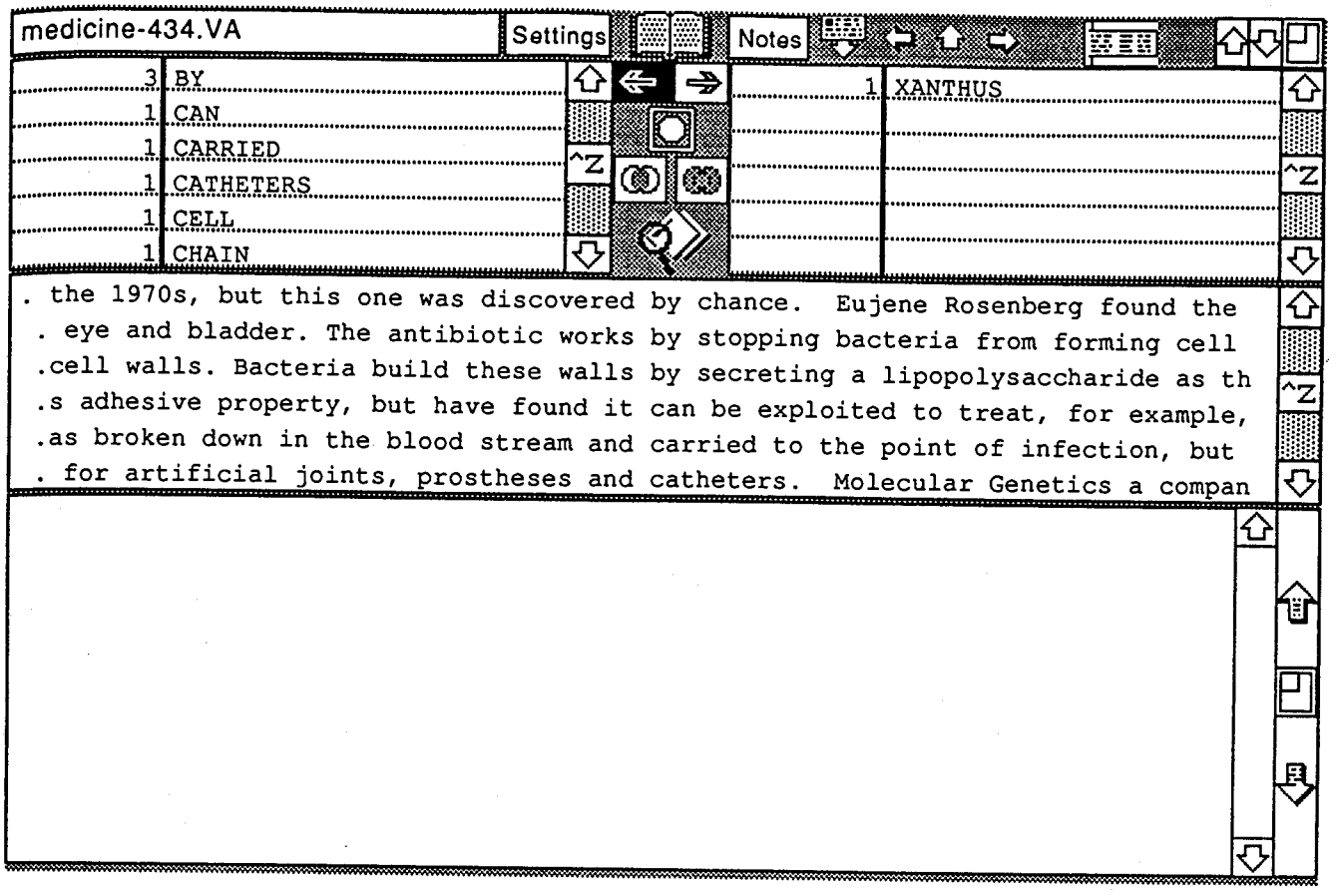


**Figure 5**

word *by* in the file, by just clicking on the **^Z** box of **Index1**, a dialogue box will pop up asking for the word one wants to find. Then one types the word *by* and a number indicating how many times the word occurs in the text will appear in the column to the left of the word[10] (See Figure 5).

If one wants to know the context in which the word appears, one just clicks on the word *by* and the listing with *by* centered in

[10] This is called Zipping to a word.

each line will appear. Thus, it was very easy to do the frequency counting of some linguistic features in this way (see Figure 5). But if the analyst wants to go further in the analysis, and wants to see the whole context where a specific word appeared, by double-clicking on the word in the middle of the screen, the text will appear at the bottom of the screen with the word highlighted. (See Figure 6) Agentless passives and nominalizations were hand counted as the

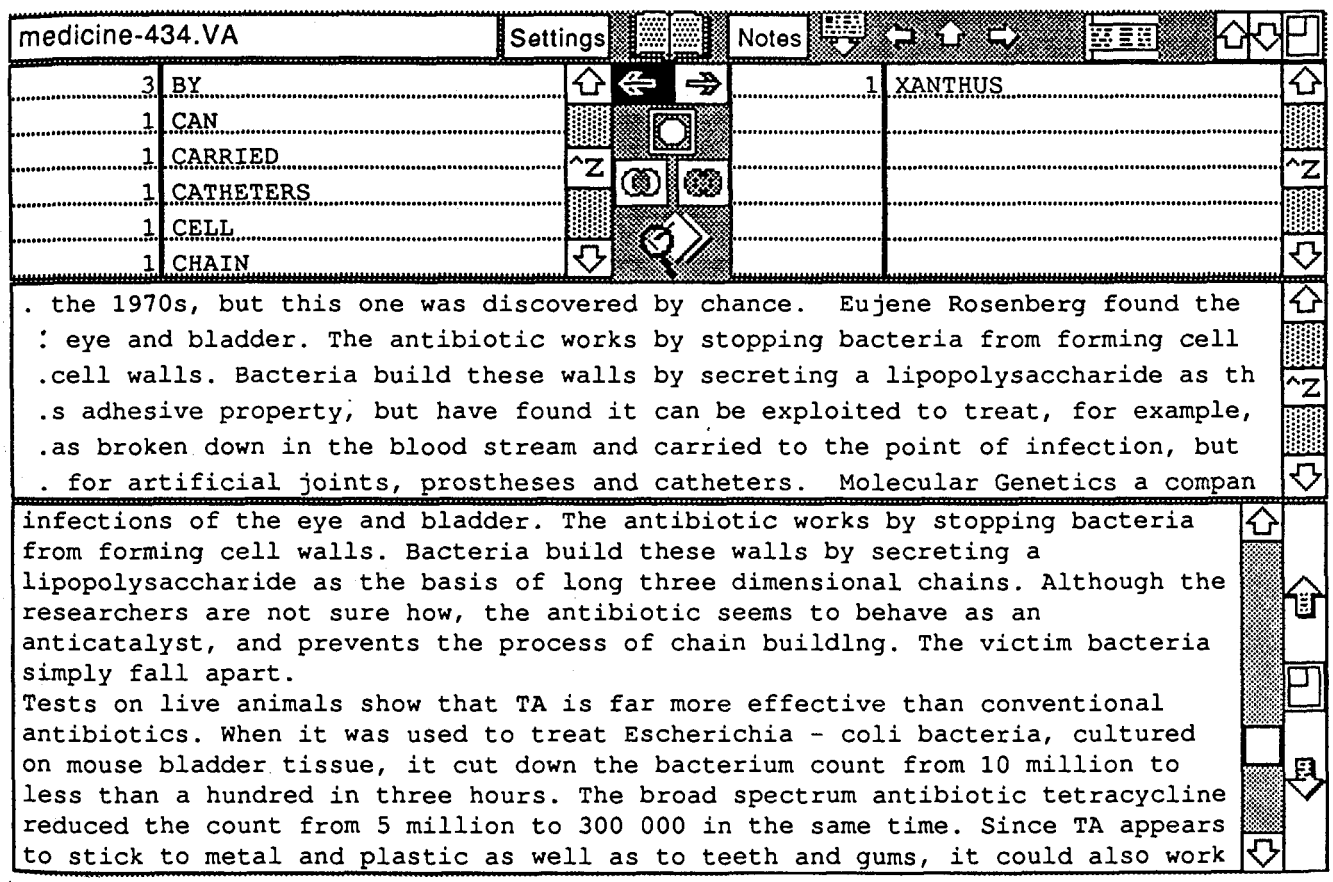


**Figure 6**

patterns of variations very much differ. In fact, it took much more time to do a proximity search of agentless passives using the computer, than counting the occurrences by hand.

After obtaining the raw number of all the occurrences of the linguistic features in each text, the counts were normalized to a text length of 1000 words. To recapitulate, when normalizing text length, one computes the number of features that would occur if the text were 1000 words.[11] Table 1.0 presents descriptive statistics for the frequencies of the linguistic features in the entire corpus of texts used in the study. Included are:

a) the mean frequency,

b) the maximum frequency,

c) the minimum frequency,

d) the range, and

e) the standard deviation

This table does not include the characterization of particular sub-genres, but provides an assessment of the overall distribution of particular features in biological science texts. Some features occur very frequently, for example nominalizations with a mean of 20.2 per 1000 words; other features occur very infrequently, for example, *by*-passives with a mean of 2.3 per 1000 words.

The variability in the frequency of features also differs from one feature to the next; some show a small difference of distribution across the corpus, such as conditional clauses. They have a maximum frequency of 8.5 per 1000 words and a minimum of 0.0 per 1000 words; other features show large differences, for

---

[11] Normalization= raw count of linguistic features ÷ length of text x 1000 words.

example predictive modals occurred 42 times in some texts but not at all in other texts. Tables 1.1 ,1.2 and 1.3 provide descriptive statistics of the frequency of each linguistic feature in each subgenre. The format for the four tables is the same.

If we were to organize the mean frequencies in Table 1.0 (for the entire corpus as a whole) according to the features having salient loadings, the order would be:

a) nominalizations
b) agentless passives
c) third person pronouns
d) demonstrative pronouns
e) perfect aspect
f) possibility modals
g) pronoun *it*
h) predictive modals
i) conditionals
j) *by*-passives.

Table 1.0 *Descriptive statistics for the corpus of biological science texts as a whole*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 12.8 | 5.7 | 30.0 | 24.3 | 10.6 |
| *by*-passives | 2.3 | 0.0 | 7.3 | 7.3 | 2.3 |
| 3rd person pronouns | 11.6 | 3.6 | 27.1 | 23.5 | 6.6 |
| pronoun *it* | 7.2 | 1.4 | 28.5 | 27.1 | 9.8 |
| demonstrative pr. | 11.1 | 2.8 | 24.2 | 21.4 | 5.9 |
| possibility modals | 9.2 | 0.0 | 24.5 | 24.5 | 6.1 |
| predictive modals | 4.2 | 0.0 | 42.0 | 42.0 | 8.3 |
| nominalizations | 20.2 | 0.0 | 38.5 | 38.5 | 10.2 |
| perfect aspect | 9.6 | 0.0 | 27.1 | 27.1 | 5.9 |
| conditionals | 2.4 | 0.0 | 8.5 | 8.5 | 2.4 |

Table 1.1 *Descriptive statistics for the corpus of the biochemistry sub-genre*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 12.7 | 5.7 | 30.0 | 24.3 | 12.2 |
| *by*-passives | 3.1 | 0.0 | 7.1 | 7.1 | 2.1 |
| 3rd person pronouns | 13.4 | 4.6 | 27.1 | 22.5 | 6.3 |
| pronoun *it* | 6.1 | 2.8 | 20.7 | 17.9 | 8.6 |
| demonstrative pr. | 9.8 | 7.1 | 14.2 | 7.1 | 2.2 |
| possibility modals | 10.6 | 4.6 | 24.5 | 17.9 | 6.5 |
| predictive modals | 7.1 | 0.0 | 42.0 | 42.0 | 13.3 |
| nominalizations | 20.4 | 0.0 | 38.5 | 38.5 | 13.2 |
| perfect aspect | 11.4 | 4.2 | 27.1 | 22.9 | 6.9 |
| conditionals | 2.4 | 0.0 | 7.1 | 7.1 | 2.4 |

Table 1.2 *Descriptive statistics for the corpus of the microbiology sub-genre*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 11.4 | 5.7 | 30.0 | 24.3 | 8.0 |
| *by*-passives | 1.4 | 0.0 | 7.3 | 7.3 | 2.3 |
| 3rd person pronouns | 7.8 | 3.6 | 15.7 | 11.9 | 3.8 |
| pronoun *it* | 7.0 | 1.4 | 28.5 | 27.1 | 11.8 |
| demonstrative pr. | 10.9 | 3.0 | 20.1 | 17.1 | 3.9 |
| possibility modals | 12.6 | 7.1 | 21.4 | 14.3 | 4.6 |
| predictive modals | 1.2 | 0.0 | 4.2 | 4.2 | 1.4 |
| nominalizations | 22.4 | 15.7 | 38.3 | 22.6 | 7.3 |
| perfect aspect | 7.0 | 0.0 | 11.6 | 11.6 | 3.7 |
| conditionals | 2.3 | 0.0 | 6.1 | 6.1 | 1.9 |

Table 1.3 *Descriptive statistics for the corpus of the biology sub-genre*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 14.2 | 7.1 | 29.8 | 22.7 | 11.2 |
| *by*-passives | 2.6 | 0.0 | 5.7 | 5.7 | 1.8 |
| 3rd person pronouns | 13.7 | 5.7 | 21.4 | 15.7 | 7.1 |
| pronoun *it* | 8.5 | 1.4 | 24.2 | 22.8 | 8.6 |
| demonstrative pr. | 12.6 | 2.8 | 24.2 | 21.4 | 9.0 |
| possibility modals | 4.5 | 0.0 | 12.8 | 12.8 | 3.6 |
| predictive modals | 4.4 | 0.0 | 11.2 | 11.2 | 3.5 |
| nominalizations | 17.8 | 2.8 | 27.3 | 24.5 | 8.6 |
| perfect aspect | 10.5 | 2.8 | 18.5 | 15.7 | 5.4 |
| conditionals | 2.6 | 0.0 | 8.5 | 8.5 | 2.8 |

The features having salient loadings in the biochemistry sub-genre would be organized as follows:

a) nominalizations

b) third person pronouns

c) agentless passives

d) perfect aspect

e) possibility modals

f) demonstrative pronouns

g) predictive modals

h) pronoun *it*

i) *by*-passives

j) conditionals.

Salient loadings in the microbiology sub-genre would fall in the following order:

a) nominalizations

b) possibility modals

c) agentless passives

d) demonstrative pronouns

e) third person pronouns

f) pronoun *it*

g) perfect aspect

h) conditionals

i) *by*-passives

j) predictive modals.

Finally, the features having salient loadings in the biology sub-genre would be organized in the following way:

a) nominalizations

b) agentless passives

c) third person pronouns

d) demonstrative pronouns

e) perfect aspect

f) pronoun *it*

g) possibility modals

h) predictive modals

i) *by*-passives

j) conditionals.

The distribution of the mean frequency of features that highly co-occur within each sub-genre and across sub-genres compared to the general corpus can be seen in Table 2.0 A cut was made in the features having salient loadings of 9.6 and over. Table 2.1 presents the results of the co-occurrence of those features having loadings of less than 9.6. As just ten linguistic features were analyzed in the study, and as the ten constitute the main reason for comparison, no exclusions of linguistic features were made in spite of the fact that some of the features had very low loadings. The abbreviations used in the tables stand for the following:

1. A.-P.: agentless passive
2. *By*-P.: *by*-passive
3. 3rd P.P.: 3rd person pronouns
4. P. *it*: pronoun *it*
5. D.: demonstrative pronouns
6. Poss. M. : Possibility modals
7. Pred. M. : Predictive modals
8. N.: Nominalizations
9. P. A. : Perfect Aspect
10. Cond. : Conditionals.

Table 2.0 *Description of the distribution of the linguistic features that had a co-occurrence of 9.6 and over within each sub-genre and across sub-genres compared to the general corpus.*

| Biochemistry | Microbiology | Biology | General |
|---|---|---|---|
| N. | N. | N | N. |
| 3rd P.P. | Poss. M. | A-P | A-P |
| A-P | A-P | 3rd P.P. | 3rd P.P. |
| P. A. | D. | D. | D. |
| Poss. M. | | P.A. | P.A. |
| D. | | | |

Table 2.1 *Description of the distribution of the linguistic features that had a co-occurrence of less than 9.6 within each sub-genre and across sub-genres compared to the general corpus.*

| Biochemistry | Microbiology | Biology | General |
|---|---|---|---|
| Pred. M | 3rd. P.P. | Pr. it | Poss. M |
| P. *it* | P. *it* | Poss. M | P. *it* |
| *By*-P. | P. A. | Pred. M. | Pred. M. |
| Cond. | Cond. | *By*-P. | Cond. |
| - | *By*- P. | Cond. | *By*-P. |
| | Pred. M. | | |

Summary

The aim of Chapter III has been to describe the way in which the data was collected.

The first part of the chapter explains the way in which the program used in the study, AnyText™, operates and gives a number of examples that substantiate the reason why the program was chosen[12].

The second part of the chapter provides the description of the data collected. Tables showing the frequency of occurrence of the linguistic features per genre are also presented.

---

12 A number of other programs were available for the study:
a) HyperRESEARCH is a HyperCard-based Macintosh application that allows quantitative and qualitative analysis. Although it was not designed for linguistic analysis, it can do word searches but the application works with one 'study' at a time. In order to do any search, one has to do the codings which can be deleted, copied or renamed. The program proves to be very useful in hypothesis testing. Researchers can express hypotheses as sets of simple "production rules," and then can electronically test the working hypothesis on any case or cases in their study. The search with AnyText™ proved to be faster.
b) WordCruncher sofware package provides a WCIndex and WCView. WCIndex enables the researcher to index DOS text files, so that they can retrieve and manipulate data from these text files with WCView. It requires a PC-compatible computer. WordCruncher enables to make sophisticated search combinations, even in large documents, with an amazing speed. However, the search with AnyText™ proved to be faster.

## Chapter IV
## Analysis of the Data

Let us remember that the objective of the present study is to search for the linguistic features which are shared specifically by texts in the field of biological science and to compare the findings to the general science corpora that have been described in Biber's analyses.

Biber uses frequency counts of particular linguistic features as a means to give exact quantitative characterization of a text; however, these counts do not identify linguistic dimensions. Linguistic dimensions are characterized on the basis of a consistent co-occurrence pattern among features; that is, the consistent co-occurrence of a cluster of features in texts define a linguistic dimension.

In order to identify the occurrences of linguistic features and genre distinctions to be included in a macro analysis, Biber makes use of microscopic analysis. This analysis, as it has been said in Chapter II, provides a functional analysis of the features, so the analyst is able to interpret textual dimensions in functional terms.

Macroscopic analyses pinpoint the underlying textual dimensions in a set of texts, enable the general description of a general account of linguistic variations among texts, and provide a

framework for the discussions of the similarities and differences among texts and genres.

Given Biber's idea that 'if certain features consistently co-occur, then it is reasonable to look for an underlying functional influence that encourages their use', I shall first compare the general results of the microscopic analysis done in the Biological Science Corpora (Table 3.0) to that of Biber's (Table 3.1), then see if the features that highly co-occur in our study coincide with Biber's (Table 3.2). I shall then compare the underlying functional dimensions in both studies from a macroscopic outlook.

Once this general comparison is done, a similar analysis with the different sub-genres would follow, keeping in mind that:

a) A macroscopic outlook analyzes the co-occurrence patterns among linguistic features, identifying textual dimensions; and

b) A microscopic analysis identifies the features and interprets the dimensions in functional terms.

Table 3.0 *Mean frequencies for the Biological Science Corpora as a whole*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 12.8 | 5.7 | 30.0 | 24.3 | 10.6 |
| *by*-passives | 2.3 | 0.0 | 7.3 | 7.3 | 2.3 |
| 3rd person pronouns | 11.6 | 3.6 | 27.1 | 23.5 | 6.6 |
| pronoun *it* | 7.2 | 1.4 | 28.5 | 27.1 | 9.8 |
| demonstrative pr. | 11.1 | 2.8 | 24.2 | 21.4 | 5.9 |
| possibility modals | 9.2 | 0.0 | 24.5 | 24.5 | 6.1 |
| predictive modals | 4.2 | 0.0 | 42.0 | 42.0 | 8.3 |
| nominalizations | 20.2 | 0.0 | 38.5 | 38.5 | 10.2 |
| perfect aspect | 9.6 | 0.0 | 27.1 | 27.1 | 5.9 |
| conditionals | 2.4 | 0.0 | 8.5 | 8.5 | 2.4 |

Table 3.1 *Mean Frequencies for the Science Corpora presented in Biber's Study*

| Linguistic feature | mean | minimum value | maximum value | range | standard deviation |
|---|---|---|---|---|---|
| agentless passives | 17.0 | 7.0 | 38.0 | 31.0 | 7.4 |
| *by*-passives | 2.0 | 0.0 | 8.0 | 8.0 | 1.7 |
| 3rd person pronouns | 11.5 | 0.0 | 46.0 | 46.0 | 10.6 |
| pronoun *it* | 5.9 | 1.0 | 16.0 | 15.0 | 3.4 |
| demonstrative pr. | 2.5 | 0.0 | 9.0 | 9.0 | 1.9 |
| possibility modals | 5.6 | 0.0 | 14.0 | 14.0 | 3.1 |
| predictive modals | 3.7 | 0.0 | 14.0 | 14.0 | 3.4 |
| nominalizations | 35.8 | 11.0 | 71.0 | 60.0 | 13.3 |
| perfect aspect | 4.9 | 0.0 | 16.0 | 16.0 | 3.5 |
| conditionals | 2.1 | 0.0 | 9.0 | 9.0 | 2.1 |

Table 3.2. *Comparison between the mean frequencies in the Biological Science Corpora and the Science Corpora presented in Biber's Study*

| Linguistic feature | mean Bio. Sc. Corpora | mean Biber's study |
|---|---|---|
| agentless passives | 12.8 | 17.0 |
| *by*-passives | 2.3 | 2.0 |
| 3rd person pronouns | 11.6 | 11.5 |
| pronoun *it* | 7.2 | 5.9 |
| demonstrative pr. | 11.1 | 2.5 |
| possibility modals | 9.2 | 5.6 |
| predictive modals | 4.2 | 3.7 |
| nominalizations | 20.2 | 35.8 |
| perfect aspect | 9.6 | 4.9 |
| conditionals | 2.4 | 2.1 |

A comparison between this study and Biber's brings out interesting results. *Nominalizations* have a high frequency of occurrence with a mean of 20.4 in this study and a mean of 35.8 in Biber's. This is the feature that most frequently occurred in the corpus analyzed. Although it is not our objective to analyze the occurrences of *nominalizations* separately; that is, reporting how many words ending in *-tion* occur in this text, and how many words ending in *-ment* occur in another; it is interesting to note that a high percentage of *nominalizations* fall into the *-tion* group. The following excerpt demonstrates this phenomenon clearly:

> "...A solution to that problem now seems to be within reach: DNA, sequences that confer high expression on globin genes, despite being located far from them, have been identified and shown still to function[13] after being engineered to be much closer to the genes. In the meantime, ironically, the need for gene therapy in thalassaemia has receded, as bone marrow transplantation has increasingly found a place in the treatment of the most serious forms of the disease. Used in that way, transplantation is a type of gene therapy, albeit one that cocks a snook at the sophistication of genetic engineering..."
>
> (Taken from *Reading Selections for Biological Science Students*, pp. 71)[14]

---

[13] Although the word *function* ends in -tion, this 'to-infinitive' is not a nominalization.

[14] The paragraph is part of one of the texts analyzed in the study.

This paragraph has 100 words, there are 5 nominalizations from which 4 fall in the *-tion* group. The rest of the 'groups' (*-ity, -ment, and -ness*) were not highly marked. Nominalizations are mostly used to indicate a referentially explicit statement intended to give information on a certain theme.

The next feature with the second highest frequency of occurrence in both studies was *agentless passives* . In the Biological Science Corpora 12.8 *agentless passives* occurred per 1000 words, while in the general Science Corpus 17 *agentless passives* occurred per 1000 words. It was observed that whenever there was a high frequency of passives there were many nominalizations, a correlation that also exists in Biber's study. The *agentless passives* are used to present propositions with no emphasis on the agent. They are used to give prominence to the patient of the verb, the entity acted upon. Agentless passives are frequently used in procedural discourse where the agent is presupposed across several clauses and the specific agent of a clause is not important to the discourse purpose. This type of discourse is typically very technical in content and formal in style.

Keeping the order of features from those which co-occurred the most to those which co-occurred the least in both studies, the third position is shared by *third person pronouns.* The word *shared* was used because the features shared the third position and almost the same mean frequencies. *Third person pronouns* have a mean frequency of 11.5 in Biber's study; and a mean frequency of 11.6 in this study. *Third person pronouns* mark reference to referents apart from the speaker and addressee. The results show that *agentless*

*passives*, *nominalizations* and *third person pronouns* highly co-occur in both studies.

The fourth feature having a salient loading in the Biological Science Corpora was the *demonstrative pronouns.* The feature was highly marked across all subgenres with a mean of 11.1. This feature was not marked at all in Biber's study. It had a mean of 2.5. Demonstrative pronouns are highly used as referential elements, a device very much used in scientific texts. It was very interesting to observe that there was a correlation between the presence of *third person pronouns* and *demonstrative pronouns* . When one of them occurred frequently, the other one did not. This does not mean that the presence of one presupposed the absence of the other. Both occurred in texts and there was always one more marked than the other.; but they do co-occur with *passives* and *nominalizations.*

The fifth more marked feature was the *perfect aspect.* The markedness of this feature in this study compared to its unmarkedness in Biber's is amazing (the same with *demonstrative pronouns*). A mean of 4.9 was reported in Biber's work whereas in this work the mean is 9.6. *Perfect aspect* proved to be very much used in the general Biological Science Corpora as the feature decribes past events.

Let us analyze the co-occurrences of these five linguistic features in the following paragraph extracted from one of the articles analyzed in the study:

> "...A solution to that problem now seems to be within reach: DNA, sequences that confer high expression

on globin genes, despite being located far from them, have been identified and shown still to function after being engineered to be much closer to the genes. In the meantime, ironically, the need for gene therapy in thalassaemia has receded, as bone marrow transplantation has increasingly found a place in the treatment of the most serious forms of the disease. Used in that way, transplantation is a type of gene therapy, albeit one that cocks a snook at the sophistication of genetic engeneering..."

(Taken from *Reading Selections for Biological Science Students*, pp. 71)

a) nominalizations 6

b) agentless passives 4

c) third person pronouns 1

d) demonstrative p. 2

e) perfect aspect 4

Although *third person pronouns* and *demonstrative pronouns* do not show a high frequency of occurrence, when one normalizes these raw counts the figures change quite a bit, so *third person pronouns* would occur 10 times per 1000 words, a figure closer to the mean in the general Biological Science Corpora.

What has been described so far is the procedure for constructing a factor. The first factor would then be the sum of the features that highly co-occurred in the study. That is:

*20.2 (nominalizations) + 12.8 (agentless passives) + 11.6 (third person pronouns) + 11.1 (demonstrative pronouns) + 9.6 (perfect aspect) = 65.3*

So, 65.3 would be the first factor in the analysis. However if the dimension underlying this factor were to be analyzed, one would first have to think of the functions of the linguistic features:

a) nominalizations: indicate a referentially explicit statement,

b) agentless passives: present propositions with no emphasis on the agent,

c) third person pronouns: mark reference to referent apart from the speaker and addressee,

d) demonstrative pronouns: mark reference to referent apart from the speaker and addressee,

e) perfect aspect: describes a past event that is psychologically relevant to the present.

When interpreting the functions of these features, and when comparing their co-occurrence underlying the dimensions presented in Biber's study, there is clear evidence that the features fall into four of Biber's dimensions:

a) *nominalizations* are related to the 'Explicit versus Situation-Dependent' dimension,

b) *agentless passives* (as well as *by*-passives) are related to the 'Abstract versus Non-Abstract Information' dimension,

c) perfect aspect and third person pronouns are related to the 'Narrative versus Non-narrative Concerns' dimension, and

d) demonstrative pronouns are related to the 'On-line Informational Elaboration' dimension.

The two dimensions in Biber study which were highly marked in the General Science Corpora were 'Explicit versus Situation

Dependent' and 'Abstract versus Non-Abstract Information'. The two mostly marked features in our study also fall into these two dimensions; however, if we take into consideration that dimensions are characterized on the basis of a consistent co-occurrence pattern among features, we cannot take Biber's dimensions as point of departure for comparison in our study as the consistent co-occurrence of the cluster of features previously analyzed are scattered in different dimensions in Biber's study. So, if we were to name this dimension in our study we would call it *'Allusion to Experimental Versus Factual Information'*. The way this dimension is labelled in this study focuses much deeper on the general function of the features in the texts included in the corpus.

Scientists are taught that empirical research is factual and agentless and this explains the use of such linguistic features in their writings. They generally write about *what was done, how it was done, reference to the studies done before, and what the results are.* But let us continue with the analysis of the remaining features.

The sixth feature having a salient loading in the study is *possibility modals*. The mean is really high, 9.2, compared to Biber's study in which the mean is 5.6. *Possibility modals* are pronouncements concerning the ability or possibility of certain events occurring, that they *can, may* or *might* occur. In the biological science world these possibilities are always present.

The pronoun *it*, in the seventh position, marks a reduced surface form that can be a noun or a phrase. In this study the mean (7.2) is close to that of Biber's (5.9).

*Predictive modals* fall in the eighth step. In Biber's study they have a mean of 3.7 not far from the mean in this study, 4.2. Predictive modals are direct pronouncements that certain events *will* occur (something always long for in natural sciences but not always achieved).

Numbers nine and ten are shared by *conditionals* (2.4) and *by*-passives (2.3) which (keeping the order) have a mean frequency in Biber's study of 2.1 and 2. Let us now see how these five linguistic features co-occur in a given text:

> "If ever a clan of flowering plants were put on this earth to help man, it would be the sunflowers, whose golden-rayed blossoms seem so symbolic of late summer. Some of them are downright amazing in their usefulness. In one species nearly every part of the plant is or has been of economic value -even its ashes. Sunflowers have been used to feed people, birds,and bees, to encourage egg laying, make paper and to serve a myriad of other purposes. Because so many sunflower seeds are purchased to feed wild birds, odd species and hybrids can pop up almost anywhere.
>
> (Taken from the book Reading Selections for Biological Science Students, pp. 12)

The distribution of raw counts of the five linguistic features in this 100-word text is the following:

a) possibility modals: 1

b) Pronoun *it* 1

c) Predictive modals 1

d) Conditionals 1

e) *by* passives 0

As can be seen, these features are not as marked as the first five features analyzed in this study. As the features unmarkedly co-occur they can be said to belong to the same factor. So the second factor in this analysis would be:

*9.2 (possibility modals) + 7.2 (pronoun **it**) + 4.2 (predictive modals) + 2.4 (conditionals) + 2.3 (**by**-passives) = 25.3*

The function of each of these linguistic features is the following:

a) possibility modals: pronounce that certain events can, may or might occur,

b) pronoun *it*: marks a reduced surface form,

c) predictive modals: pronounce that certain events will occur,

d) conditionals: specify the conditions that are required in order for certain events to occur, and

e) *by* passives: reduce the emphasis on the agent.

The functions underlying these linguistic features may be related to Biber's fourth dimension: 'Overt Expression of Persuation', as possibility modals, predictive modals, and conditionals fall into this dimension. However 'Overt Expression of Persuation' was not reported as a characteristic dimension of scientific texts in Biber's study. Biber explains that possibility modals, predictive modals and conditionals are often used to persuade; nevertheless, the analysis of these features in the corpus analyzed do not seem to indicate persuasion, but conceivable information. Hence, if this dimension

were to be named, a more comfortable expression for this analysis would be 'Conceivable Information'.

The range of variation of some linguistic features is very high in both studies:

| Features | standard deviation in this study | standard deviation in Biber's study |
|---|---|---|
| passives | 10.6 | 7.4 |
| third p.p. | 6.6 | 10.6 |
| pron. it | 9.8 | 3.4 |
| predictive modals | 8.3 | 3.4 |
| nominalizations | 10.27 | 13.3 |

The magnitude of *passives* with respect to the range of possible variation was very high in the general Biological Science corpora and across the whole study. Some other features had a very high standard deviation within the subgenres (See Tables 1.1, 1.2, and 1.3).

The co-occurrences of linguistic features having salient loadings pattern very much the same in the three subgenres compared to the the analysis just done between the General Science Corpora and Biber's study; but there are some interesting differences (See Tables 4.0, 4.1, and 4.2 )

Biochemistry sub-genre

This subgenre has a very strong first factor:

*20.4 (nominalizations) + 13.4 (third personal pronouns) + 12.7 (agentless passives) + 11.4 (perfect aspect) + 10.6 (possibility modals) + 9.8 (demonstrative pronouns) = 78.3*

The fact that possibility modals fall into this first factor does not really affect the underlying textual dimension analyzed in the general biological science corpora. When alluding to experimental information, biochemists tend to have the readers realize that what they are saying can, may or might possibly happen.

The second factor patterns with that of the general corpora.

*7.1 (predictive modals) + 6.1 (pronoun **it**) + 3.1 (**by**-passives) + 2.4 (conditionals) = 18.7*

Predictive modals in the biochemisty subgenre have the highest mean across all subgenres. Biochemistry texts tend to convince the reader of the certainty of whatever reference they make and possibility modals and predictive modals help to convey this.

Table 4. 0 *Distribution of linguistic features from those having the highest mean to those having the lowest mean in the biochemistry subgenre compared to the general biological science corpora and to Biber's study*

| *feature* | *Biochemistry* | *General Bio. Sc. C.* | *Biber's* |
|---|---|---|---|
| nominalizations | 20.4 | 20.2 | 35.8 |
| 3rd P. Pronouns | 13.4 | 11.6 | 11.5 |
| agentless passives | 12.7 | 12.8 | 17.0 |
| perfect aspect | 11.4 | 9.6 | 4.9 |
| possibility modals | 10.6 | 9.2 | 5.6 |
| demonstrative P. | 9.8 | 11.1 | 2.5 |
| predictive modals | 7.1 | 4.2 | 3.7 |
| pronoun *it* | 6.1 | 7.2 | 5.9 |
| by passives | 3.1 | 2.3 | 2.0 |
| conditionals | 2.4 | 2.4 | 2.1 |

Table 4. 1 *Distribution of linguistic features from those having the highest mean to those having the lowest mean in the microbiology subgenre compared to the general biological science corpora and to Biber's study*

| *feature* | *Microbiology* | *General Bio. Sc. C.* | *Biber's* |
|---|---|---|---|
| nominalizations | 22.4 | 20.2 | 35.8 |
| possibility modals | 12.6 | 9.2 | 5.6 |
| agentless passives | 11.4 | 12.8 | 17.0 |
| demonstrative P. | 9.8 | 11.1 | 2.5 |
| 3rd P. Pronouns | 7.8 | 11.6 | 11.5 |
| pronoun *it* | 7.0 | 7.2 | 5.9 |
| perfect aspect | 7.0 | 9.6 | 4.9 |
| conditionals | 2.3 | 2.4 | 2.1 |
| by passives | 1.4 | 2.3 | 2.0 |
| predictive modals | 1.2 | 4.2 | 3.7 |

Table 4. 2 *Distribution of linguistic features from those having the highest mean to those having the lowest mean in the biology subgenre compared to the general biological science corpora and to Biber's study*

| *feature* | *Biology* | *General Bio. Sc. C.* | *Biber's* |
|---|---|---|---|
| nominalizations | 17.8 | 20.2 | 35.8 |
| agentless passives | 14.2 | 12.8 | 17.0 |
| 3rd P. Pronouns | 13.7 | 11.6 | 11.5 |
| demonstrative P. | 12.6 | 11.1 | 2.5 |
| perfect aspect | 10.5 | 9.6 | 4.9 |
| pronoun *it* | 8.5 | 7.2 | 5.9 |
| possibility modals | 4.5 | 9.2 | 5.6 |
| predictive modals | 4.4 | 4.2 | 3.7 |
| by passives | 2.6 | 2.3 | 2.0 |
| conditionals | 2.6 | 2.4 | 2.1 |

Microbiology subgenre

The description of the underlying linguistic dimension 'Allusion to Experimental information' through the first factor just includes four features:

*22.4 (nominalizations) + 12.6 (possibility modals) + 11.4 (agentless passives) + 10.9 (demonstratives) = 57.3*

This factor does not include either *perfect aspect,* or *third person pronouns.* In fact, the loadings of these features in this subgenre were the lowest across the whole study (7.0, *perfect aspect*; 7.8 *third person pronouns*). The features were not very unmarked but appeared to be more frequent in the biochemistry (11.4, *perfect aspect*; 13.4 *third person pronouns*).) and the biology (10.5, *perfect aspect*; 13.7*third person pronouns*).) subgenres. The referential means in microbiology relies more on demonstrative pronouns (10.9). Although *present tense* was not a feature included for comparison in this study, it was noted that the microbiology subgenre makes extensive use of this feature.

The dimension underlying the second factor did again co-occur with that of the general corpora:

*7.8 (third person pronouns) + 7.0 (pronoun **it**) + 7.0 (perfect aspect) + 2.3 (conditionals) + 1.4 (**by**-passives) + 1.2 (predictive modals) = 26.7*

Again, the features having the lowest salient loadings across all subgenres are included in this factor, *conditionals* (2.3), ***by**-passives* (1.4) and *predictive modals* (1.2) -the lowest loading ever across the study. This indicates that microbiologists do not often say that 'something *will, would* or *shall* occur'. The low co-occurrence of all these features in this subgenre makes the factor

stronger. This distinctiveness of the microbiology subgenre, not patterning equally the same across the other subgenres gives substance to the importance of going beyond genres to analyze subgenres. Microbiologists tend to let the reader see that the information they put forward *may, might, can or could* be conceived, always leaving a space for a large field of probabilities but not certainties.

Biology subgenre

This subgenre patterns almost exactly the same as the general biological science corpora. The following figures gives grounds for the previous statement:

Factor 1 (general biological science corpora

*20.2 (nominalizations) + 12.8 (agentless passives) + 11.6 (third person pronouns) + 11.1 (demonstrative pronouns) + 9.6 (perfect aspect) = 65.3*

Factor 1 (biology subgenre)

*17.8 (nominalizations) + 14.2 (agentless passives) + 13.7 (third person pronouns) + 12.6 (demonstrative pronouns) + 10.5 (perfect aspect) = 68.8*

The analysis for this factor and its underlying dimension is exactly the same as that of the general science corpora previously seen.

The second factor has a certain peculiarity:

Factor 1 (general biological science corpora)

*9.2 (possibility modals) + 7.2 (pronoun **it**) + 4.2 (predictive modals) + 2.4 (conditionals) + 2.3 (**by**-passives) = 25.3*

Factor 1 (biology subgenre)

*8.5 (pronoun **it**) + 4.5 (possibility modals) + 4.4 (predictive modals) + 2.6 (**by**-passives) + 2.6 (conditionals) = 22.6*

*Possibility modals* have the lowest salient loading in the biology subgenre compared to its loadings in the other subgenres and across the study. 'Pure biologists' do not rely very much on what *may, might, can or could* be possible. They prefer to give more certainty to their statements with a more frequent use of predictive modals.

Summary

This chapter starts with a comparison of the general results of the microscopic analysis done in the Biological Science Corpora and the results in Biber's study. This comparison is followed by an analysis of the underlying functional dimensions seen in both studies. The analysis serves as a basis to establish similar comparisons between the findings in the general Biological Science Corpora and the findings in each of the subgenres included in the study, inasmuch as a comparison specifically between the cluster of features representing a factor giving way to an underlying dimension in the general Biological Science Corpora in this study did not always coincide with that of Biber's.

# Chapter V
# Conclusions

The present study offers a rationale for genre analysis. It provides support for both, the existence of genres and the importance in carrying out a study departing from simply the analysis of the markedness and/or unmarkedness of linguistic features in a typology of texts to the underlying function of their co-occurrence.

Another characteristic feature of this study is the use of a computerized text corpora and of a not-grammatically tagged computer program for the automatic identification of ten linguistic features.

Four steps were involved in the identification of linguistic features:

a) First, zipping to the word which characterized the feature,

b) Second, looking for the occurrences of the word in the whole text,

c) Third, finding out if these occurrences could be compared to the linguistic features identified, and

d) Last counting the occurrences which matched the linguistic features described.

An additional distinctive characteristic in the study is having narrowed down the analysis from the General Science Corpora, to the

Biological Science Corpora, to each of the branches of this last corpora. Such an analysis provides accurate and valuable information for future comparative analysis (See Tables 5.0, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, and 5.9).

**Table 5.0.** ***Distribution of the mean frequencies of agentless passives across the study***

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 12.7 | 11.4 | 14.2 | 12.8 | 17.0 |
| minimum | 5.7 | 5.7 | 7.1 | 5.7 | 7.0 |
| maximum | 30.0 | 30.0 | 29.8 | 30.0 | 38 |
| range | 24.3 | 24.3 | 22.7 | 24.3 | 31.0 |
| st. dev. | 12.2 | 8.0 | 11.2 | 10.6 | 7.4 |

Table 5.1. *Distribution of the mean frequencies of by-passives across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 3.1 | 1.4 | 2.6 | 2.3 | 2.0 |
| minimum | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| maximum | 7.1 | 7.3 | 5.7 | 7.3 | 8.0 |
| range | 7.1 | 7.3 | 5.7 | 7.3 | 8.0 |
| st. dev. | 2.1 | 2.3 | 1.9 | 2.3 | 1.7 |

Table 5.2. *Distribution of the mean frequencies of third person pronouns across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 13.4 | 7.8 | 13.7 | 11.6 | 11.5 |
| minimum | 4.6 | 3.6 | 5.7 | 3.6 | 0.0 |
| maximum | 27.1 | 15.7 | 21.4 | 27.1 | 46.0 |
| range | 22.5 | 11.9 | 15.7 | 23.5 | 46.0 |
| st. dev. | 6.3 | 3.8 | 7.1 | 6.6 | 10.6 |

Table 5.3. *Distribution of the mean frequencies of pronoun it across the study*

| frequencies | Bioch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 6.1 | 7.0 | 8.5 | 7.2 | 5.9 |
| minimum | 2.8 | 1.4 | 1.4 | 1.4 | 1.0 |
| maximum | 20.7 | 28.5 | 24.2 | 28.5 | 16.0 |
| range | 17.9 | 27.1 | 22,8 | 27.1 | 15.0 |
| st. dev. | 8.6 | 11.8 | 8.6 | 9.8 | 3.4 |

Table 5.4. *Distribution of the mean frequencies of demonstrative pronouns across the study*

| frequencies | Bioch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 9.8 | 10.9 | 12.6 | 11.1 | 2.5 |
| minimum | 7.1 | 3.0 | 2.8 | 2.8 | 0.0 |
| maximum | 14.2 | 20.1 | 24.2 | 24.2 | 9.0 |
| range | 7.1 | 17.1 | 21.4 | 21.4 | 9.0 |
| st. dev. | 2.2 | 3.9 | 9.0 | 5.9 | 1.9 |

Table 5.5. *Distribution of the mean frequencies of possibility modals across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 10.6 | 12.6 | 4.5 | 9.2 | 5.6 |
| minimum | 4.6 | 7.1 | 0.0 | 0.0 | 0.0 |
| maximum | 24.5 | 21.4 | 12.8 | 24.5 | 14.0 |
| range | 19.9 | 14.3 | 12.8 | 24.5 | 14.0 |
| st. dev. | 6.5 | 6.4 | 3.6 | 6.1 | 3.1 |

Table 5.6. *Distribution of the mean frequencies of predictive modals across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 7.1 | 1.2 | 4.4 | 4.2 | 3.7 |
| minimum | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| maximum | 42.0 | 4.2 | 11.2 | 42.0 | 14.0 |
| range | 42.0 | 4.2 | 11.2 | 42.0 | 14.0 |
| st. dev. | 13.3 | 1.4 | 3.5 | 8.3 | 3.4 |

Table 5.7. *Distribution of the mean frequencies of nominalizations across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 20.4 | 22.4 | 17.8 | 20.2 | 35.8 |
| minimum | 0.0 | 15.7 | 2.8 | 0.0 | 11.0 |
| maximum | 38.5 | 38.3 | 27.3 | 38.5 | 71.0 |
| range | 38.5 | 22.6 | 24.5 | 38.5 | 60.0 |
| st. dev. | 13.2 | 7.3 | 8.6 | 10.2 | 13.3 |

Table 5.8. *Distribution of the mean frequencies of perfect aspect across the study*

| frequencies | Bloch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 11.4 | 7.0 | 10.5 | 9.6 | 4.9 |
| minimum | 4.2 | 0.0 | 2.8 | 0.0 | 0.0 |
| maximum | 27.1 | 11.6 | 18.5 | 27.1 | 16.0 |
| range | 22.9 | 11.6 | 15.7 | 27.1 | 16.0 |
| st. dev. | 6.9 | 3.7 | 5.4 | 5.9 | 3.5 |

Table 5.9. *Distribution of the mean frequencies of conditionals across the study*

| frequencies | Bioch. | Microbio. | Biology | General | Biber's |
|---|---|---|---|---|---|
| mean | 2.4 | 2.3 | 2.6 | 2.4 | 2.1 |
| minimum | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| maximum | 7.1 | 6.1 | 8.5 | 8.5 | 9.0 |
| range | 7.1 | 6.1 | 8.5 | 8.5 | 9.0 |
| st. dev. | 2.4 | 1.9 | 2.8 | 2.4 | 2.1 |

The analysis presented here corroborates Biber's thesis that knowing what linguistic features co-occur in a text and across texts helps researchers understand why the linguistic features occur. This is the same as to say that this helps one to understand the underlying functional dimension of their co-occurrence.

Although this study began by investigating the co-occurrence of features in the General Biological Science Corpora in relation to the Science Corpora presented in Biber's study, the fact that the biological sciences are divided into three main branches made it almost impossible not to go deeper in the analysis. Fortunally the deeper the study went, the more interesting the results were and the more obvious the relevance of the results for curriculum developers became. For example, the fact that the study reveals what language forms tend to be used more frequently in one context than in

another, would facilitate a more scientific organization of the ESP curricula in the field of biological sciences.

The present study has identified two dimensions of variation in the general Biological Science Corpora and among three different subgenres: Biology, Microbiology and Biochemistry, which are:

a) 'Allusion to experimental information'

b) 'Conceivable information'

It has also specified the relations among subgenres with respect to those dimensions.

Additional research is required to find out the relations among texts in other fields, such as the Social Sciences, the Natural Sciences, and the Exact Sciences. The present model of genre analysis should prove useful for such related studies in ESL and it is hoped that it will provide a foundation for research to identify the relevance of genre analysis in reading and writing.

References

1. Ausubel, D. (1964). "Adults vs. children in second language learning: psychological considerations", *Modern Language Journal* No. 48, pp. 420-24.
2. Barber, C. L. (1962). "Some measurable characteristics of modern scentific prose", in *Contributions to English Syntax and Phonology*, Stockholm: Almquist and Wiksill.
3. Bates, E. (1979). *The emergence of symbols*. New York: Academic Press.
4. Ben-Amos, Dan (1976). *Folklore Genres*, Austin: University of Texas Press.
5. Besnier, Niko (1986). "Register as a sociolinguistic unit: defining formality", in *Social and Cognitive Perspectives* on Language, Los Angeles: University of Southern California.
6. Biber, Douglas (1984). A *model of textual relations within the written and spoken modes*. Ph. D. dissertation, University of Southern California.

   (1986). Spoken and written textual dimensions in English: resolving the contradictory findings, *Language* No. 62, pp. 384-414.

   (1988). Variation Across Speech and Writing, Cambridge: Cambridge University Press.
7. Biber D. and Finegan E. (1991). "On the exploitation of computerized corpora in variation studies", in *Corpus Linguistics* pp. 204-220, Longman.

8. Bloomfield, Leonard (1914). *An Introduction to the Study of Language,* New York: Holt.

9. Chomsky, Noam (1953). "System of syntactic analysis", *Journal of Symbolic Logic* No. 18, pp. 242-56.

10. Couture, Barbara (1986). *Functional approaches to writing: research perspectives*, Norwood, NJ: Ablex.

11. Crandall, J. A. (1984). "Adult ESL: The other ESP", *The ESP Journal*, No. 3, pp. 91-96.

12. Crookes, Graham (1986). "Towards a validated analysis of scientific text structure", *Applied Linguistics* No. 7, pp. 57-70.

13. Dow, R. A. and Ryan, J. T. (1990). "Preparing the language student for professional interaction" in *Interactive Language Teaching,* Cambridge: Cambridge University Press, pp. 194-210.

14. Electronic Text Corporation (1989). *WordCruncher User's Manual*, Version 4.30, Provo, UT.

15. Ellis, Rod (1986). *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

16. Ewer, J. R. and Latorre, G. (1969). *A Course in Basic Scientific English*, Longman.

17. Figueredo, M. et al. (1991). *Readings for Biological Science Students*, Havana: University of Havana Press.

18. Frow, John (1980). "Discourse genre". *The Journal of Literary Semantics* No. 9, pp. 73-9.

19. Garcia, G. and Peña, J. (1991). *Reading Selections for Foodscience Students*, Havana: University of Havana Press.

20. Gregory, M. and Carroll, S. (1978). *Language and Situation: language varieties and their social contexts*, London: Routledge and Kegan Paul.

21. Halliday, M. A. K., McIntosh, A. and Strevens, P. (1964). *The Linguistic Sciences and Language Teaching*, Longman.

22. Halliday, Michael A. K. (1973). *Explorations in the Function of Language*. London: Arnold.

(1988). "On the language of physical science" in *Registers of Written English*, London: Pinter.

23. Hesse-Biber, S., Dupui, P., and Scott, T. (1991). "HyperRESEARCH: A Computer Program for the Analysis of Quantitative Data with an Emphasis on Hypothesis Testing and Multimedia Analysis", *Qualitative Sociology*, Vol. 14, No. 4, pp. 289-306.

24. Hymes, Dell (1974). *Foundations in Sociolinguistics: An Ethnographic Approach,* Philadelphia: University of Pensilvania Press.

25. Jamieson, Katheleen M. (1975). "Antecedent genre as rhetorical constrain", *Quarterly Journal of Speech* No. 61, pp. 406-15.

26. Krashen, S. D. (1982). *Principles and practice in Second Language Acquisition*, Oxford: Pergamon Press.

27. Long, Michael H. (1990). "At Least a Second Language Acquisition Theory Needs to Explain". *TESOL Quarterly*. Vol. 24, No. 4, pp. 649-661.

28. Marshall, Stewart (1991). " A genre-based approach to the teaching of report writing", *English for Specific Purposes*, Vol. 10, No. 1, pp. 3-13.

29. Martin, James (1985). "Process and Text: Two aspects of human semiosis", in *Systemic perspectives on discourse*, Vol. 1., Norwood, NJ: Ablex.

30. Miller, Carolyn R. (1984). "Genre as social action", *Quarterly Journal of Speech* No. 70, pp. 151-67.

31. Mohan, Bernard A. (1986). *Language and Content*, Reading MA: Addison-Wesley Publishing Company.

32. Nwogy, Kevin N. (1991). "Structure of Science Popularization: A genre-analysis approach to the schema of popularized medical texts", *English for Specific Purposes*, Vol. 10, No. 2, pp. 111-123.

33. Payne, P. , Stringham, D., and Saia, M. (1991). *AnyText™ User's Manual*, Linguist's Sofware, Inc. Edmonds, U.S.A.

34. Peirce, Charles S. (1931-1935). *Collected Papers* (Vols. 1-6). Cambridge, MA: Harvard University Press.

35. Phillips, M. K. (1981). "Toward a theory of LSP methodology" in R. Mackay and J.D. Palmer (Eds.), *Languages for Specific Purposes: Program Design and Evaluation*, Rowley, MA: Newbury House.

36. Quirk, R. et al. (1985). *A comprehensive grammar of the English language*. London: Longman.

37. Reid, Joy (1991). "Responding to different topic types: a quantitative analysis from a contrastive rhetoric perspective", n *Second Language Writing: Research Insights for the Classroom,* Cambridge: Cambridge University Press.

38. Reid, Thomas B. (1956). "Linguistics, structuralism, philology", *Archivum Linguisticum* No. 8.

39. Rivers, Wilga (1973). "From linguistic competence to communicative competence", *TESOL Quarterly* No. 7, pp. 25-34. (1990)."Interaction as the key to teaching language for communication", in *Interactive Language Teaching,* Cambridge: Cambridge University Press, pp.3-16.

40. Robinson, P. (1980). *ESP (English for Specific Purposes)*, Oxford: Pergamon Press.

41. Rosenblatt, Louise M. (1989). "Writing and Reading: The Transactional Theory" in *Reading and Writing Connections,* Allyn and Bacon.

42. Rumelhart, D. E. (1981). Schemata: The building blocks of cognition. In J. T. Guthrie (ed.), Comprehension and teaching: Research reviews, Newark, DE: International Reading Association.

43. Sampson, Gloria (1990). *Transcultural Learning Everyday*, Collection of Papers reproduced at Simon Fraser University. (1992). Notes from lectures given on genre analysis in the Summer semestre at Simon Fraser University.

44. Sapir, Edward (1921). *Language*, New York: Harcourt, Brace, and World.

45. Saussure, Ferdinand de (1916). *Cours de Linguistique Générale.* (ed. Charles Bally, Albert Sechehaye and Albert Riedlinger), Lausanne: Payot.

46. Savignon, Sandra J. (1983). *Communicative Competence: Theory and Classroom Practice.* Addison-Wesley Publishing Company.

47. Saville-Troike, Muriel (1982). *The ethnography of communication*, Oxford: Basil Blackwell.

48. Schiffrin, Deborah (1981). "Tense variation in narrative". *Language* No. 57, pp. 45-62.

49. Selinker, L and Trimble, L. (1975). *Technical Communication for foreign engineering students.*,Office of Engineering Research, Seatle, University of Washington.

50. Sinclair, Johns et al. (1990). *The Collins Cobuild English Language Dictionary*, Collins: Collins, London and Glasgow.

51. Swales, John M. (1971). *Writing Scientific English*, Nelson.
(1981). *Aspects of article introductions*, Birmingham, UK: The University of Aston, Language Studies Unit.
(1984). "Research into the structure of introductions to journal articles and its application to the teaching of academic writing", in *Common ground: shared interests in ESP and communication studies*, Oxford: Pergamon Press.
(1990). *Genre Analysis. English in Academic and Research Settings*, Cambridge, Cambridge University Press.

52. Todorov, Tzvetan (1976). "The origin of genres". *New Literary History* No. 8, pp. 159-70.

53. Ure, J. and Ellis, J. (1972). "Register in descriptive linguistics and linguistic sociology". In Oscar Uribe-Villegas (ed.) *Issues in Sociolinguistics*, The Hague, Mouton.

54. Valdman, A. (1980). "Communicative ability and syllabus design for global foreign language courses", in *Studies in Second Language Acquisition*, pp. 81-96.

55. Vygotsky, Lev S. (1962). *Thought and Language* (F. Hanfmann and G. Vakar, Eds. and Trans.). Cambridge, MA: MIT Press.

Vygotsky, Lev. S. (1939). "Thought and Speech", *Psychiatry* No. 2, pp. 29-54.

56. Werner , H and Kaplan, B. (1962). *Symbol Formation*, New York: Wiley.

57. Widdowson, H. G. (1978). *Teaching language as communication*, Oxford: Oxford University Press.

(1983). *Learning purpose and language use*, Oxford: Oxford University Press.

58. Winograd, Terry (1972). *Understanding Natural Language*, New York: Academic.

(1983). *Language as a Cognitive Process*, Reading, MA: Addison-Wesley Publishing Company.

Appendix A

Title and author of original source where the texts were extracted:

1. "Understanding and Coping With Bee Poisonings" by Charles K. Zubovits, *American Bee Journal*, August, 1988.

2. "Missing Genes May 'Hold Back' Cancer" by Gail Vines, *New Scientist* No. 28, July 1988.

3. "Edging Towards Human Gene Therapy" by Peter Newmark, *Nature*, Vol. 342, No. 16, Novermber, 1989.

4. "Tests to Detect Infection" by Andrew Scott, *New Scientist*, No. 20, March, 1987.

5. "Bee Research Digest" by Dr. Edward E. Southwick, *American Bee Journal*, February, 1988.

6. "Improved Method for the Detection of Bacillus Larvae Spores in Honey" by H. Shimanuki and D. A. Knox, *American Bee Journal*, May, 1988.

7. "Inside the AIDS Virus I" by Zeda Rosenberg and Anthony Fauci, *New Scientist* No. 10, February, 1990.

8. "Inside the AIDS Virus II" by Zeda Rosenberg and Anthony Fauci, *New Scientist* No. 10, February, 1990.

9. "Inside the AIDS virus III" by Zeda Rosenberg and Anthony Fauci, *New Scientist* No. 10, February, 1990.

10. "Torn Genes" (author and publication not acknowledged).

11. "Flowers of the Sun" by Jack Sanders, *Wildflower*, Summer, 1990.

12. "Hopi Lima Beans" by Kevin Dahl, *Wildflower,* Summer, 1990.

13. "Honey Bee Mites: Comments on Recent Observations" by L. Baily, *American Bee Journal*, 1990.

14. "Flowers That Go Underground" by Allen H. Benton, *Wildflower*, Summer, 1990.

15. "Surgeons Transplant Liver From Living Donor" by Ian Anderson, *New Scientist* No. 12, August, 1990.

16. "Electrophoretic Analysis of Palm Hybrids: A Research Update" by Clifton E. Nauman, *Fairchild Tropical Garden Bulletin,* April, 1990.

18. "Don't Throw Away Your Propolis" by Warren Ogren, *American Bee Journal*, 1990.

19. "Vaccine Fights Malaria Without Antibodies" by Frank Cox, *New Scientist* No. 5, May, 1988.

20. "New Hope on the AIDS Vaccine Front" by Alexandra Levine and Jonas Salk, *Science*, Vol. 244, No. 4910, June, 1989.

21. "Fungus Breaks the Mould of Phosphate Fertilisers" (no author acknowledged), *New Scientist* No. 9, June, 1988.

22. "Medicine That Sticks to the Point" by David Nordell, *New Scientist* No. 9, June, 1988.

23. "A Switch That Sets Cancerous Cells in Motion" by Christopher Joyce, *New Scientist* No. 28, July, 1988.

24. "A Sophisticated Network of Defences" (no author acknowledged), *New Scientist* No. 7, April 1988.

Appendix B

Sample of 3 of the texts (one per sub-genre) analyzed in the study:

1. Microbiology

The best way to tell whether someone is infected with HIV is to test their blood for the presence of the virus. This can be done, but it is difficult. There is no reliable direct test for the presence of the virus that is simple enough for routine mass screening. Several biotechnology companies are trying to devise such a test. The Cetus corporation in the US, for instance, has applied several patents covering the basis of a test of this type.

At the moment the only way to test the numbers of people is to look for the indirect signs of infection that a persons immune system provides. Our immune system responds to infection with HIV by making anti-HIV antibodies, proteins that bind specifically to proteins of the virus. Usually the antibodies apparently do no do the infected individual much good. This might be because they are produced too late, after the virus has hidden away within the body's cells. Or it might be that although the antibodies bind to a virus, they cannot neutralize it.

The HIV blood tests now in use indicate that a person has been infected by detecting the presence of anti-HIV antibodies in the blood. The scientific basis of such tests is simple. First, researchers grow large quantities of HIV in the laboratory and then either purify its proteins or produce a crude cell homogenate. When the blood of an infected person is added to HIV proteins, the anti-HIV antibodies in the blood bind to the viral proteins. This binding of

antibodies to viral proteins indicates that the blood is from an infected person.

To be identified as an antibody positive in this way does not mean that a person has AIDS. Aids is the name of the disease that infection can cause. Most people who suffer from AIDS have been infected with the virus for a long time, often several years, before they show sure signs of having AIDS. The evidence we now have which comes from studies over 3 to 5 years, suggests that after 3 years, 15 to 20 per cent of the people infected with HIV developed AIDS. But clinicians have not yet been able to work out the full implications of infection. A positive result in the antibody test indicates only that a person has been infected by HIV.

It is possible that in some people the immune system can fight off, or at least contain an HIV infection. But it will be a long time before anyone can be certain and the indications so far are not good. There is evidence to suggest that the immune response generated in some people infected with HIV might protect them against further infection from other forms of the virus, even if it cannot do much against the initial infection which caused the response. Although no much confort to someone already suffering from AIDS, this brightens the prospects for finding an effective vaccine.

The development of blood tests to identify aids revolutionized investigations of the aids disease because it enabled epidemiologists to begin to chart the spread of the virus. There are however many uncertainties.

One of the worst of these is the time lag between the moment of infection and the appearance of antibodies in a person's blood.

This lag is usually between four weeks and four months, but may be longer, perhaps more than a year in some cases. There is also evidence that, very rarely, some people infected with the virus may never develop anti-HIV antibodies. One person is known to have developed full-blown AIDS despite persistent negative results in the test for anti-HIV antibodies, but this could be because the immune system functions differently in babies. A negative blood test is not, therefore, conclusive proof that you are free from infection. But in most cases it does mean that no infection has occurred. In a few cases a positive blood test can be a "false positive", indicating infection when there is none. But a second test, corrects the mistake. The arrival of direct and simple tests for the HIV virus and its genetic material should help to eliminate any uncertainty that still exists.

2. Biology

Everyone loves beautiful flowers, and knows that the most beautiful ones are usually just attempts to t rick insects into coming to the flower for nectar. If they fall for it, pollination will be achieved and the flower will produce seeds. In the struggle to produce as many progeny as possible, a beautiful flower may be an effective weapon .

Many plants depend on wind rather than insects to pollinate each other, and these plants have to use a different strategy.

A few flowers use a third strategy, the strategy of self-fertilization. One way to achieve this is to have flowers which contain both stamens and pistils (the male and female structures

respectively) and then keeping the flowers closed until the pollen has matured and fertilized the ovules. This phenomenon occurs in a variety of plants, and is quite surprising when you consider the lengths that most plants go to in trying to avoid self-fertilization.

One of the more unusual methods of achieving self-fertilization is found in plants which also produce showy cross-fertilize flowers. In a sense, this is a sort of back-up system, which allows the plant to produce more seeds than it otherwise could, or to succeed in reproducing if the flowers are not fertilized-by insects. These plants produce; special closed flowers, either underground or just at the earth's surface. After the above-ground flowers have faded and the seeds, if any, have been dispersed, these extra flowers fertilize themselves and produce a second ,crop of seeds.

These seeds will grow up to be exact genetic copies of the parent plant. Whatever benefits derive from the mixing of genes of different plants are lost to these self-fertilizing plants. Apparently, in an evolutionary sense, the benefits of producing a second crop of seeds are enough to make the lack of gene-mixing unimportant.

Among local plants, the violets are well known for producing these closed flowers. The next time you are out in a woodland where violets are blooming or have recently completed blooming, scratch away the leaf mold at the base of the plant and you will see some tiny greenish closed flowers on short stems of an inch or so in length.

There is another less common plant which has a particularly showy insect-pollinated flower, and a second set of flowers which

actually grow underground. this is the beautiful fringed milkwort Polygal paucifolia, an uncommon spring-blooming plant which has flowers like small orchids both in color and in form. They are, however not remotely related to orchids but belong to a small family of plants of which only a half dozen species occur in the Northeast.

As the flowers begin to fade, these little plants begin to produce underground closed flowers. By mid-summer, these flowers have produced their crop of seeds. At this point you may have wondered, how these underground seeds get scattered around. If they all stayed right where they were produced, they would compete with each other and with the parent plant as well.

The seeds of the underground flowers have an attached sac of oily liquid, one which has nothing to do with the growth of the seed, but is extremely attractive to ants. Being just under the ground, these seeds are easy for ants to find as they wander about looking for food. When they find these seeds the ants carry them back to their nests and eat the oil-sacs.

Having eaten the tastiest part of the seed, the ants have no further use for the seeds, so they take them out and dump them, somewhere away from the nest. This process scatters the seeds around and allows many of them to grow without competition from their siblings. Once again, the advantages of cross fertilization are surrendered in the interest of producing extra progeny.

If you ever had any doubt that cooperation is as much a law of nature as is competition, this mutually useful relationship of ants and milkworts should convince you. And if you ever doubted that the

wild is full of unexpected phenomena, the existence of flowers that bloom underground may be enough to convince you of that too.

3. Biochemistry

Introducing a session on the prospects for gene therapy in Boston, Leon Rosenberg wisely refrained from predicting when the first trials of human gene therapy would commence. The fact of the matter is that potential gene therapists have been beavering away to produce appropriate reagents and procedures in the full knowledge that, for no good reason, they will face far more barriers than are usual for experimental clinical trials.

Those interested in globin gene therapy were forced to take stock. It became clear that the control of globin gene expression is more complex than had been thought. The frustrating result was that whereas it was possible to insert globin genes into cells, insufficient globin was produced to make the cells of early potential use in therapy. But a solution to that problem now seems to be within reach: DNA sequences that confer high expression on globin genes, despite being located far from them have been identified and shown still to function after being engineered to be much closer to the genes. In the meantime, ironically, the need for gene therapy in thalassaemia has receded, as bone marrow transplantation has increasingly found a place in the treatment of the most serious forms of the disease. Used in that way, transplantation is a type of gene therapy, albeit one that cocks a snook at the sophistication of genetic engineering. Whereas its use is limited by the availability

of compatible bone marrow and the problems of graft rejection, both problems could, in principle, be solved were it possible to transplant the stem cells that are the progenitors for all blood cells rather than whole bone marrow. Such a prospect was held out by Irving Weissman. As the much-needed assay for human stem cells, he hopes to test their ability to reconstitute the haematopoietic system of the genetically immunodeficient SCID mouse. But, to judge by his experience with mouse stem cells, there would still remain the pressing problem of finding a practical way to stimulate the proliferation of isolated human stem cells if they are ever to be available in sufficient quantities for transplantalion. In principle, many genetic deficiencies could be treated by transplantation of the appropriate normal tissues but because of the practical difficulties surrogate methods are being explored. One example that excited interest at the meeting was the 'neo-organs' After several months, the implanted structures have acquired a continuous blood supply from the liver, contain structures that resemble nerves and have grown to as much as half the size of the liver. That hepatocytes in neo-organs are functional has been shown by the apparent ability of neo--organs seeded with Lewis rat hepatocytes to restore to nonnal, the elevated levels of serum bilirubin in Gunn rats.

Mice carrying foreign genes, whether introduced into fertilized eggs or embryonic stem cells, are proving to be rich sources of animal models of human diseases.

A variation on this theme is to try and interfere with the function of a gene in order to produce a model of a genetic deficiency. The transgene is a so-called dominant negative. It is the

possibility of expressing in the blood cells of AIDS patients dominant negative genes for proteins of the human immunodeficiency virus in order to interfere with human immunodeficiency virus function, that might yet drive the first successful attack on the growing battery of regulations that face anyone wishing to pioneer the use of gene therapy. For now, the only tentative step towards such therapy that has been taken is the inclusion of a genetic marker in some of the 'tumor infiltrating Lymphocytes' that are being used in an experimental procedure for cancer treatment. The marker is enabling the fate of the cells to be followed but plans are already afoot to include not just a marker gene but also a gene for interleukin-2 that may lengthen the life and efficacy of the cells. As their use will be tested only in patients for whom conventional forms of treatment have failed, there should be no more reason for excessive regulatory hurdles than is usual in such circumstances. Unfortunately, given that gene therapy is involved, even that is too much to hope for.